

THE ANALYSIS OF HUMAN SERUM ALBUMIN
PROTEOFORMS USING COMPOSITIONAL FRAMEWORK

by
Shripad Sinari

Copyright © Shripad Sinari

A thesis Submitted to the Faculty of the
MEL AND ENID ZUCKERMAN COLLEGE OF PUBLIC
HEALTH

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE
WITH A MAJOR IN BIostatISTICS

In the Graduate College
THE UNIVERSITY OF ARIZONA

2017

STATEMENT BY AUTHOR

The thesis titled **The analysis of human serum albumin proteoforms using compositional framework** prepared by **Shripad Sinari** has been submitted in partial fulfillment of requirements of a master's degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: _____
SHRIPAD SINARI

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

_____	December 01, 2016
Dean Billheimer	Date
Professor of Biostatistics	

ACKNOWLEDGMENTS

I am grateful to my advisor, Dr. Dean Billheimer, for supporting me as a full time statistician with The Statistics Consulting Laboratory while I pursued the Master's degree. His mentoring has been valuable in advancing my understanding of statistics. Lastly, I am thankful to him for introducing me to this wonderful subject of compositional data analysis.

I am thankful to all the team members on the NIDDK grant "Team Approach to Translate Novel Biomarkers for Diabetes" for their co-operation on the project. In particular, I would like to thank Dr. Olgica Trenchevska, Dr. Chad Borges, Dr. Dobrin Nedelkov and Dr. Peter Reaven for their patience in answering my questions and many valuable insights, suggestions and discussions during the project, a small part of which forms the content of this thesis. Acknowledgment is also due to National Institutes of Health (NIH) for supporting the work presented in this thesis under awards numbered R24DK090958-01A1 and P30ES006694.

Many thanks go to my committee members Dr. Edward Bedrick and Dr. Chengcheng Hu who agreed to serve on my committee.

Finally, to my wife Shama Ajgaonkar for her encouragement, support and patience.

DEDICATION

To my father.

TABLE OF CONTENTS

LIST OF TABLES	6
LIST OF FIGURES	7
ABSTRACT	8
CHAPTER 1. INTRODUCTION	9
CHAPTER 2. A SYNOPSIS ON COMPOSITIONAL FRAMEWORK	11
CHAPTER 3. MSIA AND ALBUMIN PROTEOFORMS	17
3.1. Normalization of proteomic measurements as compositions	19
3.2. Interpretation of principal component analysis	19
3.3. Relative variation biplot	20
3.4. Results without the unit sum constraint	25
CHAPTER 4. INFERENCE ANALYSIS WITH ALBUMIN COMPOSITIONS	28
4.1. Visualization of the regression estimate	30
CHAPTER 5. SOME REMARKS ON THE MSIA PROTEOFORM ANALYSIS	34
CHAPTER 6. GENERALITY OF THE COMPOSITIONAL FRAMEWORK	35
CHAPTER 7. THE PROBLEM OF ESSENTIAL ZEROS	36
CHAPTER 8. CONCLUSIONS	37
REFERENCES	38

LIST OF TABLES

TABLE 3.1.	Table of raw peak areas of a small subset of the albumin data.	17
TABLE 3.2.	Table of compositions of the small subset of the albumin data.	18
TABLE 3.3.	Table of centered log ratios of the small subset of the albumin data. . .	18
TABLE 3.4.	Table of patients with negative or non-negative loading on first principal component (PC1) by their CKD status. The values correspond to Form Biplot (Figure 3.2).	22
TABLE 3.5.	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. These are 9 separate simple linear regression models each with a single proteoform as the explanatory variable.	22
TABLE 3.6.	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. These are 9 separate simple linear regression models each with a single proteoform as the explanatory variable.	25
TABLE 4.1.	Estimate of ξ and γ from a multivariate model with additive log ratio of human serum albumin proteoforms as a response.	29
TABLE 4.2.	Estimate of deviations of γ from the identity composition O^8 . The rows are sorted by the order of the deviations. This shows that GFR is associated positively with wild type and negatively with increase in the relative abundance of cysteinylated variants. This is consistent with the conclusions of exploratory analysis done in previous chapter.	30

LIST OF FIGURES

- FIGURE 3.1. **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. 23
- FIGURE 3.2. **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. 24
- FIGURE 3.3. **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 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. 26
- FIGURE 4.1. Ternary diagram of amalgamated data. The points are samples colored by their Chronic Kidney disease (CKD) status as classified from the measurement of their glomerular filtration rate (GFR). The albumin proteoform values are amalgamated. The variables "wt.cys.gly", "wt.cys.gly.gly", "wt.cys" and "des.DA.cys" are amalgamated into the "cys" variable. The variables "wt.gly.gly" and "wt.gly" are amalgamated into the "gly" variable. The variables "wt", "des.DA.cys" and "des.D" are amalgamated into the "wt" variable. The grey dot represents the origin (O^2) of the simplex. The solid black triangle, the estimate of GFR for every 100 mL/min increase (γ^{100}). The estimate γ is from a multivariate model using the additive log ratio of compositions as response and centered GFR as a covariate (see equation 4.3). The blue contour represents the 95% confidence region for γ^{100} . The dotted lines partition the triangle into 3 regions. The relative abundance of the amalgamated proteoform represented by the vertex is the highest for points in that region. . . 32

ABSTRACT

It has been known since the days of Karl Pearson that ratios of pairwise independent random variables are correlated [23]. However, recognition of the unit sum constraint and hence appropriate methods for analysis of relative abundance have been slow to emerge [4]. Analysis of the *relative abundance* of multiple components is a characteristic of compositional data. In this thesis, we demonstrate that the compositional data analysis framework is ideally suited to exploring and analyzing the relative abundance of proteoforms measured using Mass Spectrometric Immuno Assays (MSIA). We will introduce basic concepts of compositional data and associated analysis methods. 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.

Chapter 1

INTRODUCTION

Karl Pearson [23] used the relative ratios of human bones to illustrate the issue of spurious correlation. Design or measurement procedures in several modern scientific analyses lead to relative abundance data. Examples of such data are output from high-throughput omics technologies such as Mass Spectrometric Immuno Assays (MSIA), RNA-Seq and metagenomics. Compositional data analysis is a robust framework developed to address issues in inference arising due to this induced correlation in relative abundance data [1]. Particularly, it 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 the while respecting the structural constraints inherent in the observed data. In this thesis, we will show how the MSIA compositions whose components are proteoforms can be analyzed using the compositional framework.

