Compositional covariance shrinkage and regularised partial correlations

Suzanne Jin[1,2], Cédric Notredame[1,2] and Ionas Erb[1,3]

Abstract

We propose an estimation procedure for covariation in wide compositional data sets. For compositions, widely-used logratio variables are interdependent due to a common reference. Logratio uncorrelated compositions are linearly independent before the unit-sum constraint is imposed. We show how they are used to construct bespoke shrinkage targets for logratio covariance matrices and test a simple procedure for partial correlation estimates on both a simulated and a single-cell gene expression data set. For the underlying counts, different zero imputations are evaluated. The partial correlation induced by the closure is derived analytically. Data and code are available from GitHub.

1. Introduction

The study of data variability is at the heart of data analysis. A simple way of quantifying the variability around some central tendency is given by a bilinear function defined for each pair of variables: the covariance. The distribution that makes no assumptions beyond a finite covariance matrix is the multivariate Gaussian, and it is thus optimal in a maximum-entropy sense. Covariance matrices can capture the relevant aspects of

---

[1] Centre for Genomic Regulation (CRG), Dr Aiguader, 88, 08003 Barcelona, Spain.

[2] Universitat Pompeu Fabra, Barcelona, Spain.

[3] Corresponding author: `ionas.erb@crg.eu`

variability in a wide variety of scenarios (Whittaker, 1990), and the Gaussian does not have to be the "true" underlying distribution for this[1].

Often the limitations of data acquisition make good covariance estimates difficult, and regularisation techniques are adopted to deal with small data sets. One such technique is covariance shrinkage, which is a regularisation technique that lets us invert singular covariance matrices arising from the presence of fewer samples than variables (Schäfer and Strimmer, 2005). It has recently been applied in the context of compositional data in genomics (Badri et al., 2020), where clear improvements over empirical measures were shown with a comprehensive and reproducible benchmark. Compositional data are constrained in that their samples sum to a constant, which causes biases that can be alleviated by logratio transformations (Aitchison, 1986). Badri et al. (2020) applied the standard shrinkage approach directly to logratio covariance matrices, which does not take into account the interdependence among logratio variables. Our main interest in this contribution is to provide and test an improved shrinkage procedure that is custom-made for logratio-transformed data. The properly regularized logratio covariance matrices can then be inverted to estimate logratio-based partial correlations, which we have proposed as a measure of interaction between compositional variables (Erb, 2020).

On a more general note, we aim to contribute to finding a principled way of quantifying interactions in data sets consisting of compositions or relative counts. When inferring interactions between variables, a useful distinction (see, e.g., Werhli, Grzegorczyk and Husmeier (2006) is made between relevance networks, which are based on pairwise interaction coefficients (such as Pearson correlation) and graphical models, which make use of all the variables to infer pairwise association (e.g., via partial correlation). While the former are susceptible to inferring indirect interactions that are mediated by variables that are unrelated to the pair of variables in question, the latter aim to infer the conditional independence structure of their joint distribution (Whittaker, 1990). This way they can in principle infer direct interactions between two variables. The distinction is highly relevant for compositional data, where variables are intrinsically connected due to the unit-sum constraint. Here, simple relevance networks are inadequate because of the negative bias coming from the constraint. This bias has been addressed by measures that use ratios of variables, such as Aitchison's logratio variance (Aitchison, 1986), as well as other coefficients inferring proportionality (Lovell et al., 2015; Erb and Notredame, 2016). These can be used to infer relevance networks in compositional data, and Badri et al. (2020) have shown how shrinkage improves estimates of the proportionality coefficient. However, to define graphical models for compositions similar to Kurtz et al. (2015), logratio-based partial correlations will be instrumental.

In what follows, in Section 2 we briefly review how the covariance structure and partial correlations are described in the logratio framework and outline James-Stein shrink-

---

[1]For a discussion covering this aspect of maximum-entropy approaches see van Nimwegen (2016). A more general maximum-entropy approach to compositional data has recently been proposed Weistuch et al. (2022), but it is difficult to obtain analytic results from it.

age of general covariance matrices as proposed in Schäfer and Strimmer (2005). In Section 3, the concept of logratio uncorrelated compositions introduced by Aitchison is described, and it is shown how one can make use of it to define shrinkage for logratio covariance matrices. Two equivalent approaches are proposed: 1) shrinking a so-called basis covariance matrix which is then transformed into a logratio covariance, and 2) shrinking logratio covariance matrices directly with the help of logratio uncorrelated targets. Due to its simplicity, we prefer the first approach, and both the synthetic and the experimental data benchmark of Section 4 will be based on it. In Section 5, we discuss how the use of partial correlations relates to normalisation techniques in genomics and provide an expression for the remaining partial correlation in logratio-uncorrelated compositions. A summary section and an appendix containing some of the more technical details conclude the paper.

2. Preliminaries

2.1. *Compositional data and logratio covariance*

Let us denote N compositional samples by vectors $\boldsymbol{p}_i$, $i = 1, \ldots, N$, whose D positive components (called *parts*) sum to 1. Specifically, we consider data with a large number of parts D compared with the number of samples N. Such wide compositional data are of interest in genomics (Erb, Gloor and Quinn, 2020), where they appear mainly in the form of relative counts coming from sequencing experiments (Quinn et al., 2018). In this case we would consider $\boldsymbol{p}$ the closed data obtained when dividing a count sample by its total, i.e., the compositions are the empirical parameter estimates of a multinomial model.

The counts play the role of a *basis* of the composition, a concept that was introduced by Aitchison (1982), and which we will talk about in more detail in Section 2.4. Considering the counts as relative data is justified by assigning no importance to the variations in the count totals (which correspond to the size of the basis). Compared with real numbers, relative counts constitute a discrete basis, which implies some problems of its own, especially when counts are small (Lovell, Chua and McGrath, 2020) or vanish entirely. Such problems can be alleviated with suitable weights (Law et al., 2014) or with finite-size corrections to the frequencies that can again be obtained via shrinkage (Hausser and Strimmer, 2009). The latter, known as frequency shrinkage, has the advantage that it imputes count zeros naturally. The promise that this has shown for compositional data (Quinn and Erb, 2021; Greenacre et al., 2022) will be confirmed here in our benchmarks.

Let us now consider an $N \times (D-1)$ data matrix $\boldsymbol{X}$ with real-valued elements. These elements are obtained from the logratios we evaluate from our compositional samples $\boldsymbol{p}_i$ using the additive logratio transformation (ALR, introduced in Aitchison (1982))

$$x_{ij} = \log \frac{p_{ij}}{p_{iD}}, \qquad j = 1, \ldots, D-1. \tag{1}$$

The (unbiased) sample covariance matrix $\boldsymbol{S}$ is then defined by its elements

$$s_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j), \tag{2}$$

where $\bar{x}_i$ are the components of the N-dimensional vector $\bar{\boldsymbol{x}}$ of column means of $\boldsymbol{X}$. In order to have a criterion allowing us to judge the quality of our covariance estimate based on $\boldsymbol{S}$, we need a distributional assumption. Let us assume that $\boldsymbol{X}$ was sampled from a normal distribution with population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This can be written as

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{x}$ denotes a $D-1$-dimensional random vector taking values as in the rows of $\boldsymbol{X}$. In this case the random compositions $\boldsymbol{p}$ are distributed according to Aitchison's celebrated logistic normal distribution (Aitchison, 1982). If we denote the density of the compositions themselves by $f_{\mathcal{L}}$, the Gaussian density $f_{\mathcal{N}}$ will be multiplied by an additional factor due to the resulting Jacobian:

$$f_{\mathcal{L}}(\boldsymbol{p} = \boldsymbol{p}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\prod_{j=1}^{D} p_{ij} \right)^{-1} f_{\mathcal{N}}(\boldsymbol{x} = \boldsymbol{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{x}_i$ is a function of $\boldsymbol{p}_i$ defined by (1).

