Published online 2 October 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 4 **1**
doi: 10.1093/nargab/lqaa076

Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data

Thomas P. Quinn [1,*] and Ionas Erb[2]

[1]Applied Artificial Intelligence Institute, Deakin University, 75 Pigdons Rd, WaurnPonds VIC 3216, Geelong, Australia and [2]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Carrer del Dr.Aiguader, 88, 08003, Barcelona, Spain

Received February 27, 2020; Revised July 23, 2020; Editorial Decision August 17, 2020; Accepted September 03, 2020

ABSTRACT

Many next-generation sequencing datasets contain only relative information because of biological and technical factors that limit the total number of transcripts observed for a given sample. It is not possible to interpret any one component in isolation. The field of compositional data analysis has emerged with alternative methods for relative data based on log-ratio transforms. However, these data often contain many more features than samples, and thus require creative new ways to reduce the dimensionality of the data. The summation of parts, called amalgamation, is a practical way of reducing dimensionality, but can introduce a non-linear distortion to the data. We exploit this non-linearity to propose a powerful yet interpretable dimension method called data-driven amalgamation. Our new method, implemented in the user-friendly R package amalgam, can reduce the dimensionality of compositional data by finding amalgamations that optimally (i) preserve the distance between samples, or (ii) classify samples as diseased or not. Our benchmark on 13 real datasets confirm that these amalgamations compete with state-of-the-art methods in terms of performance, but result in new features that are easily understood: they are groups of parts added together.

INTRODUCTION

Compositional data are a kind of relative data in which each part is only interpretable relative to the other parts (1,2). In the health sciences, many datasets produced by next-generation sequencing (NGS) have this property because of biological and technical factors that limit the total number of transcripts observed for a given sample (often called the 'constant-sum constraint') (3–9). As mutually dependent elements, it is not possible to interpret any component in isolation (at least without invoking the often untestable assumptions that underpin data normalization). The field of compositional data analysis (CoDA) offers an alternative way to analyze relative data by using log-ratio transforms. These transformations use one or more references to recast the data as log-contrasts (10). The log-contrasts can then be analyzed using routine statistical methods, but must get interpreted as a ratio of the numerator parts to the reference denominator parts. Example log-ratio transformations include the additive log-ratio (alr) (which uses a single component as the reference) (1), the centered log-ratio (clr) (which uses the per-sample geometric mean as the reference) (1), and the isometric log-ratio (ilr) (which uses an orthonormal basis to define a set of arbitrary log-contrasts) (11).

Compositional data exist in a simplex with one fewer dimensions than parts. The ilr offers a theoretically ideal solution because its log-contrasts move the data from the simplex into real Euclidean space (11). However, arbitrary log-contrasts lack interpretability. For example, how does an analyst make sense of the difference between the log of the product of two sets

*To whom correspondence should be addressed. Tel: +3 5227 1100; Email: contacttomquinn@gmail.com

of parts, where each part is raised to a unique power? Balances were proposed as a more interpretable log-contrast, where each balance is a log-contrast between two geometric means (12). An example 3-part balance is $\log(\sqrt{ab}) - \log(c)$. Indeed, Pawlowsky–Glahn *et al*. have shown how a set of 'principal balances' can explain an ever-decreasing portion of the variance in analogy to principal components (13) (though principal balances can be correlated). Although more complicated than a simple log-ratio, having more parts means that a single balance can describe more variance than a single log-ratio. Balances have recently become popular for the analysis and classification of microbiome compositions (14–18).

Recently, Greenacre *et al*. have challenged the interpretability of balances (19). We summarize the Greenacre *et al*. critique as follows: because the geometric mean depends on the ratios of the parts within, balances are not balances in the plain English sense of the word. Consider the balance between 'a and b' versus 'c' where $b > c$. We would expect that the 'balance' would lean toward the combined weight of 'a' and 'b'. However, with a geometric mean, the balance will tip more toward 'c' when 'a' is rare. This is because the ilr balances are defined in log space. Instead of balances, Greenacre *et al*. proposed the summed log-ratio (SLR) as a more interpretable alternative (20). An example 3-part SLR is $\log(a + b) - \log(c)$. The summation of parts is called amalgamation, and Greenacre *et al*. encourage using domain knowledge for amalgamation (i.e. *expert-driven amalgamation*) as a practical way of dealing with parts (21). However, Egozcue & Pawlowsky-Glahn have criticized SLRs because, while scale-invariant, they are 'non-linear functions in the Aitchison geometry of the simplex' and so inter-sample distances can have 'anomalous behavior' after amalgamation (10). In summary, Greenacre *et al*. argue that SLRs are an interpretable way to reduce the dimensionality of the data, while Egozcue & Pawlowsky-Glahn argue that SLRs introduce a non-linear distortion to the data. Yet, non-linearity might be advantageous for situations in which the data need to be summarized in a non-trivial way. In this case, SLRs could provide a valuable addition to the CoDA toolkit: an interpretable non-linear transform.

In this article, we propose *data-driven amalgamation* as a new method for reducing the dimensionality of compositional data. Unlike expert-driven amalgamation which uses domain knowledge, data-driven amalgamation uses an objective function. This objective function is user-defined for a given task, and combined with a search algorithm to answer the question, 'What is the best way to amalgamate the data to achieve the objective?'. We show that data-driven amalgamation can be used to find a new 3-part simplex that efficiently visualizes the data according to any user-defined objective. We benchmark data-driven amalgamation across 13 health biomarker datasets, for two separate objectives: (i) to preserve a suitable distance between samples (where we consider three different measures), and (ii) to classify samples as diseased or not. We show that the amalgamated features, which we call 'amalgams', can preserve inter-sample distances as well as principal components. Moreover, amalgams outperform principal components and principal balances as a feature reduction step before classification. We argue that amalgams are biologically meaningful concepts, and conclude the article by highlighting future areas of research.

MATERIALS AND METHODS

Motivation

Greenacre *et al*. showed that using a pairwise log-ratio selection method in the presence of SLRs does not necessarily distort inter-sample distances (20). However, their example has two limitations. First, none of the 'principal log-ratios' (i.e. the ones which explain the most variance) were SLRs. In other words, the SLRs happened to be the least important ratios. This raises the question, 'What happens when the important log-ratios are SLRs?'. Second, they only discuss expert-driven amalgamation. This raises another question, 'Is it possible to replace expert-driven amalgamation with data-driven amalgamation?'.

While we do discuss SLRs in this article, we will focus on amalgamation more generally. Our motivation is to answer two research questions:

Can we use a search heuristic to find an amalgamation that best preserves distance?

(i) Can we use a search heuristic to find an amalgamation that best preserves distance?
(ii) Can we use a search heuristic to find an amalgamation that maximizes the prediction of a dependent variable?

The first question is an unsupervised machine learning problem that seeks to find a reduced feature space (i.e. a latent space) that accurately projects the data in fewer dimensions. The second question is a supervised machine learning problem. In both cases, amalgamation adds value over traditional dimension reduction methods because it makes the lower dimension features highly interpretable.