Proteoforms are post translational modifications of proteins giving rise to new functional capabilities or regulation of the cellular environment [29, 14]. Some of these proteoforms have been implicated in diseases such as cancer [9] and age related dementia [24]. Identifying these proteoforms with sensitivity and specificity has been a challenge, especially when the abundance is low. Mass Spectrometry 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. All proteoforms that differ in their mass to charge ratio are captured using a single antibody. These proteoforms are then eluted directly on a mass spectrometer tip along with an aliquot of matrix solution. Proteoforms are then detected at their unique mass to charge ratio in MALDI time-of-flight mass spectrometry. This results in the mass spectra whose area under the curve are the raw values of relative signals of the proteoforms. Nelson, et al. [19]

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.

We begin by introducing the elements of compositional framework essential to our analysis of MSIA compositions. We also discuss the normalization scheme provided by the compositional structure. We illustrate use of this framework by exploring MSIA measurements of albumin proteoforms from patients with Type 2 diabetes mellitus. Many of these patients also have impaired renal function, and we explore and infer proteoform differences associated with chronic kidney disease (CKD). We conclude with remarks on our analysis of albumin proteoforms, and discuss emerging compositional data applications in genomics.

Chapter 2

A SYNOPSIS ON COMPOSITIONAL FRAMEWORK

A compositional data matrix may be given by:

$$M = (V_1, V_2, \dots, V_D) = (S_1, S_2, \dots, S_n)'$$

where $V_j; j = 1, 2, \dots, D$ are the components forming the columns of M and $S_i; i = 1, 2, \dots, n$ are the samples giving the rows of M . Assuming S_i to sum to 1 with non-zero components, we can see that each row, S_i , of M is a point of \mathcal{S}^d , where

$$\mathcal{S}^d = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_D)' : x_i > 0 \ (i = 1, 2, \dots, D), \sum_{i=1}^D x_i = 1 \right\} \quad (2.1)$$

Here $d = D - 1$ and $\mathbf{x} = (x_1, x_2, \dots, x_D)'$ denotes a D dimensional column vector in \mathbb{R}^D . Note that the dimension of the simplex is $d = D - 1$. We will use the convention $d = D - 1$ in the rest of the thesis. The unit sum constraint is critical and induces correlation between the parts of the composition. Thus an important assumption in applying this framework is that the relative proportions and their covariance is of interest irrespective of the total sum of the composition, which is normalized to 1. The following two principles in compositional framework ensure consistency in inference and compliance with the aforementioned requirements:

1. Scale invariance, and
2. Sub compositional coherence which means the inference from a subset of parts should be consistent with the inference from the full composition.

Modeling the data using logistic normal distributions allows for the lack of independence due to the unit sum constraint at the same time providing a sub compositionally coherent model [1].

To understand how this is done, consider an element $x = (x_1, x_2, \dots, x_D)' \in \mathcal{S}^d$. Following Aitchison [1] the additive log ratio transform is defined as:

$$\begin{aligned} \phi : \mathcal{S}^d &\rightarrow \mathbb{R}^d \\ x &\mapsto \left(\log\left(\frac{x_1}{x_D}\right), \log\left(\frac{x_2}{x_D}\right), \dots, \log\left(\frac{x_d}{x_D}\right) \right)' \end{aligned} \quad (2.2)$$

Given ϕ and a normal distribution

$$f' : \mathbb{R}^d \rightarrow \mathbb{R}$$

defined by density

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

We can define

$$f = f' \circ \phi : \mathcal{S}^d \rightarrow \mathbb{R}$$

the logistic normal density function by:

$$f(x|\mu, \Sigma) = (2\pi)^{-1/2} |\Sigma|^{-1/2} \left(\frac{1}{\prod_{i=1}^D x_i} \right) \exp \left[-\frac{1}{2} (\phi(x) - \mu)' \Sigma^{-1} (\phi(x) - \mu) \right] \quad (2.3)$$

where μ is the location parameter in \mathbb{R}^d and Σ is the $d \times d$ variance-covariance matrix. In the following, we will denote this d dimensional logistic normal distribution by \mathcal{LN}_d and A' will denote the transpose if A is a matrix.

The space \mathcal{S}^d can be given the structure of a vector space. Let C denote the *closure* operator on \mathbb{R}^{D*} , the space of all non-zero D dimensional vectors, which normalizes a non-zero vector to the unit sum simplex.

$$C : \mathbb{R}^{D^*} \rightarrow \mathbb{R}^{D^*}$$

$$x \mapsto \left(\frac{x_1}{\sum_{i=1}^D x_i}, \frac{x_2}{\sum_{i=1}^D x_i}, \dots, \frac{x_D}{\sum_{i=1}^D x_i} \right)' \quad (2.4)$$

For any two elements $x = (x_1, x_2, \dots, x_D)', y = (y_1, y_2, \dots, y_D)' \in \mathcal{S}^d$ and $\alpha \in \mathbb{R}$, we then define two operations on \mathcal{S}^d given by:

$$x \oplus y = C((x_1 \cdot y_1, x_2 \cdot y_2, \dots, x_D \cdot y_D)') \quad (2.5)$$

$$\alpha \odot x = C((x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha)') \quad (2.6)$$

The operations in equations (2.5) and (2.6) are called the perturbation and power operators, respectively. With perturbation operator as addition and power operator as the scalar multiplication, \mathcal{S}^d acquires the structure of a d dimensional vector space.

The linear transformation \log defined on \mathcal{S}^d by

$$\log : \mathcal{S}^d \rightarrow \mathbb{R}^D$$

$$x \mapsto (\log(x_1), \log(x_2), \dots, \log(x_D))' \quad (2.7)$$

maps the simplex \mathcal{S}^d into \mathbb{R}^D . Let 1_D represent the vector $(1, 1, \dots, 1)' \in \mathbb{R}^D$ and $\mathbb{1}_D$ be the one dimensional subspace of \mathbb{R}^D generated by 1_D . This gives an orthogonal decomposition of \mathbb{R}^D :

$$\mathbb{R}^D = \mathbb{1}_D \oplus \mathbb{1}_D^c \quad (2.8)$$

Notice the space $\mathbb{1}_D^c$ consists of vectors in \mathbb{R}^D with mean 0 and is isomorphic to \mathbb{R}^d as a vector space. Thus we get an invertible linear map \mathcal{L} on \mathcal{S}^d given by,

$$\mathcal{L} : \mathcal{S}^d \rightarrow \mathbb{1}_D^c$$

$$x \mapsto \left(\log\left(\frac{x_1}{g_D(x)}\right), \log\left(\frac{x_2}{g_D(x)}\right), \dots, \log\left(\frac{x_D}{g_D(x)}\right) \right)' \quad (2.9)$$

where $g_D(x)$ is the geometric mean of the components of the vector $x \in \mathcal{S}^d$. This transformation is called the *centered log ratio*. Note that $\log(g_D(x))$ is the mean of the vector $(\log(x_1), \log(x_2), \dots, \log(x_D)) \in \mathbb{R}^D$ and hence the name of the transform.

