January 2015

# Assessing Inter-rater Agreement for Compositional Data

Ningning Chen
*Purdue University*

*Graduate School Form 30*
*Updated 1/15/2015*

**PURDUE UNIVERSITY**
**GRADUATE SCHOOL**
**Thesis/Dissertation Acceptance**

This is to certify that the thesis/dissertation prepared

By Ningning Chen

Entitled
Assessing Inter-rater Agreement for Compositional Data

For the degree of Doctor of Philosophy ▾

Is approved by the final examining committee:

Bruce Craig
Chair

Jun Xie

Dabao Zhang

Lingsong Zhang

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Bruce Craig

Approved by: Jun Xie 12/4/2015
Head of the Departmental Graduate Program Date

# ASSESSING INTER-RATER AGREEMENT FOR COMPOSITIONAL DATA

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Ningning Chen

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2015

Purdue University

West Lafayette, Indiana

# ACKNOWLEDGMENTS

I would like to first and foremost, thank my advisor, Dr. Bruce Craig, for his patient guidance and mentorship throughout my PhD research. Not only has there been a substantial positive influence on my work throughout his time as my advisor, but his wisdom, knowledge, and commitment to the highest standards has inspired and motivated me to continue to make a similar mark on our field. While I will be worse off upon the completion of this dissertation simply no longer having his continual wisdom at my disposal, I don't question my ability to be a successful statistician because I was able to learn so much from him over the past three years.

Besides my advisor, I would like to thank Dr. Sameera Wijayawardana, who is a senior research scientist and our collaborator from Eli Lilly & Company, for the sharing of his research interest and topic with me. This dissertation topic originated from one of Dr. Wijayawardana's research presentations at a joint Purdue/Eli Lilly day. He has provided valuable research discussion and resources to our collaborated work. The application of this work was partially funded by Eli Lilly & Company.

Thanks also goes to my other committee members, Dr. Jun Xie, Dr. Dabao Zhang, and Dr. Lingsong Zhang, for their friendly guidance and suggestions. Similarly, I'd like to express gratitude to all the faculty and staff members in the Department of Statistics for their support, as well as my fellow graduate students for their encouragement throughout my PhD studies. Specifically, I'd like to recognize Douglas Crabill and Barret Schloerke, who helped me optimize the complicated simulation programs and my online software application.

Last but not the least, I would like to thank my parents for their unconditional support and love. No matter the situation, whether it be obstacles in life or research, the one constant in my life has always been my parents. If anything, this dissertation is the culmination of your commitment to me and my life pursuits.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

Figure Page

Figure Page

# ABBREVIATIONS

| | |
|---|---|
| GIS | Geographic information system |
| PCA | Principal component analysis |
| MLE | Maximun likelihood estimator |
| MCD | Minimum covariance determinant |
| MAR | Missing at random |
| NMAR | Not missing at random |
| ICC | Intraclass Correlation Coefficient |
| CCC | Concordance Correlation Coefficient |
| CIV | Coefficient of inter-rater variability |
| ANOVA | Analysis of variance |
| REML | Restricted Maximum Likelihood |
| MCMC | Markov chain Monte Carlo |
| BC | Bhattacharyya Coefficient |
| IHC | Immunohistochemistry |
| CDF | Cumulative density function |
| HPD | Highest posterior density |

# ABSTRACT

Chen, Ningning PhD, Purdue University, December 2015. Assessing Inter-rater Agreement for Compositional Data. Major Professor: Bruce A. Craig.

Compositional data are non-negative vectors whose elements sum to one (e.g., [0.1, 0.5, 0.4]). This type of data occurs in many research areas where the relative magnitudes between the vector's elements are of primary interest. In this dissertation we propose novel methodology for assessing inter-rate agreement based on compositional data. This is needed because existing agreement measures either involve converting the vector to a univariate value, thereby losing information, or they fail to account for the sum-to-one restriction. We propose a novel Bayesian approach, enabled by Markov chain Monte Carlo, to investigate differences in the pattern of compositional vector scores. We extend our model to handle discrete compositional scores, comparisons involving more than two raters, and studies that involve replicate scores on the same subjects. Numerous simulation studies are used to demonstrate the validity of our model and the advantages of our approach. Both simulated data and a real scoring data set are analyzed to illustrate our method and compare it to traditional agreement indices. The application of this new methodology is focused on pathology, where pathologists rate immunohistochemistry (IHC) assays using compositional scores. To enhance the use of this methodology and help with the design of future agreement studies, an R Shiny package designed for the IHC agreement analysis is developed.

# CHAPTER 1. OVERVIEW OF COMPOSITIONAL DATA

## 1.1 Introduction

Compositional data are non-negative vectors whose elements sum to one. Because each vector sums to one (or 100%), the vectors carry only relative information. This type of data arises in numerous research areas. A geologist may describe samples of rock by the proportional makeup of different minerals. A demographer may describe cities in terms of their racial breakdowns. Lastly, a forest researcher may quantify patches of forest by the relative amount of woody plants, mosses, fungi, and flowering plants.

Inference using these data has primarily focused on the comparison of group means and developing classifiers to discriminate groups. For example, geologists compare the geochemical composition of rock and soil from different locations. They also classify rocks or soil samples based on their geochemical components. For the latter, the classifier may utilize all the elements in the compositional vector or only a subset of components. Studies like these help geologists better understand different rock formations and transformation processes throughout history.

For example, in order to elucidate the nature of the petrogenetic and tectonic processes that affected the Cenozoic volcanites in Hungary, Kovács et al. (2006) applied discrimination analysis to separate alkaline basalt from calc-alkaline rocks. The compositional separation disclosed for these two types of rocks provided quantitative interpretations of stratigraphical and petrographical processes. Another study, provided by Thomas and Aitchison (2006), used variably impure metamorphosed Scottish Dalradian limestones to help correlate and discriminate lithostratigraphical sequences. That is, the succession of strata or rock layers that can be recognized and defined based on the observable rock geochemical components.

In pathology, researchers classify tissue in terms of cell abnormalities and/or deviations in their rate of growth. To classify breast cancer, special antibodies that identify the HER2/neu protein are applied to breast tissue. The antibodies are fluorescently tagged so when they attach to HER2 proteins, the cells with the protein will fluorense. The test result is typically reported as the percent of cells in a breast tissue sample that fall in different staining intensity categories. Cancers that are HER2-positive have a large amount of HER2/neu protein, resulting in a high percent of strong intensity cells. A decision rule determines whether the breast tissue sample belongs to the HER2-positive, inconclusive, or HER2-negative group.

Animal habitat or resource selection studies are particular important in providing indications of the life history, physiology, and ethological traits of a focal species. Data are frequently collected with geographic information system (GIS) and individual radio-trackers. An individual's home range is described in terms of the relative proportions of different types of habitat. An individual's habitat use over a specific time frame is also described by the relative amount of time spent in each of these habitats. There is interest in comparing the means of available habitat to the means of focal individuals' habitat use, or comparing the means of habitat use across different groups of animals. Aebischer et al. (1993), for example, considered a paired comparison compositional analysis using two data sets: 13 radio-tagged Ring-necked Pheasants in Ireland (Robertson 1986) and 17 radio-tagged Gray Squirrels in United Kingdom (Kenward 1982).

Besides classification and the comparison of means, researchers are also interested in assessing how much variability in the vector components is explainable by other factors or covariates. Aitchison (1986) investigated the relationship between lake sediment compositions and water depth. In economics, compositional data analysis is used to study how consumers allocate their budgets or expenditures among available commodities, using exogenous variables (e.g., prices). Woodland (1979), Ronning (1992) and Fry et al. (1996, 2000, 2001) proposed different methods to estimate the shares of expenditures. Details of these approaches are discussed later.

Despite the fact that compositional data are widely used by researchers, the sum-to-one constraint is often ignored when analyzing these data. This has continued despite numerous warnings by researchers about using statistical methods designed for unconstrained multivariate data (Pearson, 1897; Chayes, 1971; Rock, 1988; Rollinson 1992). Since compositional vectors are normalized to one (or 100%), element dependencies and negative correlations are introduced. With many conventional statistical analyses, it is impossible to distinguish correlation induced by the sum-to-one constraint from the natural correlation among vector components. The latter is often the purpose of the analysis. Pearson (1897) used the term "spurious correlation" to describe the correlation between ratios of absolute measurements that arises as a consequence of using ratios. In Chayes's book (1971), he discusses the implications of conventional statistical analyses in great detail.

Another key analytic issue with compositional analysis is the occurrence of zero components. For example, when counting plant species within a forest site, it could happen that one species of plant exists but is not observed in the limited sampling area. Zero components can also happen in geochemical or biological studies when the percentage of a chemical is below some detection limit. Finally, zeros are also possible when components are "truly" missing. A type of plant, for example, may not exist at a site or a geochemical component may not exist in a type of rock.

Zeros in compositional vectors make the analysis more difficult because the common distributions used to describe compositional data do not accommodate zeros. To address the zero issue, one must first determine whether the zeros are true or rounded/censored. There are imputation methods to deal with rounded zeros in compositional data, including both nonparametric (Aitchison, 1986; Martín-Fernández, 2003) and parametric (Palarea-Albaladejo et al., 2007) replacement strategies. These imputations work under different assumptions. To address true zero issues, some hierarchical models have been developed (Aitchison and Kay, 2003; Bacon-Shone, 2008). Details of these procedures are discussed later in this chapter.

For my research, I consider a different type of inference involving compositional data. In many of the examples I've previously introduced, the compositional vectors were subjectively determined by researchers. This subjectivity can be problematic. When classifying tissue samples, for example, there is not just one pathologist looking at all the samples. An agreement among pathologists is pertinent for a consistent diagnosis. At this time, there is no methodology available that assesses agreement using the compositional score vectors directly. This work fills this gap. Before describing our approach, we first use the remainder of Chapter 1 to provide a review of compositional data and its terminology and properties. Then in Chapter 2, we discuss the statistical approaches to assess agreement. We describe our modeling approach in Chapter 3 and then follow this up with some simulation studies and the analysis of real data in Chapter 4. We conclude with a summary of our approach and future directions for research.

## 1.2 Basic Concepts

John Aitchison was the first statistician to publish a book on compositional data. In his book, *The Statistical Analysis of Compositional Data*, he defined a composition as follows:

**Definition 1.1** A vector, $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, is a **D-part composition** when all its elements are positive real numbers that sum to one.

Compositional vectors are different from standard vectors in $\mathbb{R}_+^D$ due to the constraint $\sum_{i=1}^{D} x_i = 1$.

**Definition 1.2** The sample space of compositional data is:

$$\mathbb{S}^D = \Big\{ \mathbf{x} = [x_1, x_2, \ldots, x_D] \Big| x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = 1 \Big\}.$$

$\mathbb{S}^D$ is a D-dimensional simplex embedded in D-dimensional positive real space $\mathbb{R}_+^D$ so we have the subset relationship: $\mathbb{S}^D \subset \mathbb{R}_+^D \subset \mathbb{R}^D$.

Sometimes, compositions are represented as $\mathbf{w} = [w_1, w_2, \ldots, w_D]$ with $w_1, w_2, \ldots, w_D \in \mathbb{R}_+$ and $\sum_{i=1}^{D} w_i = W$. If D-1 elements of the composition as well as the constant sum $W$ are known, this composition is completely determined. Aitchison, however, did not call this a composition.

**Definition 1.3** For any vector $\mathbf{w} \in \mathbb{R}_+^D$ ($w_i > 0$ for $i = 1, 2, \ldots, D$) where $W = \sum_{i=1}^{D} w_i$, the **closure of w** is defined as:

$$\mathcal{C}[w_1, w_2, \ldots, w_D] = \left[\frac{w_1}{W}, \frac{w_2}{W}, \ldots, \frac{w_D}{W}\right].$$

Given a vector $\mathbf{w} \in \mathbb{R}_+^D$, a D-part composition $\mathbf{x}$ is simply the closure of $\mathbf{w}$.

In practice, it is sometimes computationally difficult and/or unnecessary to include all possible components in an analysis. In these cases, analysis of a subcomposition is more attractive. For example, Carr (1981) investigated a set of 102 rock specimens and reported the data in terms of the relative weights of 10 oxides. Geologists, however, are more commonly interested in just a few oxides (e.g., CaO, $Na_2O$ and $K_2O$). We can use the closure of a vector with just those three oxides to form a subcomposition.

**Definition 1.4** For a D-part composition $\mathbf{x}$ and its subvector $\mathbf{x_s} = [x_{i_1}, x_{i_2}, \ldots, x_{i_s}]$, the **subcomposition of x** with $s$ parts is defined as $\mathcal{C}[\mathbf{x_s}]$, where $i_1, i_2, \ldots, i_s$ indicates the selected indices of $\mathbf{x}$.

Inference of compositional data is different from the analysis of vectors in real space because of the unit sum constraint. Any analytic approach that ignores the unit sum constraint may result in misleading results. Aitchison (1986) proposed that compositional data analysis must meet three principles:

(i) scale invariance.

(ii) permutation invariance.

(iii) subcompositional coherence.

Scale invariance can easily be derived from the closure function (Definition 1.3). If $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ and $\mathbf{y} = \lambda\mathbf{x}$, where $\lambda$ is a positive real number, then $\mathcal{C}[\mathbf{x}] = \mathcal{C}[\mathbf{y}]$. Permutation invariance means that any function applied to compositional data yields the same result regardless of the order of the components in the composition. Subcompositional coherence means that any results found by analyzing any subcomposition should be consistent with the results when analyzing the full compositions.

In additional to these three principles, Aitchison (1986) also introduced basic operations in the D-dimensional simplex that are the analogues to operations in $\mathbb{R}^D$.

**Definition 1.5** Assume $\mathbf{x}$ is a D-part composition and let $\mathbf{u} \in \mathbb{R}_+^D$. A **perturbation** is defined as

$$\mathbf{X} = \mathbf{u} \oplus \mathbf{x} = \mathcal{C}[u_1 x_1, u_2 x_2, \ldots, u_D x_D].$$

Without loss of generality, we can restrict the perturbation vector $\mathbf{u}$ to the simplex space $\mathbb{S}^D$. This is because of the scale invariance of the closure.

We call $\ominus\mathbf{x}$ the **inverse element of x** as it undoes a perturbation. It is defined to be $\ominus\mathbf{x} = \mathcal{C}[1/x_1, 1/x_2, \ldots, 1/x_D]$. Thus,

$$\mathbf{X} \ominus \mathbf{x} = \mathcal{C}[X_1/x_1, X_2/x_2, \ldots, X_D/x_D] = [u_1, u_2, \ldots, u_D].$$

If we perturb $\mathbf{x}$ with itself, then $\mathbf{x} \oplus \mathbf{x} = \mathcal{C}[x_1^2, x_2^2, \ldots, x_D^2]$. Similarly, $\mathbf{x} \oplus \mathbf{x} \oplus \mathbf{x} = \mathcal{C}[x_1^3, x_2^3, \ldots, x_D^3]$. This procedure, perturbing a composition by itself many times, is defined as powering. A more general powering definition is as follows:

**Definition 1.6** If $\mathbf{x}$ is a D-part composition and let $\mathbf{u} \in \mathbb{R}^D$, then the **powering operation** $\odot$ is

$$\mathbf{X} = \mathbf{u} \odot \mathbf{x} = \mathcal{C}[x_1^{u_1}, x_2^{u_2}, \ldots, x_D^{u_D}].$$

In addition to powering, there is also the inner product of two compositions.

**Definition 1.7** The **inner product** of two compositions $\mathbf{x}, \mathbf{y} \in \mathbb{S}^D$ is

$$\langle \mathbf{x}, \mathbf{y} \rangle_S = \sum_{i=1}^{D} \log \frac{x_i}{g(\mathbf{x})} \log \frac{y_i}{g(\mathbf{y})},$$

where $g(\cdot)$ is the geometric mean of the composition, i.e., $g(\mathbf{x}) = (x_1 x_2 \cdots x_D)^{1/D}$.

Accordingly, the **norm** and **distance** between two vectors in $\mathbb{S}^D$ are defined as

$$\|\mathbf{x}\|_S = \langle \mathbf{x}, \mathbf{x} \rangle = \sqrt{\sum_{i=1}^{D} \left( \log \frac{x_i}{g(\mathbf{x})} \right)^2}$$

and

$$\begin{aligned} d(\mathbf{x}, \mathbf{y})_S &= \|\mathbf{x} \ominus \mathbf{y}\| \\ &= \|\mathcal{C}[x_1/y_1, \ldots, x_D/y_D]\| \\ &= \left\{ \sum_{i=1}^{D} \left[ log \left( \frac{x_i/y_i}{g(\mathbf{x})/g(\mathbf{y})} \right) \right]^2 \right\}^{1/2} \\ &= \left\{ \sum_{i=1}^{D} \left[ log \left( \frac{x_i}{g(\mathbf{x})} \right) - log \left( \frac{y_i}{g(\mathbf{y})} \right) \right]^2 \right\}^{1/2}. \end{aligned}$$

The distance above is called the **Aitchison distance** and is shown to meet all three invariance principals for compositional data analysis.

## 1.3 Modeling Compositional Data

Because of the sum-to-one constraint, there are two common approaches to describe and model compositional data. We discuss each of these in this section.

### 1.3.1 Log-ratio transformations

Log-ratio transformations provide a way to connect compositions in the $\mathbb{S}^D$ simplex space with the more familiar multivariate analyses in the $\mathbb{R}^{D-1}$ Euclidean space. There are three popular transformations available. They are the

(i) Additive log-ratio transformation (Aitchson, 1986)

(ii) Centered log-ratio transformation (Aitchson, 1986)

(iii) Isometric log-ratio transformation (Egozcue et al., 2003)

**Definition 1.8** The **additive log-ratio transformation**, $\mathrm{alr}(\mathbf{x})$, is a one-to-one transformation from $\mathbf{x} \in \mathbb{S}^D$ to $\mathbf{y} \in \mathbb{R}^{D-1}$ where

$$y_i = \log(x_i/x_D) \qquad (i = 1, 2, \ldots, D-1).$$

Under this transformation, the $\mathbf{y}$ vector no longer has the unit-sum constraint. The divisor $x_D$ doesn't necessarily need to be the last component in the composition. Different choices of the divisor will result in different additive log-ratios. Since the transformed data are all in reference to the component used in the denominator, we have to be careful in choosing metrics applied to the alr transformed data and in interpretation. For instance, a naive Euclidean distance of alr transformed data is not permutation invariant and thus should not be used.

**Definition 1.9** The **centered log-ratio transformation**, $\mathrm{clr}(\mathbf{x})$, is a one-to-one transformation from $\mathbf{x} \in \mathbb{S}^D$ to $\mathbf{z}$ where

$$z_i = \log(x_i/g(\mathbf{x})) \qquad (i = 1, 2, \ldots, D),$$

and $g(\cdot)$ is the geometric mean function. The inner product and Aitchison distance between two compositions can be expressed in forms of the clr:

$$\|\mathbf{x}\|_S = \sum_{i=1}^{D} \big(\mathrm{clr}(x_i)\big)^2,$$

$$d(\mathbf{x}, \mathbf{y})_S = \left\{ \sum_{i=1}^{D} \Big[\mathrm{clr}(x_i) - \mathrm{clr}(y_i)\Big]^2 \right\}^{1/2}.$$

The clr transformation avoids the issue of choosing an arbitrary divisor as with the alr transformation and it is symmetric in the components (permutation invariant), but the $\mathbf{z}$ vector has a zero-sum constraint $\sum_{i=1}^{D} z_i = 0$. Because of this zero-sum constraint, the covariance matrix of $\mathbf{z}$ is singular and thus eliminates the use of standard multivariate methods. Furthermore, the clr transformation doesn't preserve subcompositional coherence, because the geometric mean of the parts of a subcomposition is not necessarily equal to that of the full composition.

**Definition 1.10** The **isometric log-ratio transformation**, ilr($\mathbf{x}$), is based on the choice of an orthonormal basis (in the Euclidean sense) on the hyperplane $\mathcal{H}$ : $z_1 + \cdots + z_D = 0$ in $\mathbb{R}^D$ that is formed by the clr transformation. Egozcue et al. (2003) suggested the basis

$$\mathbf{v}_i = \sqrt{\frac{i}{i+1}} \left( \frac{1}{i}, \ldots, \frac{1}{i}, -1, 0, \ldots, \right)' \qquad (i = 1, 2, \ldots, D-1).$$

Then $\mathbf{x} \in \mathbb{S}^D$ can be transformed to $\mathbf{y} \in \mathbb{R}^{D-1}$ as

$$y_i = \text{ilr}(\mathbf{x}) = \sqrt{\frac{i}{i+1}} \log \left[ \frac{g(\mathbf{x})}{x_{i+1}} \right] \qquad (i = 1, 2, \ldots, D-1),$$

where $g(\cdot)$ is the geometric mean function.

There is a relationship between clr and ilr that can be expressed as

$$\mathbf{z} = \mathbf{V}\mathbf{y},$$

where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ is the $D \times (D-1)$ matrix with orthonormal basis vectors on the hyperplane $\mathcal{H}$. The ilr preserves all the merits of clr with the additional advantage that it avoids the singularity introduced by clr. Thus standard multivariate procedures can be used straightaway. However, the interpretation in terms of the transformed components is difficult because ilr-coordinates refer to “mixtures” of components. The clr transformation preserves the direct one-to-one relationship between the components and the clr-coordinates.

### 1.3.2 Logistic Normal distribution

In order to describe the variability of observations in the simplex sample space, a parametric class of distributions on $\mathbb{S}^D$ has to be well-defined. Aitchison (1986) proposed using the Normal distribution to model the log-ratio transformed data. Let $\mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the D-dimensional Normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

**Definition 1.11** A D-part composition $\mathbf{x}$ follows the additive logistic Normal distribution $\mathcal{L}_D(\boldsymbol{\mu}, \Sigma)$ when $\mathbf{y} = \text{alr}(\mathbf{x}) \sim \mathcal{N}_{D-1}(\boldsymbol{\mu}, \Sigma)$. Under this distribution,

$$E[y_i] = E[\log(x_i/x_D)],$$

$$\Sigma = [\sigma_{i,j}] = \text{cov}[\log(x_i/x_D), \log(x_j/x_D)],$$

where $i, j = 1, \ldots, D-1$.

It is also possible to parametrize the logistic Normal class using the clr but this requires different specifications of the mean vector and the covariance matrix. With the clr transformation, the dimension of the covariance matrix of the logistic Normal increases from $(D-1) \times (D-1)$ to $D \times D$. To avoid this unnecessary complication in matrix specification with clr, all further discussion about logistic Normal distributions consider the alr transformation.

Aitchson (1986) proved that the logistic Normal distributions have many nice properties. For example,

(i) Every subcomposition of a logistic Normal composition has a logistic Normal distribution.

(ii) A conditional subcomposition also follows a logistic Normal distribution.

(iii) Logistic Normal distributions preserve all three principles of compositional analysis: scale invariance, permutation invariance, and subcompositional coherence.

Though the logistic Normal distribution maintains nice moment properties from the Normal distribution, its inferences are all based on the ratio of two components rather than the original components. However, there is no direct transformation from the moments of the alr back to the moments of the original components. This often leads to an interpretion difficulty.

### 1.3.3 Dirichlet distributions

The log-transformation converts a compositional vector on $\mathbb{S}^D$ to the real space $\mathbb{R}^{D-1}$. Alternatively, one could consider analysis directly on the simplex space. In that situation, Dirichlet distributions are a parametric class of distributions on the simplex space $\mathbb{S}^D$ and its use in compositional data analysis dates back to 1969 when Connor and Mosimann proposed the Dirichlet distribution as a null model for rats' bone structure components and turtles' scute proportions.

**Definition 1.12** A compositional vector $\mathbf{x} \in \mathbb{S}^D$ follows the Dirichlet distribution $\mathcal{D}_D(\boldsymbol{\alpha})$ with density function

$$\frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_D)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_D)} \prod_{i=1}^{D} x_i^{\alpha_i - 1}, \text{ where all } \alpha_i > 0.$$

The support of the Dirichlet distribution is exactly the D-dimensional simplex. Letting $\alpha_0 = \alpha_1 + \alpha_2 + \cdots + \alpha_D$, this distribution has the following moment properties:

(i) $\mathrm{E}(x_i) = \alpha_i/\alpha_0$.

(ii) $\mathrm{Var}(x_i) = \alpha_i(\alpha_0 - \alpha_i)/\alpha_0^2(\alpha_0 + 1)$.

(iii) $\mathrm{Cov}(x_i, x_j) = -\alpha_i\alpha_j/\alpha_0^2(\alpha_0 + 1), \ (i \neq j)$.

A D-part composition that follows a Dirichlet distribution can be visualized as a composition formed from D independent gamma-distributed components. That is, if $w_1, w_2, \ldots, w_D$ are independent random variables from $\mathrm{Gamma}(\alpha_i, 1)$ $(i = 1, \ldots, D)$, then $\mathbf{x} = \mathcal{C}[w_1, w_2, \ldots, w_D]$ has the Dirichlet distribution $\mathcal{D}_D(\boldsymbol{\alpha})$ with the parameter vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_D)$. As a potential model for compositional data, the Dirichlet distribution is far more restrictive than the logistic Normal as it implies that the correlation between any two components of the composition are proportional to $-\alpha_i\alpha_j$.