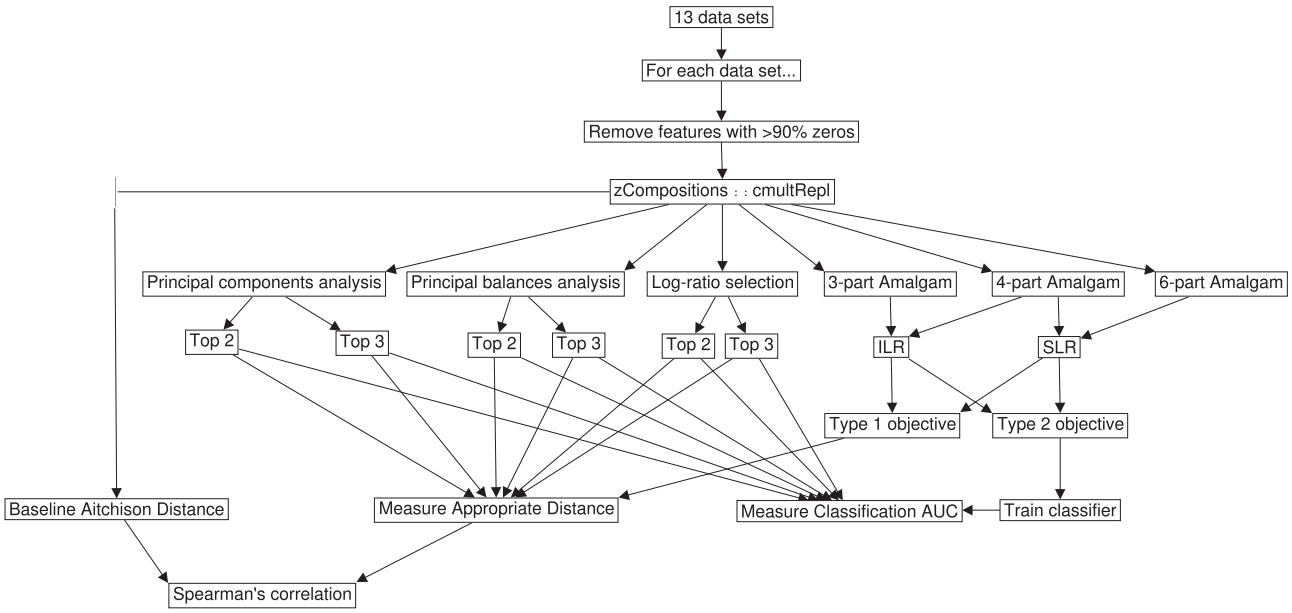

Figure 1. This figure presents an overview of the benchmark pipeline run for each dataset. After feature removal and zero replacement, each dataset underwent dimension reduction by PCA, PBA, log-ratio selection (PRA), or data-driven amalgamation. The data-driven amalgams were then analyzed directly, or first converted to SLRs. We used two criteria to benchmark the goodness of a dimension reduction method: (i) agreement between the baseline distances and the reduced-dimension distances and (ii) accuracy of the reduced-dimension classifier.

In this article, we benchmark data-driven amalgamation against the other dimension reduction methods used routinely for CoDA, including principal components analysis (PCA), principal balance analysis (PBA) (13) and pairwise log-ratio selection (PRA) (22). We define two tasks: (i) to obtain a compressed representation of inter-sample distances, and (ii) to perform a feature reduction for binary classification. Figure 1 presents a schematic overview of our benchmark procedure.

Data-driven amalgamation

The amalgamation matrix. An amalgamation is defined as the result of adding D components into $D' \leq D$ mutually exclusive subsets (1). A compositional dataset $\mathbf{X}$ describing N compositions and D components can be amalgamated into a set of D'-part compositions via an *amalgamation matrix* $\mathbf{A}$ with D rows and D' columns:

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{A}$$

where $A_{ij} \in \{0, 1\}$ and $\sum_j^{D'} A_{ij} \leq 1$ for all $i \in \{1, 2, \ldots, D\}$. In other words, the amalgamation matrix is binary and all rows (representing components) sum to 1. The intuition here is that each component D (as rows) either does or does not contribute to an amalgam D' (as columns), and that each component contributes to one amalgam at most.

The amalgamation has D' new components which we call amalgams. The *amalgamation matrix* describes to which amalgam (as a column) the original component (as a row) belongs. Since one component should never contribute to more than one amalgam, the row sums of $\mathbf{A}$ is limited to 1 (when any $\sum_j^{D'} A_{ij} = 0$, the amalgamation is also a sub-composition). Meanwhile, the column sums of $\mathbf{A}$ indicates how many components a single amalgam represents. Note that we use the term 'amalgamation' to refer both to the amalgamation of the complete composition and to the amalgamation of a sub-composition. Figure 2 shows an example of amalgamation, and illustrates how one could conceptualize amalgamation as a feed-forward network.

The objective functions. Data-driven amalgamation seeks to find the best amalgamation matrix $\mathbf{A}$ for a given dataset $\mathbf{X}$:

$$\mathbf{A}_a = f_a(\mathbf{X}) \tag{1}$$

where f_a is chosen to optimize an arbitrary objective denoted by a. Here, we consider two kinds of objectives: unsupervised (Type 1) and supervised (Type 2) objectives.

A Input Matrix

	var1	var2	var3	var4	var5
sample1	0.195	0.176	0.201	0.208	0.22
sample2	0.219	0.201	0.178	0.191	0.211
sample3	0.212	0.203	0.176	0.206	0.203
sample4	0.202	0.204	0.189	0.21	0.195

B Amalgamation Matrix

	amalg1	amalg2	amalg3
var1	0	1	0
var2	0	1	0
var3	1	0	0
var4	1	0	0
var5	0	0	1

C Amalgams Matrix

	amalg1	amalg2	amalg3
sample1	0.409	0.371	0.22
sample2	0.369	0.42	0.211
sample3	0.383	0.415	0.203
sample4	0.399	0.407	0.195

D Amalgamation as a Network

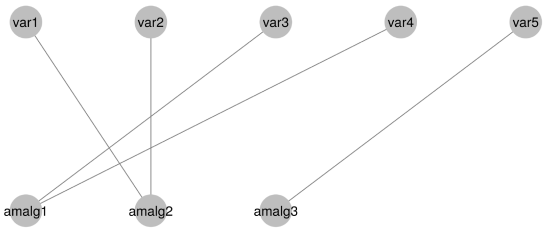

Figure 2. This figure shows an example amalgamation procedure, from the input compositions (Panel **A**) and the amalgamation matrix (Panel **B**) to the resultant amalgams (Panel **C**). The purpose of data-driven amalgamation is find the best *amalgamation matrix* for a given task. One could conceptualize amalgamation as a type of feed-forward network where each component has only one outgoing connection (Panel **D**). In this study, we search for an amalgamation matrix that maximizes an objective function; since this is equivalent to minimizing a loss, one could further conceptualize amalgams as a hidden layer in a (linearly activated) neural network. Although this suggests that we could use gradient descent, we choose to minimize the loss with a genetic algorithm because the amalgamation matrix is binary.

In all cases, the amalgamated data are first transformed using either a centered log-ratio transformation, isometric log-ratio transformation, or SLR transformation.

Type 1 objectives. Our Type 1 objectives seek to preserve the distance d between samples. This is an 'unsupervised' objective that is designed for visualization tasks. We can express this objective in terms of maximizing the Pearson's correlation ρ between the vectors of the original distances and the amalgamated distances:

$$\mathbf{A}_d = \underset{\mathbf{A}}{\operatorname{argmax}}\, \rho\left(d(\mathbf{X}), d(\mathbf{X}\cdot\mathbf{A})\right), \tag{2}$$

We consider three distances. First, we consider the log-ratio, or Aitchison, distance, which is the Euclidean distance obtained from clr-transformed data. It has a number of advantages, including scale invariance and sub-compositional dominance, that make it a preferred distance for compositional data. It can be defined as:

$$d_{\mathrm{A}}^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{D} \left(\mathrm{clr}_j(\boldsymbol{x}) - \mathrm{clr}_j(\boldsymbol{y})\right)^2. \tag{3}$$

(Here we denoted the j-th component of the clr-transformed data by clr_j.)

However, this may not be the most appropriate measure when considering amalgamations. Instead, we may want our distance measure to observe a natural continuity property: when we amalgamate parts that are identical in two samples, the distance between these samples should be unaffected by the merging of the parts. For compositional data, the notion of identity between parts across samples can be thought of loosely as a *proportionality* of the parts. Proportional parts have a vanishing log-ratio variance (i.e. they behave in an entirely coordinated way) (5). Such parts are also known as 'distributionally equivalent' (23).