Using this transform one can introduce a metric on \mathcal{S}^d given by

$$d(x, y) = \left[\sum_{i=1}^D \left(\log\left(\frac{x_i}{g_D(x)}\right) - \log\left(\frac{y_i}{g_D(y)}\right) \right)^2 \right]^{1/2} \quad (2.10)$$

for any $x, y \in \mathcal{S}^d$. This is the composition (in the sense of a function) of the usual euclidean metric on $\mathbb{1}_D^c$ induced from \mathbb{R}^D with the centered log ratio. This makes \mathcal{S}^d a complete inner product space, also called a *Hilbert space* [22, 7]. The associated inner product is given by the matrix corresponding to the projection of \mathbb{R}^D on $\mathbb{1}_D^c$, namely,

$$I_D - \frac{1}{D} \mathbf{1}_D \mathbf{1}'_D \quad (2.11)$$

where I_D is the $D \times D$ identity matrix and $\mathbf{1}_D$ is the $D \times 1$ matrix with all components equal to 1. We call the metric in equation (2.10) the Aitchison distance. By construction, \mathcal{S}^d is isomorphic to $\mathbb{1}_D^c$ as a Hilbert space.

The additive log ratio above is defined with respect to the last component as a reference component. However, a similar definition can be made by taking any component as the reference component. The additive log ratios thus defined as well as the centered log ratio give equivalent co-ordinate systems on the space \mathcal{S}^d . Additional equivalent co-ordinates on \mathcal{S}^d as well as the equivalence of the metrics are discussed in Egozcue et al. [12]. The logistic normal distribution defined by two different choices of additive log ratio are equivalent. The equivalence of metric with centered log ratio shows that the logistic normal is compatible with the Hilbert space structure of \mathcal{S}^d similar to the compatibility of the normal distribution with the Euclidean geometry of \mathbb{R}^d , i.e. the logistic normal forms a location-scale family with respect to the perturbation and power operations.

Thus we transform the simplex induced by the unit sum constraint into a sample space isomorphic to \mathbb{R}^d as a Hilbert space. The distributions such as the logistic normal which are

compatible with this new Hilbert space structure then allow us to analyze the data with the usual tools of multivariate analysis. In this sense, the logistic normal in Aitchison geometry is the counterpart of the normal theory in the Euclidean space.

We now show how the logistic normal follows the principle of subcompositional coherence.

In terms of the data matrix M , a subcomposition means a sub-matrix of M containing all its rows, but only a subset of its columns. If $U = \{i_1, i_2, \dots, i_n\}$ is such a subset then the matrix corresponding to this subcomposition is $M_U = (V_{i_1}, V_{i_2}, \dots, V_{i_n})$. If $x = (x_{i_1}, x_{i_2}, \dots, x_{i_n})$ is a row of M_U then the result of the closure operation is

$$C((x_{i_1}, x_{i_2}, \dots, x_{i_n})) = \left(\frac{x_{i_1}}{1 - \sum_{i \notin U} x_i}, \frac{x_{i_2}}{1 - \sum_{i \notin U} x_i}, \dots, \frac{x_{i_n}}{1 - \sum_{i \notin U} x_i} \right) \quad (2.12)$$

Denote

$$S_i^U = C(S_i) \quad (2.13)$$

The matrix $M_{C(U)} = \frac{1}{1 - \sum_{i \notin U} x_i} (V_u)_{u \in U} = (S_i^U)'$ is the data matrix of the subcomposition represented by U .

Note that if P is the $n \times D$ projection matrix given by

$$P = (p_{lk})$$

where

$$p_{lk} = \begin{cases} 1 & \text{if } k = i_l \text{ and } i_l \in U \\ 0 & \text{otherwise} \end{cases}$$

then

$$S_x^U = \frac{1}{1 - \sum_{i \notin U} x_i} P(x)$$

thus

$$M_{C(U)} = M(\alpha P') \quad \text{where } \alpha = \frac{1}{1 - \sum_{i \notin U} x_i}$$

Scale invariance implies that the distribution induced by P , on the simplex defined by components of U , coincides with the distribution $\mathcal{L}\mathcal{N}_{n-1}(P(\mu), P'\Sigma P)$. Thus by defining the logistic normal $\mathcal{L}\mathcal{N}_{n-1}(P(\mu), P'\Sigma P)$ as a model for $M_{C(U)}$, any inference procedure that relies on the properties of the logistic normal is assured of satisfying the principle of sub-compositional coherence [6].

We now reveal this structure in our data of proteoform abundance measured using MSIA. A comprehensive collection of analytical techniques and applications of compositional framework is available in the book Pawlowsky-Glahn et al. [21].

Chapter 3

MSIA AND ALBUMIN PROTEOFORMS

Our MSIA measurements comprise albumin proteoforms from a cross-sectional study of 283 patients with Type 2 diabetes mellitus. Glycosylation [28] and cysteinylolation [18] 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 (3.1) 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.

Table 3.1: 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

The most abundant form is called wild type 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 wild type protein. The data matrix consists of 283 rows and 9 columns. Each row being a composition and thus a point in \mathcal{S}^8 which is the space of 9 part composition given by

$$\mathcal{S}^8 = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_9)' : x_i > 0 (i = 1, 2, \dots, 9), \sum_{i=1}^9 x_i = 1 \right\} \quad (3.1)$$

This is an 8 dimensional simplex embedded in the 9 dimensional real vector space \mathbb{R}^9 .

Table (3.2) gives the compositions formed from the data subset shown in Table (3.1).

Table 3.2: 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 \mathcal{S}^8 is a Hilbert space with a metric defined by equation (2.10).