Aitchison (1986) minimized the Kullback-Leibler divergence (KL) of the logistic Normal $\mathcal{L}_D(\boldsymbol{\mu}, \Sigma)$ and the Dirichlet $\mathcal{D}_D(\boldsymbol{\alpha})$ to determine an appropriate logistic Normal distribution that approximates the Dirichlet. This KL distance is

$$K(p,q) = \int_{\mathbb{S}^D} p(\mathbf{x}|\boldsymbol{\alpha}) \log\Big(\frac{p(\mathbf{x}|\boldsymbol{\alpha})}{q(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}\Big) d\mathbf{x},$$

where $p(\mathbf{x}|\boldsymbol{\alpha})$ and $q(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the density functions of the Dirichlet $\mathcal{D}_D(\boldsymbol{\alpha})$ and the logistic Normal $\mathcal{L}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributions.

The parameters for the logistic Normal are:

$$\begin{aligned} \mu_i &= \psi(\alpha_i) - \psi(\alpha_D) \ (i = 1, \dots, D-1), \\ \sigma_{ii} &= \psi^{'}(\alpha_i) - \psi^{'}(\alpha_D) \ (i = 1, \dots, D-1), \\ \sigma_{ij} &= \psi^{'}(\alpha_D) \ (i \neq j = 1, \dots, D-1), \end{aligned} \tag{1.1}$$

where $\psi(\cdot)$ and $\psi^{'}(\cdot)$ are the first and second derivatives of the gamma function $\Gamma(\cdot)$. This approximation is particularly accurate for large $\boldsymbol{\alpha}$. In fact, when all $\alpha_i \to \infty$ at the same rate $(i = 1, \dots, D)$, $\mathcal{L}_D(\boldsymbol{\mu}, \Sigma) \to \mathcal{D}_D(\boldsymbol{\alpha})$ in distribution.

## 1.4 Statistical Inference Methods

The general approach to compositional data analysis is to apply some transformation to move the problem to the more familiar Euclidean space, and then apply standard multivariate statistical modeling or statistical testing (Filzmoser, 2012). Far fewer studies have analyzed compositional data directly in the simplex space. We describe both approaches in this section.

### 1.4.1 Log-ratio analysis

### Comparing means

An example of comparing means is a habitat study provided by Aebischer et al. (1993). They compared habitat use and availability of habitat-types for a given home range. For each individual animal $i$ ($i = 1, \ldots, n$), a compositional vector of habitat-type availability within its home range, denoted as $\mathbf{x}_{iA} = [x_{iA_1}, \ldots, x_{iA_D}]$, is compared to the proportions of sequentially collected radio locations from the tagged animal, denoted as $\mathbf{x}_{iU} = [x_{iU_1}, \ldots, x_{U_D}]$. If the habitat types are used randomly, then $\mu_{\mathbf{x}_{iA}} = \mu_{\mathbf{x}_{iU}}$. Equivalently, given the alr transformation $\mathbf{y}_{iA} = \text{alr}(\mathbf{x}_{iA})$ and $\mathbf{y}_{iU} = \text{alr}(\mathbf{x}_{iU})$, the null hypothesis then becomes a standard test of whether the mean differences between the habitat use and availability $\mathbf{d} = \boldsymbol{\mu}_{(\mathbf{y}_{iU} - \mathbf{y}_{iA})}$ follow a multivariate Normal distribution $\mathcal{N}_{D-1}(\mathbf{0}, \Sigma)$.

When the interest is in comparing average habitat use between two different animal groups, it is equivalent to constructing a multivariate two–sample test. For example, let $\mathbf{x}_{g_i}$ ($i = 1, \ldots, m$) and $\mathbf{x}_{k_i}$ ($j = 1, \ldots, n$) denote the proportions of collected radio locations from the tagged animal $g_i$ in group $G$ and the tagged animal $k_j$ in group $K$, respectively. Letting $\mathbf{y}_{g_i} = \text{alr}(\mathbf{x}_{g_i})$ and $\mathbf{y}_{k_j} = \text{alr}(\mathbf{x}_{k_j})$, a standard two–sample test such as Hotelling's $T^2$ (1931) can then be used for the null hypothesis $\mu_{\mathbf{y}_G} = \mu_{\mathbf{y}_K}$.

### Classification

The earliest study on classification using compositional data goes back to Toucher (1908), who classified boys to each county in Scotland according to various physical characteristics. Perhaps better known studies of classification with compositional data are in petrology, where the geochemical composition, e.g., relative percentage of chemical oxides, are used to classify rock samples (Thompson et al., 1972; Carr, 1981). To describe the approach, we consider Carr (1981), who collected 102 rock samples, and classified them into Permian and Post Permian rock types using just their geochemical characteristics. Carr considered logistic discriminant analysis (Cox, 1966; Anderson, 1972; Dawid, 1976), a well-established method to model type probabilities given a vector of diagnostic features. His training set consisted of 65 Permian and

37 Post-Permian rock samples and each composition $\mathbf{x} \in \mathbb{S}^{10}$, contained the relative percentages of 10 major oxides.

Aitchison (1986) argued that using the raw, untransformed compositions $\mathbf{x}$ in such logistic discriminant modeling is misleading because the covariance matrix constructed from the raw proportions has little validity in providing useful information about the nature of dependence between the components of the composition. He proposed applying the alr transformation to the raw compositions before performing the logistic discriminant analysis on the transformed vectors $\mathbf{y}$. The corresponding logistic form of the model is:

$$\log(\frac{p_1}{1-p_1}) = \beta_0 + \beta_1 y_1 + \cdots + \beta_9 y_9,$$

where $p_1$ is the probability the sample is Permian. By using standard maximum likelihood estimation methods, the estimates of $\boldsymbol{\beta}$ can be obtained.

Using the full compositions, Aitchison found three misclassifications in Permian and six misclassifications in post-Permian. He also conducted systematic subcomposition analyses to investigate which of the 10 components significantly contribute to the estimated rock type identity by applying $\chi^2$ test to each subcompostion model. The conclusion from Aitchison's subcompostion analysis is that there are two 6-part subcompostions that are adequate for classifying rock types. The misclassifications from the two subcompsitional analysis are four in Permian, seven in post-Permian, and two in Permian, six in post-Permian respectively. This contradicts Carr's (1981) result that some 2– or 3–part subcompostions would be adequate. Even though Carr's results concluded 95.1% maximum separation between these two types of rocks, Aitchison was concerned about the overfit of Carr's analysis.

### Linear modeling

In economics, estimating the relative shares of demand in total expenditure is closely related to compositional data analysis. The shares of total expenditure fall in the interval [0,1] and the sum of all shares should be equal to one. In a system of share equations, the shares are determined by two components: a deterministic component and a stochastic component. The deterministic component is derived from economic theory. A traditional approach to estimate and predict the shares equations is:

$$w_{ij} = W_i(\mathbf{z}, \boldsymbol{\beta}) + e_{ij},\ i = 1, \ldots, N-1;\ j = 1, \ldots, m,$$

where $N$ is the number of goods, $m$ is the number of individuals, $\mathbf{z}$ is a vector of exogenous variables, (i.e., prices and expenditures), $\boldsymbol{\beta}$ are the parameters, and the $W_i()$'s are the deterministic functions from economic theory which are restricted to the simplex. Due to the unit sum constraint, there are only $N - 1$ share equations to be estimated. The deviations $(\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_{N-1}})$ are typically assumed to follow a (N-1)-variate Normal distribution $\mathcal{N}_{N-1}(\mathbf{0}, \boldsymbol{\Omega})$.

Since this assumption of $\boldsymbol{e}$ results in non-zero probabilities of the estimated shares falling outside [0,1], it is obviously not an appropriate assumption. Fry et al. (1996) proposed that a logistic Normal distribution is a more realistic stochastic component. A direct approach is to model the alr transformed observed shares,

$$y_{ij} = \log(w_{ij}/w_{Nj}),$$

in terms of the parameters $\boldsymbol{\mu}$ and $\Sigma$. The vector $\boldsymbol{\mu}$ can be written in terms of the exogenous variables $\mathbf{z}$ and the parameters $\boldsymbol{\beta}$:

$$\mu_i(\mathbf{z}, \boldsymbol{\beta}) = \log(W_i(\mathbf{z}, \boldsymbol{\beta})/W_N(\mathbf{z}, \boldsymbol{\beta})).$$

This gives the following form for estimation of alr transformed observed shares:

$$y_{ij} = \log(w_{ij}/w_{Nj}) = \log(W_i(\mathbf{z}, \boldsymbol{\beta})/W_N(\mathbf{z}, \boldsymbol{\beta})) + v_{ij},\ i = 1, \ldots, N-1;\ j = 1, \ldots, m,$$

where $\mathbf{v}$ is assumed to have additive logistic Normal distribution $\mathcal{L}_N(\mathbf{0}, \boldsymbol{\Sigma})$ such that the shares $\mathbf{w}$ is distributed as logistic Normal. This assumption ensures that the estimated shares will fall into the interval [0,1]. Another advantage of assuming the logistic Normal distribution for the stochastic component is that it allows $w_i$ to be estimated as zero. For goods $i$ and a particular value of $\mathbf{z}$ and $\boldsymbol{\beta}$, if $W_i(\mathbf{z}, \boldsymbol{\beta}) = 0$, then the deterministic part of $y_i$ is $-\infty$. When we apply alr transformation on $w_i$ and any drawing of $v_i$ still gives $y_i = -\infty$, thus $w_i$ is modeled as zero. With the traditional approach, $w_i$ will always result in a non-zero estimated value.

### Principal component analysis

Compositional data are often high-dimensional and hard to visualize. Dimension-reduction techniques are often used to obtain lower-dimensional data without losing much information. Aitchison (1986) proposed the log-contrast principal component analysis as the standard principal component analysis (PCA) for compositional data. The procedure of doing log-contrast principal component analysis is the same as the standard principal component analysis in $\mathbb{R}^D$. However, the eigensolutions for the covariance structure of the clr transformed compositions should be used instead of raw compositions. The alr transformation and the ilr transformation are not preferable here because a different choice of denominator component in alr transformation results in different log-contrast and new variables are not directly interpretable by using the ilr transformation.

**Definition 1.13** Let $\lambda_1, \lambda_2, \ldots, \lambda_D$ be the D positive eigenvalues of the centered log-ratio covariance matrix $\mathbf{\Gamma}$ in a descending order, i.e., $\lambda_1 > \lambda_2 >, \cdots, > \lambda_D$ and $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_D$ are the corresponding standardized eigenvectors, satisfying

$$(\mathbf{\Gamma} - \lambda_i \mathbf{I})\boldsymbol{\alpha}_i = \mathbf{0} \;\; (i = 1, 2, \ldots, D),$$

then $\boldsymbol{\alpha'}_i \log \mathbf{x}$ is termed the $i$th logcontrast principal component.

Logcontrast principal components preserve the orthogonal and uncorrelated properties in the sense that $\boldsymbol{\alpha}_i^{'}\boldsymbol{\alpha}_j = 0$ $(i \neq j)$ and $\mathrm{Cov}(\boldsymbol{\alpha}_i^{'} \log \mathbf{x}, \boldsymbol{\alpha}_j^{'} \log \mathbf{x}) = 0$. The proportion of the total variability that is explained by the first $c$ log-contrast principal components is then

$$\sum_{i=1}^{c} \lambda_i / \sum_{i=1}^{D} \lambda_i.$$

After the introduction of the use of log-contrast principal component in PCA, Flizmoser et al.(2009) proposed a robust PCA for compositional data to handle outliers. This procedure is based on a robust covariance estimator, like the minimum covariance determinant (MCD). Since MCD only works for nonsingular data with rank equal to the number of variables, the ilr transformation is applied to the original compositional data for the use in robust PCA. Examples of classical PCA and robust PCA were provided in Flizmoser's paper (2009) using the Baltic Soil Survey (BSS) data (Reimann et al., 2003). The data consist of 769 samples of agricultural soil coming from two different layers, labeled top and bottom. All samples are represented as compositions with more than 40 chemical elements. A PCA inspection of this data set provides a better understanding of the relations between the chemical elements as well as how the geochemical processes would affect the element distribution in the survey area. In particular, visualizing the first few PCs shows the regions where certain concentrations are higher or lower due to some key geochemical processes. Because the difficulty of interpreting the results given ilr transformed space, the loadings and scores have to be back-transformed to the clr space.

### 1.4.2 Dirichlet analysis

#### Linear modeling

In the section on log-ratio linear modeling, Fry et al. (1996) proposed to use the logistic Normal distribution to model the shares in the share equations to account for the restriction that all shares have to fall in [0,1] interval. Woodland (1979) noted this restriction as well and proposed using the Dirichlet distribution.

Recall that the system of $N$ shares equations for an individual are

$$w_i = W_i(\mathbf{z}, \boldsymbol{\beta}) + e_i,\ i = 1, \ldots, N.$$

If the shares vector $\mathbf{w}$ is assumed to follow a Dirichlet distribution $\mathcal{D}_N(\boldsymbol{\alpha})$ and $\boldsymbol{\alpha}$ is defined in terms of the vector of exogenous variables $\mathbf{z}$ and $\boldsymbol{\beta}$ as $\alpha_i = kS_i(\mathbf{z}, \boldsymbol{\beta})$ $(i =$

$1, 2, \ldots, N,\ k > 0$), then immediately we have the following covariance structure from Dirichlet properties:

$\mathrm{Var}(w_i) = \alpha_i(\alpha_0 - \alpha_i)/\alpha_0^2(\alpha_0 - 1) = W_i(\mathbf{z}, \boldsymbol{\beta})(1 - W_i(\mathbf{z}, \boldsymbol{\beta}))/(\alpha_0 + 1), \ \ i = 1, 2, \ldots, N,$

$\mathrm{Cov}(w_i, w_j) = -\alpha_i\alpha_j/\alpha_0^2(\alpha_0 - 1) = -W_i(\mathbf{z}, \boldsymbol{\beta})W_j(\mathbf{z}, \boldsymbol{\beta})/(\alpha_0 + 1), \ \ i \neq j, j = 1, 2, \ldots, N,$

where $S_0 = \sum_{i=1}^{N} S_i(\mathbf{z}, \boldsymbol{\beta})$ and $\alpha_0 = \sum_{i=1}^{N} \alpha_i = kS_0$. The parameter $k$ can be viewed as a variance parameter. The bigger $k$ is, the smaller are all the elements of the covariance matrix while the mean vector of $\mathbf{w}$ is fixed. Parameters $\boldsymbol{\beta}$ and $k$ can be obtained by maximizing the log-likelihood function of the Dirichlet.

### Classification

Statistical classification with compositional data can also be done using Dirichlet regression. Consider a medical example dealing with the differential diagnosis for two diseases based on the composition of four serum protein components provided by Maier (2014). These data include 30 blood samples of diagnosed patients, and 6 more samples of patients who are undiagnosed. The purpose of the study is to classify the undiagnosed patients based on the classifier built from the 30 diagnosed samples. The Dirichlet regression for classification in this example can be set up as

$$\mathbf{y_i} \sim \mathcal{D}_4(\boldsymbol{\alpha_i}) \qquad (i = 1, \ldots, 30),$$

and

$$g(\boldsymbol{\alpha}_c) = \mathbf{X}\boldsymbol{\beta}_c \qquad (c = 1, 2, 3, 4),$$

where $\mathbf{y_1}, \ldots, \mathbf{y_{30}}$ are the 30 diagnosed blood samples and the predictor $\mathbf{X}$ is the indicator of disease type. The $g(\cdot)$ is the log link function since $\alpha_{ic} > 0$. The parameters $\boldsymbol{\beta}$ can be obtained using standard MLEs.

To make predictions for the undiagnosed patients, a likelihood-based approach can be used here. That is, calculate the Dirichlet likelihoods of the new observations given the parameters of both disease groups and assign the new observations to the disease type with the larger likelihood.

### Comparing means

Even though no explicit studies have been found using Dirichlet distributions to compare compositional means across different groups, a likelihood ratio based approach could be used to test the difference of means in this case. For example, a separate Dirichlet model can be fit within each group and an overall model is fit using all the data. One can then construct a hypothesis test for comparing means by comparing the corresponding likelihoods.

## 1.5 Compositional Analysis with Zeros

A common issue with compositional data is the presence of zeros. There are two types of zeros: essential (or true) zeros and rounded (or censored) zeros. Essential zeros mean that some elements in the composition vector are actually zero, or absent. The pattern of occurrence of true zeros should be investigated and separately modeled. Most zeros occur because of rounding/censoring. Rounded zeros mean that the value of a present component is either below a detection limit or zero due to chance variation.

All the transformation techniques we discussed in Section 1.4 can not be directly used on compositional data with zero elements. For example, if the component with a zero is chosen to be the denominator in the alr transformation, the transformed composition doesn't exist. Similarly, the geometric average $g(\cdot)$ will end up as zero if any element is zero. Therefore, both clr and ilr transformations are not applicable without some adjustments of the zeros. Finally, a D-dimensional Dirichlet distribution does not allow a zero to occur in any of the D components. Because zeros cannot be handled in a log-transformation (logistic Normal) or Dirchlet settings, it is common to treat them as censored data and impute a non-zero value prior to analysis.

### 1.5.1 Nonparametric replacement strategy for rounded zeros

Nonparametric replacement is a general strategy to replace rounded zeros. In Aitchison's book (1986), the following additive replacement strategy for rounded zeros is suggested.

$$r_j = \begin{cases} \frac{\delta(Z+1)(D-Z)}{D^2}, & \text{if } x_j = 0, \\ x_j - \frac{\delta(Z+1)(Z)}{D^2}, & \text{if } x_j > 0, \end{cases}$$

where $\mathbf{x}$ is a D-part composition with $Z$ rounded zeros and $\delta$ is a small value, less than a given threshold. Aitchison (1986) conducted sensitivity analysis and suggested the range $\frac{\delta_r}{5} \leq \delta \leq 2\delta_r$, where $\delta_r$ is the maximum rounding-off error for $\delta$. Sandford, Pierson, and Crovelli (1993) consider 0.55 as a suitable imputed value of the threshold.

Martín-Fernández (2003) argued that the Aitchison distance between two replaced compositions using additive replacement strategy is extremely sensitive to changes in $\delta$. Also if $\mathbf{x}$ has more than one zero value, then $\frac{r_j}{r_k} \neq \frac{x_j}{x_k}$, for $x_j > 0, x_k > 0$. Therefore, the covariance structure of the subcomposition on these parts is not preserved.

Alternatively, many researchers simply replace the rounded zeros in a composition $\mathbf{x}$ by a small quantity to obtain a vector of positive components, $\mathbf{w} \in \mathbb{R}^D$. Then the closure operation is used to get $\mathbf{r} = \mathcal{C}(\mathbf{w})$. This simple replacement strategy can be expressed as

$$r_j = \begin{cases} \frac{1}{1+\sum_{k|x_k=0} \delta_k} \delta_j, & \text{if } x_j = 0, \\ \frac{1}{1+\sum_{k|x_k=0} \delta_k} x_j, & \text{if } x_j > 0. \end{cases}$$

Martín-Fernández (2003) proposed the multiplicative replacement strategy:

$$r_j = \begin{cases} \delta_j, & \text{if } x_j = 0, \\ (1 - \sum_{k|x_k=0} \delta_k) x_j, & \text{if } x_j > 0. \end{cases}$$

When the percent of rounded zeros in the full data set is less than 10%, Martín-Fernández (2003) recommends imputing zeros with the values equal to 65% of the threshold using multiplicative replacement. He argues this imputation minimizes the distortion of the covariance matrix. Actually, the simple replacement strategy and the multiplicative replacement strategy are equivalent when the zero components are imputed with the same value. Using the multiplicative replacement, the imputed zero components do not depend on the amount of parts $D$ nor the number $Z$ of zeros. It is also intuitive that if $\delta_j$ is equal to the "true" detection limit or censored value, then the "true" composition can be recovered. The simple replacement strategy does not explicitly satisfy this property unless it is made to be equivalent to the multiplicative replacement. Moreover, the multiplicative replacement strategy preserves subcomposition invariance, perturbation invariance, and power transformation invariance properties. Thus it is more suitable than the additive replacement strategy and generally recommended.

### 1.5.2 Parametric replacement strategy

When the proportion of zeros is large (e.g., more than 10%), a parametric imputation strategy is recommended. Such imputation fully depends on the choice of parametric distribution. The EM algorithm (Dempster et al., 1977) is a well-known iterative procedure to impute missing data based on observed data. The standard EM can deal with values missing at random (MAR), which means the probability that a value is missing depends on the observed data but not the missing data. However, the rounded zeros in compositional data occur when they are below a detection limit, which means they are not missing at random (NMAR). Here we outline two popular EM imputations for NMAR compositional data based on the additive logistic Normal distribution and Dirichlet distributions.

***Modified EM algorithm based on additive log-ratio transformation***

Palarea-Albaladejo et al. (2007) developed a modified EM algorithm to impute rounded zeros based on the alr transformation. Let $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ be the complete compositional dataset with $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$ denoting the observed and missing compositional parts, respectively. Let $\boldsymbol{\theta}$ denote the unknown parameters of the probability distribution $P$ for the complete data. Let $L(\boldsymbol{\theta}|\mathbf{Y})$ denote the corresponding log-lilkelihood function.

*Modified E-step with alr transformation:* In a standard E-step, we compute the conditional expectation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int L(\boldsymbol{\theta}|\mathbf{Y})P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs};\boldsymbol{\theta}^{(t)})d\mathbf{Y}_{mis}.$$

With compositional data, suppose $\mathbf{y}_i = alr(\mathbf{x_i})$ where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iD}]$ and $i = 1, 2, \ldots, n$. Assume the complete transformed data $\mathbf{Y}$ follow a logistic Normal distribution $\mathcal{L}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Psi} = (\psi_{ij}) = (\log(\gamma_j/x_{iD}))$ where $\gamma_j$ is the detection limit or threshold for the component $x_j$. Note that $x_D$ has to be a component without zeros. In the modified E-step, we compute the conditional expectation incorporating the detection limit information,

$$\begin{aligned} E[y_j|\mathbf{y}_{-j}, y_j < \psi_j] &= \frac{1}{P(y_j < \psi_j)} \int_{-\infty}^{\psi_j} y_j(2\pi\sigma_j^2)^{-1/2} exp\Big[-\frac{1}{2\pi_j^2}(y_j - \mathbf{y}_{-j}^T\beta)^2\Big]dy_j \\ &= \mathbf{y}_{-j}^T\beta - \sigma_j \frac{\phi(\psi_j - \mathbf{y}_{-j}^T\beta)/\sigma_j}{\Phi(\psi_j - \mathbf{y}_{-j}^T\beta)/\sigma_j}, \end{aligned}$$

where $\mathrm{var}(y_j) = \sigma_j^2$, and $\phi$ and $\Phi$ are the density and the distribution function of the standard Normal respectively. On the $t^{th}$ iteration, we replace the values in $\mathbf{Y}$ by

$$y_{ij}^{(t)} = \begin{cases} y_{ij}, & \text{if } y_{ij} \geq \psi_{ij}, \\ E[y_{ij}|\mathbf{y}_{i,-j}, y_{ij} < \psi_{ij}; \boldsymbol{\theta}^{(t-1)}], & \text{if } y_{ij} < \psi_{ij}. \end{cases}$$

*Modified M-step with alr transformation:* This step is the same as the standard M-step by maximizing the Normal log-likelihood given the complete dataset $\mathbf{Y}$ in $t^{th}$ iteration.

### *EM algorithm based on Dirichlet distribution*

Hijazi (2011) provides the details of the EM algorithm for rounded zeros under Dirichlet models. Suppose $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$ be the complete compositional dataset which follow D-dimensional Dirichlet distribution $\mathcal{D}_D(\boldsymbol{\alpha})$. $\psi$ is the detection limit for $x_{ij}$. The E-step for this algorithm:

$$x_{ij}^{(t)} = \begin{cases} x_{ij}, & \text{if } x_{ij} \geq \psi, \\ E[x_{ij}|\mathbf{x}_{i,-j}, x_{ij} < \psi; \boldsymbol{\alpha}^{(t)}], & \text{if } x_{ij} < \psi. \end{cases}$$

The conditional expectation above can be written as:

$$\begin{aligned} E[x_{ij}|\mathbf{x}_{i,-j}, x_{ij} < \psi; \boldsymbol{\alpha}^{(t)}] &= \frac{1}{P(x_j < \psi)} \int\limits_0^{\psi} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)\Gamma(\alpha_0 - \alpha_j)} x_j^{\alpha_j - 1}(1 - x_j)^{\alpha_0 - \alpha_j - 1} dx_j \\ &= \frac{\alpha_j F_1 \psi}{\alpha_0 F_2(\psi)}, \end{aligned}$$

where $\alpha_0 = \sum_{j=1}^{D} \alpha_j$, $F_1$ and $F_2$ are the distribution functions of beta random variables with parameters $(\alpha_j + 1, \alpha_0 - \alpha_j)$ and $(\alpha_j, \alpha_0 - \alpha_j)$ respectively. This comes directly from the Dirichlet property that the marginal distribution of a Dirichlet is a beta distribution.