2.2. Covariance shrinkage

To improve estimates of $\boldsymbol{\Sigma}$ from data, adding a bit of bias to the unbiased estimator $\boldsymbol{S}$ can reduce its overall error. This is a consequence of the bias-variance decomposition of the mean-squared error and can be achieved by convexly combining $\boldsymbol{S}$ with (also known as "shrinking it towards") a suitable target matrix $\boldsymbol{T}$:

$$\hat{\boldsymbol{\Sigma}} = \lambda \boldsymbol{T} + (1 - \lambda)\boldsymbol{S}, \tag{5}$$

where the shrinkage intensity λ is between zero and one. $\hat{\boldsymbol{\Sigma}}$ is a so-called James-Stein type estimator (James and Stein, 1961). If we forget for a moment that $\boldsymbol{X}$ was obtained from logratios of compositions, a typical target matrix could be diag($\boldsymbol{S}$). While (5) tells us nothing about suitable values of λ, its optimal value can be estimated from the data via an analytic expression that minimises a mean-squared error cost-function with respect to the population covariance (Ledoit and Wolf, 2003). The derivation of this expression makes use of the fact that $\boldsymbol{S}$ is unbiased. This optimum is given by

$$\lambda^* = \frac{\sum_{j=1}^{D} \sum_{i \neq j} \left[\mathrm{var}(s_{ij}) - \mathrm{cov}(s_{ij}, \tau_{ij}) \right]}{\sum_{j=1}^{D} \sum_{i \neq j} \mathbb{E}\left[(s_{ij} - \tau_{ij})^2 \right]}, \tag{6}$$

where τ_{ij} denotes the matrix entries of $\boldsymbol{T}$. Note that here, covariance and expectation of the matrix elements refer to parameters of a distribution which can be approximated by their empirical estimates for actual calculations, see Schäfer and Strimmer (2005).

To provide further background, we also show a principled interpretation of covariance shrinkage in the Appendix. There, we sketch how covariance shrinkage is equivalent to optimising the posterior probability of $\boldsymbol{\Sigma}$ from data and prior information. Note that λ^* is inferred from data, making this a quasi-empirical Bayes approach.

The shrunk covariance estimator is useful in itself as it often by far outperforms the sample covariance unless the number of samples considerably exceeds the number of variables. Regularisation becomes a necessity for inverting the covariance matrix whenever the number of samples is smaller than the number of variables. Therefore, shrinkage improves and extends the range of applications of the analysis proposed in the next section.

2.3. Logratio-based partial correlations

Partial correlations between two variables are evaluated by controlling for the linear dependence on all the remaining variables, e.g., Whittaker (1990). This is achieved by correlating the residuals of two variables, where the residuals are with respect to the linear least squares predictors obtained from the remaining variables. Intuitively, we only correlate those contributions to our variables that are orthogonal to a subspace spanned by the variables we control for. Partial correlations can therefore be used to extract direct dependencies between variables, as opposed to the ones based on Pearson correlation, which can be dominated by indirect interactions via other variables. Partial correlations between logratio variables can be defined independently of their reference part D, so they measure association directly between parts, provided their reference is controlled for Erb (2020). Let us recall the formula for the (logratio-based) partial correlation between two parts i, j of the random composition $\boldsymbol{p}$ via their ALR covariance $\boldsymbol{\Sigma}$. Below, the backslash means taking the set difference, so the variables we control for are all the ones not belonging to the pair in question:

$$r_{ij}(\boldsymbol{p}) := \mathrm{corr}(x_i, x_j | \{x_1, \ldots, x_{D-1}\} \setminus \{x_i, x_j\}) = \frac{-\sigma_{ij}^{(-1)}}{\sqrt{\sigma_{ii}^{(-1)} \sigma_{jj}^{(-1)}}}, \tag{7}$$

where $i \neq j$ and $i, j = 1, \ldots, D-1$, and where $\sigma_{ij}^{(-1)}$ denotes the elements of the inverse of $\boldsymbol{\Sigma}$. It may appear as if the reference part p_D is "sacrificed" here, and its partial correlations cannot be determined. Note, however, that r_{ij} is invariant under permutation of the reference part: all D choices of $\boldsymbol{\Sigma}$ lead to the same partial correlations (for the parts they have in common). It follows that we can obtain the partial correlation of any pair of parts by choosing an arbitrary reference part that does not form part of the pair. We have stated our definition in terms of the ALR covariance matrix because it comes closest to what is known from unconstrained data. There is also a (computationally) more convenient representation using a covariance matrix involving all parts in the form of centred logratios (the CLR covariance introduced in Section 2.4), from which the partial correlations can be obtained via its pseudoinverse. The inverse can also be estimated with

a sparsity assumption, which has been done to create ecological networks of microbes, see Kurtz et al. (2015).

2.4. Linear independence in logratio and basis covariance matrices

For covariance matrices of unconstrained data, vanishing off-diagonal elements mean linear independence of the variables. Logratio covariance matrices cannot be linearly independent in this sense. Although logratio transformations map compositional data from the simplex to a real-valued space, they cannot remove the dependence between the parts that is introduced by the loss of one dimension. The notion coming closest to linear independence for logratios is given by *logratio uncorrelated* compositions. In the following, we state some results from Aitchison (1986), section 5.9, where this concept is introduced. A composition $\boldsymbol{p}$ is said to be logratio uncorrelated (LU) if

$$\mathrm{cov}\left(\log \frac{p_i}{p_k}, \log \frac{p_j}{p_l} \right) = 0 \tag{8}$$

for every selection of four different indices i, j, k, l from $\{1, \ldots, D\}$. This case is shown in Figure 1. The covariance structure of LU compositions can be made explicit when
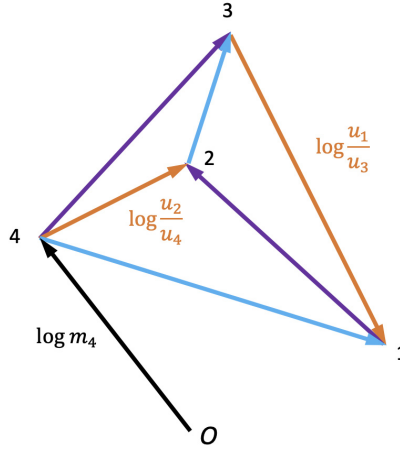

Figure 1. *Dependence structure in a logratio uncorrelated composition $\boldsymbol{u}$ with four parts. Transformed parts can be visualized as centred vectors with sampled components (labels are in random-variable notation), where length corresponds to standard deviation. The reference part is indexed by $D = 4$ and its basis $m_4 = tu_4$ is usually unknown (resulting in an unknown origin O). The vectors from the origin to the points indexed by 1 to 4 are the log-transformed basis vectors, their squared lengths correspond to the α_k defined in (9). As $\boldsymbol{u}$ is LU, they are orthogonal to each other in four dimensions. In three dimensions, only the logratio vectors of equal colour are orthogonal (e.g., the orange vectors labelled with their logratios). These correspond to vectors where all indices are different, see (8).*

defining a D-component vector $\boldsymbol{\alpha}$ with elements

$$\alpha_k = \text{cov}\left(\log\frac{p_i}{p_k}, \log\frac{p_j}{p_k}\right), \tag{9}$$

with the three indices i, j, k in $\{1,\ldots,D\}$ and all different from each other. For LU compositions, this definition makes sense because the value of α_k is constant regardless of the (unequal) indices i, j. The matrix $\boldsymbol{\Sigma}$ of an LU composition is now given by

$$\sigma_{ij} = \begin{cases} \alpha_i + \alpha_D & \text{if } i = j, \\ \alpha_D & \text{if } i \neq j. \end{cases} \qquad i, j = 1, \ldots, D-1. \tag{10}$$

(we reproduce the proof of this in the Appendix). While these definitions may seem overly abstract, they appear much more compelling in the light of Aitchison's concept of the basis of a composition – see Aitchison (1986, chapter 9). Intuitively, the basis is one of many ways our compositions can look like before the closure operation is applied. It is defined by $\boldsymbol{m} = t\boldsymbol{p}$ for some positive real t which is called the size of the basis and for which $t = \sum_{k=1}^{D} m_k$ holds (implying that it can change for each data sample). Of course, infinitely many bases of different size exist for a given unit-sum composition. Now the basis covariance matrix $\boldsymbol{\Omega}$ is defined to have the elements

$$\omega_{ij} = \text{cov}\left(\log m_i, \log m_j\right), \tag{11}$$

from which the covariance structure of a composition can be obtained unambiguously. If we assume that $\boldsymbol{\Omega}$ is diagonal, i.e., the logged components of the basis vector are mutually uncorrelated, we have the result that $\boldsymbol{\alpha}$ coincides with $\text{var}(\log\boldsymbol{m})$:

$$\alpha_i = \omega_{ii}, \qquad i = 1, \ldots, D. \tag{12}$$

This goes to show that $\boldsymbol{\alpha}$ can be seen as the variance vector of an independent distribution in a sample space that has one additional dimension (and in which the logratio sample space is embedded), see Figure 1. As $\boldsymbol{\alpha}$ was defined for logratios, this statement may appear confusing because of the apparent dependence of $\text{var}\log(t\boldsymbol{p})$ on the basis size t. However, the t-dependence drops out because $\boldsymbol{\Omega}$ is diagonal (see the proof of (12) in the Appendix).