Here $D = 9$ and $g_9(x)$ is the geometric mean of vector $x \in \mathcal{S}^8$. In our analysis we will use the centered log ratio transformation which is given by equation (2.9).

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 (3.3) gives the centered log ratios of compositions from Table (3.2). Note that centered log-ratios now sum to zero for each sample.

Table 3.3: 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 log-ratio transform (equation 2.2) may be more useful. Additive log-ratio transform

is also useful in parameter estimation in linear models when all proteoforms are included as a multivariate outcome. We illustrate below the use of this transform in our inferential analysis on the association of the albumin proteoforms with CKD (see Chapter 4). Standard multivariate methods, as well as multiple regression techniques, can now be applied to these appropriately transformed data.

3.1 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.

When a convenient reference standard exists, it may be included in the MSIA assay [19, 27]. 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 normalization scheme with noticeable reduction in the total variability of measurements. Formal comparison measures of variability are as yet unknown since the transformed values have one lower dimension than the raw values.

3.2 Interpretation of principal component 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 2.3). The centered log ratio transformed composition gives an isotropic invariant covariance structure from which a measure of total variability of a composition can be expressed as

$$\sum_{i=1}^9 \text{var} \left[\log \left(\frac{x_i}{g_9(x)} \right) \right] \quad (3.2)$$

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.

3.3 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 (2.10).

Figures (3.1) and (3.2) 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 (see Table 3.4). The plots also shows that higher proportion of wild type 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 3.1). The length of the link is approximately 0.293 whereas the actual standard deviation of the log-ratio is 0.283.

Aitchison [5] provides a good introduction and insights into numerous useful properties of the relative variation biplots and proves the equivalence of the biplots under various coordinate 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. A multivariate analysis is presented in Chapter 4. Each row of Table (3.5) gives the coefficients of the linear regression of the proteoform with GFR. Except for the double glycosylated wildtype albumin, all other forms of modified forms show a strong association with GFR.

Table 3.4: Table of patients with negative or non-negative loading on first principal component (PC1) by their CKD status. The values correspond to Form Biplot (Figure 3.2).

	PC1 Values		Total
	Negative	Non-negative	
CKD Status			
1	35 (32%)	75 (68%)	110 (39%)
2	83 (59%)	58 (41%)	141 (50%)
3	22 (69%)	10 (31%)	32 (11%)
Total	140 (49%)	143 (51%)	283

Table 3.5: 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. These are 9 separate simple linear regression models each with a single proteoform as the explanatory variable.

	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

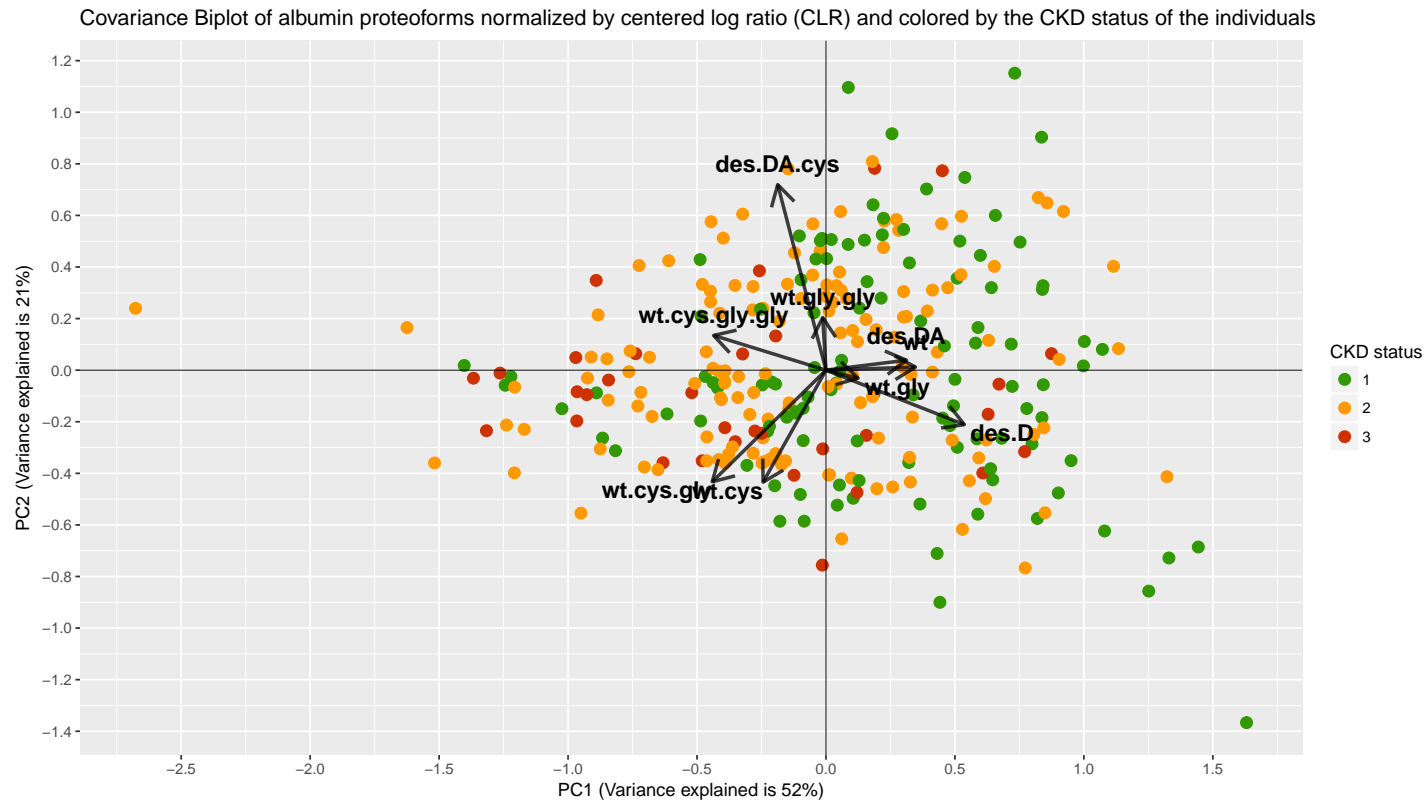


Figure 3.1: **Covariance Biplot** of individuals with Chronic Kidney disease (CKD) as classified from the measurement of their glomerular filtration rate (GFR). The albumin proteoform values are centered log ratio (CLR). The axis are the first two principal components with first component along the horizontal 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.