Second, we consider the weighted Aitchison distance which, unlike its unweighted form, has distributional equivalence (23), meaning that it is unaffected by the merging of proportional parts. It can be defined as:

$$d^2_{\mathrm{WA}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{D} \omega_j \left(\mathrm{clr}_j(\boldsymbol{x}) - \mathrm{clr}_j(\boldsymbol{y})\right)^2. \tag{4}$$

The exact form of the weights ω_j is not important, and we will use the simplest possibility by weighting each part by the total sum of all counts for that feature (i.e. the column sum). Note that, when merging parts, their weights will be added too.

Third, we consider a distance based on information-theoretic considerations. A composition is formally equivalent to a vector of discrete probabilities, and thus we could apply the Shannon index H. Advantageously, this measure is also scale-invariant when compositions are normalized to 1. Amalgamations over indices $j \in \mathcal{A}$ can then be considered a *coarse-graining* (24,25) of the parts. It is well known (25,26) that Shannon entropy can be expressed as the sum of the entropy of the parts resulting after coarse-graining and of the coarse-grained parts themselves, where the latter are renormalized and weighted by their sum:

$$H(\boldsymbol{x}) = H\left(\boldsymbol{x}_{\{1,\ldots,D\}\setminus\mathcal{A}}, \sum_{j\in\mathcal{A}} x_j\right) + H\left(\frac{\boldsymbol{x}_{\mathcal{A}}}{\sum_{j\in\mathcal{A}} x_j}\right) \sum_{i\in\mathcal{A}} x_j. \tag{5}$$

It is easy to show that the relative entropy:

$$D(\boldsymbol{x}||\boldsymbol{y}) = \sum_{j=1}^{D} x_j \log \frac{x_j}{y_j} \tag{6}$$

also remains invariant when merging distributionally equivalent parts (see Supplementary Data). For simplicity, we consider the symmetrized version:

$$d^2_{\mathrm{SR}} = \frac{D(\boldsymbol{x}||\boldsymbol{y}) + D(\boldsymbol{y}||\boldsymbol{x})}{2}. \tag{7}$$

It is well known that the maximum-likelihood estimator of entropy is negatively biased for under-sampled data, e.g. (27). To obtain a better empirical estimate of relative entropy from genomic data matrices, one could use the James-Stein type shrinkage estimator implemented in the R package entropy (28). Since it allows for an estimate of the frequencies themselves, the shrinkage estimator can be used in conjunction with the other distance measures too. This approach is potentially advantageous because it naturally imputes the zeros that present a major problem for log-ratio analysis, though more research is needed to validate amalgamation for zero-laden data.

Type 2 objectives. Our Type 2 objectives seek to maximize the percent of variance within the amalgamated data that is explained by a constraining matrix $\mathbf{L}$. This is a 'supervised' objective that is designed for prediction tasks. We can express this objective in terms of maximizing the relative size of the constrained eigenvalues of a (discriminant or) redundancy analysis (RDA) of ilr-transformed data:

$$\mathbf{A}_c = \underset{\mathbf{A}}{\mathrm{argmax}}\, \mathrm{RDA}\left(\mathrm{ilr}(\mathbf{X}\cdot\mathbf{A}) \sim \mathbf{L}\right) \tag{8}$$

One can think of RDA as a multivariable extension of a simple linear regression (29). The way dependence on variables of the external dataset $\mathbf{L}$ is evaluated is easiest understood when these variables are discrete (e.g. they are experimental groups). In this case, RDA is equivalent to a discriminant analysis. RDA finds a linear rotation of a dataset $\mathbf{X}$ such that the new coordinates partition the total variance into two fractions: (i) the *redundant axes* which can be explained by another dataset $\mathbf{L}$; and (ii) the *principal axes* which cannot be explained by another dataset $\mathbf{L}$ (these are equivalent to the ones from a principle component analysis). The fraction of the variance from (i) over the total variance provides an estimate of the 'goodness-of-fit' for $\mathbf{X} \sim \mathbf{L}$ in a regression sense, and is the value maximized by Type 2 objectives. Note that the motivation behind using an RDA instead of an ordinary regression is that this implementation can easily generalize to multivariable problems.

Note that the ilr transformation implies an unweighted Aitchison distance for the RDA. Instead, any distance measure can be used via a square matrix of pairwise distances. These can be subjected to a classical multidimensional scaling to obtain principal coordinates that can be fed to the RDA.

The summed log-ratio variant. We can create a set of SLRs from an amalgamation by a function f_{slr} that takes the log of the first column of $\mathbf{Y}$ divided by the second column, the log of the third column divided by the fourth column and so on:

$$\mathbf{S} = f_{\text{slr}}(\mathbf{Y}) = f_{\text{slr}}(\mathbf{X} \cdot \mathbf{A})$$

where the SLR matrix $\mathbf{S}$ contains $D'/2$ log-ratios.

Since SLRs have already moved the data out of the simplex, the Aitchison distance d_{A} (or its weighted version) is replaced with the Euclidean distance d_E in the Type 1 objectives:

$$\mathbf{A}_a = \underset{\mathbf{A}}{\text{argmax}}\, \rho(d_A(\mathbf{X}), d_E(\mathbf{S})) \tag{9}$$

and the ilr-transformation is not performed in the Type 2 objectives:

$$\mathbf{A}_c = \underset{\mathbf{A}}{\text{argmax}}\, \text{RDA}(\mathbf{S} \sim \mathbf{L}). \tag{10}$$

We do not use the SLRs when evaluating relative entropies.

The genetic algorithm. Since the *amalgamation matrix* is a binary matrix whose rows sum to 0 or 1, its parameters can be solved by a genetic algorithm with only $D * \text{ceiling}(\log_2(D^{'} + 1))$ bits. To implement the genetic algorithm, we used the GA package (30) with default hyper-parameters (except for the number of bits, the fitness function and the maximum number of iterations).

Note that the genetic algorithm is not required to include all components in its optimal solution. It is possible for the algorithm to learn an amalgamation that is also a sub-composition if this improves the 'fitness' of a solution (i.e. if it minimizes the loss). In practice, this happens regularly; however, the resultant amalgams still contain many components. One could force sparsity within the learned amalgamation matrix by adding a regularization penalty that decreases the 'fitness' of a solution if it contains too many components. For example, a regularized Type 1 objective might look like:

$$\mathbf{A}_d = \underset{\mathbf{A}}{\text{argmax}}\, \rho\left(d(\mathbf{X}), d(\mathbf{X} \cdot \mathbf{A})\right) - \lambda \sum_{i}^{D} \sum_{j}^{D'} A_{ij}, \tag{11}$$

where λ is a hyper-parameter that controls the sparsity of the amalgamation matrix. Higher values of λ would result in amalgams that contain fewer parts.