For completeness, let us also mention an equivalent expression to (10) in terms of the centred logratio transformation (CLR). Here the reference involves all the parts of the composition. It transforms the compositional data matrix as

$$y_{ij} = \log\frac{p_{ij}}{g(\boldsymbol{p}_i)}, \tag{13}$$

where $g(\boldsymbol{p}_i) = \prod_{k=1}^{D} p_{ik}^{1/D}$ is the geometric mean of the i-th composition. Note that the resulting data matrix $\boldsymbol{Y}$ has a constraint $\sum_{k=1}^{D} y_{ik} = 0$ acting on its rows. This constraint leads to a singular CLR covariance matrix of rank $D-1$. The covariance of $\boldsymbol{y}$ is denoted

by $\boldsymbol{\Gamma}$. The two forms of logratio covariance are equivalent and can be transformed into each other (see the first two rows in Table 1). Using the transformation from $\boldsymbol{\Sigma}$ to $\boldsymbol{\Gamma}$ on (10), we obtain the elements of $\boldsymbol{\Gamma}$ for an LU composition as

$$\gamma_{ij} = \begin{cases} \alpha_i - (2\alpha_i - \frac{1}{D}\sum_k \alpha_k)/D & \text{if } i = j, \\[2mm] -(\alpha_i + \alpha_j - \frac{1}{D}\sum_k \alpha_k)/D & \text{if } i \neq j. \end{cases} \qquad i,j = 1,\ldots,D. \qquad (14)$$

$\boldsymbol{\Sigma} \to \boldsymbol{\Gamma}$	$\gamma_{ij} = \sigma_{ij} - \frac{1}{D}\sum_{k=1}^{D}\sigma_{ik} - \frac{1}{D}\sum_{k=1}^{D}\sigma_{kj} + \frac{1}{D^2}\sum_{k,l=1}^{D}\sigma_{kl}$
$\boldsymbol{\Gamma} \to \boldsymbol{\Sigma}$	$\sigma_{ij} = \gamma_{ij} - \gamma_{iD} - \gamma_{Dj} + \gamma_{DD}$
$\boldsymbol{\Omega} \to \boldsymbol{\Sigma}$	$\sigma_{ij} = \omega_{ij} - \omega_{iD} - \omega_{Dj} + \omega_{DD}$
$\boldsymbol{\Sigma} \to \boldsymbol{\Omega}$	$\omega_{ij} = \gamma_{ij}(\boldsymbol{\Sigma}) + \beta_i + \beta_j, \qquad \text{where}$ $\beta_j = \text{cov}\,(\text{clr}_j(\boldsymbol{p}), \log g(\boldsymbol{m})) - \frac{1}{2}\text{var}\,(\log g(\boldsymbol{m})).$
$\boldsymbol{\Omega} \to \boldsymbol{\Gamma}$	$\gamma_{ij} = \omega_{ij} - \frac{1}{D}\sum_{k=1}^{D}\omega_{ik} - \frac{1}{D}\sum_{k=1}^{D}\omega_{kj} + \frac{1}{D^2}\sum_{k,l=1}^{D}\omega_{kl}$
$\boldsymbol{\Gamma} \to \boldsymbol{\Omega}$	$\omega_{ij} = \gamma_{ij} + \beta_i + \beta_j, \qquad \text{see } \boldsymbol{\Sigma} \to \boldsymbol{\Omega}.$

Table 1. *Elementwise transformations between ALR, CLR, and basis covariance matrices (collected from Aitchison (1986)). Here, indices of σ_{ij} are allowed to take the value D, in which case σ_{ij} vanishes; clr_j stands for the j-th element of the CLR transform, and g denotes the geometric mean. All these transformations can also be expressed as simple matrix operations that are not shown here.*

3. Logratio covariance shrinkage

The specific covariance structure of compositional data would suggest to use logratio uncorrelated shrinkage targets instead of the diagonal targets used for unconstrained data. We will show that it is straightforward to construct such targets. However, it is usually more convenient in practice to work with diagonal targets, for which we have to do the shrinkage on a basis of the compositions. Once the shrinkage estimate for the basis covariance is obtained, it can then be back-transformed to a logratio covariance matrix. This strategy is described in the following section.

3.1. Basis covariance shrinkage using diagonal targets

A diagonal shrinkage target may often be preferred for practical reasons. As an example, the corpcor R package (Schäfer and Strimmer, 2005) uses the diagonal of the sample covariance as a target and does not allow for alternative targets. However, we have seen that logratio covariance matrices do not allow for this simple independence structure.

Therefore, we now define the shrinkage estimator directly for the basis covariance matrix $\boldsymbol{\Omega}$:

$$\hat{\boldsymbol{\Omega}} = \lambda \boldsymbol{T}_C + (1 - \lambda)\boldsymbol{C}, \tag{15}$$

where $\boldsymbol{C}$ is the empirical basis covariance matrix, and the shrinkage target $\boldsymbol{T}_C$ could be simply its diagonal. Thus, given some compositional data, our first step to construct an uncorrelated shrinkage target is to obtain an empirical basis covariance matrix from the data. For this, it is necessary to fix the size of the basis t. Note that doing this specifies the variability of the totals across samples. Given that we will back-transform our estimator to a logratio covariance matrix, this variability is irrelevant. The logratio covariance structure will not contain information about the basis size, so it is convenient to use $t_i = 1$ for all samples $i = 1, \ldots, N$, giving a basis $\boldsymbol{m}_i = \log \boldsymbol{p}_i$. In the case of relative count data, the counts themselves can provide the basis, and $\boldsymbol{C}$ can be estimated from their logarithms (provided zeros are taken care of). Yet another possibility is to start from an empirical logratio covariance matrix and obtain $\boldsymbol{C}$ via a transformation given in Table 1. These transformations simplify for a constant basis size.

The diagonal target has the additional advantage that the expression for the optimal value of λ simplifies (Schäfer and Strimmer, 2005) compared to the more general expression (6). For diagonal targets we have

$$\hat{\lambda}^* = \frac{\sum_{j=1}^{D} \sum_{i \neq j} \widehat{\text{var}}(c_{ij})}{\sum_{j=1}^{D} \sum_{i \neq j} c_{ij}^2}. \tag{16}$$

To obtain the shrinkage estimator of a logratio covariance matrix, all we have to do now is back-transform $\hat{\boldsymbol{\Omega}}$ to an appropriate logratio covariance matrix as specified in Table 1. This will also be the strategy we use for our benchmark in Section 4.

3.2. Logratio uncorrelated targets

While we will not use them in our benchmarks, it can be of general interest to construct LU targets for direct logratio covariance shrinkage. A simple way of constructing these targets would be as follows: pick the diagonal elements of the empirical basis covariance $\boldsymbol{C}$ as our α_i and then obtain the ALR target covariance by (10) and the CLR target covariance by (14). However, we can also skip the construction of a basis covariance entirely and express the targets directly in terms of the logratio sample covariance matrices. For the ALR target given the empirical covariance $\boldsymbol{S}$ we find

$$\tau_{ij} = \begin{cases} s_{ii} - \frac{2}{D} \sum_{k=1}^{D-1} s_{ik} + \frac{2}{D^2} \sum_{k,l} s_{kl} & \text{if } i = j, \\ \frac{2}{D^2} \sum_{k,l} s_{kl} & \text{if } i \neq j. \end{cases} \quad i, j = 1, \ldots, D-1. \tag{17}$$

Similarly, using an empirical CLR covariance matrix $\boldsymbol{G}$, the corresponding target elements t_{ij} are given via the diagonal elements of $\boldsymbol{G}$ by

$$t_{ij} = \begin{cases} g_{ii} - (2g_{ii} - \frac{1}{D} \sum_k g_{kk})/D & \text{if } i = j, \\ -(g_{ii} + g_{jj} - \frac{1}{D} \sum_k g_{kk})/D & \text{if } i \neq j. \end{cases} \quad i, j = 1, \ldots, D. \tag{18}$$