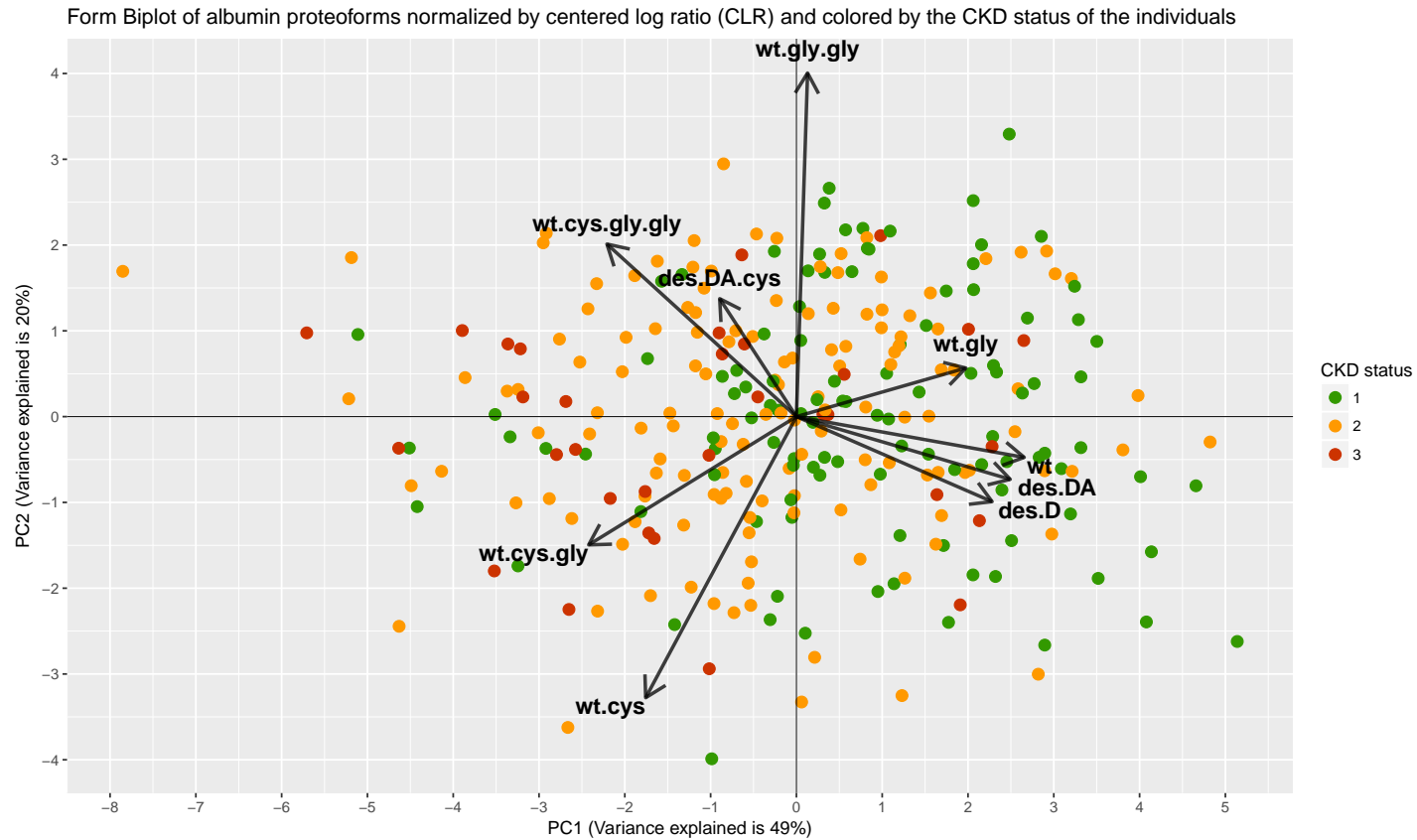


Figure 3.2: **Form Biplot** of individuals with Chronic Kidney disease (CKD) as classified from the measurement of their glomerular filtration rate (GFR). The albumin proteoform values are centered log ratio (CLR). The axis are the first two principal components with first component along the horizontal 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.

3.4 Results without the unit sum constraint

We now look at a similar analysis with log transformed raw peaks of the proteoforms. Figure (3.3) is a form biplot and Table (3.6) contains the results of regression of the log transformed raw peak areas with GFR.

Table 3.6: 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. These are 9 separate simple linear regression models each with a single proteoform as the explanatory variable.