***Bayesian replacement algorithm for zero counts***

Compositional data can also be formed by scaling counts data to sum to one. If the counts for D categories $w_1, w_2, \ldots, w_D$ contain any zero count(s), then the formed compositional vector $\mathbf{x} = \mathcal{C}(w_1, w_2, \ldots, w_D)$ will contain zero component(s). For rounded zeros due to zero counts, Daunis-i-Estadella et al.(2008) introduced a Bayesian-multiplicative approach as a replacement strategy for zero counts. Let $\mathbf{w}_i$ be the counts vector and $T_i = \sum_{j=1}^{D} w_{ij}$ be the total count. The $\mathbf{w}_i$ can be viewed as coming from a multinomial distribution with associated probabilities $\boldsymbol{\theta}_i$. The conjugate distribution of the multinomial parameters $\boldsymbol{\theta}_i$ is a Dirichlet distribution with parameter vector $\boldsymbol{\alpha}_i$, where $\alpha_{ij} = k_i p_{ij},\ j = 1, 2, \ldots, D$. From Bayes theorem, after applying such priors on $\boldsymbol{\theta}_i$, the posterior $\theta_{ij}$ becomes

$$\theta_{ij} = \frac{w_{ij} + \alpha_{ij}}{T_i + k_i}.$$

Then the replacement strategy can be written as

$$r_{ij} \quad = \quad \begin{cases} \frac{\alpha_{ij}}{T_i+k_i}, & \text{if } x_{ij} = 0, \\ x_{ij}(1 - \sum_{k|x_{ik}=0} \frac{\alpha_{ik}}{T_i+k_i}), & \text{if } x_{ij} > 0. \end{cases}$$

Notice that this strategy coincides with the multiplicative replacement when $\delta_{ij} = \frac{\alpha_{ij}}{T_i+k_i}$. Thus all desirable properties obtained by multiplicative replacement (Martín-Fernández, 2003) can be satisfied by the Bayesian-multiplicative approach.

### 1.5.3 Handling true zeros

All the strategies described above are used to replace rounded zeros. Dealing with essential zeros is more complicated and there is not a well-founded general approach to the problem. A few approaches have been proposed to handle the essential zeros when the components are either percents or counts. If we have percent essential zeros occur, an approach based on a binomial conditional logistic Normal model (Aitchison and Kay 2003) seems to be promising. On the other hand, if we have count essential zeros, an approach based on the Poisson-Log Normal distribution may be more appropriate (Bacon-Shone 2008). Both approaches are based on the idea of hierarchical modeling: first model the pattern of zeros for multiple components, then model the composition conditional on the particular pattern. This is similar to the idea of two-part models and can be viewed as an extension of the zero problem in univariate analyses cases. However, there remains many questions about these two approaches, such as the estimability of parameters and the complexity of computations, therefore they are rarely used in practice.

## 1.6 Discussion

Log-ratio transformations serve as a bridge to connect the simplex space with Euclidean space thereby allowing standard multivariate statistical techniques. However, there have been a number of criticisms towards the different forms of log-ratios. The biggest criticism about the alr transformation is in the choice of element to be the divisor. In most of the literature, it appears the divisor is chosen arbitrarily. Even though Aitchison (1986, 2000) has shown that multivariate linear regressions with compositional data as the dependent variable are invariant to the choice of divisor, it still remains problematic because the distances between alr transformed data points are not consistent given different divisors. This might lead to inconsistent conclusions when comparing compositional means across groups. That's why the clr transformation is usually used for computing the distance between two compositional vectors as clr provides a symmetric transformation method. Moreover, if a large number of zeros are present in all the components across the compositional vectors, the EM replacement strategy is not applicable because it requires at least one component without zeros to be the divisor in the alr transformation. The clr transformation avoids the need of choosing the divisor but the covariance matrix of clr is singular and clr doesn't preserve subcomposition coherence, making it difficult to adapt to standard statistical procedures without special modifications. The idea of the ilr transformation is that compositions can be represented by their coordinates in the simplex with an orthonormal basis. Using ilr avoids both the arbitrariness of divisor in alr and the singularity of covariance matrix in clr. Unfortunately, there is not a unique and simple basis as in $\mathbb{R}^D$ for ilr and the interpretation of the results may be difficult, since there is no one-to-one relation between the original components and the transformed variables.

The logistic Normal distribution and the Dirichlet distribution are two popular parametric classes in the simplex and a lot of statistical applications are developed based on these two distributions. Dirichlet distributions, indeed, have more restrictions than the logistic Normal distribution as it assumes a specific negative correlation structure among the components. Another strong restriction of the Dirichlet distributions is the complete subcompositional independence. That is, $\mathcal{C}[\mathbf{x_s}] \perp \mathcal{C}[\mathbf{x_{-s}}]$ for each possible partition of the composition. As a consequence, Dirichlet distributions are considered as the model of maximum independence compatible with unit-sum constrained random variables.

In contrast, logistic Normal distributions allow for a more flexible covariance structure, including both positive and negative correlations among the components, and more importantly, its normality assumption makes parameter estimation easier. However, it cannot handle strong forms of independence (Rayens and Srinivasan, 1994). There is also the trade-off between its flexibility and parsimony. Logistic Normal distributions require $(D-1)(D+2)/2$ parameters while Dirichlet distributions require only $D$ parameters. In practice, if we don't have enough information in the data to estimate a flexible covariance structure, Dirichlet distributions are the usual alternative. Furthermore, Berhm et al. (1998) performed a Monte Carlo simulation study comparing the performance of Dirichlet distributions and logistic Normal distributions on multivariate linear modeling with compositional data. The conclusion was when compositional distributions are influenced by common covariates, i.e., covariates that influence all the components, the Dirichlet distribution was as successful as the logistic Normal distribution. He also showed that as the correlations between log-ratios increase, both approaches gave more errors on the parameter estimates thus no obvious evidence favors the logistic Normal distribution over the Dirichlet distribution.

Another advantage of Dirichlet distributions is that they provide easy interpretation to the statistical questions in respect to the original components. Given an alternative way of parametrizing Dirichlet distribution as $\mathcal{D}_D(\boldsymbol{\mu}, k)$, where $\boldsymbol{\mu}$ is a D-dimensional vector indicating the "location" and $k$ is the "variance" parameter, it is straightforward to interpret a Dirichlet distributed composition $\mathbf{x}$ as $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$ and $k$ describes how tightly the compositional point concentrates on its mean $\boldsymbol{\mu}$. Compared to the Dirichlet covariance structure, the logistic Normal covariance structure provides an interpretation of relative information between ratios of components, but not in terms of the original components.

Several generalizations of the Dirichlet class has been proposed in the literature, e.g., the scaled Dirichlet (Aitchison, 2003), the generalized Liouville (Rayens and Srinivasan, 1994), the conditional generalized liouville (Smith, 2002), and the flexible Dirichlet (Ongaro, 2008). The purpose of these generalizations is to relax the strong independent assumption so that they can be used to model various forms of dependence structure of compositional data. However, we've seen little use of these generalized distributions in practice because they require a larger number of parameters.

## 1.7 Our Compositional Data Problem

As discussed in the introduction, there are some compositional data problems that remain open. One such statistical question is the evaluation of agreement between pairs of compositional vectors. For example, in diagnostic testing of tumors, one may want to know how well pathologists agree with each other or agree with a gold standard. Immunohistochemistry (IHC) is a staining process usually performed on cancer tissues. Pathologists give vectors of scores representing proportions of cells with different staining levels. A popular scoring vector is the percent of cells falling into negative, weak, moderate, and positive staining categories. This scoring can be viewed as counting the number of cells within the different staining intensities and then applying the closure (Definition 1.3).

The vector provided by a pathologist for a tissue sample will vary not only because of inherent variability in this counting/closure process but also because the true distribution of the cells varies slide to slide and the category cutpoints that define the intensity categories are likely subjective. Identifying and eliminating this last source of variability is important because it results in as consistent a diagnosis as possible given the inherent variability in scoring. Thus a common question in this area of research is: how to assess inter-rater agreement given pairs of compositional vectors across different slides? This thesis provides methodology to do this and that is described in Chapter 3. In Chapter 2, we provide an overview of rater agreement methodology to help set the stage for this novel work.

# CHAPTER 2. ASSESSING INTER-RATER AGREEMENT

## 2.1 Introduction

Inter-rater agreement is a measure of the similarity in ratings or scores among multiple raters or observers. Agreement in scores is very important when there is a common scoring scale and consistency in category classifications across raters is desired. A similar but distinct measure is inter-rater reliability, which assesses the relative similarity or relative order of ratings.

Two raters may have very high inter-rater reliability but very low inter-rater agreement. For example, the pairs of scores (1,2), (2,3), (3,4), and (4,5) have high reliability. However, if the scoring scale is such that 1s are classified as negative, scores of 2 and 3 as neutral, and scores of 4 and 5 as positive, the pairs of classifications are $(-,0),(0,0),(0,+)$, and $(+,+)$. Only two of the four pairs give similar classifications. To have high inter-rater agreement, the scores must be consistently the same. Inter-rater reliability allows the scores of one rater to also be consistently higher or lower than the other rater. In this chapter, we focus on the discussion of the measures of agreement, that is, the absolute differences between scores.

In medicine, raters are typically physicians or automated diagnostic devices. These raters assess a patient's severity of disease or illness. In practice, these scores can be nominal, ordinal, or continuous. For instance, nominal scores occur in diagnostic testing when raters classify patients as having or not having a certain medical condition. Ordinal scores occur when raters determine severity status or disease progression, such as the five stages of beta-cell dysfunction of Type I diabetes. Continuous scores occur in some common screening tests, including systolic blood pressure for hypertension, thyroid-stimulating hormone (TSH) for hypothyroid and hyperthyroid, and fasting blood cholesterol for heart disease.

Initial attempts to assess inter-rater agreement focused on analyzing nominal scores. For example, Goodman and Kruskal (1954) used the observed proportion of agreements as a measure of agreement. Scott (1955) introduced a chance-corrected version of this measure and this was extended by Cohen (1960) to form the kappa coefficient. The kappa coefficient and its extensions (Cohen, 1968; Fleiss, 1971; Barlow et al., 1991) are commonly used to assess agreement when scores are nominal or ordinal.

For continuous scores, the intraclass correlation coefficient (ICC) (Shrout and Fleiss, 1977) and the concordance correlation coefficient (CCC) (Lin, 1989) are two popular inter-rater agreement measures. These measures focus on the ratio of between-subject variability relative to the total variability. The ICC relies on ANOVA assumptions while the CCC does not.

More recently, the coefficient of inter-rater variability (CIV) was proposed (Haber, 2005) as an inter-rater agreement index. It looks at the ratio of between-rater variability relative to the total rater-related variability. Soon after, Barnhart (2007a) proposed another inter-rater agreement index, called the coefficient of individual agreement (CIA). The CIA can be viewed as an extension of the CIV to the cases with and without a reference/gold standard. When agreement is assessed without an existing reference/gold standard, the CIA is equivalent to the CIV. Because of their more recent development, the CIV and the CIA are far less used.

Finally, the introduction of the iota coefficient by Janson (2001) extended the assessment of agreement among multiple raters with nominal or continuous multivariate scores. In the nominal setting, the iota coefficient can also be viewed as the extension of the kappa coefficient.

In addition to agreement indices, log-linear models and latent-class models have been proposed to model agreement pattern for univariate nominal or ordinal data (Tanner and Young, 1985 a, b; Graham, 1995; Agresti, 1988, 1992). We discuss the details of these approaches later.

In the reminder of this chapter, we provide more details of these agreement models and indices. We conclude with a section on compositional data, providing justification for the need of alternative measures in this setting.

## 2.2 Inter-rater Agreement Indices

This overview will follow the development of agreement indices chronologically. This means that we start with the analysis of nominal and ordinal scores and then move to the analysis of continuous scores, both univariate and multivariate. We conclude with a discussion of the limitations with these methods and models to identify patterns of agreement.

### 2.2.1 Kappa coefficient and its extension

As mentioned in the introduction, the earliest measures of agreement were simply the proportion of observed agreements (Goodman and Kruskal, 1954). This index ranges from 0 to 1, with 1 signifying perfect agreement. When raters are uncertain about a classification, a degree of guessing may occur. This simple agreement measure does not take into account any possibility that some agreements may occur by chance. Several chance-corrected measures were proposed with the kappa coefficient being the most commonly used measure of rater agreement (Cohen, 1960).

Consider two raters and a nominal or ordinal score scale consisting of $m$ levels. We can summarize the joint evaluation of $n$ objects in a $m \times m$ contingency table, where the rows refer to the scores from Rater 1 and the columns refer to the scores from Rater 2. Each cell $c_{ij}$ in this table represents the number of objects in which Rater 1 classifies the object in level $i$ and Rater 2 classifies the object in level $j$.

Let $p_{ij} = c_{ij}/n$, $p_{i.} = \sum_{j=1}^{m} c_{ij}/n$, and $p_{.j} = \sum_{i=1}^{m} c_{ij}/n$ be the observed cell, row, and column proportions, respectively. The observed proportion of agreement between the two raters is $p_0 = \sum_{i=1}^{m} p_{ii}$. Assuming independence among evaluations and raters, the expected proportion of agreement by chance is $p_c = \sum_{i=1}^{m} p_{i.}p_{.i}$. Given these two proportions, the kappa coefficient is defined to be

$$\hat{\kappa} = \frac{p_0 - p_c}{1 - p_c}.$$

Theoretically, the range of $\hat{\kappa}$ is from -1 to 1, though it is usually observed between 0 and 1. A value of 1 represents perfect agreement (i.e., $p_0 = 1.0$) while a value of 0 represents pure chance agreement. The value, $p_c$, is the proportion of times raters would agree if they randomly assign a score on every case with the probabilities that match their marginal proportions of ratings. Landis and Koch (1977) categorized the degree of agreement based on different ranges of kappa values. These categorizations are shown in Table 2.1. While commonly used, this interpretation is subjective and by no means universally accepted.

Table 2.1.
Interpretation of $\kappa$ by Landis and Koch (1977)

| $\kappa$ | Interpretation |
|---|---|
| $< 0$ | Poor agreement |
| $(0.01 - 0.20)$ | Slight agreement |
| $(0.21 - 0.40)$ | Fair agreement |
| $(0.41 - 0.60)$ | Moderate agreement |
| $(0.61 - 0.80)$ | Substantial agreement |
| $(0.81 - 1.00)$ | Almost perfect |

Some researchers (e.g., Uebersax, 1987) argued that the kappa coefficient is not a “true” chance-corrected measure as claimed because it supposes that raters simply assign a score at random on every case when not completely certain. A more effective way to account for this is done by modeling rater agreement (e.g., Agresti, 1992). We will discuss such models later in next section.

The kappa coefficient treats all disagreements between categories equally. However, this is not always desired. When considering an ordinal score, disagreements between extreme categories may be considered more severe than disagreements between adjacent categories. For example, in a cancer diagnostic study, a disagreement when the classifications are benign and cancerous, is far more egregious than a disagreement when the classifications are neutral and benign.

In 1968, Cohen proposed a weighted kappa coefficient, which incorporates subjective disagreement weightings. Suppose $w_{ij}$ represents the weight assigned to the $(i, j)$th cell, the weighted kappa coefficient is then

$$\hat{\kappa}_w = \frac{\sum_{i=1}^m \sum_{j=1}^m w_{ij} p_{ij} - \sum_{i=1}^m \sum_{j=1}^m w_{ij} p_{i.} p_{.j}}{1 - \sum_{i=1}^m \sum_{j=1}^m w_{ij} p_{i.} p_{.j}}, \; (w_{ij} \in \mathbb{R}^+).$$

When $w_{ij} = 1$ for $i = j$ and $w_{ij} = 0$ for $i \neq j$, the weighted kappa $\hat{\kappa}_w$ is equal to the kappa coefficient $\hat{\kappa}$.

Studies have showed that the kappa coefficient is sensitive to different populations of subjects and different marginal distributions of ratings (Feinstein and Cicchetti, 1990; Byrt et al., 1993). For example, the kappa coefficient can only reach its theoretical maximum value of 1 when both raters have the same marginal distribution of ratings. This sensitivity leads to an interpretation difficulty of the kappa coefficient and further complicates the interpretation of $\hat{\kappa}$ (Table 2.1). We can calculate the maximum value kappa can achieve given unequal marginal distributions to help interpret the kappa value obtained. Often tests for marginal homogeneity of ratings are suggested before considering the kappa coefficient. It is also recommended to avoid comparing kappa coefficients across different studies and populations.

Under the assumption of homogenous marginal distributions, Bloch and Kraemer (1989) introduced an alternative version of Cohen's kappa, called the intraclass kappa. The intraclass kappa is defined for data consisting of dichotomous scores on each of $n$ subjects sampled from a population. It is assumed that the two scores for each subject are interchangeable, i.e., in the population of subjects, the two scores for each subject have a distribution that is invariant under permutations of the raters.

Let $x_{ij}$ denote the score for the $i$th subject by Rater $j$ and $p_i = \mathrm{P}(x_{ij} = 1)$ denote the probability that subject $i$ is a "success". Over the population of subjects, let $\mathrm{E}(p_i) = P$ and $\mathrm{var}(p_i) = \sigma_P^2$, then the intraclass kappa can be expressed as

$$\kappa_I = \frac{\sigma_P^2}{P(1-P)}.$$

The expected probability of each of the cell frequencies based on our model is listed below:

Table 2.2.
The probability model for the joint responses

| $x_{i1}$ | $x_{i2}$ | Obs. freq. | Expected probability |
|---|---|---|---|
| 1 | 1 | $c_{11}$ | $P^2 + \sigma_P^2$ |
| 1 | 0 | $c_{10}$ | $P(1-P) - \sigma_P^2$ |
| 0 | 1 | $c_{01}$ | $P(1-P) - \sigma_P^2$ |
| 0 | 0 | $c_{00}$ | $(1-P)^2 + \sigma_P^2$ |

The MLE of the intraclass kappa can be obtained as

$$\hat{\kappa}_I = \frac{4(c_{00}c_{11} - c_{01}c_{10}) - (c_{01} - c_{10})^2}{(2c_{00} + c_{01} + c_{10})(2c_{11} + c_{01} + c_{10})}.$$

This estimator is identical to the estimator of an intraclass correlation coefficient (ICC) for dichotomous data. Different types of ICC will be discussed in the next subsection.

If the marginal distributions of ratings are not homogenous and depend on some covariates, then the kappa coefficient needs to be investigated for different covariates. Barlow et al. (1991) proposed a stratified kappa to deal with non-homogenous marginal distributions given categorical covariates. Suppose that a covariate has $m$ different strata, and $\hat{\kappa}_1, \ldots, \hat{\kappa}_m$ donate the kappa coefficients for each of these strata. The stratified kappa is simply the weighted average of these kappa coefficients $\hat{\kappa}_s = \sum_{i=1}^m w_i \hat{\kappa}_i$ $(i = 1, \ldots, m)$ where $w_i$ is the weight for each stratum.

Barlow et al. (1991) considered three weighting schemes: 1) equal weighting, 2) weighting by the relative sample size of each stratum, and 3) weighting by the inverse variance, and compared them to the non-stratified kappa coefficient. Simulations show the estimator using stratum sample size as weights minimizes the mean square error among these three weighting options. However, with $m$ and/or the number of covariates increasing, there are often only a few observations in each stratum, resulting in poor estimates of stratified kappa. Also the stratified kappa is not invariant to different populations. If the subjects in different stratum are from different populations, it is inappropriate to average stratum-specific kappa. Donner (1996) discussed a method using large-sample variance of kappa to test the homogeneity of kappa across populations, and also proposed a goodness-of-fit test as an alternative test if the sample size is relatively small.

Fleiss's kappa (1971) is a generalized kappa for more than two raters. Fleiss's kappa calculates the degree of agreement in classification over that which would be expected by chance, thus it is not simply a weighted average of pairwise kappas. The Fleiss's kappa is defined as

$$\kappa_F = \frac{\bar{p}_0 - \bar{p}_e}{1 - \bar{p}_e},$$

where $\bar{p}_0$ and $\bar{p}_e$ represent the observed agreement and expected agreement by chance among $k$ raters. In terms of a contingency table, we can present the data as follows:

| Subject | Category 1 | $\cdots$ | Category $m$ | |
|---|---|---|---|---|
| 1 | $k_{11}$ | $\vdots$ | $k_{1m}$ | $p_{1.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $k_{n1}$ | $\vdots$ | $k_{nm}$ | $p_{n.}$ |
| | $p_{.1}$ | $\cdots$ | $p_{.m}$ | |

In this table, $k_{ij}$ is the number of raters who assigned the $i$th subject to the $j$th category, $(i = 1, \ldots, n;\ j = 1, \ldots, m)$, and the proportion of all assignments which were to the $j$th category is

$$p_{.j} = \frac{1}{nk} \sum_{i=1}^{n} k_{ij}.$$

The proportion of rater pairs that are in agreement on subject $i$, relative to the number of all possible rater pairs is

$$p_{i.} = \frac{1}{k(k-1)} \sum_{j=1}^{m} k_{ij}(k_{ij} - 1).$$

Then the observed agreement and expected agreement by chance are

$$\bar{p}_0 = \frac{1}{n} \sum_{i=1}^{n} p_{i.}$$

and

$$\bar{p}_e = \sum_{j=1}^{m} p_{.j}^2.$$

Note that different from Cohen's kappa, Fleiss kappa can be used only with binary or nominal-scale scores. It can be interpreted as the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their scores randomly. Often when the Fleiss kappa is not satisfactory, pairwise kappa will be investigated.

### 2.2.2 Correlation coefficients of rater agreement

We now move to the discussion of continuous scores, moving from indices to correlation coefficients.

### Intraclass correlation coefficient (ICC)

The ICC is a measure that quantifies how strongly subjects in the same group or scores of the same subject resemble each other. The ICC is now commonly described within the random effects model framework. The most basic model is

$$y_{ij} = \mu + b_i + \epsilon_{ij},$$

where $y_{ij}$ is the $j^{th}$ observation on the $i^{th}$ subject, $\mu$ is an overall mean, $b_i$ is a random effect due to subject $i$, and $\epsilon_{ij}$ is a random error term. Then the corresponding theoretical formula for ICC is

$$\rho_I = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2}, \tag{2.1}$$

where $\sigma_b^2$ is the variance of $b_i$ and $\sigma_\epsilon^2$ is the variance of $\epsilon_{ij}$. This can be interpreted as the proportion of total variance due to subject differences.

One prominent application of ICC is to measure inter-rater reliability or agreement of univariate continuous ratings from multiple raters on the same set of subjects, with different forms for reliability and agreement study purposes. Since the ICC is defined as the between-subject reliability relative to the total variability, from a design point of view, this means that data used in calculating ICCs require multiple measurements on these subjects. However, given different types of designs and study purposes, modifications of the basic ICC are needed. Shrout and Fleiss (1979) discussed the ICC under three different study designs. McGraw and Wong (1996) proposed more versions of the ICC, which were not defined by Shrout and Fleiss (1979), and distinguished ICCs for rater–reliability versus rater–agreement. Chen and Harnhart (2008) provided a summary of different versions of ICC for both data with and without replicates. We give a summary of ICC formulas for inter-rater agreement under the three typical cases.

Case 1: A random set of $n$ subjects is selected from a population. Each subject is rated by a different set of $k$ raters, randomly selected from a larger population of raters. In this design raters are nested within subjects (Table 2.3). ICC is calculated based on the one-way random effect ANOVA. Let $y_{ij}$ denote the $j^{th}$ rating ($j = 1, \ldots, k$) on subject $i$ ($i = 1, \ldots, n$). The linear model $y_{ij} = \mu + \alpha_i + e_{ij}$ is assumed, where $\mu$ is the overall mean, $\alpha_i$ is the random effect of subject $i$ assumed to be Normally distributed as $N(0, \sigma_\alpha^2)$, and $e_{ij}$ is a residual component equal to the sum of the inseparable effects of the rater and the measurement error, which is assumed to be i.i.d. Normal $\mathcal{N}(0, \sigma_e^2)$. The ICC and its corresponding estimate are

$$ICC_1 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}, \qquad \widehat{ICC_1} = \frac{BMS - WMS}{BMS + (k-1)WMS}.$$

Table 2.3.
Sources of variance for Case 1

| Source of Variance | *df* | Mean Squares |
|---|---|---|
| Between subjects | $n-1$ | BMS |
| Raters(subjects) | $n(k-1)$ | WMS |

Table 2.4.
Sources of variance for Case 2 & Case 3

| Source of Variance | *df* | Mean Squares |
|---|---|---|
| Between subjects | $n-1$ | BMS |
| Within subjects | $n(kr-1)$ | WMS |
| Between raters | $(k-1)$ | RMS |
| Interaction | $(k-1)(n-1)$ | IMS |
| Error | $kn(r-1)$ | EMS |

Case 2: Randomly choose $k$ raters from the population of raters and $n$ subjects from the population of subjects. Each rater then scores each of the $n$ subjects $r$ times.