4. A benchmark of logratio covariance shrinkage

We verify our estimators of covariance and partial correlation on both simulated data as well as single-cell gene expression data. The latter are subsampled from a "ground truth" of greater sample size (this strategy has also been used for microbiome data in Badri et al. (2020)). Throughout, the corpcor R package from Schäfer and Strimmer (2005) is used for covariance matrix shrinkage. Note that we use the default implemented there, which consists in separately shrinking the diagonal elements of the covariance matrix with another λ parameter, see Opgen-Rhein and Strimmer (2007). The aim of the benchmark is the evaluation of our proposed procedure to first shrink a basis covariance and then transform the result to a logratio covariance. This is compared with a naive approach of direct logratio covariance shrinkage (with a diagonal target containing variances that inevitably mix parts) as well as no shrinkage at all. In the first part of this section, the logratios of our compositions are taken for granted and we just sample them from a normal distribution of rather moderate dimension ($D = 40$). In this case we do not have to be concerned about zeros in the basis samples. A scenario typical for genomics data sets is treated in the second part of this section, where the basis of our compositions are high-dimensional relative counts taken from single-cell gene expression data ($D = 500$). There we also test various ways of dealing with zeros as well as a finite-size correction for low counts in form of frequency shrinkage.
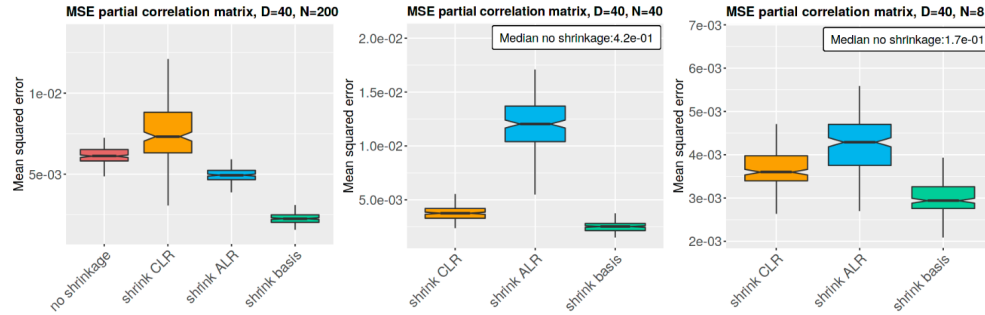

Figure 2. *Mean squared error of partial correlation matrix for sample sizes $N = 200$, $N = 40$, and $N = 8$. Colours indicate different estimation procedures (no shrinkage, naive shrinkage of CLR / ALR covariance and basis covariance shrinkage). Each boxplot contains the results of 200 simulations. Whenever estimates without shrinkage are not shown, their median value is given in the legend.*

4.1. Synthetic logistic normal data

To have reasonably realistic population parameters for the normal distribution, they are inferred from a single-cell gene expression data set (Riba, Oravecz and Durik et al.,

2022), where a subset of 240 genes that are non-zero everywhere across 5637 cells are selected. From these, in each of the 200 repetitions of the simulation, in step 1) $D = 40$ parts are chosen randomly, from which ALR and CLR covariance matrices as well as the partial correlation matrix (via the pseudoinverse of the CLR covariance) are constructed across the 5K+ cells. This is our ground truth. Then 2) the ALR mean vector $\boldsymbol{\mu}$ and the ALR covariance matrix $\boldsymbol{\Sigma}$ are used to produce N multivariate normal samples, where $N = 8$, 40, and 200. 3) The N samples are backtransformed to compositions. From these, sample estimates of CLR, ALR and (constant-size) basis covariance matrices as well as the partial correlation estimates are produced. This is done in three different ways: without shrinkage, with direct shrinkage of the logratio covariance matrices (using a diagonal target for the logratios), and via shrinkage of the basis covariance matrix (i.e., using a diagonal target for the log-transformed basis). 4) These sample estimates are compared to the ground truth established in the first step using an element-wise mean squared error. Note that after basis covariance shrinkage, partial correlations are identical if the basis is transformed to an ALR or a CLR covariance matrix (but it is convenient to use the CLR to obtain them).

The results of this procedure are seen in Figure 2 for partial correlation and Supplementary Figure S.1 for covariance. There is no doubt that shrinkage leads to substantial improvements when estimating covariation, even for $N = 5D$. In this regime, naive CLR shrinkage fares worse than using no shrinkage at all when estimating partial correlations. Interestingly, the error is small for the covariance matrix itself but seems to be compounded for inference of its inverse. In all cases we see a clear advantage when using shrinkage of the basis covariance, especially when comparing with naive ALR covariance shrinkage. We conclude that we achieve substantial improvements of our estimators when tailoring the shrinkage to the needs of logratio covariance matrices.

4.2. Single-cell gene expression data

Instead of simulating samples from a (realistic) covariance matrix, we now draw our covariance matrices directly from the data. Since these are sample covariance matrices, the benchmark now consists in recovering them as well as possible with undersampled data. While this is not ideal (the test case could in principle be as close or closer to the population covariance than the sample covariance that serves as ground truth), it seems a reasonable assumption that the degradation effect due to undersampling is greater than the deviation of the full sample covariance from the population covariance. In the previous section we could see that shrinking the covariance still has positive effects even for $N = 5D$. To avoid circularity, however, we prefer to not shrink the ground truth covariance matrix here. Our benchmark procedure is summarized in Figure 3. Single-cell gene expression data usually contain a very large number of zero counts, which would require strategies (e.g., pooling several cells into a single sample) that are not explored here. Instead, we concentrate on a core set of genes that are highly expressed and thus yield only a few zero counts. These can still be used to test a number of imputation strategies. We test two basic procedures implemented in the R package zCompositions
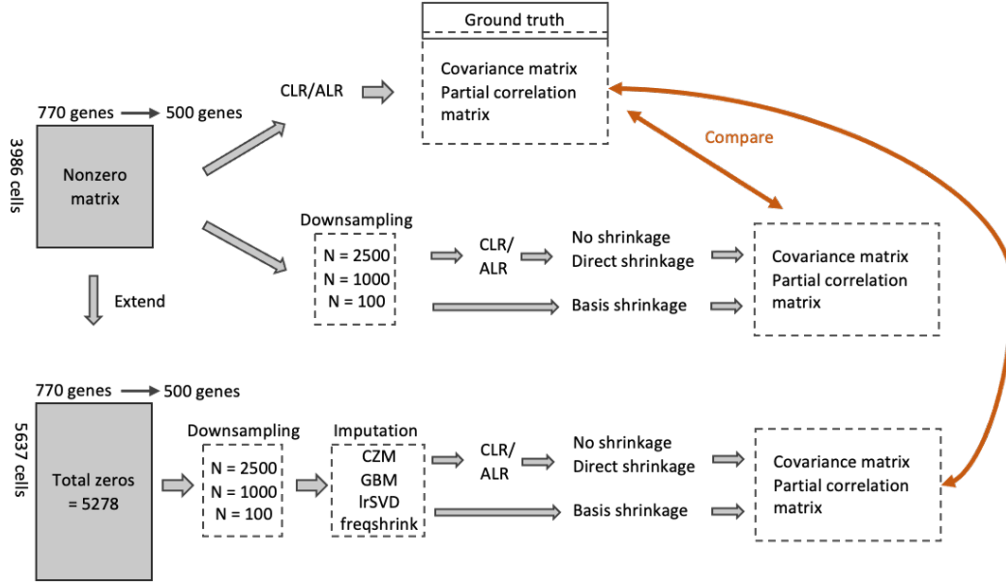
Figure 3. *Benchmark procedure for single-cell gene expression data.*

(Palarea-Albaladejo and Martín-Fernández, 2015) (Count Zero Multiplicative replacement, or CZM, and Geometric Bayesian Multiplicative replacement, or GBM), as well as one more sophisticated strategy that takes covariance structure into account (lrSVD), which was added to zCompositions recently (Palarea-Albaladejo et al., 2022). We also test frequency shrinkage as implemented in the R package entropy (Hausser and Strimmer, 2009).