Implementation. Here we present the amalgam package for the R programming language which solves the aforementioned objectives as a reproducible and easy-to-use software tool. Below, we show an example of its use for mock data.

```
# install from GitHub

devtools::install_github(''tpq/amalgam'')

# Load package and sample data

library(amalgam)

data(iris)

# find best amalgamation

A <- amalgam(x = iris[,1:4],

             n.amalgams = 3, # how many amalgams to return

             maxiter = 50, # how long to run genetic algorithm
```

Table 1. This table describes the datasets used to benchmark dimension reduction

Study Code	Source	Type	Features	Group 1	Size	Group 2	Size
1a	selbal	16s	48	CD	662	HC	313
1b	selbal	16s	60	MSM	73	Non-MSM	55
2a	Franzosa *et al.*	Shotgun	153	IBD	164	HC	56
2b	Franzosa *et al.*	Shotgun	158	CD	88	UC	76
2c	Franzosa *et al.*	Metabolites	885	IBD	164	HC	56
2d	Franzosa *et al.*	Metabolites	885	CD	88	UC	76
3a	MicrobiomeHD	16s	278	C. diff	93	Diarrhea	89
3b	MicrobiomeHD	16s	610	C. diff	93	HC	154
3c	MicrobiomeHD	16s	1133	CRC	120	HC	172
3d	MicrobiomeHD	16s	1302	CRC	120	Adenoma	198
4a	TCGA	microRNA	188	Tumor	1078	Non-Tumor	104
4b	TCGA	microRNA	188	Her2	77	Non-Her2	927
4c	TCGA	microRNA	188	LumA	524	LumB	194

CD: Crohn's disease; HC: Healthy control; MSM: Men who have sex with men; UC: Ulcerative colitis; IBD: Inflammatory bowel disease; CRC: Colorectal cancer. This table is reproduced from (18).

```
                objective = objective.keepDist, # if preserving distance

                # objective = objective.keepWADIST # another distance

                # objective = objective.keepSKL # another distance

                # objective = objective.maxRDA, # if maximizing RDA

                z = iris[,5], # only needed if maximizing RDA

                asSLR = FALSE, # if TRUE, n.amalgams must be even

                shrink = FALSE) # toggles James-Stein type shrinkage

# visualize results

plot(A, col = iris[,5])
```

The x argument defines the input data, the n.amalgams argument sets the number of amalgams, the maxiter argument sets the number of genetic algorithm iterations, the objective argument defines the objective, the z argument defines the constraining matrix, the asSLR argument toggles whether to convert the amalgams into SLRs, and the shrink argument toggles whether to use James–Stein type shrinkage. This package depends on the GA (30), compositions (31) and vegan (32) packages.

Benchmark evaluation

Competing compositional methods. We benchmark amalgamation and SLRs against competing dimension reduction methods designed for compositional data. This includes (i) a PCA of clr-transformed data (33) [implemented in compositions (31)], (ii) a PBA (using the log-ratio variance clustering heuristic) (13) [implemented in balance (34)] and (iii) the pairwise log-ratio selection method proposed by (22) (PRA) [implemented in propr (35)]. For each dimension reduction technique, we consider the best two dimensions and the best three dimensions separately. Note that using three amalgams only occupies two dimensions because of the simplex.

Data acquisition. We use the same 13 health biomarker datasets previously used to benchmark binary classification pipelines for compositional data (18). These datasets were acquired from multiple sources (17,36–43) and span several difficult-to-study NGS data types (including 16s, metagenomics, metabolomics and microRNA). The number of samples, number of features, and outcomes-of-interest are described in Table 1. To facilitate between-study comparisons, these data underwent the same pre-processing steps as in (18): for all datasets, we removed features that had more than 90% zeros; for the metabolomic and microRNA datasets, we only included features in the top decile of total abundance. The data are available already pre-processed for immediate use from https://zenodo.org/record/3378099.

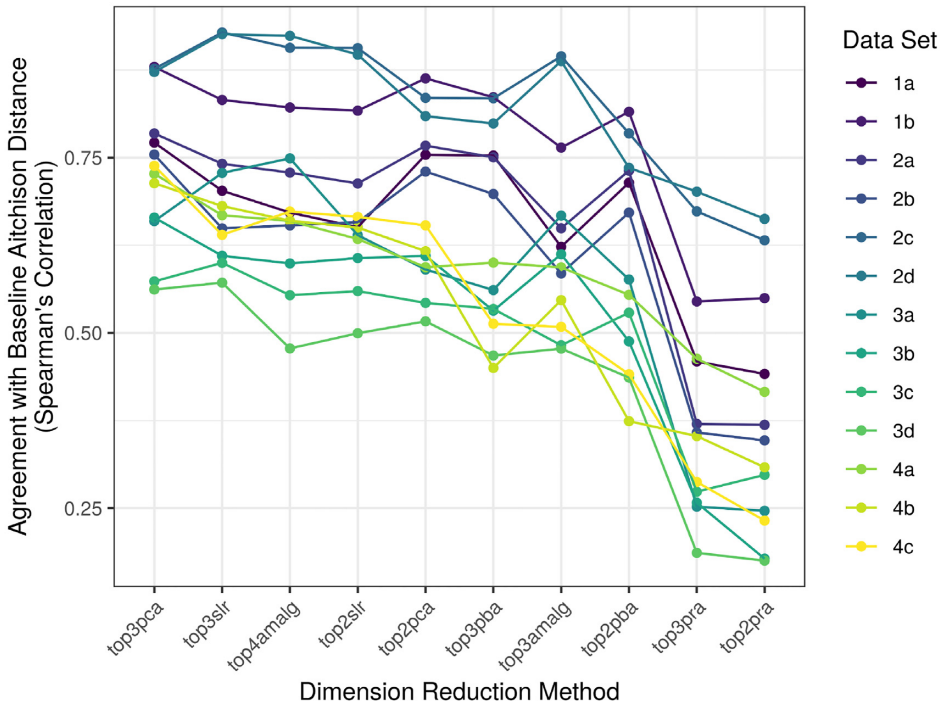

Figure 3. This figure shows the agreement between the baseline Aitchison distance and the distance computed on the dimension-reduced data (y-axis) for each method (x-axis), as grouped by the dataset studied (color). Methods toward the left of the x-axis agree more with the baseline on average. Here, we see that using four amalgams or three SLRs can actually preserve the inter-sample distances quite well. A statistical analysis of the differences is presented in Supplementary Table S1.

Zero replacement. All competing log-ratio methods fail in the presence of zeros. To address this impediment, we first replace zeros using the cmultRepl function from the zCompositions package (44). Although we do not necessarily need zero replacement when using amalgamation or SLRs, we use the zero-replaced data to make a fair comparison.

Amalgams for preserving distance. The 'gold standard' distance for compositions is the Aitchison distance (45,46). As such, we can evaluate the quality of a dimension reduction method based on how well the dimension-reduced distances agree with the true Aitchison distances (20). For each dataset, we measure this agreement as the Spearman's correlation between the inter-sample distances from the dimension-reduced data and those from the full data. Note that when the reduced dimensions are already in log space, we compute a Euclidean distance instead.

We also compare how well each of our three dimension-reduced distances agrees with the corresponding baseline distances after 100 iterations, with and without James–Stein type shrinkage. This agreement is measured as Pearson's correlation (as defined in the Type 1 objectives).

Amalgams for classification. In a supervised setting, a good dimension reduction method should help classify a withheld test data. As such, we can evaluate the goodness of a dimension reduction method in terms of classification accuracy. For each dataset, we measure classification accuracy by cross-validation. Model training occurs in three steps: (i) we use data-driven amalgamation to reduce the dimensionality of the training set; (ii) we CLR- or SLR-transform the resultant amalgams; and (iii) we fit a logistic regression classifier. During model deployment, the test set has its dimensions reduced *according to the training set rule*, thus ensuring test set independence. We repeat this procedure on 20 separate 67–33% training-test set splits, and report the 'out-of-the-box' performance without any hyper-parameter tuning because of the small sample sizes. The workflow is arranged using the exprso package (47).

