Chapter 1
The analysis of human serum albumin proteoforms using compositional framework

Shripad Sinari, Dobrin Nedelkov, Peter Reaven and Dean Billheimer

Abstract Mass Spectrometric Immuno Assays (MSIA) can now measure multiple modified forms of a protein in large cohorts of patients. These measurements consist of the relative abundances of proteoforms, and are well-suited for the compositional data analysis statistical framework. In this article, we describe an approach to the analysis of relative abundance of proteoforms from MSIA data using the compositional framework. We demonstrate the application of these concepts by exploring the association of human serum albumin's post translational modifications and kidney function in patients with Type 2 diabetes mellitus. Finally, we discuss the pitfalls of ignoring the compositional nature of such data, and highlight emerging applications demonstrating the generality of the framework.

1.1 Introduction

A relatively small number of genes yield the enormous diversity in the eukaryotes by means of post translational modifications of proteins. These modified proteins, also called proteoforms, give rise to new functional capabilities and regulate the

Shripad Sinari, MS
Shripad Sinari, BIO5 Institute, The University of Arizona, e-mail: shripad@email.arizona.edu

Dobrin Nedelkov, PhD
Dobrin Nedelkov, Molecular Biomarkers Laboratory, Biodesign Institute, Arizona State University, e-mail: dobrin.nedelkov@asu.edu

Peter Reaven, MD
Peter Reaven, Phoenix VA Health Care System, e-mail: Peter.Reaven@va.gov

Dean Billheimer, PhD
Dean Billheimer, Epidemiology and Biostatistics, BIO5 Institute, The University of Arizona, e-mail: dean.billheimer@arizona.edu

cellular environment [27, 13]. They have been implicated in diseases such as cancer [8] and age related dementia [22].

Identifying these proteoforms with sensitivity and specificity has been a challenge, especially when the abundance is low. Mass Spectromety Immuno Assay (MSIA) is an approach developed to address these challenges. MSIA combines the sensitivity of the immuno assay based approaches with the specificity of detection from mass spectroscopy. Moreover, MSIA allows the simultaneous detection of multiple proteoforms in a single assay. Nelson, et.al [18] is a useful reference for the details of this approach.

The use of immuno based assays to enrich for proteoforms imposes a constraint on the measurement of their concentrations. Thus the resulting peak areas capture information on relative abundances of the proteoforms rather than their absolute concentration. Measurements of the *relative abundance* of multiple components is a characteristic of compositional data [1]. In this paper we demonstrate that the compositional data analysis framework is ideally suited to exploring and analyzing MSIA data. Particularly, the framework allows interpretation of complicated covariance structure, guarantees consistency between analyses of a part and the whole composition, and permits the use of standard multivariate statistical methods, all while respecting structural constraints inherent in the observed data.

We begin by exploring the compositional nature of MSIA data in our example of albumin proteoforms. We show how the compositional framework allows for reference free normalization of this data. Explore the association of glycosylated and cysteinylated albumin proteoforms with prognosis of chronic kidney disease CKD in patients with Type 2 diabetes mellitus. Finally, we discuss the role of compositional data in our application as well as in genomics and conclude with remarks on the exploratory analysis of albumin proteoforms.

1.2 MSIA and albumin proteoforms

Our MSIA measurements comprise albumin proteoforms from 283 patients with Type 2 diabetes mellitus. Glycosylation [26] and cysteinylation [17] of albumin are two important post translational modifications that have been associated with advanced chronic kidney disease (CKD). Here we explore the association of these proteoforms with chronic kidney disease (CKD).

Table [1.1] shows a small subset of the raw data. The first column is the sample identifier and the remaining 9 columns represent the raw peak areas of the 9 albumin post translational modifications.

The most abundant form is called wildtype and is denoted by "wt". The cysteinylated proteoforms are annotated with ".cys" and the glycosylated proteoforms with ".gly". The proteoforms annotated with "des" are truncated forms of wildtype protein. The data matrix consists of 283 rows and 9 columns. Each row being a composition and thus a point in $\mathscr{S}^8$ which is the space of 9 part composition given by

Table 1.1: Table of raw peak areas of a small subset of the albumin data.

ID	des.DA	des.D	des.DA.cys	wt	wt.cys	wt.gly	wt.cys.gly	wt.gly.gly	wt.cys.gly.gly
546101	4969.01	6021.65	3318.04	68552.31	55486.38	27058.15	15544.28	9834.52	4291.44
546103	7272.77	6704.79	8614.81	98730.16	134177.76	35674.16	28190.84	10905.35	5562.96
546104	6589.51	5673.29	8419.23	107413.43	104393.18	40453.52	33830.79	15787.60	10996.31
546105	7119.19	6802.57	8144.94	98650.74	90278.81	29793.74	17440.50	8504.92	3608.09
546106	5880.67	4774.71	6249.13	67389.77	89762.67	25233.69	23333.77	8539.34	5624.68

$$\mathscr{S}^8 = \left\{\mathbf{x} = (x_1, x_2, \ldots, x_9) : x_i > 0 (i = 1, 2, \ldots, 9), \sum_{i=1}^{9} x_i = 1\right\} \tag{1.1}$$

This is an 8 dimensional simplex embedded in the 9 dimensional real vector space $\mathbb{R}^9$. Table [1.2] gives the compositions formed from the data subset shown in Table [1.1].