To separate the effects of shrinkage and imputation, we defined two scenarios: 1) imputation is avoided by using a large subset of cells with only nonzero counts, and 2) the effect of shrinkage and imputation are jointly evaluated by including all the cells. More precisely, we used 770 genes for which 3986 cells (out of a total of 5637 cells) have no zeros, from which a ground-truth can be constructed. For each of the 200 re-samplings, 500 genes are chosen randomly from the 770 genes. Covariance and partial correlation matrices computed on the 3986 $\times$ 500 nonzero matrix define our ground truth. Then, in scenario 1, this matrix is subsampled to have *N*=2500, *N*=1000, and *N*=100 cells. Covariance and partial correlations computed on each of these matrices, and under the different shrinkage schemes are compared to the ground truth using mean squared error. In the second scenario, the same ground truth as defined before is used, while the subsampled data (*N*=2500, *N*=1000, and *N*=100) are drawn from the extended matrix of 5637 cells x 500 genes. Here, on average about 1500 zeros are imputed for the 2500 samples, 600 zeros are imputed for the 1000 samples and 60 for the 100 samples. Having additional samples that contain zeros is not ideal because these samples could in principle induce a slightly different covariance structure. The benefit, however, is that it provides a natural set of zero counts when otherwise zeros would have to be
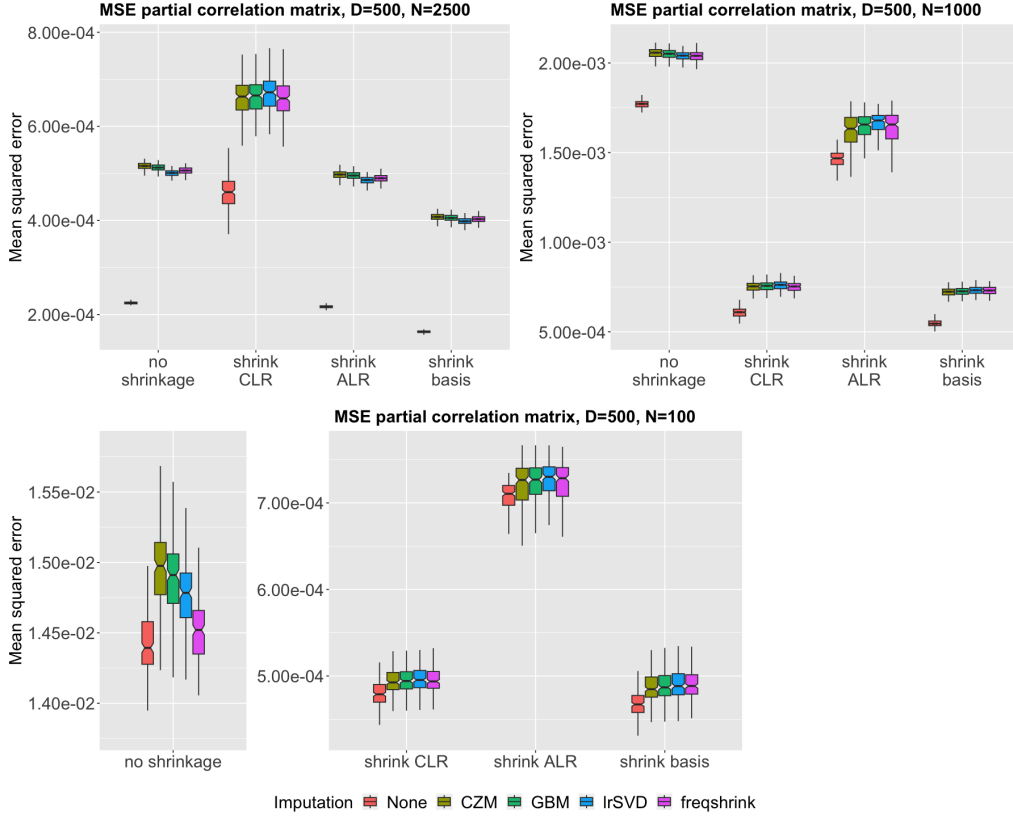
Figure 4. *Mean squared error of partial correlation matrices computed on three different sizes of subsamples (N=2500, 1000, 100) of single-cell gene expression data. Boxplots contain 200 samples each using different estimation procedures (no shrinkage, naive ALR/CLR shrinkage, basis covariance shrinkage). Their colour indicates the type of zero imputation used.*

artificially introduced. Now the subsampled matrix is first imputed using four different methods, and then covariance and partial correlations are again computed under the different shrinkage schemes.

The results for partial correlations are shown in Figure 4, and for the covariance matrices in Supplementary Figure S.2. Two main conclusions can be drawn from the results for such high-dimensional data. First, the effect of using the basis covariance shrinkage is strong compared with naive ALR covariance shrinkage, but it is much weaker when comparing with naive CLR covariance shrinkage. Naive shrinkage of the CLR covariance only leads to an important deterioration for the partial correlations at higher sample sizes, in concordance with what we observed in the synthetic benchmark. Apart from that, the improvement is mainly observed when no zeros are imputed, which hints at a flattening of the effect caused by the imputation. The advantage of basis covariance shrinkage becomes entirely invisible when directly evaluating the covariance matrices, whose inference seems to be less prone to error than their inverse. A second main con-

clusion is that imputation of zeros leads to considerable increase in mean squared error. All imputation schemes lead to a comparable loss in accuracy, with only slight differences between methods.

5. Logratio-based partial correlations for wide data

We have addressed some questions regarding covariance structure that remained unanswered in Erb (2020), especially with respect to the application to wide data sets like the ones occurring in genomics. We think it is beneficial to also discuss some questions regarding the interpretation of logratio-based partial correlations both in CoDA and genomics. In genomics, partial correlation analysis is common, but it has rarely been based on logratio analysis (Kurtz et al., 2015). In CoDA, on the other hand, logratio analysis is the dominant paradigm, but partial correlations have been introduced quite recently (Erb, 2020). In the following sections we will discuss the relationship between classical partial correlations and their logratio counterparts.

5.1. Normalisations to avoid compositional bias

We can think of two strategies to avoid compositional bias when analyzing covariance structure: trying to recover the original (absolute) signal (as contained in one particular basis covariance $\boldsymbol{\Omega}$) or using only that part of the signal that absolute and relative data have in common (which is contained in the logratio covariance). The first ("normalisation") strategy relies on assumptions because the closure operation discards a part of the signal and to recover the original totals up to a constant factor, additional information is needed. The second ("CoDA") strategy voluntarily discards this information and doing so introduces dependencies between the variables, but it is assumption-free. Interestingly, the distinction between the two approaches is not all that clear-cut because the logratio transformations Aitchison introduced can be used to pursue the "normalisation" strategy of re-opening the data with additional assumptions. This was discussed in Quinn et al. (2018), see especially the Supplementary Material. In the following, we revisit these arguments from the perspective of logratio covariance matrices and their basis covariance $\boldsymbol{\Omega}$. Let us start with the ALR covariance $\boldsymbol{\Sigma}$. It can be obtained from the basis covariance by

$$\sigma_{ij} = \omega_{ij} - \omega_{iD} - \omega_{jD} + \omega_{DD} \tag{19}$$

(see Table 1). Note that if the reference variable m_D remains unchanged across samples in the basis, the three last terms on the right-hand side vanish, and the two covariances coincide. This is the reason behind the widespread normalisation procedure in genomics where (absolutely) unchanged reference genes are used to "normalize" the data to avoid compositional bias. Similarly, for the CLR covariance $\boldsymbol{\Gamma}$ we have the relationship

$$\omega_{ij} = \gamma_{ij} + \beta_i + \beta_j \tag{20}$$

(see again Table 1). From the dependence of β_i on $g(\boldsymbol{m})$ we see that this implies that if the geometric mean of the basis remains unchanged across samples, the β terms vanish and

the CLR covariance is identical with the basis covariance. This is in agreement with the notion of effective library size normalisation, where a reference that remains unchanged on the absolute data is constructed to avoid compositional bias. The geometric mean can remain unchanged across samples before closure when the majority of genes is not subject to systematic change. Under this assumption the CLR transformed data have an identical correlation structure as the log-transformed absolute data.

We can see that the CoDA strategy has the advantage that, while it is on the safe side if the assumptions aren't met, it also recovers the original signal if they happen to be fulfilled. In the latter case, however, a specific reference and thus logratio covariance would be picked out because it recovers the basis covariance. While this is necessary for evaluating both covariance and correlation, it is unnecessary for partial correlation. The results remain invariant under permuting the reference in $\mathbf{\Sigma}$ and they are also identical when using the pseudoinverse of $\mathbf{\Gamma}$ for their evaluation, see also Erb (2020).