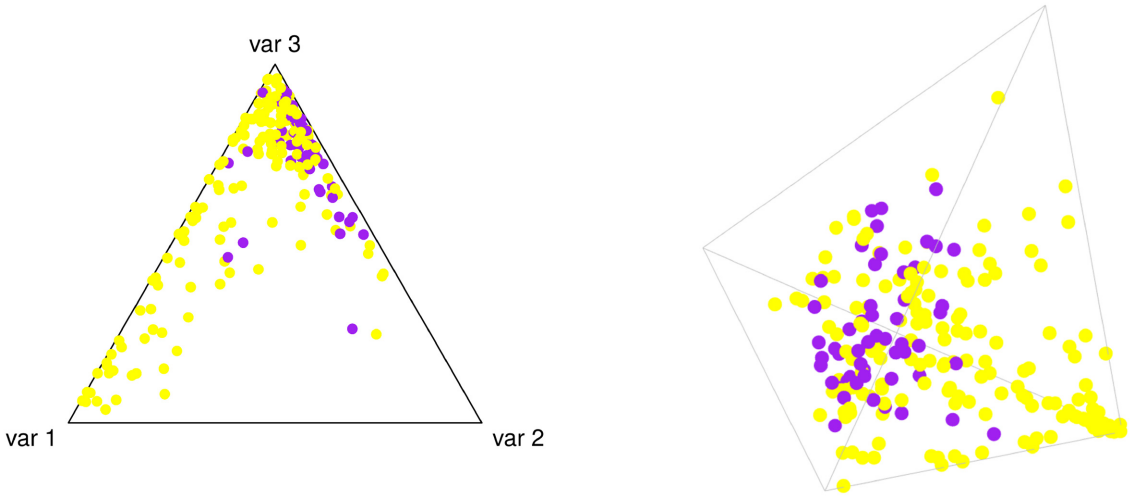

Figure 4. This figure projects the Franzosa *et al*. microbiome data across three amalgams (left panel) and four amalgams (right panel), chosen to preserve inter-sample distances. Light yellow dots show samples with inflammatory bowel disease, while dark purple dots show healthy samples. From Figure 3, we know that the distances in these figures are as coherent as a PCA plot of equal dimension. Yet, the variables of the simplex (i.e. the corners of the triangle) are easily understood, and allow the analyst to visualize the data in the same space that they exist: a simplex.

RESULTS AND DISCUSSION

Amalgams can preserve Aitchison distances

For NGS health biomarker data, each clinical sample is a composition. We can calculate inter-sample distances using the Aitchison distance. One critique against the use of amalgamation for dimension reduction is that it distorts inter-sample distances (i.e. it is not 'sub-compositionally dominant') (10,12,48). A more relevant property of distance measures in this context, however, seems to be information monotonicity (24), i.e. the fact that distances do not increase when parts are amalgamated. While this is a known property of relative entropy, it has been shown recently to also hold true for Aitchison distance (49). Although amalgamation *can* distort distances, we perform data-driven amalgamation with an objective that preserves the Aitchison distances. Figure 3 shows the agreement between the baseline Aitchison distance and the distances computed using the top-2 or top-3-reduced dimensions. The x-axis presents 10 methods ranked from highest-to-lowest based on the (geometric) average agreement. Here, we see that the use of amalgamations (or SLRs) preserves distances as well as a PCA, and both do better than a PBA of equal dimension. Supplementary Table S1 shows the 95% confidence interval for the median of the differences between each method, computed using the Wilcoxon Rank-Sum test.

Although data-driven amalgamation does not outperform PCA, it is arguably more interpretable. Although PCA is just a linear rotation of the data, its application to compositional data requires the use of clr-coordinates. Consequently, the coefficients of each principal component actually form a complex log-contrast where each variable (e.g. gene or microbe) is raised to an arbitrary power, then multiplied together. On the other hand, each amalgamation is a simple sum of parts, and therefore exists as a pooled construct that is intuitive to biologists (indeed, one might relate each amalgam to a 'gene module' or a 'bacteria community'). Advantageously, amalgamation allows the analyst to visualize the data in the same space that they exist: a simplex. As an example, Figure 4 shows the 3-part (2D) and 4-part (3D) amalgams computed for the Franzosa *et al*. microbiome dataset. Based on Figure 3 and Supplementary Table S1, we know that the distances in these figures are as coherent as a PCA plot of equal dimension. Yet, the variables of the simplex (i.e. the corners of the triangle) are easily understood: they are groups of bacteria added together.

Importantly, the amalgamations learned by our software do not appear arbitrary. Supplementary Figure S4 shows a heatmap of amalgam membership for each taxa from the Franzosa *et al*. case study, after 20 replications of n.amalgams = 2 with a Type 1 objective function. A clear pattern emerges: several dozen taxa consistently cluster together across all 20 random seeds, while many more switch only sparingly. This suggests that the amalgam memberships learned by the genetic algorithm carry meaning beyond chance occurrence. On the other hand, consensus clustering across replications might lead to more robust results, especially when the analyst plans to interpret amalgam membership directly.

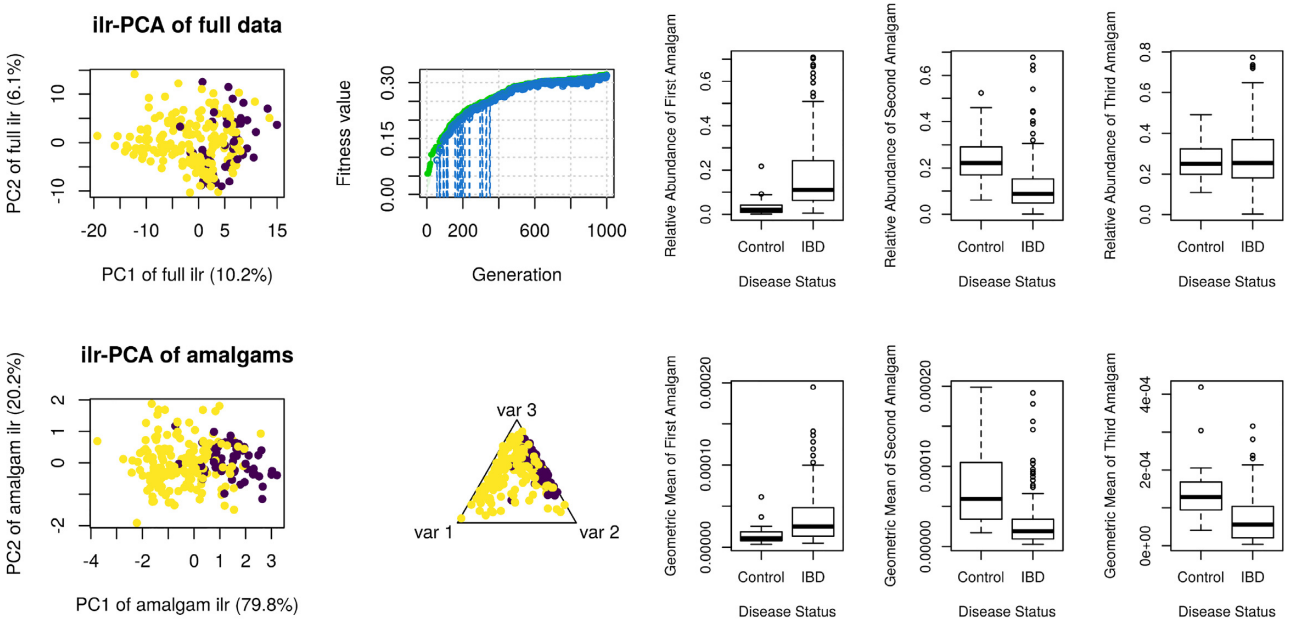

Figure 5. This figure shows the four-panel output from the amalgam software, next to boxplots of the amalgam sums and their corresponding geometric means. The top-left panel shows a PCA of the ilr-coordinates. The top-right panel shows the 'fitness' over each generation of the algorithm. The bottom-left panel shows a PCA of the ilr-coordinates of the amalgams. The bottom-right panel shows a ternary plot of all samples along the three amalgams selected to maximize the separation of sick guts (light yellow) and healthy guts (dark purple). The boxplots refer to these amalgams, where the per-sample sums are computed without a re-closure of the data (i.e. they are summed using the raw proportions). By only comparing the geometric means, one would miss the exciting insight that there exists a 'sick gut community' signature that uniquely occupies up to one half of the sick gut.