Table 1.2: Table of compositions of the small subset of the albumin data.

ID	des.DA	des.D	des.DA.cys	wt	wt.cys	wt.gly	wt.cys.gly	wt.gly.gly	wt.cys.gly.gly
546101	0.03	0.03	0.02	0.35	0.28	0.14	0.08	0.05	0.02
546103	0.02	0.02	0.03	0.29	0.40	0.11	0.08	0.03	0.02
546104	0.02	0.02	0.03	0.32	0.31	0.12	0.10	0.05	0.03
546105	0.03	0.03	0.03	0.36	0.33	0.11	0.06	0.03	0.01
546106	0.02	0.02	0.03	0.28	0.38	0.11	0.10	0.04	0.02

Typically additional information may be present that provides clinical status associated with each sample. In our data, we have 2 additional columns. One gives the CKD status of the patient and the other gives the value of the glomerular filtration rate (GFR) for the patient. GFR is used to determine the health of the kidney and classify the patient in one of the three CKD status (low, medium or high).

The simplex $\mathscr{S}^8$ is a Hilbert space with a metric defined by

$$d(x,y) = \left[\sum_{i=1}^{9} (\log(\frac{x_i}{g_9(x)}) - \log(\frac{y_i}{g_9(y)}))^2\right]^{1/2} \tag{1.2}$$

Here $g_9(x)$ is the geometric mean of vector $x \in \mathscr{S}^8$. Multiple co-ordinate systems exists on this Hilbert space. Details of these co-ordinates as well as their equivalence can be found in Egozcue et al. [11]. In our analysis we will use the centered log ratio transformation which is given by

$$\begin{aligned} &\mathscr{L} : \mathscr{S}^8 \to \mathbb{R}^9 \\ &\qquad x \longmapsto (\log(\frac{x_1}{g_9(x)}), \log(\frac{x_2}{g_9(x)}), \ldots, \log(\frac{x_D}{g_9(x)})) \end{aligned} \tag{1.3}$$

The centered log ratio allows us to look at all the proteoforms and gives a covariance matrix that is more interpretable than the original composition, and is suitable for exploration using the principal component analysis (PCA). Table [1.3] gives the centered log ratios of compositions from Table [1.2]. Note that centered logratios now sum to zero for each sample.

Table 1.3: Table of centered log ratios of the small subset of the albumin data.

ID	des.DA	des.D	des.DA.cys	wt	wt.cys	wt.gly	wt.cys.gly	wt.gly.gly	wt.cys.gly.gly
546101	-0.91	-0.72	-1.31	1.71	1.50	0.78	0.23	-0.23	-1.06
546103	-0.97	-1.05	-0.80	1.64	1.95	0.62	0.39	-0.56	-1.23
546104	-1.17	-1.31	-0.92	1.63	1.60	0.65	0.47	-0.29	-0.65
546105	-0.79	-0.83	-0.65	1.84	1.75	0.64	0.11	-0.61	-1.47
546106	-0.91	-1.12	-0.85	1.53	1.82	0.55	0.47	-0.54	-0.95

For some proteins a naturally occurring native or highly abundant form exists. In applications where there is such a highly abundant form, it is the ratio with this form that is often of most interest. In such cases, an alternate co-ordinate system named the additive logratio transform (equation [1.9] in appendix) may be more useful. Additive logratio transform may also be useful in parameter estimation in linear models when all proteoforms are included as a multivariate outcome. Standard multivariate methods, as well as multiple regression techniques, can now be applied to these appropriately transformed data.

1.2.1 Normalization of proteomic measurements as compositions

In addition to providing a convenient structure to apply the usual multivariate analyses methods, the co-ordinate transformations in the compositional setting also performs a normalization of the data. Figure [1.1] shows that the variability as well as skewness in each variable is reduced by the centered log ratio transform.

When a convenient reference standard exists, it may be included in the MSIA assay [18, 25]. The reference standard is used to determine the *absolute* concentrations of the proteoforms from a calibration curve. Typically. it is a modified version of the protein with a known mass to charge ratio. If poorly matched to the target protein, however, the reference standard may increase the variability in the calibrated data. For our dataset, no reference standard was used. In this situation, taking compositional nature of the data into account and applying appropriate transformations provides a good normalization scheme with reduction in the total variability of measurements.