These benefits of logratio-based partial correlation are lost when using more naive approaches, like deriving partial correlations from the covariance matrix of $\log \boldsymbol{p}$, or from a *relative* count basis using $\log \boldsymbol{m}$ directly. Formally, partial correlations derived when not specifying a reference do not coincide with their logratio counterparts.
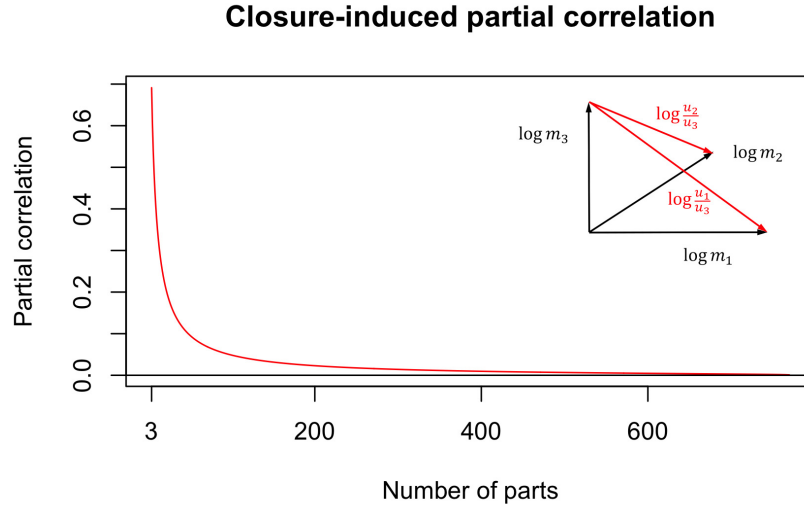
Closure-induced partial correlation

Figure 5. *A worst-case scenario for spurious logratio-based partial correlation of a fixed pair of parts in artificially uncorrelated data (see text). Successive increase of the number of parts from three (PC=0.69) to 770 (PC=0.001) shows how the closure-induced spurious signal tends to zero with the number of parts. Inset: The case $D = 3$. The vanishing correlation between the first two parts could be recovered from the logratios with respect to the third part (indicated in red) by correlating their projections into the plane that is orthogonal to $\log m_3$. This would just be the correlation between residuals when controlling for $\log m_3$. However, the information about $\log m_3$ is lost, so partial correlation between the logratios remains biased.*

5.2. What is the effect of the closure?

Partial correlations based on logratios can emulate their absolute counterparts extremely well (a benchmark of which will be shown in another contribution). However, the closure operation makes us lose one dimension which cannot be recovered. So there remains a difference between partial correlations derived from absolute (using logarithms only) versus relative data (using logratios). Or, to put it differently, partial correlations can change when we derive them from a logratio covariance as opposed to the original (usually unknown) basis covariance. Using the relation between $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ of Table 1, the respective covariance inverses are related by

$$\mathbf{\Omega}^{-1} = (\mathbf{\Gamma} + \mathbf{B})^{-1}, \tag{21}$$

where the $D \times D$ matrix B has elements $\beta_i + \beta_j$. We see that we can neatly separate the covariance signal introduced by the variability of the basis size in the matrix $\mathbf{B}$, but taking the inverse will mix these signals again. To further simplify the relationship of the inverses, we need a more specific expression for the covariance. A simple way to extract the partial correlation that is only due to the closure operation is to consider a diagonal basis covariance, as it has no partial correlation by definition. We know from section 2.4 that linear independence corresponds to a logratio uncorrelated (LU) covariance in the embedded logratio sample space (where one dimension is lost). In other words, the closure applied to an uncorrelated basis leads to the dependence structure of LU compositions[2].

The elements of $\mathbf{\Sigma}$ for an LU composition are given by (10) and (9). The fact that $\mathbf{\Sigma}$ has nonzero off-diagonal elements for LU compositions immediately shows that the partial correlations of LU compositions do not vanish. It is not difficult to determine the precise expression. For $\boldsymbol{u}$ an LU composition, its (logratio-based) partial correlations are given by

$$r_{ij}(\boldsymbol{u}) = \sqrt{\frac{\alpha_i^{-1}\alpha_j^{-1}}{\left(\sum_{k \neq i} \alpha_k^{-1}\right)\left(\sum_{k \neq j} \alpha_k^{-1}\right)}}, \tag{22}$$

for all $i \neq j$, and where $\boldsymbol{\alpha}$ is given by (9) (see the Appendix for a proof).

To get an idea how strong this "spurious" signal is, we make the following experiment. To obtain an (artificial) system of linearly independent variables, we take the diagonal of the basis covariance matrix of the 770 genes considered in the previous section. This diagonal corresponds to $\boldsymbol{\alpha}$, and we can calculate the logratio-based partial correlations between all pairs using (22). For a worst-case scenario, we select the pair of parts for which we obtain the strongest partial correlation. With this pair fixed, we repeat calculating its partial correlation after eliminating an arbitrary part (and thus one component from the vector $\boldsymbol{\alpha}$). We repeat this process until we have only one additional part

---

[2]Dirichlet compositions are LU [Aitchison 1986], and this (often undesired) property gave rise to Aitchison's introduction of the logistic normal. The richer covariance structure of the latter enables modelling that goes beyond unit-sum induced interactions. Of course, a logistic normal with an LU covariance matrix also models LU compositions.

left with our original pair. The result of this process is shown in Figure 5 (going from right to left along the horizontal axis). We see a similar "dilution" of the compositional effect as for the negative correlations between multinomial counts, see Greenacre et al. (2022). In our case, the spurious correlation turns out to be positive, but it is evaluated between residuals of logratios, not compositional parts.

6. Conclusion

We have proposed a simple way of obtaining shrinkage-based estimators for logratio covariance matrices and their inverse. As shown in our benchmark, these estimators clearly improve on the (still widely used) empirical estimators and also outperform estimators that shrink logratio covariance matrices towards a diagonal target as proposed in Badri et al. (2020). The use of Aitchison's notion of logratio uncorrelated compositions is crucial to obtain the improved estimates.

With our contribution we also aim to promote the use of partial correlations for the analysis of compositional data sets. We see three advantages in using them: 1) They take all parts into account for the evaluation of pairwise interaction (and thus can factor out indirect interactions). 2) Their evaluation is reference-independent. 3) Negative correlations can be meaningfully evaluated with respect to the set of variables that are controlled for. While partial correlations change under taking subcompositions, this is expected because it reflects the effect of the removal of variables that are partialled out. However, we have also shown that a "spurious" positive signal remains that is due to the loss of the D-th dimension in compositional data sets. We have derived the expression for this remaining partial correlation induced by the closure operation and have shown how it "dilutes out" with a growing number of parts. In this sense, partial correlations between parts are expected to show little spurious signal whenever there are sufficiently many parts, a common occurrence in genomic data.

For genomic data sets, we have also drawn some connections with the problem of their normalisation. Sequencing data are relative counts, and as such their analysis within the scale-free logratio framework can be beneficial. The use of various normalisation schemes to control for sequencing depth can essentially be circumvented in this way. We see much promise in the use of logratio-based partial correlations especially for situations when normalisation assumptions fail.

7. Appendix

7.1. Bayesian interpretation of covariance shrinkage

In the Bayesian view of statistical inference, the parameters of a distribution are considered to follow distributions themselves. The posterior distribution of the parameter in question incorporates information from both the data (via the likelihood) and the prior distribution of the parameter. Shrinkage estimators also can incorporate prior informa-

tion, and while a suitable empirical estimator can maximise likelihood, optimization of shrinkage is related to maximizing the posterior of the parameter[3]. In the following we want to make this more precise for Gaussian covariance shrinkage.