Amalgams can improve disease prediction

In the previous section, we show that data-driven amalgamation can successfully represent inter-sample distances in a coherent way (analogous to a principal coordinate analysis). By changing the objective function, we can instead search for an amalgamation that maximizes the separation between a binary class (analogous to a discriminant analysis). Supplementary Figures S1 and 2 show the area under the receiver operating curve (AUC) for classifiers trained on the top-2 or top-3-reduced dimensions, (respectively), where each boxplot shows the distribution of AUCs across 20 randomly selected test sets. In both figures, we see the same trend: amalgams perform as well as or better than PCA, principal balances, and select log-ratios. Amalgams (and SLRs) only under-perform on the Franzosa *et al*. data. Supplementary Table S2 shows the 95% confidence interval for the median of the differences between the AUCs for each method, computed using the Wilcoxon Rank-Sum test. Interestingly, the use of just three amalgams outperforms the use of three principal balances, the latter being a higher-dimensional representation.

As an example, Figure 5 shows the four-panel output from the amalgam software. Notably, the bottom-right panel shows the distribution of samples across the 3-part simplex designed to maximize class separation. Visually, we can confirm that the amalgams do separate healthy guts from unhealthy guts based on the microbiome composition. However, amalgamation gives us a unique insight into the underlying process: the first amalgam not only associates with a sick gut, but makes up *most of the sick gut*. This perspective is reinforced by a boxplot of the per-sample pre-closure amalgam sums: the first amalgam takes up ∼10–60% of the entire gut composition of sick patients, compared with only ∼5% of healthy guts. Therefore, we can interpret the first amalgam as a kind of 'sick gut community' whose members hardly ever appear in healthy patients. Interestingly, the 53 taxa that belong to this amalgam all tend to have low average abundance, suggesting that data-driven amalgamation can identify useful signals even among rarer taxa (see Supplementary Figure S3).

Below the boxplots of the per-sample amalgams, we see boxplots of the per-sample geometric means. Just as amalgams are the building blocks of SLRs, geometric means are the building blocks of balances. Although amalgamations and geometric means do not have to agree (as Greenacre *et al*. show in the beer-wine-spirit data), they do in this example. As such, misinterpreting a balance as if it were an amalgamation is of little consequence. However, by only comparing the geometric means, one would miss the exciting insight that there exists a 'sick gut community' signature that uniquely occupies up to one half of the sick gut.

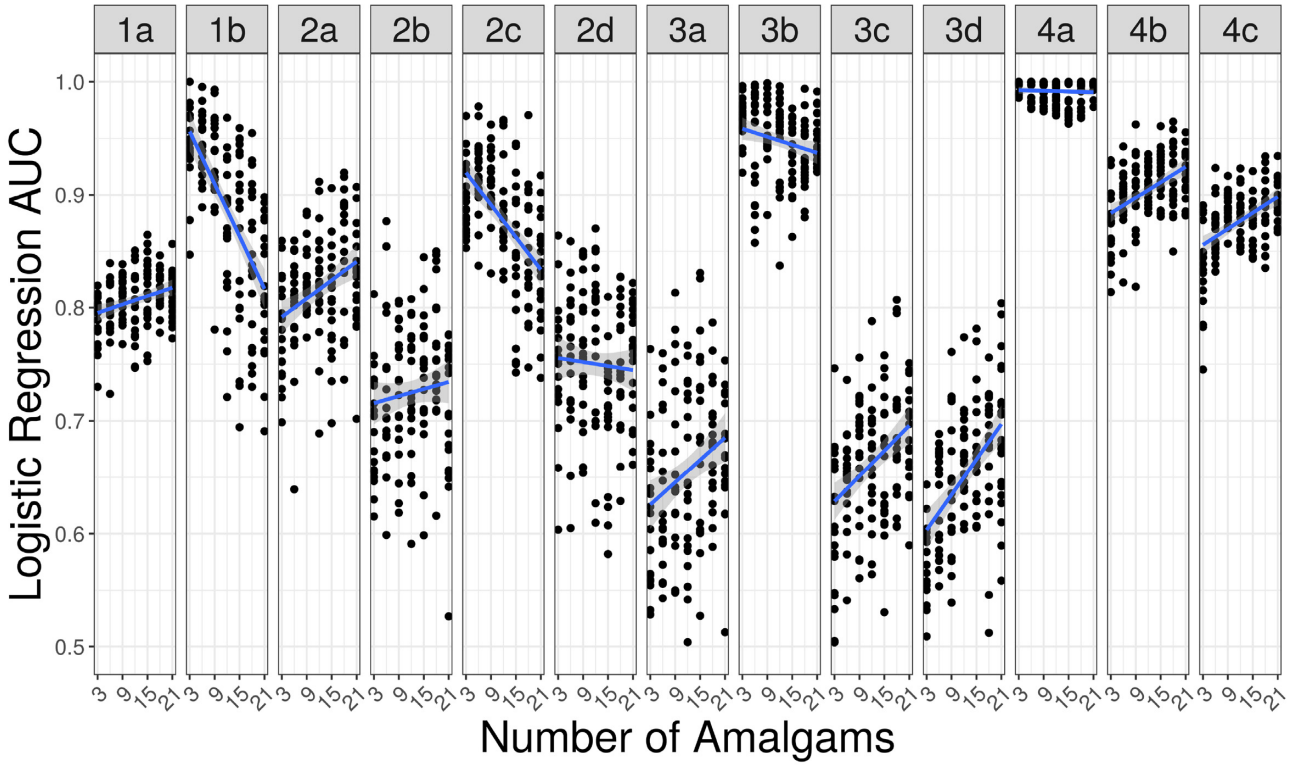

Figure 6. This figure shows the logistic regression classification AUC (y-axis) based on the number of amalgams (x-axis) used to train the model, organized by the dataset under study (facet). Each point describes a different training-test set split, and a line is drawn to show the trend. Here, we see that 8 of the 13 datasets have a better classification AUC when using more amalgams. This might suggest that using three amalgams under-fit these data. Interestingly, three datasets do markedly worse with more amalgams. This might suggest that using more amalgams over-fit these data.

Table 2. This table reports the 95% confidence interval for the median of the difference between the amalgam-based logistic regression classifier AUCs and the other procedures benchmarked in (18)

	top3amalg versus	top4amalg versus	top2slr versus	top3slr versus
Selbal	−0.0357 to 0.0033	−0.0330 to 0.0048	−0.031 to 0.007	−0.023 to 0.014
PBA	−0.055 to −0.014	−0.053 to −0.013	−0.051 to −0.011	−0.0413 to −0.0031
ABA	−0.053 to −0.012	−0.052 to −0.010	−0.0501 to −0.0096	−0.0401 to −0.0014
RBA	−0.057 to −0.016	−0.055 to −0.014	−0.053 to −0.013	−0.043 to −0.005
DBA	−0.07 to −0.03	−0.069 to −0.027	−0.065 to −0.027	−0.056 to −0.017
ACOMP	−0.018 to 0.024	−0.018 to 0.026	−0.016 to 0.027	−0.007 to 0.035
CLR	−0.064 to −0.023	−0.062 to −0.021	−0.059 to −0.020	−0.049 to −0.012

Here, we see that amalgam-based classifiers perform as well as the balance selection method selbal (17), but under-performs when compared to using all ilr-transformed or clr-transformed features. SLR: Summed log-ratios; PBA: Principal balance analysis; ABA: Anti-principal balance analysis; RBA: Random balance analysis; DBA: Distal balance analysis; ACOMP raw proportions; CLR: Centered log-ratio transformed data.