This is a block design with subjects as blocks (Table 2.4). Under Case 2, the ICC is calculated based on a two-way random effects ANOVA. The assumed linear model is $y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl}$, where $\mu$, $y_{ijl}$, and $\alpha_i$ have the same assumptions and interpretations as in Case 1, and $\beta_j$ is the random effect from Rater $j$ which is assumed to be i.i.d Normal $\mathcal{N}(0, \sigma^2_\beta)$. The interaction $(\alpha\beta)_{ij}$ is estimable when $r \geq 2$ and are assumed to be i.i.d Normal $\mathcal{N}(0, \sigma^2_{\alpha\beta})$. Without repeated ratings, the effects due to components $(\alpha\beta)_{ij}$ and $e_{ij}$ can not be separated. The ICC and its estimate for Case 2 without replicates are:

$$ICC_2 = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta} + \sigma^2_e},$$

$$\widehat{ICC_2} = \frac{BMS - EMS}{BMS + (k-1)EMS + k(RMS - EMS)/n}.$$

When $r > 1$, the interaction term $(\alpha\beta)_{ij}$ can be estimated. Thus the estimate for $ICC_2$ becomes

$$\widehat{ICC_2} = \frac{BMS - IMS}{BMS + k(r-1)EMS + (k-1)IMS + k(RMS - IMS)/n}.$$

Case 3: Each subject is rated by each of the fixed $k$ raters. Under Case 3, the ICC is calculated based on a two-way mixed effects ANOVA. The linear model for Case 3 has the same form as Case 2 but now $\beta_j$ is a fixed effect from Rater $j$ subject to some constraint (e.g., $\sum \beta_j = 0$). With repeated ratings, the interaction $(\alpha\beta)_{ij}$ is still a random effect subject to the some constraint (e.g., $\sum_{j=1}^k (\alpha\beta)_{ij} = 0$). Again $(\alpha\beta)_{ij}$ and $e_{ij}$ cannot be separated without repeated ratings. The ICC for Case 3 is

$$ICC_3 = \frac{\sigma^2_\alpha - \sigma^2_{\alpha\beta}/(k-1)}{\sum_j^k \beta_j^2 + \sigma^2_\alpha + \sigma^2_{\alpha\beta} + \sigma^2_e}$$

The estimate without replicates is reduced to $\widehat{ICC_2}$:

$$\widehat{ICC_3} = \frac{BMS - EMS}{BMS + (k-1)EMS + k(RMS - EMS)/n} = \widehat{ICC_2},$$

and the estimate with $r > 1$ repeated ratings is

$$\widehat{ICC_3} = \frac{BMS - IMS}{BMS + k(r-1)EMS + (k-1)IMS + k(RMS - IMS)/n}.$$

Under the three cases discussed above, all the ICC denominators include rater variability. This is why they are considered measures of inter-rater agreement instead of reliability. If the relative standing of subjects is of interest, it does not matter that Rater 1 consistently assigns relatively higher scores than Rater 2. Therefore, the rater variability is deemed to be an irrelevant source of variance and is excluded from the denominators of the ICC. If any difference between raters are considered to be disagreements, the denominator of the ICC should include the total score variability.

The range of these ICC's is usually (0, 1) but it can be negative under Case 3 due to the negative correlation between any two interaction components $(\alpha\beta)_{ij}$ and $(\alpha\beta)_{ij'}$. An alternative is to use REML estimates of mixed effects models and REML method can avoid interpreting negative values of ICC. An ICC value close to 1 can be interpreted as high agreement among raters while a smaller ICC means less agreement.

Since the ICC is calculated based on the ANOVA model assumption, that is, data follow the one-way random effect model or two-way mixed-effect model Normality and constant variance assumptions, we have to use it carefully because the violation of ANOVA assumptions will result in a serious bias of the estimates of ICC. An alternative coefficient that can be used to assess inter-rater agreement without the ANOVA assumptions is called the Concordance correlation coefficient (CCC), first proposed by Lin (1989).

### Concordance correlation coefficient (CCC)

The original CCC (Lin, 1989) was used to evaluate the agreement between two univariate continuous measurements. Suppose two vectors of measurements from two raters are $X_1$ and $X_2$, then the CCC is expressed as

$$\mathrm{CCC} = 1 - \frac{\mathrm{E}\{(X_1 - X_2)^2\}}{\mathrm{E}\{(X_1 - X_2)^2 | X_1, X_2 \text{ are uncorrelated}\}} = \frac{2\rho_{12}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}.$$

Similar to the Pearson correlation coefficient, the CCC ranges from -1 to 1 with 1 indicating perfect agreement. If $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$, the CCC reduces to the Pearson correlation coefficient.

When there are replications from the raters, $\sigma_j^2$ can be further decomposed into between and within subject variability from Rater $j$ (i.e., $\sigma_j^2 = \sigma_{jW}^2 + \sigma_{jB}^2$). The CCC with replications (Barnhart 2005) can be further expressed as

$$\text{CCC} = \frac{2\rho_{12}\sigma_{1B}\sigma_{2B}}{2\sigma_{1B}\sigma_{2B} + (\mu_1 - \mu_2)^2 + (\sigma_{1B} - \sigma_{2B})^2 + \sigma_{1W}^2 + \sigma_{2W}^2}.$$

Barnhart (2002) also extended Lin's CCC (1989) to the case of multiple ($k > 2$) raters. Here the CCC is

$$\begin{aligned}\text{CCC}_o &= 1 - \frac{\text{E}\big\{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(X_i - X_j)^2\big\}}{\text{E}\big\{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(X_i - X_j)^2|X_1,\ldots,X_k \text{ are uncorrelated}\big\}}\\ &= \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\text{CCC}_{ij}\xi_{ij}}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\xi_{ij}},\end{aligned}$$

where $\xi_{ij} = \text{E}\big\{(X_i - X_j)^2|X_1,\ldots,X_k \text{ are uncorrelated}\big\} = \sigma_i^2 + \sigma_j^2 + (\mu_i - \mu_j)^2$. This $\text{CCC}_o$ can be interpreted as the weighted average of all pairwise CCCs, where higher weights are assigned to the pairs of raters whose ratings have higher variances and larger mean differences.

Barnhart (2002, 2007a) discussed the relationship between the ICC and the CCC. The differences between them are 1) the ICC are proposed for both random and fixed raters, while the CCC usually treats the raters as fixed; and 2) the ICC requires ANOVA model assumptions, while the CCC does not. However, he showed that if the ANOVA model assumptions are met, the CCC equals the ICC in specific cases. Even though Lin (1989) objected to the use of the ICC as a way of assessing agreement between methods of measurement, there are similarities between certain specifications of the ICC and the CCC. Chen and Barnhart (2008) provided a detailed discussion about the comparison of the ICC and the CCC for data with and without replicates. Moreover, the limitation of comparability of populations are present in both the ICC and the CCC. This means, the ICC and the CCC are strongly influenced by the variance of the trait in the sample/population in which it is assessed (Muller 1994). We investigate this limitation later in this section.

### 2.2.3 Iota coefficient and its extension

All the agreement measures discussed so far are for univariate data. Suppose a rater gives ratings on multiple features of a subject. One way to deal with multivariate ratings is to convert them into a univariate measure. This might result in some inconsistencies, however, due to the compression, and therefore, possible loss of information. Janson (2001) extended Cohen's kappa to a general case where the ratings from multiple raters are multivariate nominal or interval data. The iota coefficient is defined as

$$\iota = 1 - \frac{d_o}{d_e},$$

where $d_o$ is the observed disagreement between raters and $d_e$ is the expected disagreement by chance. Next, we follow with the detailed calculations of $d_o$ and $d_e$ for both continuous and nominal data.

For continuous data, suppose $x_{ijl}$ is the rating from Rater $j$ ($j = 1, \ldots, k$) on Subject $i$ ($i = 1, \ldots, n$) for Feature $l$ ($l = 1, \ldots, m$). The observed disagreement, $d_o$, is the average of the squared Euclidean distances between raters' ratings of the same subjects. This disagreement is

$$d_o = \left[n\binom{k}{2}\right]^{-1} \sum_{j<j'} \sum_{i=1}^{n} \sum_{l=1}^{m} (x_{ijl} - x_{ij'l})^2.$$

Similarly, the expected disagreement, $d_e$, is the average of the squared Euclidean distances between one rater's rating of a subject and any other rater's rating of any subject. This disagreement $d_e$ is

$$d_e = \left[n^2\binom{k}{2}\right]^{-1} \sum_{j<j'} \sum_{i'=1}^{n} \sum_{i=1}^{n} \sum_{l=1}^{m} (x_{ijl} - x_{i'j'l})^2.$$

Since the calculation of the iota coefficient is based on squared Euclidean distances, there is an equivalency to the ANOVA sum squares from a two-way layout. Based on the decomposition of the sums of squares in the two-way ANOVA (Table 2.4 with $r$=1), let the total sum of squares for the $l$th feature $SS_{Tl}$ be decomposed into the sum of squares between subjects $SS_{Bl}$ and the sum of squares within subjects $SS_{Wl}$, and $SS_{Wl}$ is further decomposed into the sum of squares between raters $SS_{Rl}$ and the residual term $SS_{el}$. Then $d_o$ and $d_e$ can also be expressed as

$$d_o = \left[n\binom{k}{2}\right]^{-1} k \sum_{l=1}^{m} SS_{Wl},$$

$$d_e = \left[n\binom{k}{2}\right]^{-1} \sum_{l=1}^{m} [(k-1)SS_{Tl} + SS_{Rl}].$$

With univariate interval data ($m$=1) and two raters ($k$=2), the iota coefficient reduces to the Cohen's weighted kappa $\kappa_w$ with the weights inversely proportional to the squared Euclidean distances between ratings.

With multivariate ratings, one important consideration is whether each feature contributes equally to the observed and expected distances. One reason that different features may contribute differently to the distances could be that features are measured on different scales and/or have unequal variances. Another reason is that disagreements for some features are more important than that from other features. Adding weights to the disagreements will take this into account. Suppose $w_l$ is the assigned weight for the $l$th feature that is incorporated into the calculations of $d_0$ and $d_e$, then the weighted observed and expected disagreement become:

$$d_o = \left[n\binom{k}{2}\right]^{-1} \sum_{j<j'} \sum_{i=1}^{n} \sum_{l=1}^{m} w_l (x_{ijl} - x_{ij'l})^2,$$

and

$$d_e = \left[n^2\binom{k}{2}\right]^{-1} \sum_{j<j'} \sum_{i'=1}^{n} \sum_{i=1}^{n} \sum_{l=1}^{m} w_l (x_{ijl} - x_{i'j'l})^2.$$

Correspondingly, the calculations of iota coefficient based on ANOVA decomposition are

$$d_o = \left[n\binom{k}{2}\right]^{-1} k \sum_{l=1}^{m} w_l SS_{Wl},$$

and

$$d_e = \left[n\binom{k}{2}\right]^{-1} \sum_{l=1}^{m} w_l \left[(k-1)SS_{Tl} + SS_{Rl}\right].$$

For nominal data, Janson (2001) suggested to simply sum the number of disagreements over features. The calculations for observed and expected disagreement are then

$$d_o = \left[n\binom{k}{2}\right]^{-1} \sum_{j<j'} \sum_{i=1}^{n} \sum_{l=1}^{m} \mathbb{1}\{x_{ijl} \neq x_{ij'l}\},$$

$$d_e = \left[n^2\binom{k}{2}\right]^{-1} \sum_{j<j'} \sum_{i'=1}^{n} \sum_{i=1}^{n} \sum_{l=1}^{m} \mathbb{1}\{x_{ijl} \neq x_{i'j'l}\}.$$

Similar to the interval data case, weights can be assigned to the disagreements from different features by including $w_l$. This weighted iota coefficient can be viewed as an extension of the kappa coefficient to the multivariate nominal data. When there are only two raters and one nominal-scale feature, the iota coefficient reduces to Cohen's kappa (1960). The interpretation of the iota coefficient is similar to the interpretation of the $\kappa$. A value of 1 indicates perfect agreement and the lower limit of $\iota$ is $-1/(k-1)$.

### 2.2.4 Limitations of the agreement indices for continuous data

A common limitation of all the agreement indices discussed so far is the difficulty in comparing indices across different population or studies. The reason is that they all depend on between-subject variability (Vangeneugden et al., 2004, 2005; Molenberghs et al., 2007; Barnhart et al., 2007). To demonstrate this, we consider two populations where Population 1 has true subject scores $\mathbf{x} \sim N(30, 5)$ and Population 2 has true subject scores $\mathbf{y} \sim \mathcal{N}(30, 10)$. For both populations, we consider two raters rating 50 subjects whose scores are distributed Gamma$(\mathbf{x}, 0.5)$ and Gamma$(\mathbf{y}, 0.5)$, respectively. This setup indicates Population 2 has larger between-subject variability but regardless of population, both raters give unbiased scores with the same precision. Because the rater distributions for a given subject are identical, we expect to see a good agreement between these two raters. Table 2.5 summarizes the average of ICC$_2$, ICC$_3$ and CCC for each population setting. The kappa coefficient is used for nominal or ordinal data so it is not calculated here. The iota coefficient reduces to the CCC when there is a single response feature. These averages are based on 100 simulated data sets.

Table 2.5.
ICC and CCC for two raters assessing 50 subjects from two populations

| Population | ICC$_2$ | ICC$_3$ | CCC |
|---|---|---|---|
| $\mathbf{x} \sim N(30, 5)$ | 0.291 | 0.292 | 0.287 |
| $\mathbf{y} \sim N(30, 10)$ | 0.610 | 0.610 | 0.605 |

Even though the rater distributions are identical in each population setting, all agreement indices are much smaller in Population 1, where there is less variability among subjects. Such a phenomenon will also occur with multivariate data because the calculation is based on the same ANOVA decomposition.

It is easy to understand this limitation by investigating the basic form of chance-corrected agreement indices:

$$\frac{\text{Between subjects variation}}{\text{Between subjects variation} + \text{Within subjects variation}}.$$

As long as the between subjects variation gets bigger, the index value will increase. Due to this fact, some researchers argue that such indices should be interpreted as a reliability measure that assesses the degree of differentiation of subjects from a population, rather than agreement (Vangeneugden et al., 2004, 2005; Molenberghs et al., 2007).

### 2.2.5 Coefficient of inter-rater variability (CIV)

Haber et al. (2005) proposed an approach to evaluate inter-rater agreement that does not have this limitation. Their coefficient of inter-rater variability (CIV) is defined as the ratio of the between-rater variability to the total rater variability. This coefficient compares the rater difference component relative to the total rater-related components (i.e., inter-rater component + intra-rater component), thus it is considered to be a more appropriate measure for inter-rater agreement.

In order to compare the CIV to the ICC, we follow the same notation of the ICC study designs (Section 2.2.2, page 37-40), that is, $\sigma_\alpha^2$, $\sigma_\beta^2$, $\sigma_{\alpha\beta}^2$, and $\sigma_e^2$ represent the variabilities due to subjects, raters, subject by rater interactions, and within rater error, respectively. CIV is defined as

$$\frac{\sigma_\beta^2 + \sigma_{\alpha\beta}^2}{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma_e^2}.$$

The CIV is an index between 0 and 1, and 1-CIV is interpreted as an inter-rater agreement measure with a value of 1 signifying perfect agreement and a value of 0 signifying complete disagreement.

For the Case 1 ICC calculation scenario, the study design for the CIV is altered to be a random set of $k$ raters is selected from a large rater population and each rater rates a random set of $n$ subjects from a large subject population. This is a more realistic design in rater agreement studies than the design presented in Table 2.3 because there is usually limited raters that can be randomly selected. The design layout is shown in Table 2.6. The inter-rater agreement index based on the CIV is $\psi = 1 - \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_e^2} = \frac{\sigma_e^2}{\sigma_e^2 + \sigma_\beta^2}$

Table 2.6.
Sources of variance for Case 1

| Source of Variance | *df* | Mean Squares |
|---|---|---|
| Between raters | $k-1$ | BMS |
| Subjects(rater) | $k(n-1)$ | WMS |

For the Case 2 and Case 3 scenarios, the CIV study designs are the same but the agreement index is calculated as $\psi = 1 - \frac{\sigma_\beta^2 + \sigma_{\alpha\beta}^2}{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma_e^2}$.

Barnhart et al. (2007a) proposed a coefficient of individual agreement (CIA), which emphasizes the interchangeability or switch-ability of multiple raters (or methods). He compared the CCC, the CIV, and the CIA in detail, given the scenarios with and without a reference/gold standard. In fact, when there is no gold standard, the CIA is equivalent to the CIV. In this chapter we focus on the discussion of agreement measures without references. To compare the CIV/CIA to the CCC, we assume there are two raters and each rater rates $n$ subjects $r > 1$ times. The CCC and the 1-CIV/CIA index $\psi$ can be written as:

$$\text{CCC} = \frac{2\rho_{12}\sigma_{1B}\sigma_{2B}}{2\sigma_{1B}\sigma_{2B} + (\mu_1 - \mu_2)^2 + (\sigma_{1B} - \sigma_{2B})^2 + \sigma_{1W}^2 + \sigma_{2W}^2},$$

$$\psi = \frac{(\sigma_{1W}^2 + \sigma_{2W}^2)}{2(1-\rho_{12})\sigma_{1B}\sigma_{2B} + (\mu_1 - \mu_2)^2 + (\sigma_{1B} - \sigma_{2B})^2 + \sigma_{1W}^2 + \sigma_{2W}^2}.$$

Both coefficients decreases when the correlation decreases. In contrast to the CCC, $\psi$ decreases when within-subject variability decreases and the between-subject variability increases. Moreover, $\psi$ is shown to depend less on the magnitude of $\sigma_B^2/\sigma_W^2$ than the CCC (Barnhart et al. 2007).

Compared to the ICC and the CCC, the CIV has a simple intuitive definition in terms of the difference between the scores assigned by different raters to the same subject while the ICC and the CCC use correlations to evaluate rater agreement. In Haber's paper (2005), a non-parametric estimation approach was proposed, thus the CIV is not subject to the ANOVA assumptions as the ICC is.

## 2.3 Modeling Patterns of Rater Agreement

The agreement measures discussed in the previous section are all single indices that focus on the degree of agreement. However, how raters differ from each other is also an important aspect in assessing rater agreement. For example, is there a systematic bias in one of the raters? Log-linear models and latent-class models are approaches that can assess the patterns of rater agreement for nominal or ordinal scale data.

### 2.3.1 Log-linear models

Assuming there are $k$ raters who categorize $n$ subjects into $m$ nominal categories, Tanner and Young (1985a) modeled the agreement structure using the following log-linear model:

$$\log(c_{ij\ldots l}) = \mu + \mu_i^{R_1} + \mu_j^{R_2} + \cdots + \mu_l^{R_k} + \delta_{ij\ldots l}, \ (\underbrace{i, j, \ldots, l}_{k} = 1, \ldots, m),$$

where $c_{ij\ldots l}$ is the expected count in the $(ij\ldots l)^{th}$ cell of the joint k-dimensional cross-classificiation of the ratings, $\mu$ is the overall effect, $\mu_h^{R_r}$ is the effect due to categorization by the $r$th rater in the $h$th category $(r = 1, \ldots, k; h = 1, \ldots, m)$, and $\sum_{i=1}^m \mu_i^{R_1} = \cdots \sum_{i=1}^m \mu_i^{R_k} = 0$. The additional term $\delta_{ij\ldots l}$ indicates agreement beyond chance for the $(ij\ldots l)^{th}$ cell.

When modeling ordinal data, Agresti (1988) argued that ordinal scale ratings always exhibit a positive association between ratings. That is, there is a tendency for high (low) ratings by one rater to be accompanied by high (low) ratings by another rater. He proposed a log-linear model with linear-by-linear association, a combination of Tanner and Young's (1985a) log-linear model with the uniform association model (Goodman 1979). Assuming only two raters for simplicity, the corresponding log-linear model is

$$\log(c_{ij}) = \mu + \mu_i^{R_1} + \mu_j^{R_2} + \beta\lambda_i\lambda_j + \delta_{ij},$$

where $\lambda_1 < \cdots < \lambda_m$ are fixed scores assigned to the response categories.

To investigate different patterns of agreement beyond chance, $\delta_{ij\ldots l}$ can be specified and tested. For example, $\delta_{ij\ldots l} = \delta_i I_{\{i=j=\cdots=l\}}$, where $I$ is an indicator function, assumes non-homogeneous pattern of agreement by response category. The specification $\delta_{ij\ldots l} = \delta I_{\{i=j=\cdots=l\}}$ assumes homogeneous agreement among raters and the specification $\delta_{ij\ldots l} = 0$ assumes nothing beyond chance agreement. These three model representations are nested so likelihood ratio tests can be constructed to compare them.

### 2.3.2 Latent-class models

Latent-class models were proposed to investigate inter-rater agreement (Aickin, 1990; Uebersax and Grove, 1990; Agresti, 1992) using unobserved (latent) variables. It is assumed that there is an unobserved categorical scale $X$, with $V$ categories, such that subjects in each category of $X$ are homogeneous. The basic latent-class model applied to nominal scale data is:

$$\log(c_{ij\ldots lv}) = \mu + \mu_i^{R_1} + \mu_j^{R_2} + \cdots + \mu_l^{R_k} + \mu_v^X + \mu_{iv}^{R_1X} + \mu_{jv}^{R_2X} + \cdots + \mu_{lv}^{R_kX},$$

$$\underbrace{i, j, \ldots, l}_{k} = 1, \ldots, m, \text{ and } v = 1, \ldots, V.$$

Latent-class models applied to ordinal data of raters' ratings treat the unobserved variable $X$ as ordinal and assume a linear-by-linear association between each classification and $X$, assigning scores for both observed scale and unobserved scale (Agresti and Lang, 1993). For simplicity, the basic latent-class model for ordinal data from two raters is

$$\log(c_{ijv}) = \mu + \mu_i^{R_1} + \mu_j^{R_2} + \mu_v^X + \beta^{R_1X}\lambda_i x_v + \beta^{R_2X}\lambda_j x_v,$$

where $\{\lambda\}$ and $\{x\}$ are the assigned scores to response categories and latent categories respectively.

A strong agreement, in terms of relatively high probability of identical ratings, requires both similar marginal distributions and a strong positive association. For example, in a simple case that the ordinal response is binary (1 and 2) and there are 2 raters ($R_1$ and $R_2$), one can simply compare the marginal distributions using odds ratios:

$$\frac{P(R_1 = 1|X = v)/P(R_1 = 2|X = v)}{P(R_2 = 1|X = v)/P(R_2 = 2|X = v)} = \exp(\delta_{R_1} - \delta_{R_2}),\ v = 1, \ldots, V,$$

where $\delta_{R_1} = \mu_1^{R_1} - \mu_2^{R_1}$ and $\delta_{R_2} = \mu_1^{R_2} - \mu_2^{R_2}$. The variation in marginal distributions can be addressed by variation in the $\delta$'s parameters. On the other hand, the strength of association is induced by the association between each rater and the latent variable.

Thus, the strength of agreement improves in the two-way tables as $\{\delta\}$ move toward uniformity and the association between each rater and $X$ increases. This model can be expanded to higher orders and standard likelihood ratio tests can be constructed to compare models, but we do not discuss the details here.

## 2.4 Agreement with Compositional Data

The existing agreement indices and methods are for univariate or multivariate nominal, ordinal, or continuous data. However, not a single agreement index can be directly used to assess agreement of compositional data. Since compositional data contain D components ($D > 2$), an agreement measure for multivariate data is needed.

Based on the idea of the iota coefficient, if there is an appropriate function to measure the distance between two compositional vectors, then this distance measure can replace the squared Euclidean distance to calculate the observed and expected disagreement. Aitchison distance and Mahalanobis distance with clr transformation are two candidates that meet the three principals in compositional data analysis (Aitchison 1986). With this in mind, we investigate two questions.

***Question 1: How does the iota coefficient behave using clr transformed compositional data?***

Suppose the true compositional scores of 10 subjects $\boldsymbol{\mu}_i\, (i = 1, \ldots, 10)$ follow a Dirichlet distribution $\mathcal{D}(\boldsymbol{\mu}, k_s)$. Rater scores are obtained assuming they are $\mathcal{D}(\boldsymbol{\mu}_i, k)$, $(i = 1, \ldots, 10)$, where $k$ is the intra-rater variability parameter. Ratings coming from the same Dirichlet distribution indicate a perfect agreement between the two raters. When $k_s$ gets bigger, the variability among subjects tend to be smaller.

The observed and expected disagreement is calculated based on the squared Aitchison distance and squared Mahalanobis distance. Mahalanobis distance for two compositional vectors is $d_M(\mathbf{x}, \mathbf{y}) = \sqrt{[\mathrm{clr}(\mathbf{x}) - \mathrm{clr}(\mathbf{y})]' \mathbf{S}^{+} [\mathrm{clr}(\mathbf{x}) - \mathrm{clr}(\mathbf{y})]}$ where $\mathbf{S}^{+}$ is the Moore-Penrose pseudoinverse inverse of the data covariance matrix. The clr transformation results in a singular covariance matrix of transformed data, thus a generalized inverse of $\mathbf{S}$ is needed. Table 2.7 summarizes the limitation of the existing agreement indices: variability among subjects is needed to get a proper assessment of similarity among raters. When there is little or no variability among the slides, the intra-rater variability dominates and the iota coefficient is small.

Table 2.7.
Iota coefficient based on Aitchison distance and Mahalanobis distance

| Distributions of Subjects | Iota(Squared Aitchison) | Iota(Squared Mahalanobis) |
|---|---|---|
| $\mathbf{x} \sim \mathcal{D}_{10}((0.3, 0.4, 0.3), k_s = 10)$ | $0.818_{(0.07)}$[a] | $0.800_{(0.07)}$ |
| $\mathbf{x} \sim \mathcal{D}_{10}((0.3, 0.4, 0.3), k_s = 50)$ | $0.478_{(0.15)}$ | $0.409_{(0.16)}$ |

[a] values were calculated based on 1000 simulations and presented as $\text{mean}_{(sd)}$

***Question 2: Does a null distribution of distances exist when raters agree perfectly?***

To answer this question, we address whether the Aichison distance and Mahalanobis distance are invariant to the mean and variance of the Dirichlet distribution. The hope is that if a null distribution can be found invariant to the mean, we can then compare the observed distance distribution or average distance to the null distribution to assess the degree of agreement.