Let us start with the Gaussian conjugate prior distribution of the precision matrix $\boldsymbol{\Sigma}^{-1}$ in the case of a fixed mean $\boldsymbol{\mu}$. This is known as the Wishart (or multivariate Gamma) distribution with density

$$f_{\mathcal{W}}(\boldsymbol{\Sigma}^{-1}|\boldsymbol{V},v) = \frac{|\boldsymbol{\Sigma}^{-1}|^{(v-D)/2}e^{-\frac{1}{2}\mathrm{tr}\left(V^{-1}\boldsymbol{\Sigma}^{-1}\right)}}{2^{v(D-1)/2}|\boldsymbol{V}|^{v/2}\Gamma_{D-1}(\frac{v}{2})}, \tag{23}$$

where the number of degrees of freedom v is a positive integer, the parametric matrix $\boldsymbol{V}$ is positive definite of order $(D-1) \times (D-1)$, and Γ denotes the multivariate Gamma function. The particular form of this density can be understood better when decomposing the precision matrix as

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{U}^{T}, \tag{24}$$

where $\boldsymbol{U}$ is of order $(D-1) \times v$. If the v vectors in $\boldsymbol{U}$ are independently drawn from a normal distribution with mean zero and covariance $\boldsymbol{V}$ (where $v \geq D-1$), then the precision matrix follows a Wishart distribution with density (23).

To estimate a Gaussian covariance matrix, we need more than a single sample of data. Let us now consider a data matrix $\boldsymbol{X}$ where the N rows were sampled according to (3). The joint likelihood of these samples can be written in terms of their sample parameter estimates $\boldsymbol{S}$ and $\bar{\boldsymbol{x}}$ as

$$f_{\mathcal{N}}(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{N}f_{\mathcal{N}}(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) =$$

$$\left(\frac{(2\pi)^{D-1}}{|\boldsymbol{\Sigma}|}\right)^{\frac{N}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left((N-1)\boldsymbol{S}+N(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^{T}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\right)\right]\right\}. \tag{25}$$

Let us now assign the conjugate priors as follows:

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}^{-1} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{1}{\kappa}\boldsymbol{\Sigma}\right), \tag{26}$$

$$\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\boldsymbol{V}, v). \tag{27}$$

The joint posterior density then factorizes as $P(\boldsymbol{\mu}|\boldsymbol{\Sigma}^{-1})P(\boldsymbol{\Sigma}^{-1})$, where the marginal $P(\boldsymbol{\Sigma}^{-1})$ is again Wishart. More precisely

$$P(\boldsymbol{\Sigma}^{-1}|\boldsymbol{S},\bar{\boldsymbol{x}},\boldsymbol{\mu}_0,\kappa,\boldsymbol{V},v) = f_{\mathcal{W}}(\boldsymbol{\Sigma}^{-1}|\boldsymbol{V}^*,v+N), \tag{28}$$

where

$$\boldsymbol{V}^* = (N-1)\boldsymbol{S}+\boldsymbol{V}+\frac{\kappa N}{\kappa+N}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^{T}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0) \tag{29}$$

---

[3]As an example, in the case of shrinkage estimates of multinomial frequencies, the target frequencies are renormalized pseudocounts to the count data, and the shrinkage intensity λ is the prior sample size of the conjugate Dirichlet prior that has the pseudocounts as its parameters (Hausser and Strimmer, 2009).

(see DeGroot (2004), p.178)[4]. It follows that under the Bayesian model, an improved covariance estimator will involve an additive correction to the sample covariance. We can now write (29) in terms of the shrinkage estimator when taking into account the posterior sample size:

$$(\nu + N)\hat{\boldsymbol{\Sigma}} = \boldsymbol{V}^*. \tag{30}$$

Identifying (30) with (5), we obtain the identities

$$(1 - \lambda) = (N - 1)/(\nu + N), \tag{31}$$

$$\lambda \boldsymbol{T} = \frac{1}{\nu + N} \left(\boldsymbol{V} + \frac{\kappa N}{\kappa + N} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \right), \tag{32}$$

from which it follows that

$$\lambda = \frac{\nu + 1}{\nu + N}, \tag{33}$$

$$\boldsymbol{T} = \frac{1}{\nu + 1} \left(\boldsymbol{V} + \frac{\kappa N}{\kappa + N} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \right). \tag{34}$$

Optimization of λ thus boils down to a tuning of the prior degrees of freedom ν relative to the sample size N, while the target $\boldsymbol{T}$ defines a prior covariance structure.

7.2. Proof of (10) for LU compositions

We start with the case $i \neq j$ (see [Aitchison 1986], section 5.9):

$$\sigma_{ij} = \mathrm{cov}\left(\log \frac{p_i}{p_D}, \log \frac{p_j}{p_D} \right) = \mathrm{cov}\left(\log \frac{p_i}{p_k} + \log \frac{p_k}{p_D}, \log \frac{p_j}{p_D} \right) = 0 + \sigma_{kj} \tag{35}$$

because of (8). It is therefore clear that σ_{ij} is constant for unequal indices and the definition (9) makes sense for LU compositions. So $\sigma_{ij} = \alpha_D$ if $i \neq j$. Similarly, in the case $i = j$ we have

$$\sigma_{ii} = \mathrm{var}\left(\log \frac{p_i}{p_D} \right)$$
$$= \mathrm{cov}\left(\log \frac{p_i}{p_D}, \log \frac{p_i}{p_D} \right) = \mathrm{cov}\left(\log \frac{p_i}{p_k} + \log \frac{p_k}{p_D}, \log \frac{p_i}{p_D} \right)$$
$$= \mathrm{cov}\left(\log \frac{p_k}{p_i}, \log \frac{p_D}{p_i} \right) + \mathrm{cov}\left(\log \frac{p_k}{p_D}, \log \frac{p_i}{p_D} \right) = \alpha_i + \alpha_D, \tag{36}$$

which concludes the proof.

---

[4]For the sake of completeness, it should perhaps be mentioned that $P(\boldsymbol{\mu}|\boldsymbol{\Sigma}^{-1}) = f_{\mathrm{N}}\left((\kappa \boldsymbol{\mu}_0 + N\bar{\boldsymbol{x}})/(\kappa + N), \boldsymbol{\Sigma}/(\kappa + N)\right)$.

7.3. Proof of (12) for an uncorrelated basis

Let the composition $\boldsymbol{u}$ have a basis $t\boldsymbol{u}$ with $\mathrm{cov}(\log(tu_i), \log(tu_j)) = 0$ for all $i \neq j$. We show that it follows that

$$\mathrm{var}\left(\log(tu_k)\right) = \mathrm{cov}\left(\log \frac{u_i}{u_k}, \log \frac{u_j}{u_k}\right) \tag{37}$$

where none of the indices i, j, k are equal. We write the left-hand side as

$$\mathrm{cov}\left(\log(tu_k), \log(tu_k)\right) = \mathrm{cov}\left(\log \frac{tu_k}{tu_i} + \log(tu_i), \log \frac{tu_k}{tu_j} + \log(tu_j)\right)$$

$$= \mathrm{cov}\left(\log \frac{tu_k}{tu_i}, \log \frac{tu_k}{tu_j}\right) + \mathrm{cov}\left(\log \frac{tu_k}{tu_i}, \log(tu_j)\right)$$

$$+ \mathrm{cov}\left(\log(tu_i), \log \frac{tu_k}{tu_j}\right) + \mathrm{cov}\left(\log(tu_i), \log(tu_j)\right)$$

$$= \mathrm{cov}\left(\log \frac{u_i}{u_k}, \log \frac{u_j}{u_k}\right) + \mathrm{cov}\left(\log(tu_k) - \log(tu_i), \log(tu_j)\right) + \ldots \tag{38}$$

The terms after the first one are all zero because of the condition that $\mathrm{cov}(\log(tu_i), \log(tu_j)) = 0$ for any indices $i \neq j$, which concludes the proof.