Amalgamation as an information bottleneck

The data benchmarked in this study were also benchmarked for other CoDA classification procedures, including balance selection. Compared to these, we find that data-driven amalgamation performs as well as the balance selection method selbal (17), but under-performs when compared to using *all* ilr-transformed or clr-transformed features (see Table 2). This is not surprising when we consider that using only three (or four) amalgams would have a limited capacity to explain the total structure of the data. In other words, it is possible that classifiers trained on so few amalgams *under-fit* the data. To test this hypothesis, we also trained logistic regression classifiers for $k = [3, 6, ..., 18, 21]$ amalgams. Figure 6 shows that 8 of the 13 datasets have a better classification AUC when using more amalgams. These findings reinforce the intuition that amalgams act as an 'information bottleneck' when reducing the dimensionality of the data. If there are too few amalgams, the model will under-fit.

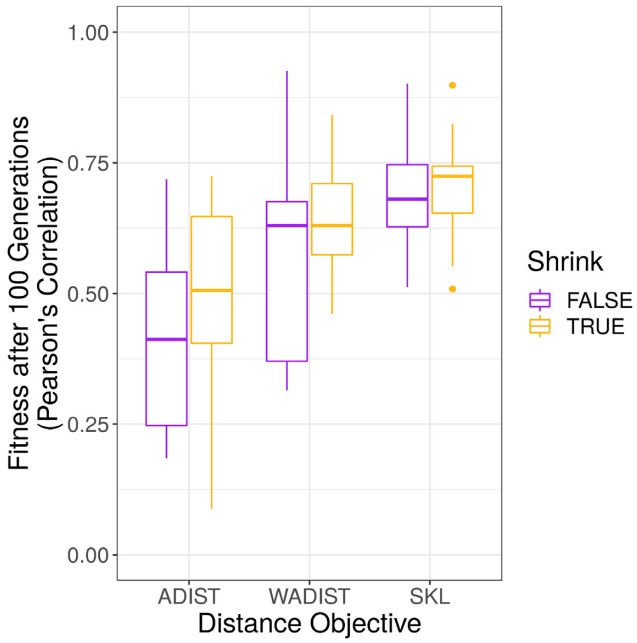

Figure 7. This figure shows how well the Aitchison (ADIST), weighted Aitchison (WADIST) and relative entropy (SKL) dimension-reduced distances agree with the corresponding baseline distances for 13 datasets (after 100 iterations, with and without James–Stein type shrinkage). Here, we see that the distributionally equivalent distances tend to have better agreement after amalgamation. We see further improvement with James–Stein type shrinkage.

Alternatives to the Aitchison distance

Although the Aitchison distance has a number of advantages, including scale invariance and sub-compositional dominance, it may not be the most appropriate measure when considering amalgamations. Instead, we may want our distance measure to observe a natural continuity property: when we amalgamate parts that are identical in two samples, the distance between these samples should be unaffected by the merging of the parts. This property, called *distributional equivalence*, is found in the weighted version of the Aitchison distance (23), and also in relative entropy (see Supplementary Data). Figure 7 shows how well each of these three dimension-reduced distances agrees with the corresponding baseline distances for 13 datasets. Here, we see that the distributionally equivalent distances tend to have better agreement after amalgamation. Better agreement means that the closest samples will remain close together while the furthest samples will remain far apart (though our use of Pearson's correlation allows for any scale factor). This finding supports our hypothesis that distributionally equivalent distances are appropriate for amalgamated data. Interestingly, we see further improvement for all distances with James–Stein type shrinkage. This is noteworthy because for non-zero shrinkage, zeros are imputed directly, making it potentially useful for zero-laden compositional count data like those encountered in microbiome or single-cell research.

LIMITATIONS AND FUTURE DIRECTIONS

A major critique of amalgamation focuses on its non-linear behavior. However, we acknowledge this when designing our search heuristic, and instead use the non-linearity to our advantage. Still, data-driven amalgamation has some limitations. First, genetic algorithms, although faster than an exhaustive search, are still quite slow (especially when compared with a PCA). For example, the Franzosa *et al*. microbiome data starts to converge after ∼1500 iterations, which takes ∼3.5 min on an Intel i7 laptop computer. A PCA of an ilr of the same dataset takes ∼50 ms. Second, data-driven amalgamation appears less effective than other simple-but-fast methods for binary classification (e.g. the distal discriminative balance analysis method described in (18)), but easily generalized to multivariable regression. Third, data-driven amalgamation requires the user to select 'hyper-parameters' to guide the dimension reduction, for example the number of amalgams. We see the importance of this hyper-parameter in Figure 6, where using too few (or too many) amalgams can impair classification accuracy. Fourth, amalgamation assumes that the relationship between the parts is explained by addition (a logical OR); as such, amalgamation would miss relationships explained by multiplication (a logical AND). On the other hand, balances would capture AND relationships but miss OR relationships.

It may be possible to resolve the first two limitations by relaxing the definition of the amalgamation matrix, for example by allowing the amalgamation matrix to become non-binary. This would allow an amalgam to equal the sum of parts of parts (not just the sum of parts). Although this expands the search space, it may enable the use of a faster search algorithm, such as gradient descent. Indeed, a total relaxation of all amalgamation matrix constraints would leave us with a single-layer neural network where the hidden layer is analogous to the amalgam set. We expect that re-factoring data-driven amalgamation as a neural network would improve the representative power of the amalgams and also decrease the run-time. However, care is needed to define the weight constraints in a way that maintains the interpretability of the hidden layer. For example, the subtraction of parts would work arithmetically, but it is unclear what interpretation this would imply. One might maintain some interpretability by introducing explicit neural network constraints. For example, one might require that a component never contributes more than itself (i.e. $\sum_j^{D'} A_{ij} \leq 1$), or that a component always contributes positively (i.e. $0 \leq A_{ij'}$).

An important property that we have only mentioned in passing is the implicit handling of zeros that is achieved by amalgamation. We could envision objective functions that remove zeros across samples by specifically merging zero-laden parts with other parts, though more research is needed to validate amalgamation for zero-laden data. Another problem that can be alleviated by amalgamation is under-sampling: the merging of parts is a way of reducing dimensions without discarding data and can reduce the number of variables such that an inversion of their covariance matrix becomes possible. This enables, for example, an evaluation of partial correlations (50) on the new variables without having to resort to regularization.

SUMMARY

In this report, we present data-driven amalgamation as a new method and conceptual framework for reducing the dimensionality of compositional data. Although amalgamation is criticized for distorting inter-sample distances, we show that data-driven amalgamation can preserve inter-sample distances as well as PCA when guided by an objective function. We also show that data-driven amalgamation can outperform both PCA and principal balances as a feature reduction method for classification, and performs as well as a supervised balance selection method called selbal.

Amalgamation not only allows the analyst to visualize the data in a lower-dimensional simplex (resembling the one in which the data naturally exist), but can also reveal interesting patterns about the relative abundances of the compositions. We demonstrate this through the discovery of a 'sick gut community' bacterial signature that occupies more than one half of the sick gut, but is rarely found in healthy samples. We encourage principled research into data-driven amalgamation as a tool for understanding high-dimensional compositional data, especially zero-laden count data for which standard log-ratio transforms fail.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

No external funding.
Conflict of interest statement. None declared.

REFERENCES

1. Aitchison,J. (1986) *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., London.
2. van den Boogaart,K.G. and Tolosana-Delgado,R. (2013) Introduction. In: *Analyzing Compositional Data with R Use R!* Springer, Berlin, Heidelberg, 1–12.
3. Fernandes,A.D., Macklaim,J.M., Linn,T.G., Reid,G. and Gloor,G.B. (2013) ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS One*, **8**, e67019.
4. Fernandes,A.D., Reid,J.N., Macklaim,J.M., McMurrough,T.A., Edgell,D.R. and Gloor,G.B. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, **2**, 15.
5. 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.*, **11**, e1004075.
6. Gloor,G.B., Wu,J.R., Pawlowsky-Glahn,V. and Egozcue,J.J. (2016) It's all relative: analyzing microbiome data as compositions. *Ann. Epidemiol.*, **26**, 322–329.