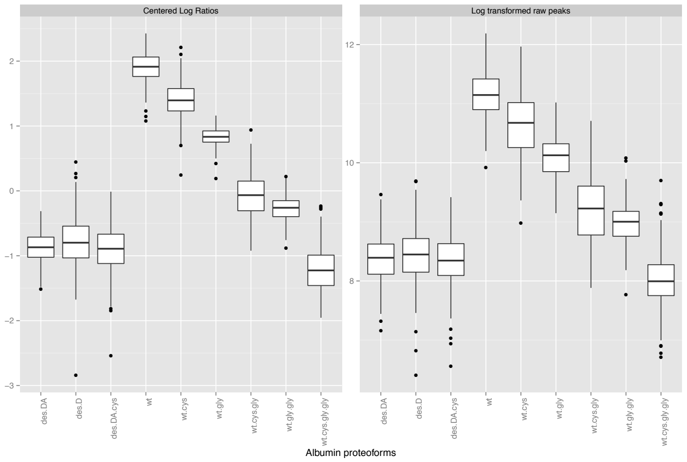

Fig. 1.1: Box plots of the log transformed raw values and the centered log ratios of albumin proteoforms. The box plots for each proteoform indicates a decrease in overall variability and skewness in the centered log ratios.

1.2.2 Interpretation of multivariate proteoform analysis

Aitchison (see [2]) discusses the difficulty of interpretation of the principal component analysis (PCA) on the raw data. In particular, issues arise due to lack of spherically symmetric distributions. This is dealt by the use of logistic normal distribution (equation [1.4] in appendix). The centered log ratio transformed composition gives an isotropic invariant covariance structure from which a measure of total variablility of a composition can be expressed as

$$\sum_{i=1}^{9} var\left[\log(\frac{x_i}{g_9(x)})\right] \tag{1.4}$$

and the principal components become orthonormal log linear contrasts. Subcompositional coherence is compatibility of inferences between the full and a subset of the proteoforms. Such coherence in inference is guaranteed by the compositional framework [3]. We demonstrate this coherence in the analysis of our example below. It is also shown how ignoring the constraint can lead to misleading interpretation of the data.

1.2.3 Relative variation biplot

A biplot is a visual aid to understand and interpret the results of a PCA. Biplots show the structure of variables in terms of major axes of variability (principal components). The horizontal axis is first principal component (PC1), while the vertical axis is the second (PC2). Each point represents an individual sample. Variables are denoted by arrows. Points may be colored, shaped or labeled by a classification or for identification.

A biplot resulting from the PCA of a covariance matrix of variables is called a *relative variation biplot*. Any biplot can be displayed in two forms. One is called the covariance biplot where distances between the variables are approximations of the standard deviations of the corresponding log-ratios and angle cosines between links estimate the correlations between log-ratios. Links are the difference vectors connecting the tips of the arrows representing the variables. The second is called a form biplot where distance between points are approximation of the distances given by the metric in equation [1.2].

Figures [1.2] and [1.3] are the covariance and the form relative variation biplots respectively, for the albumin dataset. Points are colored by the CKD status of the individual from whom the sample was obtained. The biplots indicate that the relative proportions of the albumin proteoforms in the sample can distinguish between the higher CKD status (encoded as 3 and color coded as red) and the lowest CKD status (encoded as 1 and color coded as green). The samples with lower CKD status are mostly to the right and those with higher CKD status mostly to the left. The plots also shows that higher proportion of wildtype albumin is associated with lower CKD status as one would expect. Higher proportions of the cysteinylated versions of albumin proteoforms are associated with poor CKD status. The proportion of variance explained by the covariance and the form biplots are `73%` and `70%` respectively.

These plots provide an approximation to the covariance structure of the albumin proteoforms. An example is that of the link between the proteoforms wt and des.D in the covariance biplot (Figure 1.2). The length of the link is approximately `0.293` whereas the actual standard deviation of the log-ratio is `0.283`.

Aitchison [4] provides a good introduction and insights into numerous useful properties of the relative variation biplots and proves the equivalence of the biplots under various co-ordinate systems.

The associations seen in the biplots can also be confirmed in the linear regression of the proteoforms with the continuous measurement of CKD status, GFR. Each row of Table [1.4] gives the coefficients of the linear regression of the proteoform with GFR.

Table 1.4: Table of coefficients of linear regression with GFR. The model uses the centered log ratios of the proteoform. Each individual proteoform was regressed against GFR.