Consider two populations, where Population 1 has 50 pairs of scores from the Dirichlet $\mathcal{D}((0.3, 0.4, 0.3), k_s = 50)$ and Population 2 has 50 pairs of scores from $\mathcal{D}((0.1, 0.1, 0.8), k_s = 50)$. For each population, squared Aitchison distances and squared Mahalanobis distances are calculated between pairs of scores and the averages of the distances are stored. Repeat this procedure 1000 times within each population to get a distribution of averaged distances.

Comparisons between the two different distributions of subjects reveal significant differences in distributions of distances (Table 2.8). This suggests that Aitchison distance and Mahalanobis distance based on compositional data are not invariant to the compositional means of the Dirichlet distribution.

Table 2.8.
Distributions of squared Aitchison distance and squared Mahalanobis distance between compositional pairs

| Distributions of Subjects | Squared Aitchison | Squared Mahalanobis |
|---|---|---|
| $\mathbf{x} \sim \text{Dir}_{10}((0.3, 0.4, 0.3), k_s = 50)$ | $0.267_{(0.09)}$[a] | $2.356_{(0.64)}$ |
| $\mathbf{x} \sim \text{Dir}_{10}((0.1, 0.1, 0.8), k_s = 50)$ | $0.665_{(0.28)}$ | $2.181_{(0.65)}$ |

[a] mean$_{(sd)}$ based on 1000 simulations.

***Potential solution: Can we fit the observed compositional ratings with some parametric distribution?***

In the rater agreement problem, if we assume the compositional rating(s) from each rater on each subject comes from a certain distribution, then the distance between two compositional vectors can be viewed as the probabilistic distance between two distributions. As we discussed in Chapter 1, there are two parametric distributions, the logistic Normal and the Dirichlet, that are available for modeling compositional data. However, to be able to estimate the parameters of the distribution, we need replicate observations of each subject for each rater.

We consider the Dirichlet distribution primarily because it has fewer parameters. A D-part compositional rating $\mathbf{x}_{ij} \sim \mathcal{D}(\boldsymbol{\mu}_i, k)$ involves $D$ parameters. The vector $\boldsymbol{\mu}_i$ is the underlying mean of subject $i$ from Rater $j$. When $k$ increases, the variance within the rater, as well as the negative correlations between pairs of vector elements gets smaller. That is, the rater is more consistent in rating the same subject multiple times.

Rauber et al. (2008) discussed different distance measures for probability distributions and concluded that only the Chernoff distance is an appropriate metric to measure the distance between two Dirichlet distributions. The definition of the Chernoff distance is:

$$D_c(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b) = -\log \Big\{ \int_{\mathbf{x}} p_a^{1-\lambda}(\mathbf{x}; \boldsymbol{\theta}_a) p_b^{\lambda}(\mathbf{x}; \boldsymbol{\theta}_b) d\mathbf{x} \Big\}$$

where $0 < \lambda < 1$. Bhattacharyya distance $D_B$ is a special case of Chernoff distance when $\lambda = 1/2$. The Bhattacharyya coefficient $\rho_B$ is defined as the logarithm of the negative Bhattacharyya distance so that $\rho_B$ is between 0 and 1. The form of $\rho_B$ between two Dirichlet distributions is:

$$\rho_B(\mathcal{D}_a, \mathcal{D}_b) = \frac{\prod_{c=1}^{D} \Gamma((k_a\mu_{ac} + k_b\mu_{bc})/2)}{\Gamma(\frac{1}{2}(k_a + k_b))} \times \frac{\sqrt{\Gamma(k_a)\Gamma(k_b)}}{\sqrt{\prod_{c=1}^{D} \Gamma(k_a\mu_{ac})}\sqrt{\prod_{c=1}^{D} \Gamma(k_b\mu_{bc})}}$$

In general, the Bhattacharyya coefficient $\rho_B$ is a measure of the amount of overlap between two statistical samples or populations. If $\rho_B = 1$, the two samples overlap perfectly, while $\rho_B = 0$ means there is no agreement between two raters. This means, if we know the distributions of raters' scores, Bhattacharyya coefficient is a potential candidate to measure overall agreement between raters. We will investigate in detail the behavior of the Bhattacharyya coefficient in Chapter 3.

## 2.5 Discussion

In Chapter 1, we reviewed what common compositional data problems have been addressed in the literature. In this chapter, we reviewed the existing indices and methods that can be used to assess inter-rater agreement with nominal, ordinal, and continuous data, for both univariate and multivariate responses. We pointed out the strengths and limitations of these approaches. Our goal is to develop a method to assess inter-rater agreement with compositional data. Some preliminary investigations could not find a simple adjustment of a current approach. However, the idea of modeling the compositional vectors using an appropriate parametric distribution and then comparing the probabilistic distances (discussed in Section 2.4) is very appealing and similar in flavor to the CIV. The difficulty we have to overcome is the lack of replicate ratings on each subject from each rater. This puts heavy restrictions on the parametric distributions we can consider and how we can model the relationship between scores from two raters. In the next chapter, we detail our use of Dirichlet distributions and the Bhattacharyya coefficient as a means to assess agreement of compositional data.

# CHAPTER 3. ASSESSING INTER-RATER AGREEMENT FOR COMPOSITIONAL DATA

## 3.1 Motivation

All the popular agreement indices and methods described in Chapter 2 are designed either for univariate data or for unconstrained multivariate data. The need for agreement methodology designed for compositional data arises primarily within pathology and the medical sciences, where similarity in compositional scores is crucial for consistent prognosis and treatment.

We start this chapter with a brief description of our motivating application, immunohistochemistry (IHC) assays, and an overview and critique of the current agreement measures used by researchers to compare IHC scores. We then provide a description of our proposed methodology, both in terms of model concept/structure and approach to inference. We conclude the chapter with a few simulation studies to demonstrate the benefits of our approach relative to the currently-used agreement methodology.

### 3.1.1 Immunohistochemistry (IHC) assays

IHC is a process of detecting targeted antigens through their interaction with tagged antibodies. An antigen is any substance (e.g., protein, chemical, pollen, bacteria) that causes one's immune system to produce antibodies against it. To visualize an antibody-antigen interaction, the antibody of a target antigen is tagged with fluorescein or other enzyme that will catalyze a color-producing reaction.

IHC staining is widely used in the diagnosis of cancerous tumors, where overexpression (or underexpression) of certain proteins predicts disease status. Tagged antibodies bind to these proteins so higher (lower) intensity of the color-reaction indicate cancer is present.

An IHC assay typically involves a sample of tissue, which consists of a very large number of cells. After the staining process, a trained pathologist exams the assay slide under a microscope and provides a compositional vector score. This vector represents the percent of cells in the sample that fall in each of four ordered staining categories, traditionally labeled negative, weak, moderate, and positive. Since different pathologists will examine and score different assays/slides, agreement between pathologists is very important for consistency in prognosis and therapy. Pathologists are trained how to score and numerous studies are performed to assess the agreement among pathologists.

### 3.1.2 Current IHC agreement methods

The design of a typical agreement study between two pathologists, or a pathologist and an automated reader, is shown in Table 3.1. The two raters score each of the $n$ slides once. Each slide is typically a different tissue sample with a different mean, so there are no replicates to assess rater consistency (i.e., test-retest reliability or intra-rater variability).

This setup is the same as the study design described in Case 3 on page 36 (Table 2.3) for assessing the ICC. The difference here is that the response is a compositional vector instead of a univariate continuous score.

Table 3.1.
Basic layout for an IHC agreement study

| Slide | Rater A | Rater B |
|---|---|---|
| 1 | $\mathbf{x}_1 = (x_{11}, x_{12}, x_{13}, x_{14})$ | $\mathbf{y}_1 = (y_{11}, y_{12}, y_{13}, y_{14})$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $\mathbf{x}_n = (x_{n1}, x_{n2}, x_{n3}, x_{n4})$ | $\mathbf{y}_n = (y_{n1}, y_{n2}, y_{n3}, y_{n4})$ |

Most agreement assessments involve converting the composite vectors into univariate responses and then using the ICC (or CCC). The most popular conversion is called the H-score. The H-score is a weighted sum of the percent components and ranges between 0 and 300. The weights are simply the whole numbers 0, 1, 2 and 3, respectively,

$$\text{H-score} = 0 \times \%\ \text{negative} + 1 \times \%\ \text{weak} + 2 \times \%\ \text{moderate} + 3 \times \%\ \text{positive}. \quad (3.1)$$

The H-score method was first introduced by McCarty et al. (1985) and quickly grew in popularity (e.g., Michelle, 1999; Flanagan et al., 2008; Bhargava et al. 2009).

There are, however, potential drawbacks with the H-score. First, the conversion to an H-score is not one-to-one and results in some loss of information (Etzioni et al. 2005). For example, suppose the two scores for a sample are: $(20\%, 70\%, 10\%, 0\%)$ and $(40\%, 35\%, 20\%, 5\%)$. The first vector suggests the cells are predominately weak, whereas the second vector suggests a more uniform mixture of the first three types. Their H-scores, however, are the same (H-score=90), suggesting these two vectors are comparable.

The second drawback with the H-Score is not particular to the H-score but rather the ICC (or CCC) methodology. This was discussed in Chapter 2 and involves the interpretability of the ICC (or CCC) across studies. That is, they depend on the between-slide variability. The larger the variability among slides, the closer the ICC (CCC) index is to 1. Although the CIV/CIA is insensitive to this, we have not seen any application of the CIV/CIA using the H-score.

Despite these concerns with the H-score, or any other univariate conversion, there has been no literature on the analysis of the score vectors. This may, in part, be due to the fact that the one multivariate approach, the iota coefficient, is not designed to handle the sum-to-one restriction that these vectors have. The iota coefficient also suffers from the same drawback as the ICC in regards to its dependence on the variability among slides.

Our goal is to fill this void and propose a multivariate approach that explicitly considers both the frequency and intensity of tissue staining. We also want an approach that is insensitive to the variability among slides, thereby allowing one to make comparisons across studies. To do this, we first need to consider, conceptually, how a rater scores a slide and how raters will differ in scoring. That is the focus of the next section.

### 3.1.3 Latent model for determining average percents

To conceptualize how a rater assigns percents to each category, we assume that the rater visualizes a continuous spectrum of intensities (low to high) and determines "cutpoints" along this spectrum to define the categories. It is our belief that these cutpoints are rater-specific and that differences in these cutpoints are what cause differences in vector scores.

Consider the two red (low intensity) to yellow (high intensity) images pictured in Figure 3.1. Both images span a rater's spectrum of intensities. The rate at which each image changes from red to yellow is related to the CDF of cell intensities on the slide. The first image represents a slide that has a relatively uniform distribution of intensities and the second image represents a slide with larger high intensity staining (i.e., more yellow than red).

Now suppose that Rater A and B are asked to apportion the spectrum to "No yellow", "Moderate Yellow" and "Strong Yellow" categories. The marks labeled $(A_1, A_2)$ and $(B_1, B_2)$ along the bottom of each spectra represent the chosen breakpoints, which in turn define the mean response of each rater for that image.

To handle the unit-sum constraint, we consider describing the distribution of cell intensities on each slide using logistic distributions. Rather than varying the logistic distributions across slides, we consider the standard logistic distribution and vary Rater A's cutpoints for each slide to alter the mean. Figure 3.2 represents the same two images/mean responses in Figure 3.1 under this construction. No information regarding Rater A is lost in doing this. The red solid lines represent the cutpoints of

Rater A with the percents defined as the area under the curve, which defines Rater A's percent of each category.

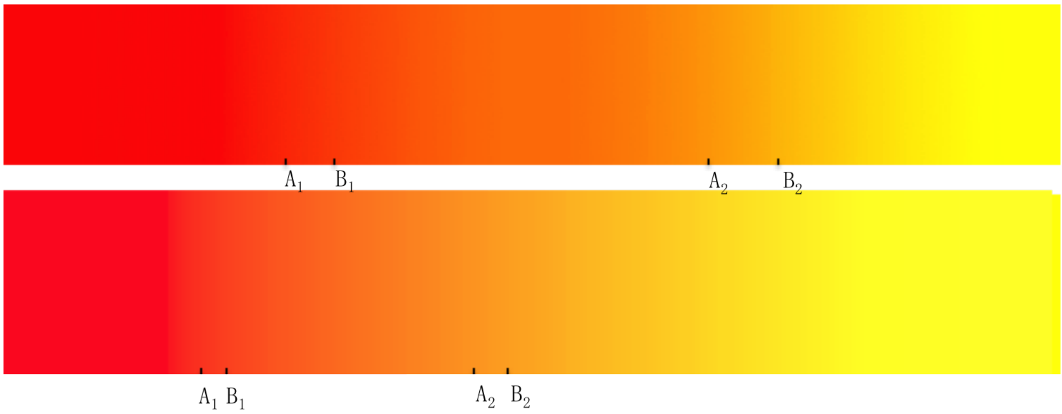


Fig. 3.1. Red-yellow spectrum

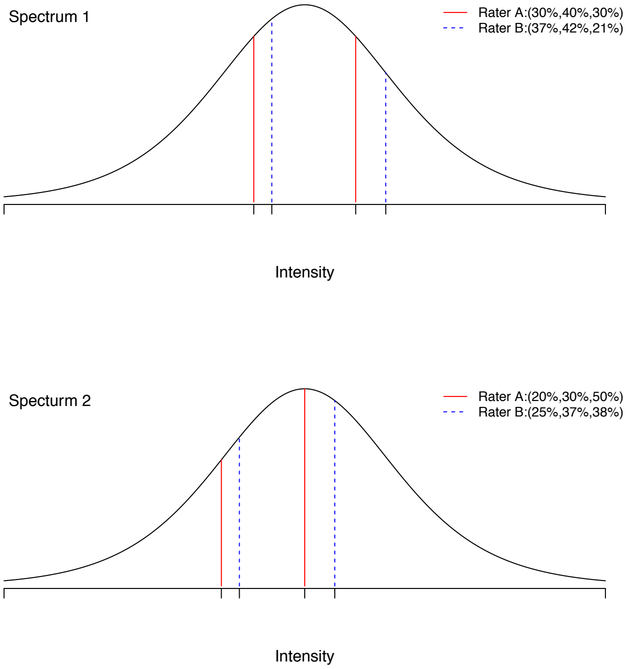


Fig. 3.2. Logit link of slide means and rater-specific cutpoints

If Rater B were to agree perfectly with Rater A, the cutpoints would match and the shifts would be zero. To link Rater B's cutpoints to Rater A's, we allow for rater differences by considering shifts in the cutpoints. For each of the two distributions (slides) Rater B's cutpoints, and thus percents, are defined by adding these shifts to Rater A's cutpoints. These are represented by the blue dashed lines.

The most general model would allow these shifts to vary slide to slide. We, however, keep them constant. The reasoning for this is two-fold. First, we expect there to be some consistency across slides. If Rater B scores one image on average to have more yellow, all images on average should be scored to have more yellow. Second, we simply don't have enough information to consider more general models.

Given the use of the standard logistic distribution and constant shifts to describe the mean score vectors, we're imposing a proportional odds relationship between the cumulative percents of the two raters. Let $A$ and $B$ denote the events that Rater A and Rater B, respectively, categorizes a randomly chosen cell from a slide. The proportional odds relationship means that for any slide,

$$\frac{\text{odds}(B \leq C_k)}{\text{odds}(A \leq C_k)} = e^{\delta_k},\ k = \{1, 2\},$$

where $\delta_k$ is defined as the shift between the two raters at the boundary between category $C_k$ and $C_{k+1}$. Other distributions, such as the standard Normal distribution, could be used in place of the standard logistic. While this switch would eliminate this proportional odds relationship, we do not expect the choice to have a substantial impact on the shift parameters.

Given these additive shifts, an additional caveat is that Rater B's cutpoints cannot cross. In other words, we need to avoid the situation when $\log\big[\text{odds}(A \leq C_k)\big] + \delta_k > \log\big[\text{odds}(A \leq C_{k+1})\big] + \delta_{k+1}$. In a 3-dimensional compositional vector, this would most likely happen when the two cutpoints are very close. However, we wouldn't expect a slide to be bimodal in intensity so this event will be rare. More concerns would be when a slide has one dominating category. This is where we have to look out for crossing cutpoints.

Our approach links the two rater means using constant shifts on the logit scale. An alternative approach would be to consider constant shifts on the intensity scale. In other words, assuming Rater's A and B have fixed cutpoints on the intensity spectrum and the logistic distribution associated with each slide is changing in shift and scale (Figure 3.3). With only 3 categories, this model involves 4 cutpoints (instead of 2 cutpoints for each slide and 2 shift parameters) but involves the estimation of two logistic distribution parameters for each slide. Thus, with 3 categories, it has the same number of parameters as our proposed model. When the number of categories is over 3, however, it results in fewer parameters than our proposed proportional odds model and thus is more restrictive.

With this approach, the log odds ratio is:

$$\frac{\text{odds}(B \leq C_k)}{\text{odds}(A \leq C_k)} = e^{\frac{\eta_{Ak}-\eta_{Bk}}{s_i}},\ k = \{1, 2\},$$

where $s_i$ is the scale parameter for the logistic distribution of slide $i$, and $\eta_A$ and $\eta_B$ are the fixed cutpoints for Rater A and B, respectively. Based on this model assumption, when one category dominates on a slide (i.e., $s_i$ is very small), the corresponding log odds ratio tends to be really big.

In practice, we often observe that two raters score consistently when one category dominates. This deviates from the expectation under this model thereby supporting the proportional odds model. If the two raters were to become more disparate when one category dominates, then the fixed cutpoint model would be the better choice.

To illustrate the differences between these two models, we use a simplex plot assuming the scores are 3-part compositions. Figure 3.4 plots six pairs of means from Rater A and Rater B, where "+" represents Rater A's mean percents on six different slides. The "1" and "2" are Rater B's means assuming proportional odds and assuming fixed cutpoints, respectively. We considered shifts of (0.6, 0.3) on both the intensity and logit scales. Thus, if the slide had a standard logistic distribution of intensities, the two approaches would result in the same means for Rater B.

The slide distributions were obtained by sampling the location parameter from a Normal $\mathcal{N}(0, 1)$ and the scale parameter from a Gamma(2, 0.5). On this simplex, we

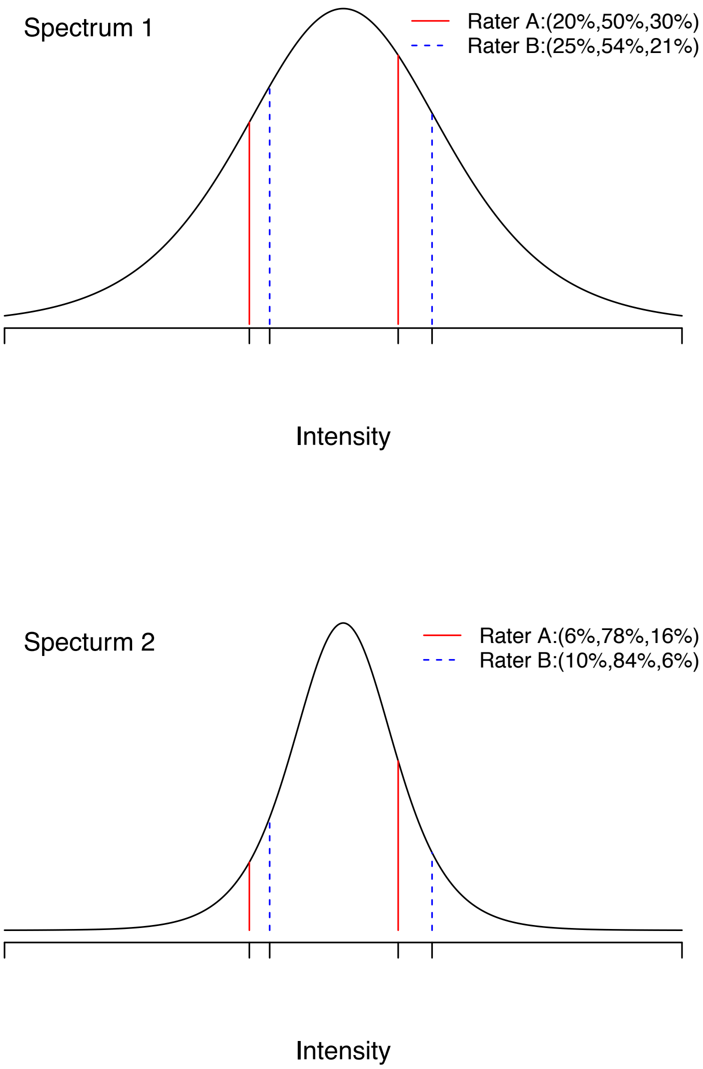


Fig. 3.3. Fixed cutpoints of raters on different slides

can see that all Rater B's means are shifted to the lower left corner due to the fact that the two shift parameters are positive.

For most of the means, the differences between "1" and "2" are not that substantial. There are two means, however, where the difference is more profound. These are both cases where Rater's A mean is dominated by the second category.

Through discussions with pathologists and examination of real data, we feel the proportional odds model is more realistic. As a result, we will focus the remainder of this chapter on just the proportional odds model.

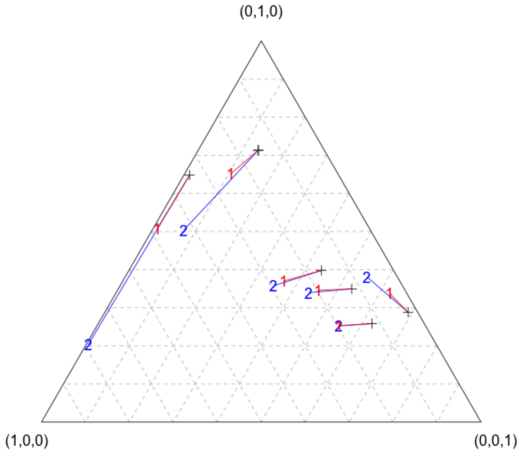


Fig. 3.4. Six pairs of 3-part compositions

### 3.1.4 Candidate distributions for response scores

Now that we have a model that links the two mean vectors for each slide, we need a model to describe the variation in response about these means. In Chapter 1, we discussed two distributions used to describe compositional vectors. These two distributions will be used here to model the inherent intra-rater variability.

Recall if a D-dimensional vector $\mathbf{x}$ follows a Dirichlet $\mathcal{D}_D(\boldsymbol{\mu}, k)$, it has the moment properties:

$$\mathrm{E}(x_i) = \mu_i,\ \mathrm{Var}(x_i) = \frac{\mu_i(1-\mu_i)}{k+1},\ \mathrm{Cov}(x_i, x_j) = \frac{-\mu_i\mu_j}{k+1},\ (i \neq j = 1, \ldots, D).$$

The dispersion parameter $k$ describes the intra-rater variability. For any $\boldsymbol{\mu}$ a bigger $k$ means less intra-rater variability. If we use the Dirichlet distribution to describe the variation about the mean, it only requires one parameter or $n$ parameters if we were to vary it across slides, thus 2 or $2n$ if we allow it to vary across two raters or across raters and slides.

The other distribution is the logistic Normal. Recall if a D-dimensional vector $\mathbf{x}$ follows a logistic Normal distribution $\mathcal{L}_D(\mathbf{m}, \Sigma)$, it has the following moment properties:

$$\mathrm{E}\left(\log(\frac{x_i}{x_D})\right) = m_i,\ \mathrm{Var}\left(\log(\frac{x_i}{x_D})\right) = \sigma_{ii},\ \mathrm{Cov}\left(\log(\frac{x_i}{x_D}), \log(\frac{x_j}{x_D})\right) = \sigma_{ij},$$

$$(i \neq j = 1, \ldots, D-1).$$

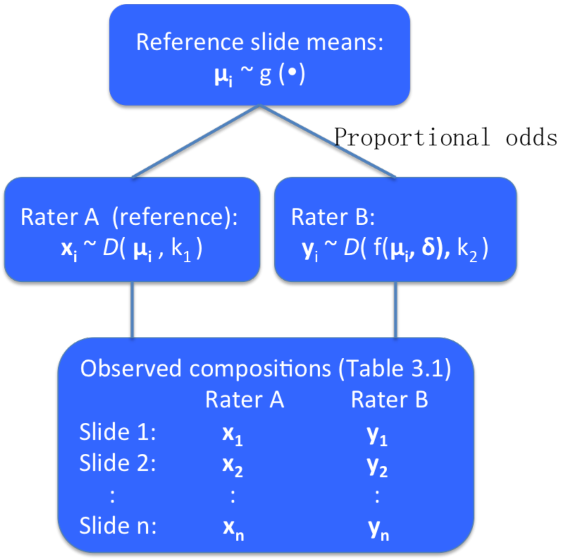
For a given mean, the logistic Normal distribution requires $D(D-1)/2$ parameters to describe the intra-rater variability, thus $nD(D-1)/2$ or $nD(D-1)$ if we allow it to vary across two raters or across raters and slides. Compared to the Dirichlet distribution, this distribution is more flexible and can describe both positive and negative covariances among score components. Its drawbacks are the increased number of parameters and the fact that there are no closed-form solutions for $\mathrm{E}(x_i)$ and $\mathrm{Var}(x_i)$. This greatly increases the computational complexity when using this distribution.