	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

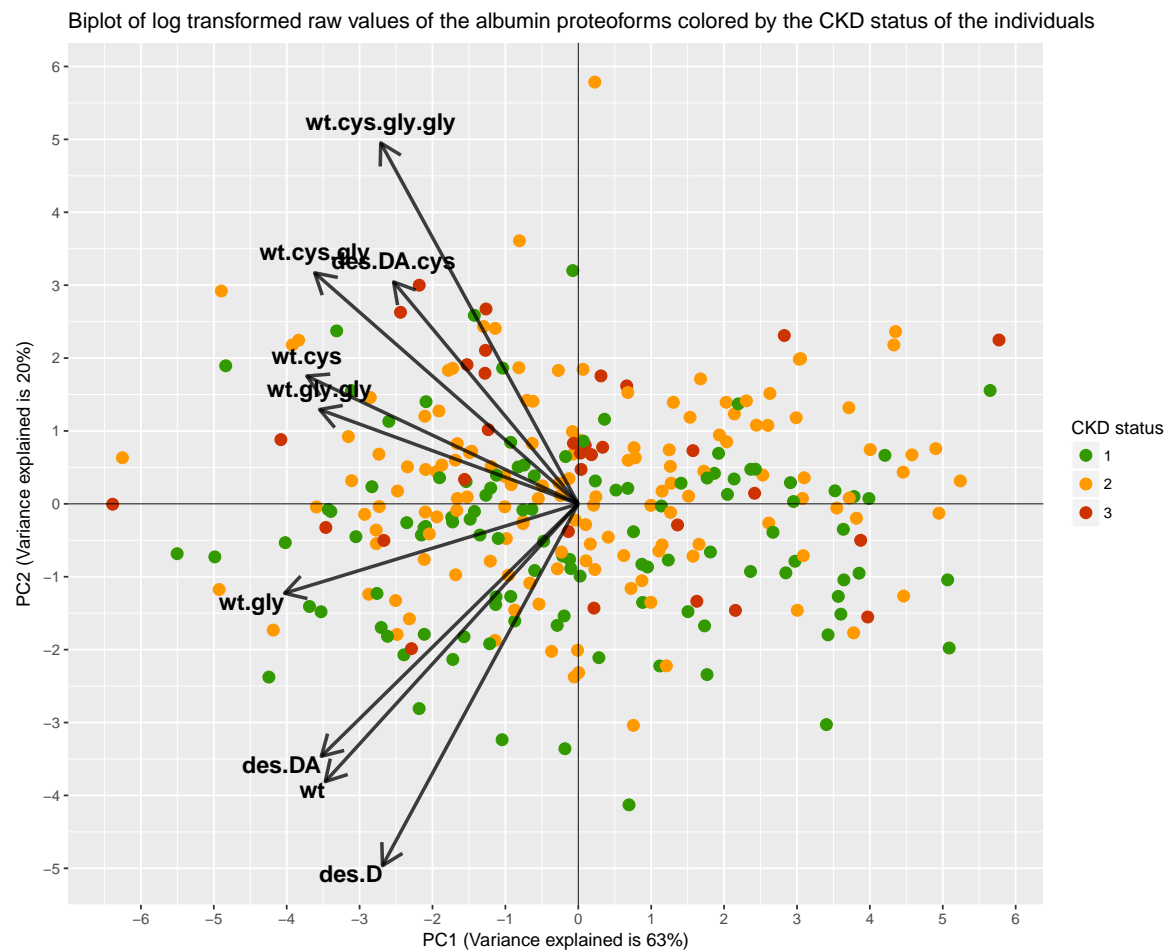


Figure 3.3: **Form Biplot** of individuals with Chronic Kidney disease (CKD) as classified from the measurement of their glomerular filtration rate (GFR). The albumin proteoform values are log transformed raw peak areas. The axis are the first two principal components with first component along the horizontal 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 (3.3) 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 3.5), and the whole composition, i.e. the principal component analysis using all proteoforms (Figure 3.2), 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 3.6) and the corresponding PCA (Figure 3.3) 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. [16], that use of compositional framework may not lead to dramatically different results across the board but the application of Aitchison distance (equation 2.10) provides more meaningful insights.

Chapter 4

INFERENCEAL ANALYSIS WITH ALBUMIN COMPOSITIONS

In order to understand the change in the nature of albumin compositions that are associated with CKD status, we perform a multivariate linear regression of the compositions with continuous measure of CKD status, namely GFR. Additive log ratio with respect to wild type albumin is used as a response. GFR has been measured in units of volume per time (mL/min). We will follow Billheimer et al. [7], in conducting and visualizing this multivariate linear regression analysis.

The regression equation is given by:

$$y = \beta_0 + \beta_1 (x_{\text{GFR}} - \bar{x}_{\text{GFR}}) + e \quad (4.1)$$

Here β_0 and β_1 are vectors in \mathbb{R}^8 and \bar{x}_{GFR} is the mean of GFR. The inverse of additive log ratio (equation 2.2) is:

$$\begin{aligned} \phi_{x_D}^{-1} : \mathbb{R}^d &\rightarrow \mathcal{S}^d \\ x &\mapsto \left(\frac{e^{x_1}}{1 + \sum_{i=1}^d e^{x_i}}, \frac{e^{x_2}}{1 + \sum_{i=1}^d e^{x_i}}, \dots, \frac{e^{x_d}}{1 + \sum_{i=1}^d e^{x_i}}, \frac{1}{1 + \sum_{i=1}^d e^{x_i}} \right)' \end{aligned} \quad (4.2)$$

Applying such an inverse transform ϕ_{wt}^{-1} , i.e. inverse of the additive log ratio with respect to wildtype, to equation (4.1) gives:

$$\psi = \xi \oplus \gamma^{u_{\text{GFR}}} \oplus \phi_{wt}^{-1}(e) \quad (4.3)$$

where

$$\psi = \phi_{wt}^{-1}(y)$$

$$\xi = \phi_{wt}^{-1}(\beta_0)$$

$$\gamma = \phi_{wt}^{-1}(\beta_1)$$

$$u_{\text{GFR}} = x_{\text{GFR}} - \bar{x}_{\text{GFR}}$$

Thus the equation (4.3) shows that ξ which is the location vector of the composition is perturbed by γ for each unit increase of GFR. This provides a compositional interpretation of our regression. The uncertainty in these estimates can be represented by the image under ϕ_{wt}^{-1} of the confidence regions estimated in \mathbb{R}^8 using the multivariate normal distribution. We display such a confidence region for γ computed on amalgamated data in Figure (4.1). The following table gives the estimates of ξ and γ from the multivariate regression model.

Table 4.1: Estimate of ξ and γ from a multivariate model with additive log ratio of human serum albumin proteoforms as a response.

	ξ	γ
wt.cys.gly	0.0576	0.1107
wt.cys.gly.gly	0.0182	0.1108
wt.cys	0.2467	0.1108
des.DA.cys	0.0245	0.1108
wt.gly.gly	0.0466	0.1111
wt.gly	0.1410	0.1112
des.DA	0.0258	0.1114
wt	0.4119	0.1115
des.D	0.0278	0.1116

The magnitude and direction of change is given by the difference of γ with origin of the \mathcal{S}^8 , namely the 9 dimensional vector $O^8 = (\frac{1}{9}, \dots, \frac{1}{9})$

Table 4.2: Estimate of deviations of γ from the identity composition O^8 . The rows are sorted by the order of the deviations. This shows that GFR is associated positively with wild type and negatively with increase in the relative abundance of cysteinylated variants. This is consistent with the conclusions of exploratory analysis done in previous chapter.

	γ/O^8
wt.cys.gly	0.997
wt.cys.gly.gly	0.997
wt.cys	0.997
des.DA.cys	0.998
wt.gly.gly	1.000
wt.gly	1.001
des.DA	1.003
wt	1.003
des.D	1.004

The deviations in Table (4.2) show that GFR is positively associated with wild type and decreases as relative abundance of the cysteinylated variants increases. Note that GFR is inversely related to CKD status. This analysis demonstrates the consistency in inference between different co-ordinate systems and across the dimensions (univariate vs multivariate) when using compositional framework.