	Estimate	Std. Error	t value	Pr(>|t|)
des.DA	27.34	5.65	4.84	0.00
des.D	14.78	3.45	4.29	0.00
des.DA.cys	-10.41	3.80	-2.74	0.01
wt	27.91	5.26	5.30	0.00
wt.cys	-17.95	4.77	-3.77	0.00
wt.gly	34.51	9.93	3.48	0.00
wt.cys.gly	-16.07	3.94	-4.08	0.00
wt.gly.gly	3.41	6.72	0.51	0.61
wt.cys.gly.gly	-12.77	3.79	-3.37	0.00

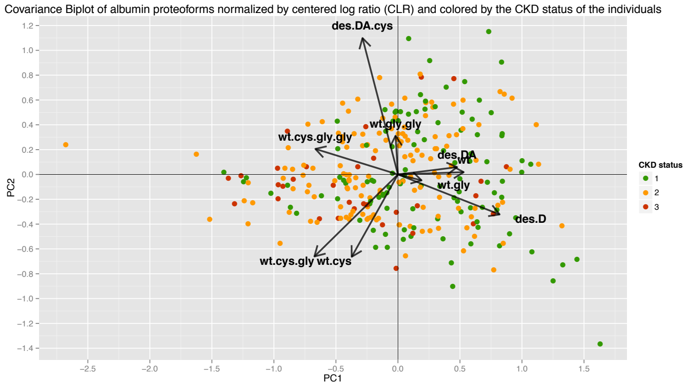

Fig. 1.2: **Covariance Biplot** of individuals with Chronic Kidney disease (CKD) as classified from the measurement of their glomerular filtration rate (GFR). The albumin proteoform values are centered log ratio (CLR). The axis are the first two principal components with first component along the horizontal and the second along the vertical axis. The points represent individuals (about 283) whose samples contributed to the MSIA measurements and whose CKD status was known at the start of the study.

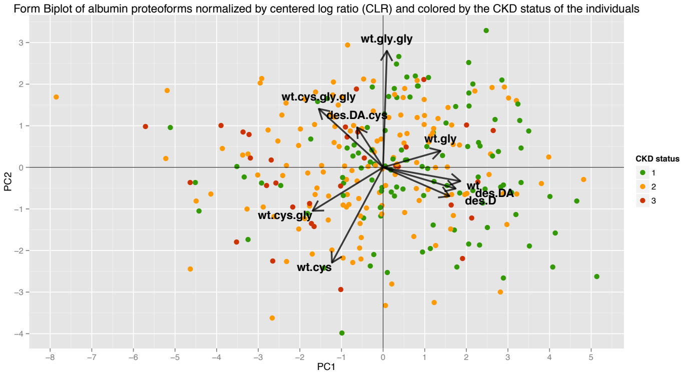

Fig. 1.3: **Form Biplot** of individuals with Chronic Kidney disease (CKD) as classified from the measurement of their glomerular filtration rate (GFR). The albumin proteoform values are centered log ratio (CLR). The axis are the first two principal components with first component along the horizontal and the second along the vertical axis. The points represent individuals (about 283) whose samples contributed to the MSIA measurements and whose CKD status was known at the start of the study.

1.2.4 *Results without the unit sum constraint*

We now look at a similar analysis with log transformed raw peaks of the proteoforms. Figure [1.4] is a form biplot and Table [1.5] contains the results of regression of the log transformed raw peak areas with GFR.

Table 1.5: Table of coefficients of linear regression with GFR. The model uses the log transformed raw peak areas of the proteoform. Each individual proteoform was regressed against GFR.

	Estimate	Std. Error	t value	Pr(>|t|)
des.DA	10.33	3.52	2.93	0.00
des.D	9.66	2.80	3.45	0.00
des.DA.cys	-6.39	3.03	-2.11	0.04
wt	11.34	3.42	3.31	0.00
wt.cys	-4.53	2.46	-1.84	0.07
wt.gly	5.07	3.79	1.34	0.18
wt.cys.gly	-4.96	2.25	-2.20	0.03
wt.gly.gly	1.53	4.09	0.37	0.71
wt.cys.gly.gly	-6.77	2.81	-2.41	0.02

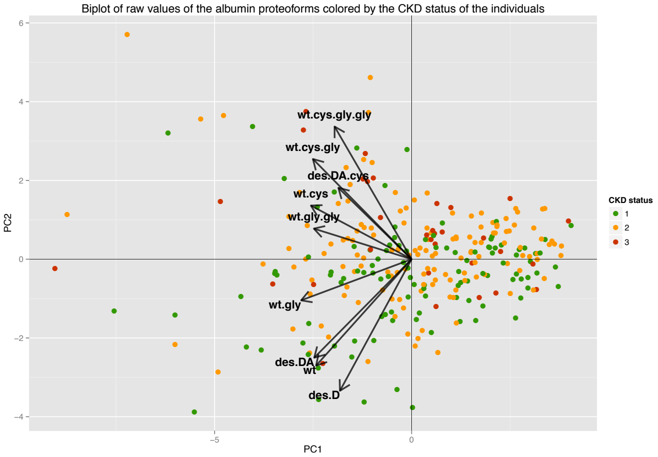

Fig. 1.4: **Form Biplot** of individuals with Chronic Kidney disease (CKD) as classified from the measurement of their glomerular filtration rate (GFR). The albumin proteoform values are log transformed raw peak areas. The axis are the first two principal components with first component along the horizontal and the second along the vertical axis. The points represent individuals (about 283) whose samples contributed to the MSIA measurements and whose CKD status was known at the start of the study.