## 3.2 Hierarchical Model & Notation

We use Figure 3.5 to describe the hierarchical structure of our complete model. We assume the $n$ slides in the study are sampled randomly from a population distribution $g(\cdot)$. This distribution can be any distribution that accommodates compositional data. We will specify choices for this distribution in the next section. Two raters score each of $n$ slides once. Thus, we have $n$ pairs of IHC scores $\{\mathbf{x}_i, \mathbf{y}_i\}$, each score a $D$-part compositional vector. The observed data set for a typical agreement study is shown in Table 3.1 (page 56).

For each slide $i$, the score mean $\boldsymbol{\mu_i}$ represents the mean score vector for Rater A. If the Dirichlet distribution is used to describe the observed score vectors, we assume

$$\mathbf{x}_i \sim \mathcal{D}_D(\boldsymbol{\mu}_i, k_1)\,\text{and}\,\mathbf{y}_i \sim \mathcal{D}_D(f(\boldsymbol{\mu}_i, \boldsymbol{\delta}), k_2),$$



| | Rater A | Rater B |
|---|---|---|
| Slide 1: | $\mathbf{x}_1$ | $\mathbf{y}_1$ |
| Slide 2: | $\mathbf{x}_2$ | $\mathbf{y}_2$ |
| : | : | : |
| Slide n: | $\mathbf{x}_n$ | $\mathbf{y}_n$ |

Fig. 3.5. Hierarchical model structure

where $k_1$ and $k_2$ are the intra-rater variability parameters and $f(\boldsymbol{\mu}_i, \boldsymbol{\delta})$ are Rater B's means determined from Rater A's means and the shift parameters. We describe $f(\cdot)$ for slide $i$ as follows:

(i) Determine the logistic cutpoints of Rater A:

$$\eta_{iJ} = \log\left(\frac{\sum_{j=1}^{J} \mu_{ij}}{1 - \sum_{j=1}^{J} \mu_{ij}}\right), \ (J = 1, \ldots, D-1).$$

(ii) Shift the cutpoints:

$$\eta_{yiJ} = \eta_{iJ} + \delta_J, \ (J = 1, \ldots, D-1),$$

making sure the $\delta_J \in \mathbb{R}$ do not result in flipping cutpoints.

(iii) Back-transform to obtain the mean vector of Rater B.

To summarize, the unknown parameters in this model include:

(i) the $D-1$ shift parameters $\delta_1, \ldots, \delta_{D-1}$.

(ii) the intra-rater variability parameters $k_1$ and $k_2$.

(iii) the $n$ reference mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$.

### 3.2.1 Frequentist Estimation

A typical agreement study involves a single pair of scores from each of $n$ slides. We assume these observations are Dirichlet distributed but we do not know their underlying means (nor the intra-rater variability parameters) that generate the scores. Our model, however, assumes that for each pair, the difference in the logits of the cumulative means is $\delta_J$. In other words,

$$\delta_J = \text{logit}\left(\sum_{j=1}^{J} \mu_{yij}\right) - \text{logit}\left(\sum_{j=1}^{J} \mu_{xij}\right),\ J = 1, \ldots, D-1,$$

This means we can estimate the $\delta$'s using the average difference in observed logits. In other words, we use $\sum_{j=1}^{J} x_{ij}$ and $\sum_{j=1}^{J} y_{ij}$ as our estimates for the cumulative means of each slide and compute the average difference in their logits. The standard errors of these $\hat{\delta}$'s, however, depend on the unknown means and $k$'s so we cannot easily compute them.

We can estimate the intra-rater variabilities only if we assume $k_1 = k_2 = k$. For example, we can take an MLE approach for the observed $\mathbf{y}$ using the observed $\mathbf{x}$ as our estimates for the unknown Rater A's means and compute Rater B's means using the estimated $\delta$'s. This estimate tends to underestimate the true variability parameter (i.e., estimates there to be more intra-rater variability than there truly is). We will compare this frequentist estimation to our Bayesian estimation later in the simulation studies.

When there are replicates, we can take a method of moments approach using each slide's $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ for our estimates of the cumulative means and use the average difference in their logits as the estimate for $\delta$'s. Similarly the variances of each cumulative mean or the variances of each cumulative logit can be used to estimate $k$. For example, in a simple Beta distribution case (i.e., $x \sim \text{Beta}\left(k\mu, k(1-\mu)\right)$), the mean of the cumulative logit is $\text{E}(\log(x/(1-x)) = \digamma(k\mu) - \digamma(k(1-\mu))$ and the variance is $\text{Var}(\log(x/(1-x)) = \digamma_1(k * \mu) - \digamma_1(k * (1-\mu))$, where $\digamma$ and $\digamma_1$ are the digamma and trigamma functions.

The inclusion of replicates also provides the opportunity to assess the proportional odds assumption because we can separately estimate distributions of $\mathbf{x}$ and $\mathbf{y}$ for each slide. Our focus, however, is on a typical study that has no replicates so we leave this discussion to the future research section (Chapter 5).

### 3.3 Bayesian Inference via MCMC

Based on our hierarchical model structure, we are interested in the posterior distribution $\pi(\boldsymbol{\delta}, k_1, k_2|\text{data})$. To simplify the calculations, we consider the complete posterior

$$\pi(\boldsymbol{\delta}, \boldsymbol{\mu_1}, \cdots, \boldsymbol{\mu_n}, \mathbf{k}|\text{data}) \propto \pi(\text{data}|\boldsymbol{\delta}, \boldsymbol{\mu_1}, \cdots, \boldsymbol{\mu_n}, \mathbf{k})\pi(\boldsymbol{\delta}, \boldsymbol{\mu_1}, \cdots, \boldsymbol{\mu_n}, \mathbf{k}),$$

where $\boldsymbol{\mu_1}, \cdots, \boldsymbol{\mu_n}$ are the latent means. Bayesian inference provides a natural framework to incorporate these latent variables in our analysis. In addition, it allows us to borrow information across pairs of scores (Bayesian shrinkage) to improve the precision of our estimates. Finally, the posterior samples allow us to directly assess the uncertainty in the parameters.

Shrinkage estimators are commonly used in situations with a lack of replication. The idea is to move the Dirichlet mean estimates closer to a provided constant value (in our case, provided by the data) so that the resulting parameter estimates (both means, shifts, and dispersion parameters) have improved mean square error. A well-known example of this approach, and similar in flavor to our problem, is the James-Stein estimator for a set of Normally distributed random variables, each with an unknown mean. The raw estimator of each mean would be the observed value. James-Stein suggested shrinking these means towards a common value based on a ratio of the variability between observations versus the variability of an observation. We consider this shrinkage approach in our Bayesian inference and describe it in next section.

### 3.3.1 Priors

As described in Section 1.1.4, there are two common distributions used to model compositional data. The Dirichlet distribution describes the variability about the mean using a single parameter $k$. The logistic Normal, on the other hand, requires a variance-covariance matrix that can involve up to six parameters when modeling 4-part compositional vectors.

In our model, we need to describe a distribution for the reference means and a distribution that describes the observed scores given a mean vector. For the latter, we've chosen to use the Dirichlet distribution because we simply do not have enough information to consider the more flexible logistic Normal distribution. For the distribution of reference means $g(\cdot)$, either distribution can be considered.

For the unknown reference means $\boldsymbol{\mu}_i$ $(i = 1, \ldots, n)$, we consider them coming from a Dirichlet prior $\mathcal{D}(\boldsymbol{\mu}_p, k_p)$, where $\boldsymbol{\mu}_p$ and $k_p$ are population-level parameters. We take an empirical Bayes approach and use the observed data $\mathbf{x}$ to estimate these two hyperparameters,

$$\hat{\boldsymbol{\mu}}_p = \mathrm{E}(\boldsymbol{\mu}) = \bar{\mathbf{x}}, \tag{3.2}$$

$$\hat{k}_p = \max\left(\frac{\hat{\boldsymbol{\mu}}_p(1-\hat{\boldsymbol{\mu}}_p)}{\mathrm{var}(\mathbf{x})} - 1\right). \tag{3.3}$$

This prior, in essence, shrinks each estimated slide mean towards the population mean. The degree of shrinkage depends on the estimate $\hat{k}_p$, which is calculated based on the conditional expectation rule,

$$\begin{aligned}\mathrm{Var}(x) &= \mathrm{E}\left(\mathrm{Var}(x|\mu, k)\right) + \mathrm{Var}\left(\mathrm{E}(x|\mu, k)\right)\\ &= \mathrm{E}\left(\frac{\mu(1-\mu)}{k+1}\right) + \frac{\mu_p(1-\mu_p)}{k_p+1}.\end{aligned}$$

The term $\mathrm{E}\left(\frac{\mu(1-\mu)}{k+1}\right)$ can be estimated if we have an estimate of $k$. Since we ignore this term in our estimate, we underestimate $k_p$ especially when $k$ is small. Thus, we use the maximum function in (3.3) to somewhat account for the underestimation. A simulation study was performed to investigate the impact of this underestimation by comparing the results using our estimated $k_p$ and the results when we use the true $k_p$. The only noticeable difference is that when $k \leq 10$, we see some overestimation of $k$ when using (3.3).

For the intra-rater variability parameters $k_1$ and $k_2$, we consider uninformative priors $U(0, 150)$. These uniform priors span a wide range for the intra-rater variability. The shift parameters $(\delta_1, \ldots, \delta_{D-1})$ are Normally distributed with mean zero, that is, $\delta_J \sim N(0, \sigma^2_{\delta_J})$ $(J = 1, \ldots, D-1)$, where $\sigma^2_{\delta_J}$ is a rater-level parameter. In an ordered rating system, there is often more consistency in scoring the extremes compared to the intermediate categories. That means our prior belief is that $\sigma_1$ and $\sigma_{D-1}$ are often smaller than the other $\sigma$'s. Therefore, we use $\sigma_{\delta_1} = \sigma_{\delta_3} = 3$ and $\sigma_{\delta_2} = 4$ in the simulations and analyses later in the case of 4-dimensional compositions.

### 3.3.2 Estimating the posterior distribution via Markov chain Monte Carlo

The posterior distribution of the unknown parameters given the observed data can be expressed as

$$\begin{aligned}
&\pi(\delta_1, \ldots, \delta_{D-1}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n, k_1, k_2 | \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n) \\
&\propto \prod_{i=1}^{n} \mathcal{D}_D(\mathbf{x}_i | \boldsymbol{\mu}_i, k_1) \mathcal{D}_D(\mathbf{y}_i | \boldsymbol{\mu}_i, \boldsymbol{\delta}, k_2) \mathcal{D}_D(\boldsymbol{\mu}_i | \hat{\boldsymbol{\mu}}_p, \hat{k_p}) \\
&\times N_{\delta_1}(0, \sigma^2_{\delta_1}) \cdots N_{\delta_{D-1}}(0, \sigma^2_{\delta_{D-1}}). \qquad (3.4)
\end{aligned}$$

Since the posterior distribution is a multidimensional mixture distribution with no closed form, we construct a Markov chain to draw samples from the posterior distribution and approximate the quantities of interest using

$$\mathrm{E}(f(\Theta)) \approx \frac{1}{N} \sum_{t=1}^{N} f(\Theta_t),$$

where $\Theta$ is the unknown parameter set.

The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) is a very useful tool for drawing samples from a multi-dimensional stationary distribution.

This approach involves a proposal and acceptance-rejection step. In the proposal step, a candidate value $z^*_{t+1}$ is drawn from a proposal distribution, e.g., a Gaussian distribution, with probability $g(z^*_{t+1}|z_t)$. In the acceptance-rejection step, we calculate the conditional probability to accept the proposed candidate at state $t+1$. Based on the detailed balance:

$$\pi(z_t)g(z^*_{t+1}|z_t) = \pi(z^*_{t+1})g(z_t|z^*_{t+1}),$$

the accept conditional probability is then derived to be

$$\min(1, \frac{\pi(z^*_{t+1})g(z_t|z^*_{t+1})}{\pi(z_t)g(z^*_{t+1}|z_t)}).$$

If accepted, set $z_{t+1} = z^*_{t+1}$. Otherwise, set $z_{t+1} = z_t$. In our case, the algorithm consists of a series of successive single Metropolis-Hastings steps that are detailed below.

### 3.3.3 Metropolis—Hastings sampling implementation

The MCMC algorithm to estimate $\Theta = (\delta_1, \ldots, \delta_{D-1}, k_1, k_2, \boldsymbol{\mu})$ can be summarized as follows:

(i) Initialize the parameters: $\boldsymbol{\mu}^{\mathbf{0}} = \mathbf{x},\ \delta^0_1 = \cdots = \delta^0_{D-1} = 0,\ k^0_1 = k^0_2 = 50.$

(ii) Compute model hyperparameters based on (3.2) and (3.3).

(iii) Iterate through the following three updates for $T$ iterations.

a) Update $\delta_1, \ldots, \delta_{D-1}$ through a series of single parameter updates. These shifts need to be monitored to make sure that all cutpoints remain in the proper order. That is, for a proposed $\delta^*_1$

$$\eta_{i1} + \delta^*_1 < \eta_{i2} + \delta_2, \text{ for all } i = 1, \ldots, n; \tag{3.5}$$

For $J = 2, \ldots, D-2$, the check is

$$\eta_{i,J-1} + \delta_{J-1} < \eta_{iJ} + \delta^*_J < \eta_{i,J+1} + \delta_{J+1}, \text{ for all } i = 1, \ldots, n; \tag{3.6}$$

And for $J = D-1$, the check is

$$\eta_{iJ} + \delta_J^* > \eta_{i,J-1} + \delta_{J-1}, \text{ for all } i = 1, \ldots, n; \tag{3.7}$$

To satisfy these restrictions, the proposal distributions are truncated Normals. For the $t^{th}$ iteration:

i) When $J = 1$, propose $\delta_1^* \sim N(\delta_1^t, \sigma_1^2)$, where $\delta_1^* < \min(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) + \delta_2^t$.

ii) When $J = 2, \ldots, D-2$, propose $\delta_J^* \sim N(\delta_J^t, \sigma_J^2)$, where $\max(\boldsymbol{\eta}_{J-1} - \boldsymbol{\eta}_J) + \delta_{J-1}^{t+1} < \delta_J^* < \min(\boldsymbol{\eta}_{J+1} - \boldsymbol{\eta}_J) + \delta_{J+1}^t$.

iii) When $J = D-1$, propose $\delta_{D-1}^* \sim N(\delta_{D-1}^t, \sigma_{D-1}^2)$, where $\delta_{D-1}^* > \max(\boldsymbol{\eta}_{D-2} - \boldsymbol{\eta}_{D-1}) + \delta_{D-2}^{t+1}$.

Note that the best choices of $\sigma_1, \ldots, \sigma_{D-1}$ depend on the data and often require some fine-tuning. In our case we chose $\sigma_1 = \cdots = \sigma_{D-1} = 0.1$.

The acceptance rate $r$ is:

$$\begin{aligned} r &= \min\left(1, \prod_{i=1}^{n} \frac{\mathcal{D}(\mathbf{y_i}|\boldsymbol{\mu_i}, \delta_J^*, k_2)}{\mathcal{D}(\mathbf{y_i}|\boldsymbol{\mu_i}, \delta_J^t, k_2)} \frac{N_{\delta_J^*}(0, \sigma_{\delta_J}^2)}{N_{\delta_J^t}(0, \sigma_{\delta_J}^2)} \frac{TN_{\delta_J}(L_J, U_J, \delta_J^*, \sigma_J^2)}{TN_{\delta_J^*}(L_J, U_J, \delta_J^t, \sigma_J^2)}\right) \\ &= \min\left(1, \prod_{i=1}^{n} \frac{\mathcal{D}(\mathbf{y_i}|\boldsymbol{\mu_i}, \delta_J^*, k_2)}{\mathcal{D}(\mathbf{y_i}|\boldsymbol{\mu_i}, \delta_J^t, k_2)} \frac{N_{\delta_J^*}(0, \sigma_{\delta_J}^2)}{N_{\delta_J^t}(0, \sigma_{\delta_J}^2)} \frac{\Phi(\frac{U_J - \delta_J^t}{\sigma_J}) - \Phi(\frac{L_J - \delta_J^t}{\sigma_J})}{\Phi(\frac{U_J - \delta_J^*}{\sigma_J}) - \Phi(\frac{L_J - \delta_J^*}{\sigma_J})}\right), \end{aligned} \tag{3.8}$$

where $TN(a, b, \mu, \sigma)$ is the truncated Normal function with $a$ and $b$ as the lower and upper truncated boundaries, and $U_J$ and $L_J$ representing the upper and lower boundaries of the corresponding proposed shift $\delta_J^*$ (3.4-3.6). When $J = 1$, $L_J = -\infty$, and when $J = D-1$, $U_J = \infty$.

Generate $u \sim U(0, 1)$ and accept $\delta_J^{t+1} = \delta_J^*$ if $u < r$. Otherwise set $\delta_J^{t+1} = \delta_J^t$.

b) Update intra-rater variabilities $k_1$, $k_2$:

Propose $k_1^* \sim N(k_1^t, \sigma_k^2)$, where $k_1^* \in (0, 150)$. Thus the proposal distribution is a truncated Normal. The variance $\sigma_k^2$ is again fine-tuned. In our case, we chose $\sigma_k = 5$.

Calculate the acceptance ratio $r$:

$$r = \min\left(1, \frac{\pi(k_1^*|\boldsymbol{\delta}^{t+1}, \boldsymbol{\mu}^t, \boldsymbol{x}, \boldsymbol{y})}{\pi(k_1^t|\boldsymbol{\delta}^{t+1}, \boldsymbol{\mu}^t, \boldsymbol{x}, \boldsymbol{y})}\right)$$
$$= \min\left(1, \prod_{i=1}^{n} \frac{\mathcal{D}(\mathbf{x}_i|k_1^*, \boldsymbol{\mu}_i^t)}{\mathcal{D}(\mathbf{x}_i|k_1^t, \boldsymbol{\mu}_i^t)} \frac{\Phi(\frac{150-k_1^t}{\sigma_k}) - \Phi(\frac{0-k_1^t}{\sigma_k})}{\Phi(\frac{150-k_1^*}{\sigma_k}) - \Phi(\frac{0-k_1^*}{\sigma_k})}\right). \tag{3.9}$$

Generate $u \sim U(0,1)$ and accept $k_1^{t+1} = k_1^*$ if $u < r$. Otherwise set $k_1^{t+1} = k_1^t$.

The update procedure of $k_2$ is the same as above with the acceptance rate $r$ calculated as:

$$r = \min\left(1, \frac{\pi(k_2^*|\boldsymbol{\delta}^{t+1}, \boldsymbol{\mu}^t, \boldsymbol{x}, \boldsymbol{y})}{\pi(k_2^t|\boldsymbol{\delta}^{t+1}, \boldsymbol{\mu}^t, \boldsymbol{x}, \boldsymbol{y})}\right)$$
$$= \min\left(1, \prod_{i=1}^{n} \frac{\mathcal{D}(\mathbf{y}_i|\boldsymbol{\delta}^{t+1}, k_2^*, \boldsymbol{\mu}_i^t)}{\mathcal{D}(\mathbf{y}_i|\boldsymbol{\delta}^{t+1}, k_2^t, \boldsymbol{\mu}_i^t)} \frac{\Phi(\frac{150-k_2^t}{\sigma_k}) - \Phi(\frac{0-k_2^t}{\sigma_k})}{\Phi(\frac{150-k_2^*}{\sigma_k}) - \Phi(\frac{0-k_2^*}{\sigma_k})}\right). \tag{3.10}$$

c) Update reference means $\boldsymbol{\mu}_i$ for slide $i = 1, \ldots, n$:

Propose $\boldsymbol{\mu}_i^* \sim \mathcal{D}_D(\boldsymbol{\mu}_i^t, V)$. We chose $V$ to be relatively big ($V$=80) such that it is less likely to propose unrealistic means. In other words, $\boldsymbol{\mu}_i$ that do not meet the constraints (3.4 - 3.6), given the current $\boldsymbol{\delta}$. When we get a proposed $\boldsymbol{\mu}_i^*$ that is unrealistic, we skip this update round.

Calculate the acceptance rate $r_i$ for each slide mean:

$$r_i = \min\left(1, \frac{\pi(\boldsymbol{\mu}_i^*|\boldsymbol{\delta}^{\boldsymbol{t+1}}, k_1^{t+1}, k_2^{t+1}, \boldsymbol{x}, \boldsymbol{y}, \mu_p, k_p)\mathcal{D}(\boldsymbol{\mu}_i^t|\boldsymbol{\mu}_i^*, V)}{\pi(\boldsymbol{\mu}_i^t|\boldsymbol{\delta}^{\boldsymbol{t+1}}, k_1^{t+1}, k_2^{t+1}, \boldsymbol{x}, \boldsymbol{y}, \mu_p, k_p)\mathcal{D}(\boldsymbol{\mu}_i^*|\boldsymbol{\mu}_i^t, V)}\right)$$
$$= \min\left(1, \frac{\mathcal{D}(\mathbf{y}_i|\boldsymbol{\delta}^{t+1}, k_2^{t+1}, \boldsymbol{\mu}_i^*)\mathcal{D}(\mathbf{x}_i|k_1^{t+1}, \boldsymbol{\mu}_i^*)\mathcal{D}(\boldsymbol{\mu}_i^t|\boldsymbol{\mu}_i^*, V)}{\mathcal{D}(\mathbf{y}_i|\boldsymbol{\delta}^{t+1}, k_2^{t+1}, \boldsymbol{\mu}_i^t)\mathcal{D}(\mathbf{x}_i|k_1^{t+1}, \boldsymbol{\mu}_i^t)\mathcal{D}(\boldsymbol{\mu}_i^*|\boldsymbol{\mu}_i^t, V)} \frac{\mathcal{D}(\boldsymbol{\mu}_i^*|\boldsymbol{\mu}_p, k_p)}{\mathcal{D}(\boldsymbol{\mu}_i^t|\boldsymbol{\mu}_p, k_p)}\right). \tag{3.11}$$

Generate $u \sim U(0,1)$ and accept $\boldsymbol{\mu}_i^{t+1} = \boldsymbol{\mu}_i^*$ if $u < r$. Otherwise, set $\boldsymbol{\mu}_i^{t+1} = \boldsymbol{\mu}_i^t$.

(iv) Set $t = t + 1$ and repeat Step (iii) until $T$ samples are drawn.

## 3.4 Model Interpretation

Our model is set up to compare one rater to a reference rater. The shift parameters estimated from our model are with respect to the reference rater. Based on the

MCMC algorithm, we obtain a set of correlated draws from the posterior distribution of interest. We extract every $10^{th}$ posterior sample with a 500 sample burn-in and use the posterior means and credible intervals based on these remaining posterior samplings $\Theta = (\delta_1, \ldots, \delta_{D-1}, k_1, k_2)$ to interpret the inter-rater agreement between two raters.

### 3.4.1 Interpretation of shift parameters

Posterior means of shifts $\delta_1, \ldots, \delta_{D-1}$ are the quantities of prime interest because they indicate how Rater B differs from the reference rater (Rater A). For example, a positive shift, $\delta_J > 0$ means Rater B has higher cutpoint value when distinguishing between category $J$ and category $J + 1$, and vice versa. A shift $\delta_J$ close to zero then indicates Rater B has a good agreement with the reference rater on that cutpoint. By constructing credible intervals of posterior shift parameters, we can make an inference that whether one or more shifts are significantly different from zero or not. Note, however, that the category means typically depend on two cutpoints so agreeing on a single cutpoint does not imply agreement on category percents.

### 3.4.2 Overall agreement index

While the shift parameters give information on the pattern of differences, they do not give a measure of overall agreement. We propose the use of an index that incorporates both the shifts and the intra-rater variabilities. Given that we use Dirichlet distributions to describe intra-rater variability, a natural probabilistic distance measure is the Bhattacharyya coefficient ($\rho_B$ or BC), which we briefly discussed in Chapter 2. We now investigate how BC behaves when measuring the amount of overlap between two Dirichlet distributions and discuss how it can be used as an agreement index in our case.

For demonstration purposes, we revert back to a 3-dimensional vector. In Figure 3.6, we assume two raters have the same intra-rater variability parameter but different means. Suppose $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_5$ are Rater A's means for five different slides. Red points are simulated from $\mathcal{D}(\boldsymbol{\mu}_i, 100)\,(i = 1, \ldots, 5)$ to represent possible scores of Rater A. Blue points were simulated from $\mathcal{D}(\boldsymbol{\mu}_i^*, 100)\,(i = 1, \ldots, 5)$ where $\boldsymbol{\mu}_i^*$ was computed based on $\boldsymbol{\mu}_i$ and the shifts $(\delta_1, \delta_2) = (0.3, 0.3)$. The corresponding true BC values are shown for these five pairs of Dirichlet distributions.

It is clear that in each case there is a difference in the average score between Raters A and B but there is also some overlap in scores suggesting the two raters are somewhat similar in scores. The BCs range from 0.714 to 0.825, which are fairly close to one, suggesting good agreement between the two raters. The shape of the Dirichlet distribution depends on the mean so even though the same $\boldsymbol{\delta}$ and $k$ were used, the BC is not the same for each of the means. When the distribution is concentrated in a corner, the overlapped portion is higher. Therefore, given the same shifts $(0.3, 0.3)$, $\text{BC}_1$, $\text{BC}_3$, and $\text{BC}_5$ are slightly bigger than $\text{BC}_2$ and $\text{BC}_4$. Because of this, we suggest reporting the BC for some reference mean so it can be compared across studies.