4.1 Visualization of the regression estimate

The Table (4.2) shows that there are groups of albumin proteoforms that are influenced similarly by changes in GFR. There are 3 such groups, namely, cysteinylated variants, only glycated variants and the third consisting of wild type and truncated variants. Amalgamating the variants in these three groups gives us a convenient visualization scheme in terms of the ternary diagram [13].

The variables "wt.cys.gly", "wt.cys.gly.gly", "wt.cys" and "des.DA.cys" are amalgamated into the "cys" variable. The variables "wt.gly.gly" and "wt.gly" are amalgamated into the "gly" variable. The variables "wt", "des.DA.cys" and "des.D" are amalgamated into the "wt" variable. Multivariate regression is then performed using the additive log

ratio (equation (2.2) of "cys" and "gly" variables with respect to the "wt" and GFR as covariate. The estimate is mapped back to simplex, $\gamma = (0.332, 0.333, 0.334)$ for the cys, gly and wt variables, using the inverse additive log ratio (equation 4.2). This results in the ternary diagram (Figure 4.1) where we simultaneously visualize the sample compositions as well as the regression estimate. Here we will plot the estimate for 100 mL/min change for purposes of better illustration.

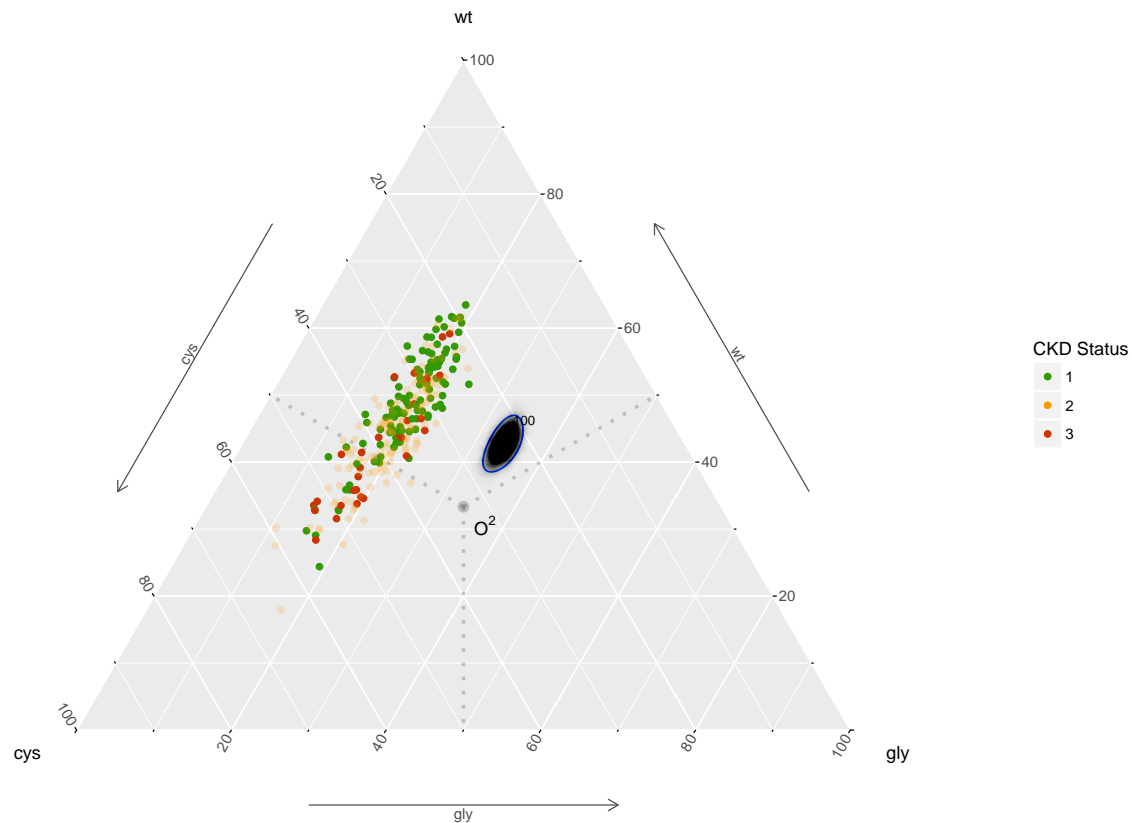


Figure 4.1: Ternary diagram of amalgamated data. The points are samples colored by their Chronic Kidney disease (CKD) status as classified from the measurement of their glomerular filtration rate (GFR). The albumin proteoform values are amalgamated. The variables "wt.cys.gly", "wt.cys.gly.gly", "wt.cys" and "des.DA.cys" are amalgamated into the "cys" variable. The variables "wt.gly.gly" and "wt.gly" are amalgamated into the "gly" variable. The variables "wt", "des.DA.cys" and "des.D" are amalgamated into the "wt" variable. The grey dot represents the origin (O^2) of the simplex. The solid black triangle, the estimate of GFR for every 100 mL/min increase (γ^{100}). The estimate γ is from a multivariate model using the additive log ratio of compositions as response and centered GFR as a covariate (see equation 4.3). The blue contour represents the 95% confidence region for γ^{100} . The dotted lines partition the triangle into 3 regions. The relative abundance of the amalgamated proteoform represented by the vertex is the highest for points in that region.

The ternary diagram (Figure 4.1) shows that for every 100 mL/min of increase in GFR, the relative ratio of cysteinylated variants decrease, that of only glycated variants have little change and the overall relative ratio of the wild type forms increase, consistent with previous analyses.

Chapter 5

SOME REMARKS ON THE MSIA PROTEOFORM ANALYSIS

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 constraint into account. Recently, Borges et al. [8] 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. In the albumin proteoforms this means that the variability such as batch effects due to antibody used for immuno affinity capture is normalized. In general, all non-proteoform specific variability is reduced.

Chapter 6

GENERALITY OF THE COMPOSITIONAL FRAMEWORK

Gene expression as measured in RNA-Seq is a relative abundance of the transcript counts. The total counts are constrained by the sequencing capacity. Thus a gene transcript expressed at the same level, i.e. no up or down regulation between treatments, may yield different counts in the two treatments depending on the relative expression of other genes, resulting in a composition. Housekeeping genes that are not expected to differ between treatments are often used to normalize such data, yielding relative ratios which are compositional by definition. Other normalization schemes such as RPKM or FPKM, result in compositions as well. See Lovell et al. [16] for detailed exposition on this issue. If the compositional nature is not taken into account, correlation induced due to the sum constraint can lead to misleading interpretation.