Although the regression tables indicate similar relationships between each proteoform and GFR, the strength of the association is reduced or deemed insignificant. The picture in the biplots however is less clear. In Figure [1.4] all proteoforms point to the left indicating association of the total signal of albumin with CKD status. However the association of cysteinylated proteoforms with poor CKD status is lost.

Consistency between the part, i.e the univariate analysis (Table [1.4]), and the whole composition, i.e. the principal component analysis using all proteoforms (Figure [1.3]), is evident for analysis done with centered log ratios. The proteoform "wt.gly.gly" does not carry information about the health of the kidney of the patient. This consistency is absent between the univariate analysis (Table [1.5]) and the corresponding PCA (Figure [1.4]) done with the log transformed raw peaks of the proteoforms.

As seen in this example, consideration of compositional structure brings added insights into the covariance structure of the components and vindicate the use of standard analytical tools. This is consistent with observation in Lovell et.al. [15], that use of compositional framework may not lead to dramatically different results across the board but the application of Aitchison distance (equation [1.2]) provides more meaningful insights.

1.3 Discussion

The multivariate exploration of albumin proteoforms highlights the importance of the cysteinylated proteoforms of albumin in the prognosis of diabetic patients with CKD in our data. Such insights are absent from the analysis that does not take the compositional contraint into account. Recently, Borges et al. [7] have shown that cysteinylation of albumin can result from sample storage or handling. In such cases, the consideration of compositional framework can reflect on the quality of the data. Thus such analysis brings about better understanding of the roles of cysteinylated versus glycosylated proteoforms of albumin in the prognosis of CKD or serves to provide a quality check on samples. The compositional framework also provides a convenient and interpretable normalization scheme. The normalization constant is the mean of log transformed components. Hence the name centered log ratio. In the albumin proteoforms this means that the variability such as batch effects due to antibody used for immunoaffinity capture is normalized. In general, all non-proteoform specific variability is reduced.

Many widely used high throughput technologies produce similar data that are inherently compositional. Two prominent ones being quantification of gene expression using RNA-Seq and metagenomics, which is the study of the composition of microbial genomes in a sample. In RNA-Seq the counts of mRNA are often reported as a composition. The extraction of mRNA from a fixed volume of starting material puts the unit sum constraint on the counts observed in the experiment. Differential expression of genes between conditions is the question about change in relative abundance with respect to a reference, called a housekeeping gene. It is important to realize that information on absolute abundance is lost in these measurements. Two genes may have the same level of mRNA production but one might be differentially expressed while the other is not. However correlation induced due to the sum constraint can lead to misleading interpretation in case of such data, if the compositional nature is not taken into account. See Lovell et al. [15] for detailed exposition on this issue.

Similarly, applications in metagenomics involve comparison of the compositions within or between different conditions of genetically diverse microorganisms. The compositional structure of microbial community here is more evident. Community composition analysis is employed in diverse applications such as exploring the biodiversity of habitat [23], common pathogens in clinical settings [19] or classification of the microbes into genus [10] and phylogeography [9]. Cell fractionation techniques or size selection similar to proteomics is often used in sample collection to create homogeneuous populations of cells and enrichment of the target DNA [24]. These methods impose a compositional contraint. Statistical analysis of such data can benefit from the use of compositional framework [14].

One limitation of compositional approach is worth mentioning. This is the problem of *essential zeros*. Essential zeros arise when zero is valid value for some parts of the composition. This is distinct from the inability to detect a signal due to the signal being lower than the limit of detection. Such below the detection limit zeros are called *rounded zeros* in the compostional literature. An example of essential zero arises in a compositional data consisting of family budgets. Some families may not consume alcohol and hence the money allocated to this expenditure may be zero.

In proteomic applications, zeros are often treated as rounded zeros (e.g., below detection limit). Thus rounded zeros are often replaced by multiplicative strategy. In this strategy, the zeros in a composition are replaced by small non-zero values. To maintain unit sum constraint, the non-zero components are multiplied by a suitable value. In our data set, we replaced the zero values in raw peak areas with half of the lowest non-zero terms for that proteoform, before computing the centered log ratios or the log transformations. A detailed discussion on zeros as well as the several methods of dealing with rounded zeros can be found in Martín-Fernández et al. [16]. An important point to note is that, samples with a part value as zero lie on the edge of the simplex which is excluded from our definition of composition for mathematical convenience. Problems arise in extending the metric as well as the full logistic normal distribution to the edges.