In Figure 3.7, it is assumed Raters A and B have the same Dirichlet means, i.e., $\delta_1 = \delta_2 = 0$, but different intra-rater variability parameters. Here we only show two means but consider two sets of dispersions parameters. One mean is located in the center and the other one is located in a corner. By varying Rater B's variability parameter from $k = 100$ to $k = 160$ with Rater A's variability parameter fixed at $k = 50$, we can see that the overlapped portion between these two raters decreases. This results in an decreased BC from around 0.94 to 0.85. Again, the BC calculated from a mean in the center overlapped slightly less than those from a mean in the corner. This indicates that even if two raters agree perfectly on the means, the BC will not necessarily be 1 and it decreases when there is a bigger deviance between the two intra-rater variabilities.

One way to view BC is as a conditional agreement index, that is, how close two raters' distributions are conditional on a given slide (or mean). Thus, this BC index is similar in flavor to the CIV/CIA and the between-slide variability doesn't affect it. Notice that if two raters have exactly the same mean and intra-rater variability, the

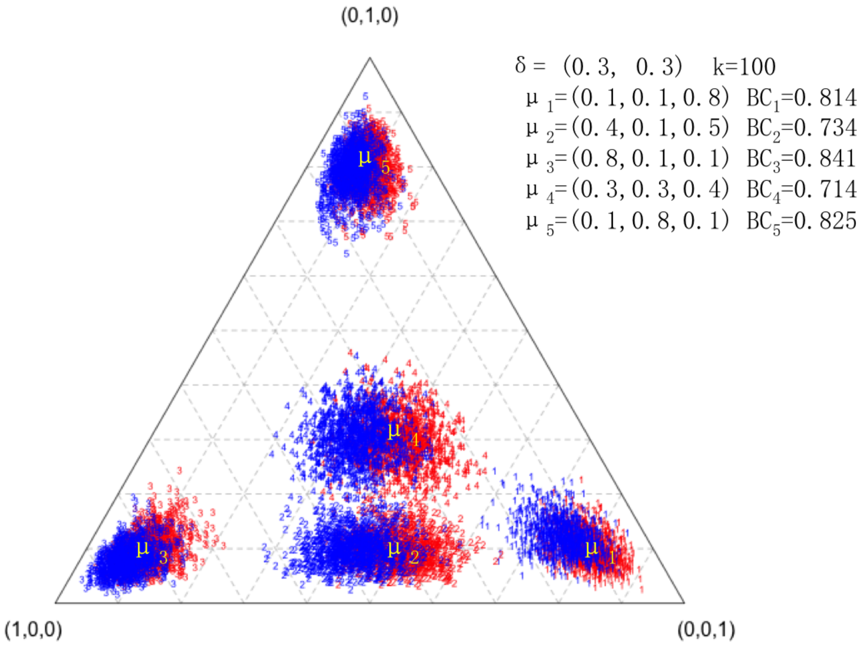


Fig. 3.6. BCs based on different means

BC is equal to 1 regardless of the value of the intra-rater variability, while ICC and CCC are likely to be smaller given larger intra-rater variability. Thus, we propose to take a look at both BC and $k$ (or $k_1$ and $k_2$) together as measures of the overall agreement.

## 3.5 An Illustrative Example

### 3.5.1 Data Simulation

To demonstrate our model and estimation approach, we consider two scenarios. Scenario 1 is set up such that Rater A and Rater B have good agreement while Scenario 2 indicates Rater A and Rater B are different, especially in distinguishing Category 2 and 3. For illustration purpose, we set $k_1 = k_2 = 50$ and update $k$ as a

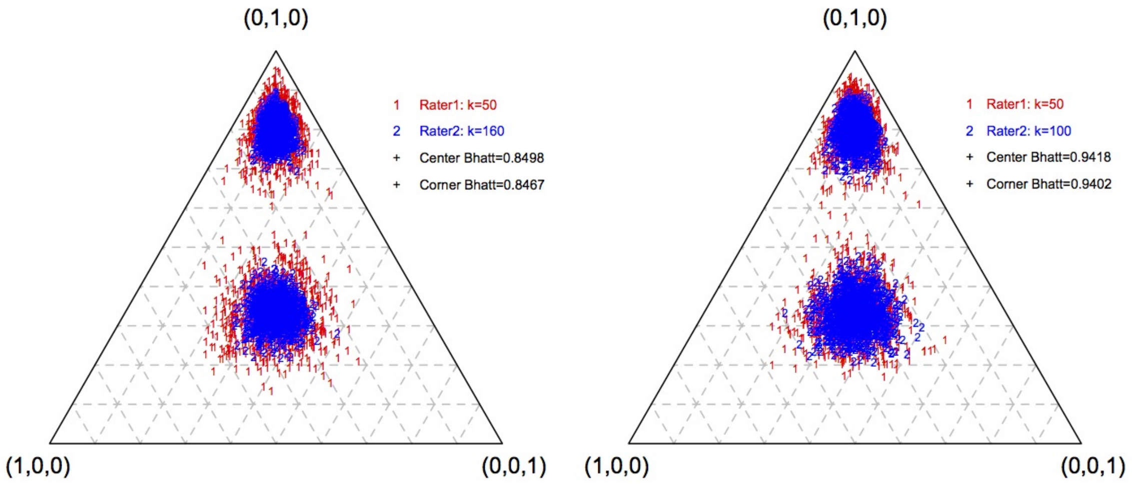


Fig. 3.7. BCs based on different intra-rater variabilities

single intra-rater variability parameter. The Scenario 1 data are simulated using the following steps:

(i) Simulate 50 Rater A (i.e., the reference rater) means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{50}$ from a Dirichlet prior $\mathcal{D}((0.3, 0.4, 0.3), 10)$.

(ii) Generate Rater A's scores using $\mathbf{x}_i \sim \mathcal{D}(\boldsymbol{\mu}_i, 50)$.

(iii) Using the shifts $(\delta_1, \delta_2) = (-0.1, 0.1)$, compute the Rater B means $\boldsymbol{\mu}_{yi} = f(\boldsymbol{\mu}_i, \delta_1, \delta_2)$ using the logit transformation.

(iv) Simulate Rater B scores using $\mathbf{y}_i \sim \mathcal{D}(\boldsymbol{\mu}_{yi}, 50)$.

This gives us 50 pairs of scores $(\mathbf{x}_i, \mathbf{y}_i)$, which we use to assess agreement. For Scenario 2, we repeat steps (iii)-(iv) above with $(\delta_1, \delta_2) = (0.3, 0.8)$.

### 3.5.2 Data Visualization

Figure 3.8 contains two ternary plots to help visualize the two scenarios. Due to the larger absolute shifts (+0.3, +0.8) in Scenario 2, Rater B's scores are further away from Rater A's scores (right plot). The plots do not link pairs of scores but it is clear in the right figure that the average score is different. For the left plot, there is much more overlap.

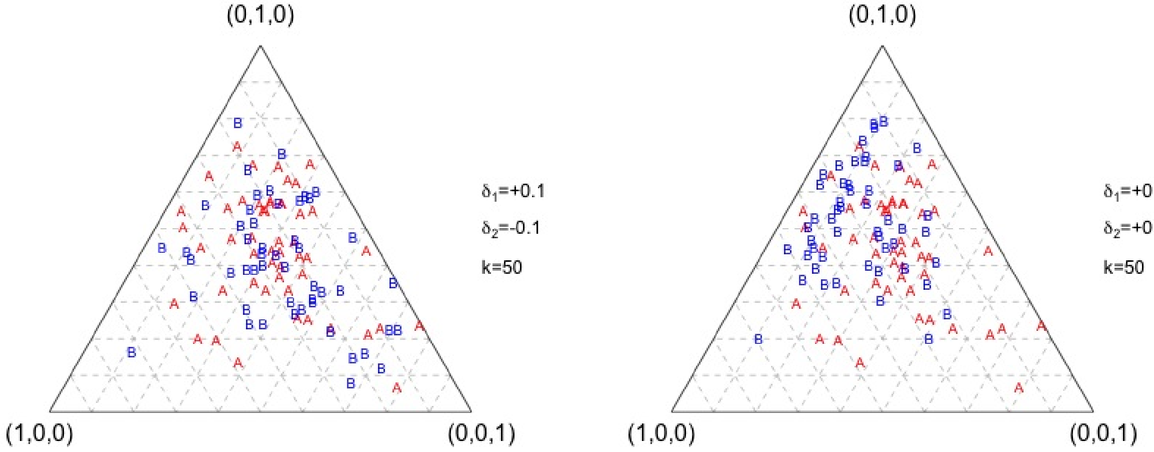


Fig. 3.8. Simulated data on the simplex

### 3.5.3 MCMC output

To demonstrate that our algorithm converges on a single posterior distribution (single mode), we performed two runs with different starting values (Table 3.2) for each of these two examples. Run 1 starts at shift values representing perfect agreement and small intra-rater variability. Run 2 starts at shift values representing substantial deviations in means and relatively large intra-rater variability.

Figure 3.9 displays the posterior samples of $\delta_1$, $\delta_2$ and $k$ for Scenario 1 and Scenario 2 under both starting values. It is clear that both chains converge quickly to the same

Table 3.2.
Starting values for parameters in two runs of the MCMC

| Parameter | Run 1 | Run 2 |
|---|---|---|
| $\delta_1$ | 0 | -1 |
| $\delta_2$ | 0 | 1 |
| $k$ | 100 | 10 |
| $\boldsymbol{\mu}_x$ | $\mathbf{x}$ | $\mathbf{x}$ |

posterior. We discard the first 500 iterations (10%) as burn-in and extract every $10^{th}$ iteration to compute posterior means and variances of parameters as well as compute 95% credible intervals of parameters combining both chains. Table 3.3 summarizes the results from the runs in Figure 3.8. All 95% credible intervals contain the true parameters and the posterior means are very close to the true values. Except for the 95% credible interval of $\delta_1$ under Scenario 1, all other shift credible intervals do not contain zero.

Table 3.4 summarizes the BC index for several different reference means. For Scenario 1, all the BC estimates are between 0.94 and 0.95 and very close to the truth. For Scenario 2, because of the bigger shift parameters, the BC estimates are reduced to around 0.6. The posterior mean estimates are bigger than the true values because of the underestimation of $k$ and $\boldsymbol{\delta}$ in this case.

Table 3.3.
Posterior summary based on the combination of two chains of 5000 draws

| True Parameter | Posterior Mean | 95% Credible Interval |
|---|---|---|
| $\delta_1 = +0.1$ | 0.0 | (-0.129, 0.116) |
| $\delta_2 = -0.1$ | -0.161 | (-0.273, -0.052) |
| $k = 50$ | 51.99 | (37.86, 67.65) |
| $\delta_1 = +0.3$ | 0.254 | (0.117, 0.377) |
| $\delta_2 = +0.8$ | 0.709 | (0.574, 0.842) |
| $k = 50$ | 42.49 | (31.11, 54.20) |

Table 3.4.
BCs computed based on posterior estimates

| Reference Mean | $\delta_1 = +0.1, \delta_2 = -0.1$ | | $\delta_1 = +0.3, \delta_2 = +0.8$ | |
|---|---|---|---|---|
| | Posterior Mean(Truth) | 95% CI | Posterior Mean(Truth) | 95% CI |
| $(0.3, 0.4, 0.3)$ | 0.945(0.951) | (0.883, 0.996) | 0.609(0.490) | (0.481, 0.749) |
| $(0.1, 0.1, 0.8)$ | 0.940(0.944) | (0.869, 0.998) | 0.527(0.387) | (0.367, 0.683) |
| $(0.5, 0.3, 0.2)$ | 0.951(0.945) | (0.892, 0.996) | 0.692(0.593) | (0.585, 0.797) |

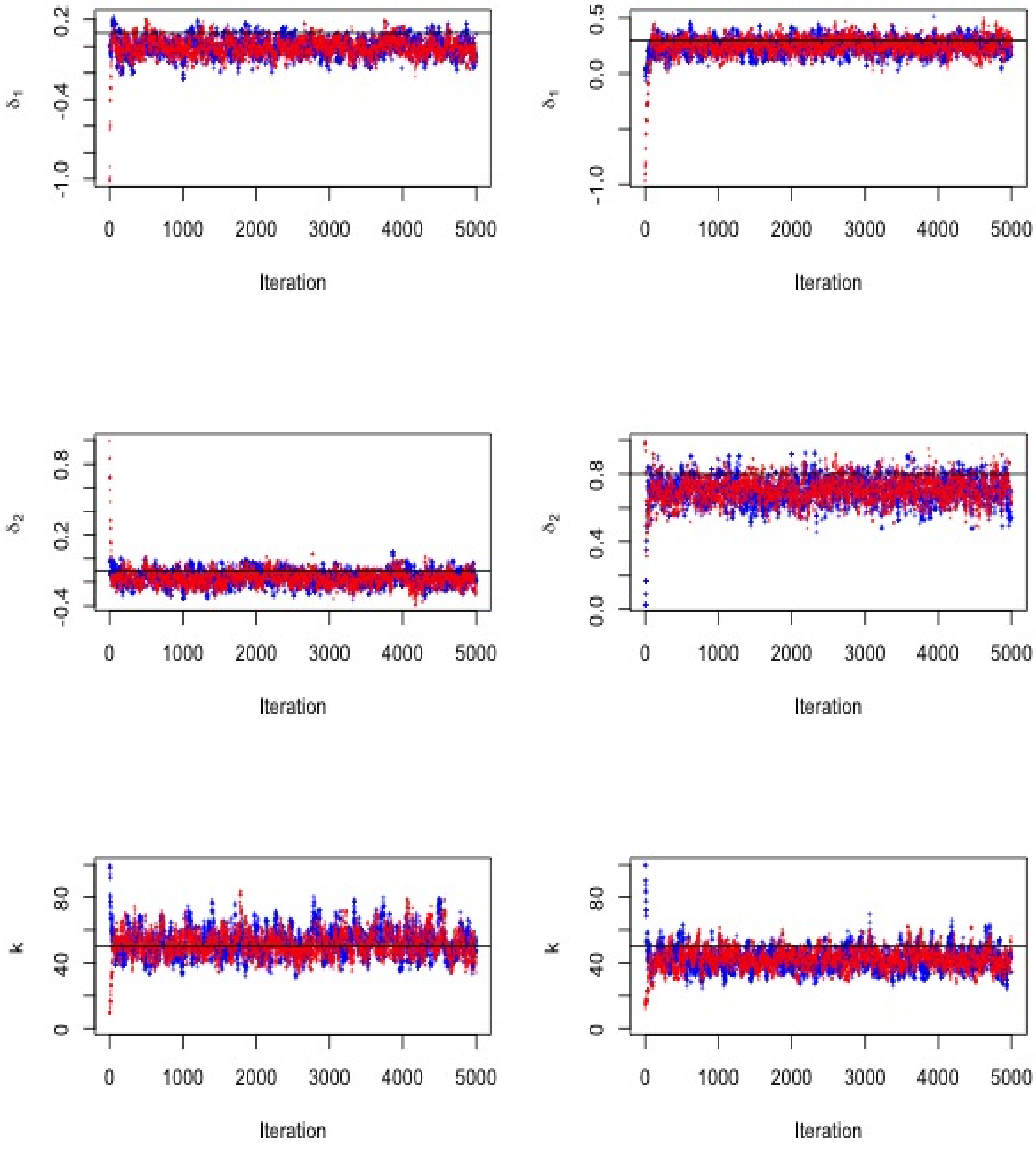


Fig. 3.9. Sampled values for $\delta_1$, $\delta_2$, and $k$ from two runs of the MCMC

This simulation demonstrates the ability of our model to obtain reasonable estimates for standard data sets without replicates. More intensive simulation studies are needed to investigate bias. We use the rest of this chapter for these investigations.

### 3.6 Simulation Studies on Continuous Data

To investigate possible bias of the shift and intra-rater variability estimates, we performed simulation studies following the general data simulation procedure described in the previous section. For each data set, we run a MCMC chain with 5000 iterations. Since typical IHC scores are 4-part compositional vectors, these simulations involve 4 dimensional compositional data. We consider two settings and simulate 50 data sets of 50 slides from each setting. There are no replicate scores for any of the slides. For Setting 1, the population of means is $\mathcal{D}_4(\boldsymbol{\mu}_p = (0.25, 0.25, 0.25, 0.25), k_p = 10)$. For each set of 50 slides, the three shift parameters $\delta$'s are randomly chosen and the intra-rater variability parameter ranges between 10 and 100. For Setting 2, we use the same shift parameters and $k$ as Setting 1 but the slide mean distribution is $\mathcal{D}_4(\boldsymbol{\mu}_p = (0.1, 0.1, 0.2, 0.6), k_p = 80)$. This setting involves a non-central and compact population of slides while the first setting involves a disperse central population. The thought here is to investigate the robustness of the methodology under non-ideal (Setting 2) conditions. For both simulation settings, the posterior means of $\delta_1, \delta_2, \delta_3$ and $k$ are used as our point estimates. For each simulation, absolute deviances between the true shift parameters and the posterior estimates are calculated (i.e, deviance = estimated - true ). To better investigate the pattern of $k$ estimates, the % deviances between the true $k$ and the posterior estimates are calculated( i.e., (estimated - true)/true ).

Figure 3.10 displays these deviances versus the true values for all simulations from Setting 1. We label the 50 simulated data set with numbers from 1 to 50 to be able to pair the deviances across a single data set and across settings. In general, there appears to be no bias from shift estimates as the deviances bounce above and below the zero reference lines. Data sets with a low number (e.g., 1, 2, 4, and 5) tend to have larger deviances. A likely reason for this is that these runs involved cases with a large intra-rater dispersion parameter ($k = 10$), which in general mean more uncertainty. We observe some overestimation of $k$ especially when $k = 10$. Further investigations suggest that this overestimation coming from the underestimation of $k_p$ (not shrinking enough towards the overall mean).

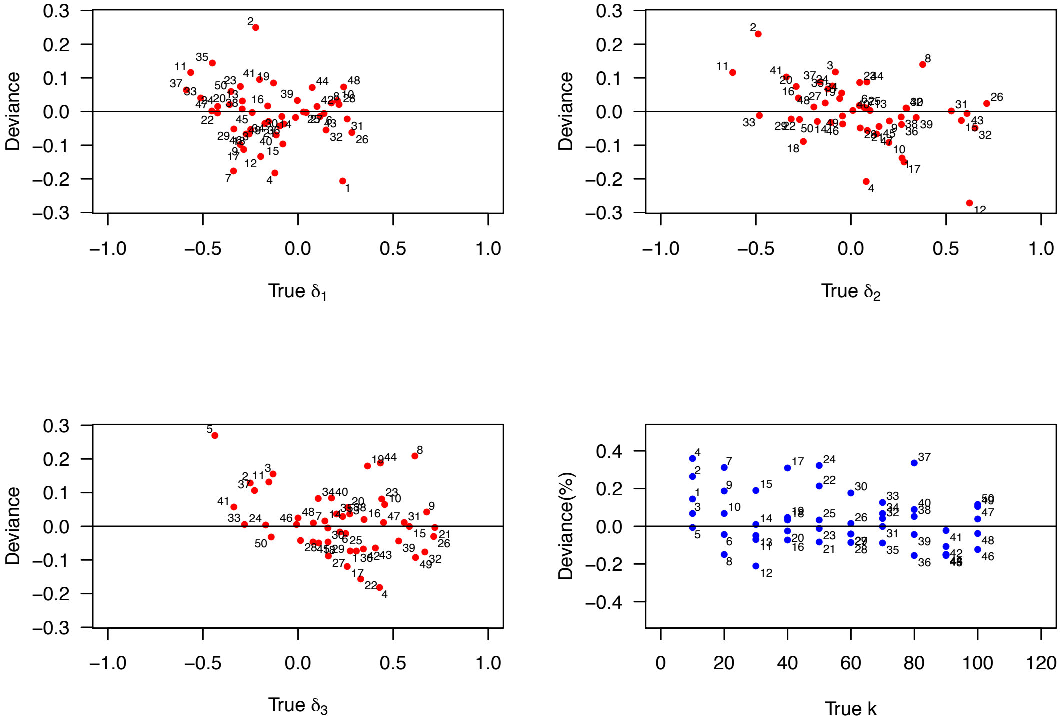


Fig. 3.10. Deviances of model estimates given continuous data sets from a disperse, central mean population

We also use the frequentist estimation approach described in Section 3.2.1 to estimate our model parameters. We can still get unbiased estimates of $\delta$'s. However, $k$ gets severely underestimated using the MLE (Figure 3.11), which demonstrates the strength of our Bayesian approach to improve the model estimation.

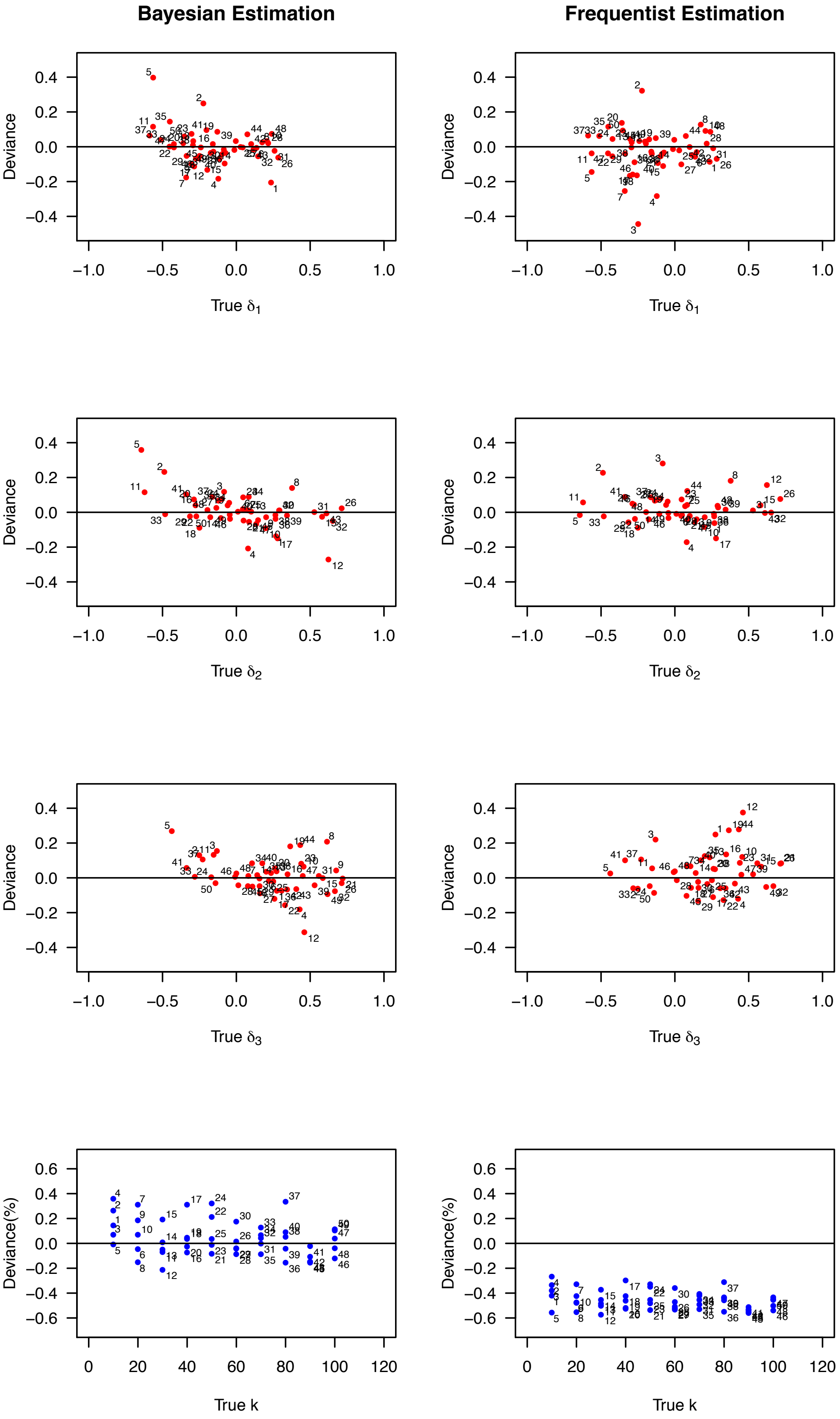


Fig. 3.11. Comparison of estimated model parameters

Figure 3.12 displays the deviances versus the true parameters for the 50 simulations under Setting 2. Each data set number here indicates the same shift and intra-rater variability parameters used in Setting 1. One difference between the shift estimates we notice immediately is that the deviances of $\delta_3$ are generally smaller than those in Figure 3.10. This is due to the skewness of the slide mean population in this setting. The slide means are skewed to the positive category so there is a lot of information about the cutpoint between the moderate and positive categories. Conversely, there is less information regarding the first shift, and while it is not as strongly apparent, we see more dispersion in the $\delta_1$ estimates under the second setting. Besides, the estimates of $k$ get slightly bigger compared to those from Setting 1. More simulations have been done by varying population mean $\boldsymbol{\mu_p}$ and population dispersion $k_p$ and it suggests that less variation in slide means results in an overestimation of $k$ (i.e., less intra-rater variability) especially for small $k$.