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 [25], common pathogens in clinical settings [20] or classification of the microbes into genus [11] and phylogeography [10]. Cell fractionation techniques or size selection similar to proteomics is often used in sample collection to create homogeneous populations of cells and enrichment of the target DNA [26]. These methods impose a compositional constraint due to scale invariance. Statistical analysis of such data can benefit from the use of compositional framework [15].

Chapter 7

THE PROBLEM OF ESSENTIAL ZEROS

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 compositional 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. [17]. An important point to note is that, samples with a part value as zero lie on a face of the simplex. The support of the d dimensional logistic normal distribution, \mathcal{LN}_d excludes these lower dimensional faces. Thus problems arise in extending the metric from the compositional Hilbert space to these faces.

Chapter 8

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.

REFERENCES

- [1] J Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177, 1982.
- [2] J Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983.
- [3] J Aitchison. Simplicial inference. In Marlos A G Viana and Donald St P Richards, editors, *Algebraic Methods in Statistics and Probability*, volume 287 of *Contemporary Mathematics*. American Mathematical Society, Providence, Rhode Island, 2001.
- [4] J Aitchison and J J Egozcue. Compositional data analysis: where are we and where should we be heading? *Mathematical Geology*, 37(7):829–850, 2005.
- [5] J Aitchison and M Greenacre. Biplots of compositional data. *Applied Statistics*, 51(4):375–392, 2002.
- [6] J Aitchison and S M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.
- [7] D Billheimer, P Guttorp, and W F Fagan. Statistical Interpretation of Species Composition. *Journal of the American Statistical Association*, 96(456):1205–1214, December 2001.
- [8] C R Borges, D S Rehder, S Jensen, M R Schaab, N D Sherma, H Yassine, B Nikolova, and C Breburda. Elevated plasma albumin and apolipoprotein A-I oxidation under suboptimal specimen storage conditions. *Molecular & cellular proteomics : MCP*, 13(7):1890–1899, July 2014.
- [9] R Chammas, J L Sonnenburg, N E Watson, T Tai, M G Farquhar, N M Varki, and A Varki. De-N-acetyl-gangliosides in humans: unusual subcellular distribution of a novel tumor antigen. *Cancer Research*, 59(6):1337–1346, March 1999.
- [10] G Chanturia, D N Birdsell, M Kekelidze, E Zhgenti, G Babuadze, N Tsertsvadze, S Tsanova, P Imnadze, S M Beckstrom-Sternberg, J S Beckstrom-Sternberg, M D Champion, S Sinari, M Gyuranecz, J Farlow, A H Pettus, E L Kaufman, J D Busch, T Pearson, J T Foster, A J Vogler, D M Wagner, and P Keim. Phylogeography of *Francisella tularensis* subspecies holarctica from the country of Georgia. *BMC microbiology*, 11(1):139, June 2011.
- [11] The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, June 2012.

- [12] J J Egozcue and C Barcelo-Vidal. Elements of simplicial linear algebra and geometry. In Vera Pawlowsky-Glahn and Antonella Buccianti, editors, *Compositional Data Analysis*, pages 141–156. John Wiley & Sons, 2011.
- [13] N Hamilton. *ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams*, 2016. R package version 2.1.1.
- [14] T M Karve and A K Cheema. Small Changes Huge Impact: The Role of Protein Post-translational Modifications in Cellular Homeostasis and Disease. *Journal of Amino Acids*, 2011(2):1–13, 2011.
- [15] H Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2(1):73–94, 2015.
- [16] D Lovell, W Müller, J Taylor, and A Zwart. Proportions, percentages, ppm: do the molecular biosciences treat compositional data right. In Vera Pawlowsky-Glahn and Antonella Buccianti, editors, *Compositional Data Analysis*. John Wiley & Sons, 2011.
- [17] J A Martín-Fernández, J Palarea-Albaladejo, and R A Olea. Dealing with zeros. In Vera Pawlowsky-Glahn and Antonella Buccianti, editors, *Compositional Data Analysis*, pages 43–58. John Wiley & Sons, 2011.
- [18] K Nagumo, M Tanaka, V Tuan Giam Chuang, H Setoyama, H Watanabe, N Yamada, K Kubota, M Tanaka, K Matsushita, A Yoshida, H Jinnouchi, M Anraku, D Kadowaki, Y Ishima, Y Sasaki, M Otagiri, and T Maruyama. Cys34-Cysteinylated Human Serum Albumin Is a Sensitive Plasma Marker in Oxidative Stress-Related Chronic Diseases. *PloS one*, 9(1):e85216–9, January 2014.
- [19] R W Nelson, J R Krone, A L Bieber, and P Williams. Mass-Spectrometric Immunoassay. *Analytical Chemistry*, 67(7):1153–1158, 1995.
- [20] M J Pallen. Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology*, 141(14):1856–1862, February 2014.
- [21] V Pawlowsky-Glahn and A Buccianti. *Compositional Data Analysis. Theory and Applications*. John Wiley & Sons, September 2011.
- [22] V Pawlowsky-Glahn and J J Egozcue. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5):384–398, October 2001.
- [23] K Pearson. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, 60(359-367):489–498, 1896.

- [24] S Peleg, F Sananbenesi, A Zovoilis, S Burkhardt, S Bahari-javan, R Carlos Agis-Balboa, P Cota, J L Wittnam, A Gogol-Doering, L Opitz, G Salinas-Riester, M Dettenhofer, H Kang, L Farinelli, W Chen, and A Fischer. Altered Histone Acetylation Is Associated with Age-Dependent Memory Impairment in Mice. *Science*, 328(5979):753–756, May 2010.
- [25] H Teeling and F O Glockner. Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. *Briefings in Bioinformatics*, 13(6):728–742, November 2012.
- [26] T Thomas, J Gilbert, and F Meyer. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1):3, February 2012.
- [27] O Trenchevska, M R Schaab, R W Nelson, and D Nedelkov. Development of multiplex mass spectrometric immunoassay for detection and quantification of apolipoproteins C-I, C-II, C-III and their proteoforms. *Methods*, pages 1–7, March 2015.
- [28] F E Vos, J B Schollum, and R J Walker. Glycated albumin is the preferred marker for assessing glycaemic control in advanced chronic kidney disease. *Clinical Kidney Journal*, 4(6):368–375, November 2011.
- [29] C T Walsh and S G Tsodikova. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angewandte Chemie International Edition in English*, 2005.