1.4 Conclusions

The results of the exploratory analysis of albumin data using compositional data framework shows that changes in the proportions of the cysteinylated albumin proteoforms can reveal information about the status of the chronic kidney disease in an individual, or indicate issues with data storage and handling ex vivo. This analysis implies that MSIA assays can be used to explore the clinical role of post translational modifications of a protein. Compositional framework is essential in inference related to such relative proportions data. The framework provides for normalization of data and also validate the application of conventional multivariate analysis techniques. It provides for consistency between analysis of the part and the whole composition through the principle of subcompositional coherence. Ignoring the limitation imposed by the summation constraint in these relative proportions data, as is often the case, can result in loss of valuable insights or worse, lead to misleading conclusions.

Acknowledgements The authors would like to thank Dr. Borges for guidance on various mechanisms that can lead to higher proportions of cysteinylated albumin proteoforms in MSIA data.

Appendix

A synopsis on compositional framework

Compositional data describe the proportion that each of D components contributes to the whole. Coherence of inference between subset of components and the whole composition, also called subcompositional coherence, is an important feature of the analysis. Scale invariance is essential for such analysis.

A D part composition is defined as an element of the d-dimensional positive simplex

$$\mathscr{S}^{\mathrm{d}} = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_D) : x_i > 0 (i = 1, 2, \dots, D), \sum_{i=1}^{D} x_i = 1 \right\} \tag{1.5}$$

where $d = D - 1$. We will use the convention $d = D - 1$ in the rest of the appendix. Thus the sample space is a subset of the space $\mathscr{S}^{\mathrm{d}}$.

Let C denote the *closure* operator on $\mathbb{R}^{\mathrm{D}}$ which normalizes the vector to a unit sum. That is, for $z \in$,

$$C(z) = \left(\frac{z_1}{\sum_{i=1}^{D} z_i}, \frac{z_2}{\sum_{i=1}^{D} z_i}, \dots, \frac{z_D}{\sum_{i=1}^{D} z_i} \right) \in \mathscr{S}^{\mathrm{d}} \tag{1.6}$$

Also, we define two additional operations. For any two elements $x = (x_1, x_2, \dots, x_{\mathrm{D}})$ and $y = (y_1, y_2, \dots, y_{\mathrm{D}}) \in \mathscr{S}^{\mathrm{d}}$ and for $\alpha \in \mathbb{R}$, we define:

$$x \oplus y = C((x_1 \cdot y_1, x_2 \cdot y_2, \dots, x_D \cdot y_D)) \tag{1.7}$$

$$\alpha \odot x = C((x_1^{\alpha}, x_2^{\alpha}, \dots, x_D^{\alpha})) \tag{1.8}$$

The operations in equations [1.7] and [1.8] are called the pertubation and power operators, respectively. With pertubation operator as addition and power operator as the scalar multiplication, $\mathscr{S}^{\mathrm{d}}$ acquires the structure of a d dimensional Hilbert space [6, 21] with a metric given by [1.2].

The additive log ratio transform defined as:

$$\begin{aligned} &\phi : \mathscr{S}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{d}} \\ &\quad x \longmapsto (\log(\frac{x_1}{x_D}), \log(\frac{x_2}{x_D}), \dots, \log(\frac{x_d}{x_D})) \end{aligned} \tag{1.9}$$

and the centered log ratio [1.3] are alternative co-ordinate systems on this space.

The dependence structure induced by the unit sum (compositional) constraint is often addressed by using the class of logistic normal distributions as appropriate models of the data. For illustration, consider an element $x = (x_1, x_2, \dots, x_{\mathrm{D}}) \in \mathscr{S}^{\mathrm{d}}$. Following Aitchison [1] the pullback of the multivariate normal distribution using ϕ from equation [1.9] is the logistic normal density function given by:

$$f(x|\mu,\Sigma) = (2\pi)^{d/2}|\Sigma|^{-1/2}\left(\frac{1}{\prod_{i=1}^{k} x_i}\right) exp\left[-\frac{1}{2}(\phi(x)-\mu)'\Sigma^{-1}(\phi(x)-\mu)\right]$$

where μ is the location parameter in $\mathbb{R}^{\mathrm{d}}$ and Σ is the $d \times d$ variance-covariance matrix, $(\prod_{i=1}^{D} x_i)^{-1}$ is the Jacobian of the transformation. In the following, we will denote this d dimensional logistic normal distribution by $\mathscr{L}\mathscr{N}_{\mathrm{d}}$ and A' will denote the transpose if A is a matrix.