7.4. Proof of (22) for LU compositions

Let $\boldsymbol{\Sigma}$ be given by the elements (10). The following are expressions for its determinant (here given without proof):

$$|\boldsymbol{\Sigma}| = \prod_{i=1}^{D} \alpha_i \sum_{i=1}^{D} \frac{1}{\alpha_i} = \sum_{i=1}^{D} \prod_{k \neq i} \alpha_k, \tag{39}$$

where the second equality is found by evaluating the greatest common denominator of the α_i^{-1}. We show that the matrix inverse of $\boldsymbol{\Sigma}$ is given by the elements

$$\sigma_{ij}^{(-1)} = \frac{1}{\sum_{k=1}^{D} \prod_{l \neq k} \alpha_l} \begin{cases} \sum_{k \neq i} \prod_{l \neq k, i} \alpha_l & \text{if } i = j, \\ -\prod_{k \neq i, j} \alpha_k & \text{if } i \neq j. \end{cases} \quad i, j = 1, \ldots, D-1. \tag{40}$$

That this is the inverse can be easily proven when multiplying the resulting matrix with the matrix whose elements are given by (10). The off-diagonal elements (where $i \neq j$) of the result are

$$\sum_{m=1}^{D-1} \sigma_{im} \sigma_{mj}^{(-1)} = \sum_{m \neq i, j, D} \sigma_{im} \sigma_{mj}^{(-1)} + \sigma_{ii} \sigma_{ij}^{(-1)} + \sigma_{ij} \sigma_{jj}^{(-1)} =$$

$$\frac{1}{\sum_{k=1}^{D} \prod_{l \neq k} \alpha_l} \left[-\alpha_D \sum_{m \neq i, j, D} \prod_{k \neq m, j} \alpha_k - (\alpha_i + \alpha_D) \prod_{k \neq i, j} \alpha_k + \alpha_D \sum_{k \neq j} \prod_{l \neq k, j} \alpha_l \right]$$

The square bracket evaluates to

$$- \alpha_D \sum_{m \neq j, D} \prod_{k \neq m, j} \alpha_k - \alpha_i \prod_{k \neq i, j} \alpha_k + \alpha_D \sum_{k \neq j} \prod_{l \neq k, j} \alpha_l$$

$$= -\alpha_D \sum_{m \neq j, D} \prod_{k \neq m, j} \alpha_k - \alpha_i \prod_{k \neq i, j} \alpha_k + \alpha_D \sum_{k \neq j, D} \prod_{l \neq k, j} \alpha_l + \alpha_D \prod_{l \neq D, j} \alpha_l = 0. \quad (41)$$

Thus the off-diagonal elements vanish. For the diagonal elements we have

$$\sum_{m=1}^{D-1} \sigma_{im} \sigma_{mi}^{(-1)} = \sum_{m \neq i, D} \sigma_{im} \sigma_{mi}^{(-1)} + \sigma_{ii} \sigma_{ii}^{(-1)} =$$

$$\frac{1}{\sum_{k=1}^{D} \prod_{l \neq k} \alpha_l} \left[-\alpha_D \sum_{m \neq i, D} \prod_{k \neq m, i} \alpha_k + (\alpha_i + \alpha_D) \sum_{k \neq i} \prod_{l \neq k, i} \alpha_l \right]. \quad (42)$$

The square bracket evaluates to

$$- \alpha_D \sum_{m \neq i, D} \prod_{k \neq m, i} \alpha_k + \alpha_i \sum_{k \neq i} \prod_{l \neq k, i} \alpha_l + \alpha_D \sum_{k \neq i, D} \prod_{l \neq k, i} \alpha_l + \alpha_D \prod_{l \neq D, i} \alpha_l$$

$$= \alpha_i \sum_{k \neq i} \prod_{l \neq k, i} \alpha_l + \alpha_D \prod_{l \neq D, i} \alpha_l = \sum_{k \neq i} \prod_{l \neq k} \alpha_l + \prod_{l \neq i} \alpha_l = \sum_{k} \prod_{l \neq k} \alpha_k.$$

Inserting this back into (42) shows that the diagonal elements are 1, proving the inverse. We can now evaluate the partial correlation coefficient (7) for an LU composition $\boldsymbol{u}$ by inserting the expression for the inverse (40) and (to obtain the third equality below) making use of the second equality in (39):

$$r_{ij}(\boldsymbol{u}) = \frac{-\sigma_{ij}^{(-1)}}{\sqrt{\sigma_{ii}^{(-1)} \sigma_{jj}^{(-1)}}} = \frac{\prod_{k \neq i, j} \alpha_k}{\sqrt{\left(\sum_{k \neq i} \prod_{l \neq k, i} \alpha_l \right) \left(\sum_{k \neq j} \prod_{l \neq k, j} \alpha_l \right)}}$$

$$= \frac{\prod_{k \neq i, j} \alpha_k}{\sqrt{\left(\prod_{l \neq i} \alpha_l \sum_{k \neq i} \alpha_l^{-1} \right) \left(\prod_{l \neq j} \alpha_l \sum_{k \neq j} \alpha_l^{-1} \right)}}$$

$$= \frac{\prod_{k \neq i, j} \alpha_k}{\left(\prod_{k \neq i, j} \alpha_k \right) \sqrt{\left(\alpha_j \sum_{k \neq i} \alpha_l^{-1} \right) \left(\alpha_i \sum_{k \neq j} \alpha_l^{-1} \right)}}$$

$$= \sqrt{\frac{\alpha_i^{-1} \alpha_j^{-1}}{\left(\sum_{k \neq i} \alpha_k^{-1} \right) \left(\sum_{k \neq j} \alpha_k^{-1} \right)}}, \quad (43)$$

which is the expression given in (22).

References

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society, Series B*, 44(2), 139–160.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall.

Badri, M., Kurtz, Z.D., Bonneau, R. and M͗uller, C.L. (2020). Shrinkage improves estimation of microbial associations under different normalization methods, *NAR Genomics and Bioinformatics*, 2(4), lqaa100.

DeGroot, M.H. (2004). *Optimal Statistical Decisions*, Wiley.

Erb, I. (2020). Partial correlations in compositional data analysis *Applied Computing and Geosciences*, 6, 100026.

Erb, I. and Notredame, C. (2016). How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, 135(1-2), 21–36.

Erb, I., Gloor, G.B. and Quinn, T.P. (2020). Editorial: Compositional data analysis and related methods applied to genomics – a first special issue from NAR Genomics and Bioinformatics. *NAR Genomics and Bioinformatics*, 2(4) lqaa103.

Erb, I., Quinn, T.P., Lovell, D. and Notredame, C. (2017). Differential Proportionality – A Normalization-Free Approach To Differential Gene Expression. *Proceedings of CoDaWork 2017, The 7th Compositional Data Analysis Workshop*; available under bioRxiv, 134536.

Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I. and Quinn, T. (2022). Aitchison's Compositional Data Analysis 40 years On: A Reappraisal. arXiv:2201.05197

Hausser, J. and Strimmer, K. (2009). Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks *Journal of Machine Learning Research*, 10, 1469-1484.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In: *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, 1 361–379, Univ. California Press.

Kurtz, Z.D., M͗uller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J. and Bonneau, R.A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Comput Biol*, 11(5):e1004226.

Law, C.W., Chen, Y., Shi, W. and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15:R29.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10, 603–621.

Lovell, D.R., Chua, X.-Y. and McGrath, A. (2020). Counts: an outstanding challenge for log-ratio analysis of compositional data in the molecular biosciences, *NAR Genomics and Bioinformatics*, 2(2), lqaa040.

Lovell, D. Pawlowsky-Glahn, V. Egozcue, J.J,, Marguerat, S. and Bähler, J. (2015). Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol.*, Mar 16;11(3) :e1004075.

Opgen-Rhein, R. and Strimmer, K. (2007). Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Palarea-Albaladejo, J. and Martín-Fernńdez, J.A. (2015). zCompositions – R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143, 85–96.

Palarea-Albaladejo, J., Martín-Fernández, J., Ruiz-Gazen, A. and Thomas-Agnan, C. (2022). lrSVD: An efficient imputation algorithm for incomplete high-throughput compositional data. *Journal of Chemometrics*, e3459.

Quinn, T.P. and Erb, I. (2021). Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data. *NAR Genomics and Bioinformatics*, 2(4) lqaa076.

Quinn, T.P., Erb, I., Richardson, M.F. and Crowley, T.M. (2018). .Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16), 2870–2878.

Quinn, T.P., Richardson, M.F., Lovell, D. and Crowley, T.L. (2017). propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Scientific Reports*, 7(1) 16252.

Riba, A., Oravecz, A., Durik, M., Jiménez, S., Alunni, V., Cerciat, M., Jung, M., Keime, C., Keyes, W.M. and Molina, N. (2022). Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nat Commun*, 13, 2865.

Schäfer, J. and Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 32.

van Nimwegen, E. (2016). Inferring Contacting Residues within and between Proteins: What Do the Probabilities Mean? *PLoS Comput Biol*, 12(5), e1004726.

Weistuch, C., Zhu, J., Deasy, J.O. and Tannenbaum, A.R. (2022). The maximum entropy principle for compositional data. *BMC Bioinformatics*, 23:449.

Werhli, A.V., Grzegorczyk, M. and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks, *Bioinformatics*, 22(20), 2523–2531.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.