7. Gloor,G.B., Macklaim,J.M., Pawlowsky-Glahn,V. and Egozcue,J.J. (2017) Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.*, **8**, 2224.
8. Quinn,T.P., Erb,I., Richardson,M.F. and Crowley,T.M. (2018) Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, **34**, 2870–2878.
9. Calle,M.L. (2019) Statistical analysis of metagenomics data. *Genomics Inform.*, **17**, e6.
10. Egozcue,J.J. and Pawlowsky-Glahn,V. (2019) Compositional data: the sample space and its structure. *TEST*, **28**, 599–638.
11. Egozcue,J.J., Pawlowsky-Glahn,V., Mateu-Figueras,G. and Barceló-Vidal,C. (2003) Isometric logratio transformations for compositional data analysis. *Math. Geol.*, **35**, 279–300.
12. Egozcue,J.J. and Pawlowsky-Glahn,V. (2005) Groups of parts and their balances in compositional data analysis. *Math. Geol.*, **37**, 795–828.
13. Pawlowsky-Glahn,V., Egozcue,J.J. and Tolosana Delgado,R. (2011) Principal balances. In: *Proceedings of CoDaWork 2011, The 4th Compositional Data Analysis Workshop*, pp. 1–10.
14. Silverman,J.D., Washburne,A.D., Mukherjee,S. and David,L.A. (2017) A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, **6**, e21887.
15. Washburne,A.D., Silverman,J.D., Leff,J.W., Bennett,D.J., Darcy,J.L., Mukherjee,S., Fierer,N. and David,L.A. (2017) Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, **5**, e2969.
16. Morton,J.T., Sanders,J., Quinn,R.A., McDonald,D., Gonzalez,A., Vázquez-Baeza,Y., Navas-Molina,J.A., Song,S.J., Metcalf,J.L., Hyde,E.R. *et al.* (2017) Balance trees reveal microbial niche differentiation. *mSystems*, **2**, e00162-16.
17. Rivera-Pinto,J., Egozcue,J.J., Pawlowsky-Glahn,V., Paredes,R., Noguera-Julian,M. and Calle,M.L. (2018) Balances: a new perspective for microbiome analysis. *mSystems*, **3**, e00053-18.
18. Quinn,T.P. and Erb,I. (2020) Interpretable log contrasts for the classification of health biomarkers: a new approach to balance selection. *mSystems*, **5**, e00230–19.
19. Greenacre,M., Grunsky,E.C. and Bacon-Shone,J. (2019) A comparison of amalgamation and isometric logratios in compositional data analysis. ResearchGate doi: https://www.researchgate.net/publication/332656109, May 2019, preprint: not peer reviewed.
20. Greenacre,M. (2020) Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation. *Appl. Comput. Geosci.*, **5**, 100017.
21. Greenacre,M. (2019) Comments on: compositional data: the sample space and its structure. *TEST*, **28**, 644–652.
22. Greenacre,M. (2019) Variable selection in compositional data analysis using pairwise logratios. *Math. Geosci.*, **51**, 649–682.
23. Greenacre,M. and Lewi,P. (2009) Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *J. Classif.*, **26**, 29–54.
24. Amari,S.-I. (2016) *Information Geometry and its Applications*. Springer, Vol. **194**, pp. 1373.
25. DeDeo,S. (2018) Information theory for intelligent people. http://santafe.edu/ simon/it.pdf.
26. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
27. Chao,A. and Shen,T.-J. (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.*, **10**, 429–443.
28. Hausser,J. and Strimmer,K. (2009) Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, **10**, 1469–1484.
29. Paliy,O. and Shankar,V. (2016) Application of multivariate statistical techniques in microbial ecology. *Mol. Ecol.*, **25**, 1032–1057.
30. Scrucca,L. (2013) GA: a package for genetic algorithms in R. *J. Stat. Softw.*, **53**, 1–37.
31. van den Boogaart,K.G. and Tolosana-Delgado,R. (2008) A unified R package to analyze compositional data. *Comput. Geosci.*, **34**, 320–338.
32. Oksanen,J., Blanchet,F.G., Friendly,M., Kindt,R., Legendre,P., McGlinn,D., Minchin,P.R., O'Hara,R.B., Simpson,G.L., Solymos,P. *et al.* (2019) vegan: community ecology package.
33. Aitchison,J. and Greenacre,M. (2002) Biplots of compositional data. *J. R. Stat. Soc. C*, **51**, 375–392.
34. Quinn,T.P. (2018) Visualizing balances of compositional data: a new alternative to balance dendrograms. *F1000Res.*, **7**, 1278.
35. Quinn,T.P., Richardson,M.F., Lovell,D. and Crowley,T.M. (2017) propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.*, **7**, 16252.
36. Gevers,D., Kugathasan,S., Denson,L.A., Vázquez-Baeza,Y., Van Treuren,W., Ren,B., Schwager,E., Knights,D., Song,S.J., Yassour,M. *et al.* (2014) The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe*, **15**, 382–392.
37. Noguera-Julian,M., Rocafort,M., Guillén,Y., Rivera,J., Casadellà,M., Nowak,P., Hildebrand,F., Zeller,G., Parera,M., Bellido,R. *et al.* (2016) Gut microbiota linked to sexual preference and HIV infection. *EBioMedicine*, **5**, 135–146.
38. Schubert,A.M., Rogers,M.A.M., Ring,C., Mogle,J., Petrosino,J.P., Young,V.B., Aronoff,D.M. and Schloss,P.D. (2014) Microbiome data distinguish patients with clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls. *mBio*, **5**, e01021-14.
39. Baxter,N.T., Ruffin,M.T., Rogers,M. A.M. and Schloss,P.D. (2016) Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.*, **8**, 37.
40. Duvallet,C., Gibbons,S.M., Gurry,T., Irizarry,R.A. and Alm,E.J. (2017) Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.*, **8**, 1784.
41. Franzosa,E.A., Sirota-Madi,A., Avila-Pacheco,J., Fornelos,N., Haiser,H.J., Reinker,S., Vatanen,T., Hall,A.B., Mallick,H., McIver,L.J. *et al.* (2018) Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.*, **4**, 293–305.
42. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
43. Netanely,D., Avraham,A., Ben-Baruch,A., Evron,E. and Shamir,R. (2016) Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. *Breast Cancer Res.*, **18**, 74.
44. Palarea-Albaladejo,J. and Martín-Fernández,J.A. (2015) zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometr. Intell. Lab.*, **143**, 85–96.
45. Martín-Fernández,J., Barceló-Vidal,C., Pawlowsky-Glahn,V., Buccianti,A., Nardi,G. and Potenza,R. (1998) Measures of difference for compositional data and hierarchical clustering methods. In: *Proceedings of IAMG*. Vol.**98**, pp. 526–531.
46. Aitchison,J., Barceló-Vidal,C., Martín-Fernández,J.A. and Pawlowsky-Glahn,V. (2000) Logratio analysis and compositional distance. *Math. Geol.*, **32**, 271–275.
47. Quinn,T., Tylee,D. and Glatt,S. (2017) exprso: an R-package for the rapid implementation of machine learning algorithms. *F1000Res.*, **5**, 2588.
48. Filzmoser,P. and Hron,K. (2019) Comments on: compositional data: the sample space and its structure. *TEST*, **28**, 639–643.
49. Erb,I. and Ay,N. (2020) The information-geometric perspective of Compositional Data Analysis. arXiv doi: https://arxiv.org/abs/2005.11510, 23 May 2020, preprint: not peer reviewed.
50. Erb,I. (2020) Partial correlations in compositional data analysis. *Appl. Comput. Geosci.*, **6**, 100026.