The part S composition is orthogonal projection of the full composition, with respect to the inner product on $\mathscr{S}^{\mathrm{d}}$ [12]. The class preserving property of logistic normal distributions, i.e., if $x \in \mathscr{L}\mathscr{N}_D(\mu,\Sigma)$ and A is a $n \times D$ matrix then $Ax \in \mathscr{L}\mathscr{N}_n(A\mu, A\Sigma A')$ [5], ensures that the orthogonal projections satisfy the property of subcompositional coherence.

A comprehensive review of analytical techniques and applications of compositional framework are available in the book Pawlowsky-Glahn et al. [20].

References

1. Aitchison, J.: The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological) pp. 139–177 (1982). URL http://www.jstor.org/stable/10.2307/2345821
2. Aitchison, J.: Principal component analysis of compositional data. Biometrika **70**(1), 57–65 (1983). URL http://biomet.oxfordjournals.org/content/70/1/57.short
3. Aitchison, J.: Simplicial inference. In: M.A.G. Viana, D.S.P. Richards (eds.) Algebraic Methods in Statistics and Probability, *Contemporary Mathematics*, vol. 287. American Mathematical Society, Providence, Rhode Island (2001). DOI 10.1090/conm/287. URL http://www.ams.org/conm/287/
4. Aitchison, J., Greenacre, M.: Biplots of compositional data. Applied Statistics **51**(4), 375–392 (2002). DOI 10.1111/1467-9876.00275. URL http://dx.doi.org/10.1111/1467-9876.00275

5. Aitchison, J., Shen, S.M.: Logistic-normal distributions: Some properties and uses. Biometrika **67**(2), 261–272 (1980). URL http://biomet.oxfordjournals.org/content/67/2/261.short
6. Billheimer, D., Guttorp, P., Fagan, W.F.: Statistical Interpretation of Species Composition. Journal of the American Statistical Association **96**(456), 1205–1214 (2001). DOI 10.1198/016214501753381850. URL http://www.tandfonline.com/doi/abs/10.1198/016214501753381850
7. Borges, C.R., Rehder, D.S., Jensen, S., Schaab, M.R., Sherma, N.D., Yassine, H., Nikolova, B., Breburda, C.: Elevated plasma albumin and apolipoprotein A-I oxidation under suboptimal specimen storage conditions. Molecular & cellular proteomics : MCP **13**(7), 1890–1899 (2014). DOI 10.1074/mcp.M114.038455. URL http://www.mcponline.org/cgi/doi/10.1074/mcp.M114.038455
8. Chammas, R., Sonnenburg, J.L., Watson, N.E., Tai, T., Farquhar, M.G., Varki, N.M., Varki, A.: De-N-acetyl-gangliosides in humans: unusual subcellular distribution of a novel tumor antigen. Cancer Research **59**(6), 1337–1346 (1999). URL http://cancerres.aacrjournals.org/content/59/6/1337.full
9. Chanturia, G., Birdsell, D.N., Kekelidze, M., Zhgenti, E., Babuadze, G., Tsertsvadze, N., Tsanava, S., Imnadze, P., Beckstrom-Sternberg, S.M., Beckstrom-Sternberg, J.S., Champion, M.D., Sinari, S., Gyuranecz, M., Farlow, J., Pettus, A.H., Kaufman, E.L., Busch, J.D., Pearson, T., Foster, J.T., Vogler, A.J., Wagner, D.M., Keim, P.: Phylogeography of Francisella tularensis subspecies holarctica from the country of Georgia. BMC microbiology **11**(1), 139 (2011). DOI 10.1186/1471-2180-11-139. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21682874&retmode=ref&cmd=prlinks
10. Consortium, T.H.M.P.: A framework for human microbiome research. Nature **486**(7402), 215–221 (2012). DOI 10.1038/nature11209. URL http://dx.doi.org/10.1038/nature11209
11. Egozcue, J.J., Barcelo-Vidal, C.: Elements of simplicial linear algebra and geometry. In: V. Pawlowsky-Glahn, A. Buccianti (eds.) Compositional Data Analysis, pp. 141–156. John Wiley & Sons (2011). DOI 10.1002/9781119976462.ch4. URL http://dx.doi.org/10.1002/9781119976462.ch4
12. Egozcue, J.J., Pawlowsky-Glahn, V.: Simplicial geometry for compositional data. Geological Society, London, Special Publications **264**(1), 145–159 (2006). URL http://sp.lyellcollection.org/content/264/1/145.short
13. Karve, T.M., Cheema, A.K.: Small Changes Huge Impact: The Role of Protein Posttranslational Modifications in Cellular Homeostasis and Disease. Journal of Amino Acids **2011**(2), 1–13 (2011). DOI 10.4061/2011/207691. URL http://www.hindawi.com/journals/jaa/2011/207691/
14. Li, H.: Microbiome, metagenomics, and high-dimensional compositional data analysis. Annual Review of Statistics and Its Application **2**(1), 73–94 (2015). DOI 10.1146/annurev-statistics-010814-020351. URL http://dx.doi.org/10.1146/annurev-statistics-010814-020351
15. Lovell, D., Müller, W., Taylor, J., Zwart, A.: Proportions, percentages, ppm: do the molecular biosciences treat compositional data right. In: V. Pawlowsky-Glahn, A. Buccianti (eds.) Compositional Data Analysis. John Wiley & Sons (2011). URL http://books.google.com/books?hl=en&lr=&id=Ggpj3QeDoKQC&oi=fnd&pg=PT215&dq=Proportions+Percentages+PPM+Do+the+Molecular+BiosciencesTreat+Compositional+Data+Right&ots=cII3kxnfSb&sig=icwOFojg2zPXj2WPUj9IQ2K4MCk
16. Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A.: Dealing with zeros. In: V. Pawlowsky-Glahn, A. Buccianti (eds.) Compositional Data Analysis, pp. 43–58. John Wiley & Sons (2011). DOI 10.1002/9781119976462.ch4. URL http://dx.doi.org/10.1002/9781119976462.ch4
17. Nagumo, K., Tanaka, M., Chuang, V.T.G., Setoyama, H., Watanabe, H., Yamada, N., Kubota, K., Tanaka, M., Matsushita, K., Yoshida, A., Jinnouchi, H., Anraku, M., Kadowaki, D., Ishima,

Y., Sasaki, Y., Otagiri, M., Maruyama, T.: Cys34-Cysteinylated Human Serum Albumin Is a Sensitive Plasma Marker in Oxidative Stress-Related Chronic Diseases. PloS one **9**(1), e85,216–9 (2014). DOI 10.1371/journal.pone.0085216. URL http://dx.plos.org/10.1371/journal.pone.0085216

18. Nelson, R.W., Krone, J.R., Bieber, A.L., Williams, P.: Mass-Spectrometric Immunoassay. Analytical Chemistry **67**(7), 1153–1158 (1995). DOI 10.1021/ac00103a003. URL http://pubs.acs.org/doi/abs/10.1021/ac00103a003
19. PALLEN, M.J.: Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. Parasitology **141**(14), 1856–1862 (2014). DOI 10.1017/S0031182014000134. URL http://www.journals.cambridge.org/abstract_S0031182014000134
20. Pawlowsky-Glahn, V., Buccianti, A.: Compositional Data Analysis. Theory and Applications. John Wiley & Sons (2011). URL http://books.google.com/books?id=Ggpj3QeDoKQC&printsec=frontcover&dq=intitle:Compositional+Data+Analysis+Theory+and+Applications&hl=&cd=1&source=gbs_api
21. Pawlowsky-Glahn, V., Egozcue, J.J.: Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment **15**(5), 384–398 (2001). DOI 10.1007/s004770100077. URL http://link.springer.com/10.1007/s004770100077
22. Peleg, S., Sananbenesi, F., Zovoilis, A., Burkhardt, S., Bahari-javan, S., Agis-Balboa, R.C., Cota, P., Wittnam, J.L., Gogol-Doering, A., Opitz, L., Salinas-Riester, G., Dettenhofer, M., Kang, H., Farinelli, L., Chen, W., Fischer, A.: Altered Histone Acetylation Is Associated with Age-Dependent Memory Impairment in Mice. Science **328**(5979), 753–756 (2010). DOI 10.1126/science.1186088. URL http://www.sciencemag.org/content/328/5979/753.full
23. Teeling, H., Glockner, F.O.: Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective. Briefings in Bioinformatics **13**(6), 728–742 (2012). DOI 10.1093/bib/bbs039. URL http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbs039
24. Thomas, T., Gilbert, J., Meyer, F.: Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation **2**(1), 3 (2012). DOI 10.1186/2042-5783-2-3. URL http://www.microbialinformaticsj.com/content/2/1/3
25. Trenchevska, O., Schaab, M.R., Nelson, R.W., Nedelkov, D.: Development of multiplex mass spectrometric immunoassay for detection and quantification of apolipoproteins C-I, C-II, C-III and their proteoforms. Methods pp. 1–7 (2015). DOI 10.1016/j.ymeth.2015.02.020. URL http://dx.doi.org/10.1016/j.ymeth.2015.02.020
26. Vos, F.E., Schollum, J.B., Walker, R.J.: Glycated albumin is the preferred marker for assessing glycaemic control in advanced chronic kidney disease. Clinical Kidney Journal **4**(6), 368–375 (2011). DOI 10.1093/ndtplus/sfr140. URL http://ckj.oxfordjournals.org/cgi/doi/10.1093/ndtplus/sfr140
27. Walsh, C.T., Tsodikova, S.G.: Protein posttranslational modifications: the chemistry of proteome diversifications. Angewandte Chemie International Edition in English (2005). URL http://onlinelibrary.wiley.com/doi/10.1002/anie.200501023/full