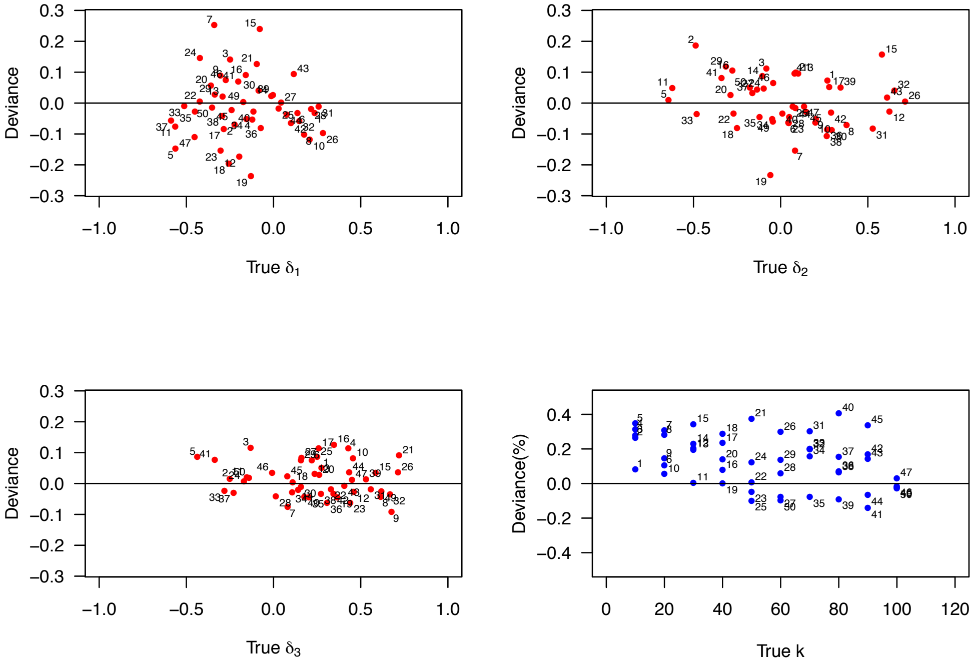


Fig. 3.12. Deviances of model estimates given continuous data sets from a skewed slide population

We also investigated the posterior distributions of the estimated parameters for the simulated data sets. Since our imposed priors of the parameters are pretty flat, most information is coming from the data. Note that the 50 simulations we display in Figures 3.10 and 3.12 assumed one intra-rater variability parameter $k$ for both raters. When there are no replicates from raters, allowing $k_1$ and $k_2$ to be estimated separately would result in flat posterior distributions of $k_1$ and $k_2$, because there is no information of each intra-rater variability provided by data.

In contrast, when the true $k_1$ and $k_2$ are different but we update them as if they were the same in our MCMC procedure, we end up getting a "pooled" estimate of $k_1$ and $k_2$, that is, the posterior distributions of $k$ would be between the two truths. We've run a couple chains with data simulated from different $k_1$ and $k_2$ but only update one $k$, and this is the only noticeable result. There were negligible changes in the estimates of $\delta$'s.

## 3.7 Model Extensions: Discrete Compositional Scores with Rounded Zeros

The previous model assumes the scores are on the simplex spaces. In practice, however, pathologists give compositional scores using deciles values between 0 and 1. We conceptualize this as pathologists rounding percentages to the nearest decile when scoring slides. Since the Dirichlet distributions don't accommodate vectors with 0 elements, a modification of our approach is needed.

### 3.7.1 Modified model and MCMC implementation

In the modified model, we consider the $\mathbf{x}$ and $\mathbf{y}$ to still be on the simplex space but they are now latent scores. Figure 3.13 adds this latent layer to the hierarchical model structure. The observed data $\mathbf{x}^{'}$ and $\mathbf{y}^{'}$ are the rounded decile compositional vectors of these latent scores. To guarantee these vectors sum to one, we assume the lowest, highest, and the second lower categories are rounded. The third category is then one minus the sum of these rounded values. To include and update the latent scores, we need to add an additional step to our MCMC algorithm. We propose two modifications to the Metropolis-Hastings sampling implementation described in Section 3.3.3.

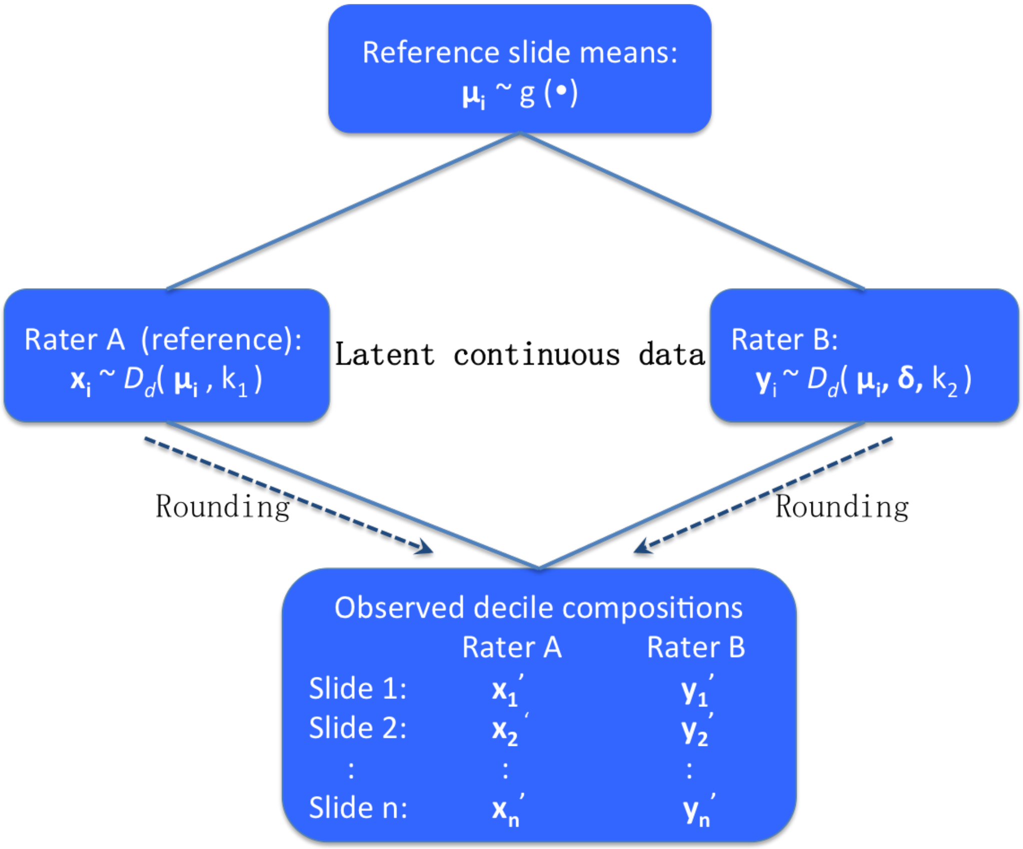


Fig. 3.13. Hierarchical model structure for discrete data

*Modified Initialization Step*

First, apply the multiplicative replacement strategy (Martín-Fernández, 2003) described in Chapter 1 (Section 1.5.1, page 20) to replace the observed zeroes in compositional scores of both raters with a small value. Let $\mathbf{x}'_{(r)}$ and $\mathbf{y}'_{(r)}$ be the discrete observed data with the replacement. Then initialize the parameters using $\boldsymbol{\mu}^{\mathbf{0}} = \mathbf{x}'_{(\mathbf{r})}$, $\delta_1^0 = \cdots = \delta_{D-1}^0 = 0$, $k_1^0 = k_2^0 = 50$, and $t = 0$.

*Propose Continuous Compositional Vectors*

Before updating the shift parameters in the MCMC step, an extra step is implemented to update the latent scores. In this step, we propose:
for $i = 1, \ldots, n$,

$$x^*_{ij} \sim \begin{cases} U(x^{'}_{ij} - 0.05, x^{'}_{ij} + 0.05), & \text{if } x^{'}_{ij} > 0, \\ U(0, 0.05), & \text{if } x^{'}_{ij} = 0, \\ U(0.95, 1), & \text{if } x^{'}_{ij} = 1, \end{cases} \qquad (j = 1, \ldots, D).$$

To guarantee $\sum_{j=1}^{D} x^*_{ij} = 1$, we set $x^*_{i3} = 1 - \sum_{j \neq 3}^{D} x^*_{ij}$. Then we calculate the acceptance ratio:

$$r_i = \min\left(1, \frac{\mathcal{D}(\mathbf{x}^*_i | \boldsymbol{\mu}^t_i, k_1)}{\mathcal{D}(\mathbf{x}^t_i | \boldsymbol{\mu}^t_i, k_1)}\right).$$

Generate $u \sim U(0, 1)$ and accept $\mathbf{x}^{t+1}_i = \mathbf{x}^*_i$ if $u < r_i$. Otherwise set $\mathbf{x}^{t+1}_i = \mathbf{x}^t_i$. Similarly, we update Rater B's latent continuous compositional vectors $\mathbf{y}^t_i$ in the same way as described above.

Besides these two modified steps, all other update steps are the same as described in Section 3.3.3, with $\mathbf{x}_i$ replaced by the updated $\mathbf{x}^t_i$ in each iteration.

### 3.7.2 Simulation studies for discrete data

To investigate how our model performs on discrete compositional data, we rounded the 50 continuous data sets simulated from Setting 1 of the previous section to obtain decile score vectors. We then implemented our modified MCMC algorithm to estimate the shift and the intra-rater variability parameters. Figure 3.14 shows that the deviances for the shift and the intra-rater variability parameters are still bouncing around the zero reference lines.

To better compare the estimates from each continuous data set and its paired discrete data set, Figure 3.15 displays the differences of the model estimates for each data set. Each red point represents the posterior mean estimate from a continuous data set minus the posterior mean estimate from the corresponding rounded decile data set. Each blue point represents the relative difference between $k$ estimates (i.e., diff($\hat{k}$)/$k$). There are no obvious patterns in the shift estimates and the differences are relatively small (between -0.1 and 0.1). However, we do observe that the estimated $k$ from the 50 continuous data sets tends to be bigger than those from the 50 rounded data sets. This might be coming from the rounding procedure because it introduces more artificial intra-rater variability.

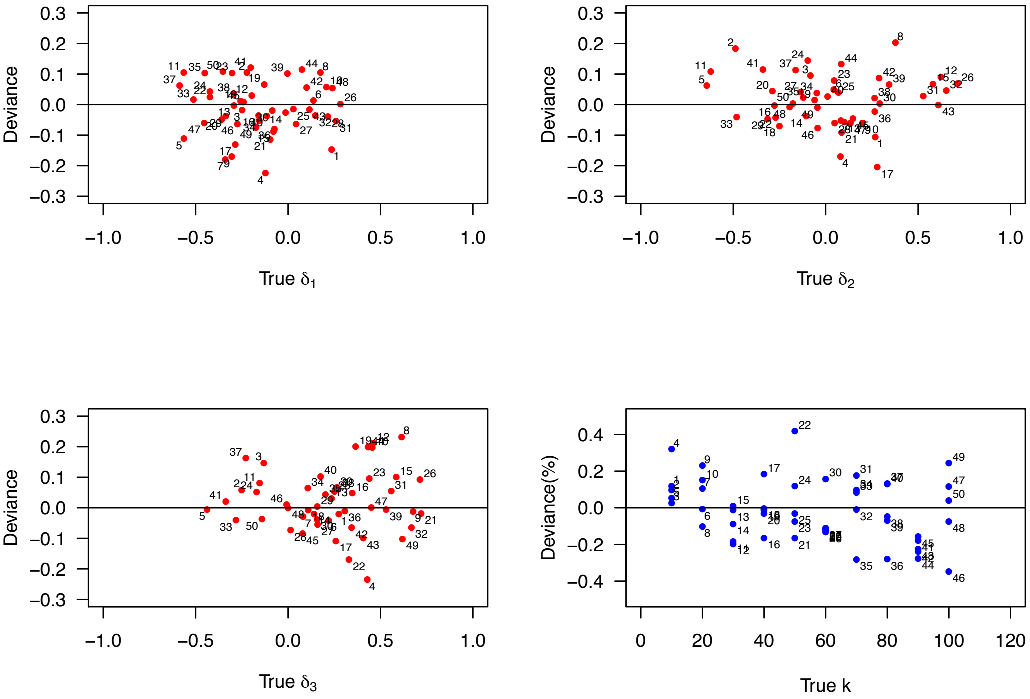


Fig. 3.14. Deviances of model estimates given decile data sets from a disperse, central mean population

We also summarize the standard deviation of posterior samples within a simulation in Table 3.5. Even though we don't see a pattern between the dispersion of posterior means of $\delta$'s versus the value of $k$, we do see bigger uncertainty of $\hat{\delta}$'s within a MCMC chain when $k$ is smaller, which is also expected.

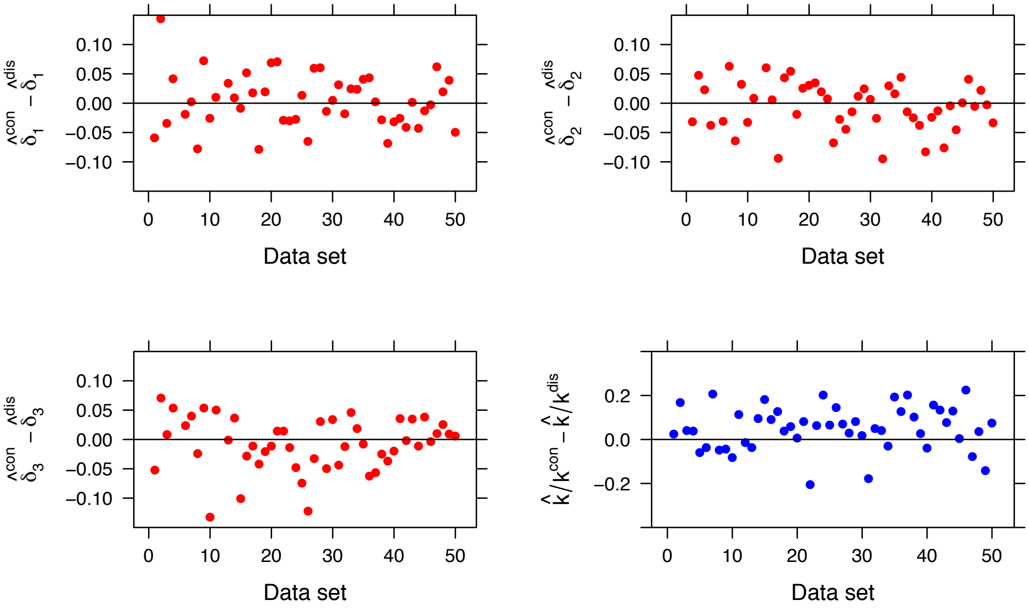


Fig. 3.15. Comparison: continuous and discrete data sets

Table 3.5.
SD of posterior samples based on different $k$'s

| $k$ | SD($\hat{\delta_1}$) | SD($\hat{\delta_2}$) | SD($\hat{\delta_3}$) |
|---|---|---|---|
| 10 | 0.13 | 0.11 | 0.12 |
| 50 | 0.07 | 0.07 | 0.07 |
| 100 | 0.05 | 0.05 | 0.06 |

## 3.8 Model Extension: Replicate Scores on All or a Subset of Slides

All simulations so far assumed that raters only score each slide once. In those cases, we restricted the intra-rater variabilities for both raters to be the same (i.e., $k_1 = k_2$). When there are replicate scores from raters on all or a subset of the slides, we can relax the restriction $k_1 = k_2$ and update $k_1$ and $k_2$ separately in the MCMC algorithm.

To investigate how the replication impacts our model estimates, we take the 50 simulated decile data sets from Setting 1 in the previous section, and simulated one extra replicate score on each of the 50 slides from each rater. We perform the same estimation approach and summarize the means and standard deviations of the 50 deviances in Table 3.6. The first column, as the reference, is the one without any replication. The second column represents the data set with two replicates on all 50 slides from both raters. Columns 3 and 4 represent the same data set as the second column except that they have replicate scores on a subset of 25 slides and replicate scores on a subset of 10 slides, respectively. Column 5 represents a data set including no replicates for the reference rater but two replicates on all 50 slides from Rater B. In terms of the estimates of $\delta$'s, as the number of slides with replicates decreases, the standard deviations of the 50 deviances increase slightly but it is almost negligible. In terms of the estimates of $k$, we only compare the column 2-5 because we estimate a single $k$ when there is no replication. There is an increase of the observed deviations as the number of slides with replicates decreases.

Table 3.6.
Summary of 50 deviances for shifts and intra-rater variabilities

| Parameter | No rep | Rep on 50 slides | Rep on 25 slides | Rep on 10 slides | Unbalance |
|---|---|---|---|---|---|
| $\delta_1$ | $+0.01(0.08)$ | $0.0(0.07)$ | $+0.01(0.08)$ | $+0.02(0.09)$ | $+0.02(0.08)$ |
| $\delta_2$ | $-0.01(0.08)$ | $-0.01(0.06)$ | $0.00(0.07)$ | $0.00(0.08)$ | $-0.01(0.07)$ |
| $\delta_3$ | $-0.02(0.10)$ | $-0.01(0.07)$ | $-0.01(0.09)$ | $-0.01(0.09)$ | $-0.01(0.08)$ |
| $k_1$ | $+2.5(11.4)$ | $+2.6(8.64)$ | $+0.83(12.40)$ | $-1.55(16.3)$ | $-3.76(14.9)$ |
| $k_2$ | | $+3.6(10.4)$ | $+2.35(13.8)$ | $+1.00(15.9)$ | $+4.49(11.3)$ |

## 3.9 Model Extension: More Than Two Raters

When there are more than two raters, we simply modify the MCMC update Step (v) in Section 3.3.3 to accommodate this. One of the raters has to be chosen as a reference rater, then we estimate the shifts from each of the other raters compared to the reference.

*Modified Update Reference Means Step*

Let $\mathbf{x}, \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(R)}$ denote the observed data from the reference rater, Rater 1 to Rater R ($R \geq 2$), and $k_{(0)}, k_{(1)}, \ldots, k_{(R)}$ denote their corresponding intra-rater variability parameters. For notation simplicity we assume the observed data are continuous. For each of the R raters, their shifts are denoted as $\boldsymbol{\delta}_{(1)}, \ldots, \boldsymbol{\delta}_{(R)}$. Propose $\boldsymbol{\mu}_i^* \sim \mathcal{D}(\boldsymbol{\mu}_i^t, V)$. Calculate the acceptance rate $r_i$ for slide $i$,

$$\begin{aligned} r_i &= \min\left(1, \frac{\pi(\boldsymbol{\mu}_i^* | \boldsymbol{\delta}_{(1)}^{t+1}, \cdots, \boldsymbol{\delta}_{(R)}^{t+1}, k_{(0)}^{t+1}, \cdots, k_{(R)}^{t+1}, \boldsymbol{x}, \boldsymbol{y}_{(1)}, \cdots, \boldsymbol{y}_{(R)}, \mu_p, k_p)\mathcal{D}(\boldsymbol{\mu}_i^t | \boldsymbol{\mu}_i^*, V)}{\pi(\boldsymbol{\mu}_i^t | \boldsymbol{\delta}_{(1)}^{t+1}, \cdots, \boldsymbol{\delta}_{(R)}^{t+1}, k_{(0)}^{t+1}, \cdots, k_{(R)}^{t+1}, \boldsymbol{x}, \boldsymbol{y}_{(1)}, \cdots, \boldsymbol{y}_{(R)}, \mu_p, k_p)\mathcal{D}(\boldsymbol{\mu}_i^* | \boldsymbol{\mu}_i^t, V)}\right) \\ &= \min\left(1, \prod_{r=1}^{R} \frac{\mathcal{D}(\mathbf{y}_{(r)i} | \boldsymbol{\delta}_{(r)}^{t+1}, k_{(r)}^{t+1}, \boldsymbol{\mu}_i^*)\mathcal{D}(\mathbf{x}_i | k_{(0)}^{t+1}, \boldsymbol{\mu}_i^*)\mathcal{D}(\boldsymbol{\mu}_i^t | \boldsymbol{\mu}_i^*, V)}{\mathcal{D}(\mathbf{y}_{(r)i} | \boldsymbol{\delta}_{(r)}^{t+1}, k_{(r)}^{t+1}, \boldsymbol{\mu}_i^t)\mathcal{D}(\mathbf{x}_i | k_{(0)}^{t+1}, \boldsymbol{\mu}_i^t)\mathcal{D}(\boldsymbol{\mu}_i^* | \boldsymbol{\mu}_i^t, V)} \frac{\mathcal{D}(\boldsymbol{\mu}_i^* | \boldsymbol{\mu}_p, k_p)}{\mathcal{D}(\boldsymbol{\mu}_i^t | \boldsymbol{\mu}_p, k_p)}\right). \end{aligned} \tag{3.12}$$

Generate $u \sim U(0, 1)$ and accept $\boldsymbol{\mu}_i^{t+1} = \boldsymbol{\mu}_i^*$ if $u < r_i$. Otherwise, set $\boldsymbol{\mu}_i^{t+1} = \boldsymbol{\mu}_i^t$.

### 3.10 More Simulations with Different Sample Size

To better understand the sampling distributions of the posterior means of our model parameters based on different sample sizes, we provide two more simulation cases. For each case described below, 80 decile data sets are generated from a population $\mathcal{D}([0.25, 0.25, 0.25, 0.25], 10)$ for different numbers of slides $n$. Assume there are two replicates on each slide from both raters, thus $k_1$ and $k_2$ can be estimated separately.

**Case 1:** Two raters are relatively consistent in scoring all levels and both raters have large intra-rater variabilities. Set $\delta_1 = -0.1,\ \delta_2 = 0.2,\ \delta_3 = 0.1,\ k_1 = 10,\ k_2 = 20$, and $n = 20, 40, 60, 80, 100$. The distribution of 80 posterior means is displayed using boxplots in Figure 3.16.

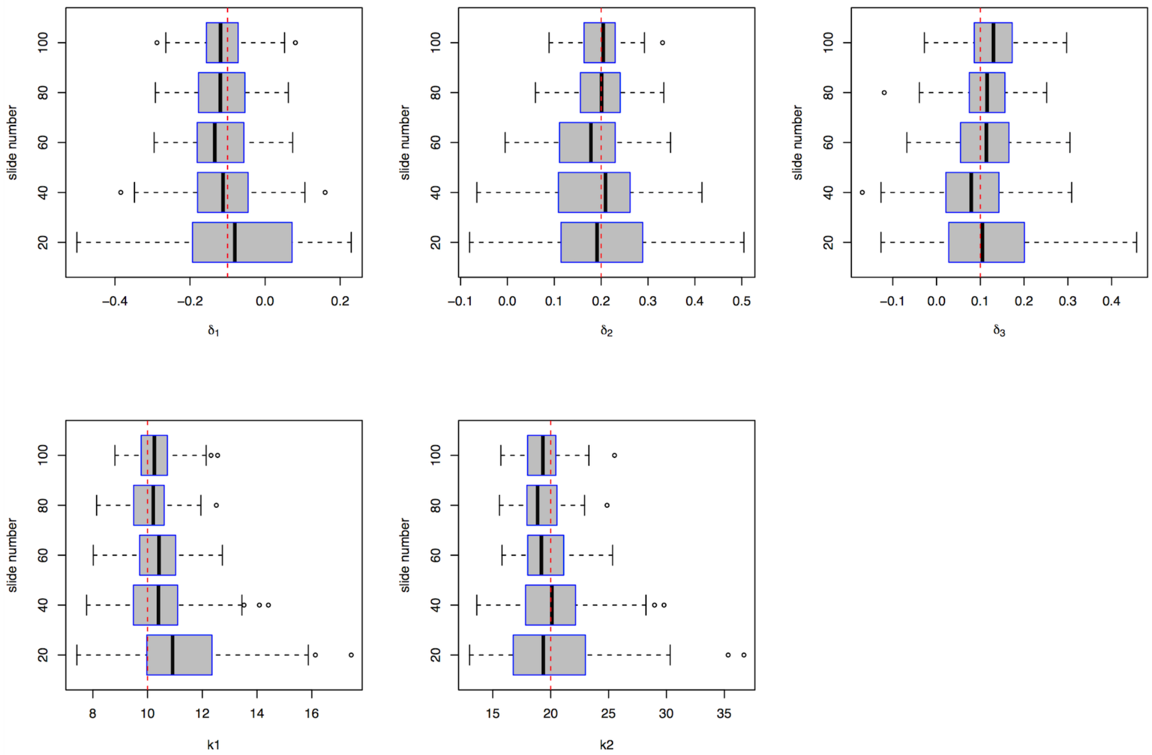


Fig. 3.16. Distribution of posterior means for case 1

**Case 2:** Two raters are relatively consistent in scoring the negative and the positive levels but vary quite a bit in distinguishing between the middle two levels. Rater 2 is relatively consistent when scoring the same slide. Set $\delta_1 = -0.1,\ \delta_2 = -0.6,\ \delta_3 = -0.2,\ k_1 = 30,\ k_2 = 60$, and $n = 20, 40, 60, 80, 100,$.

In Figure 3.16 and 3.17, again, no evidence of bias is shown in the estimates. Generally, when the number of slides increases, the uncertainty of posterior means becomes smaller. Data sets from a skewed population were also simulated and we observe a similar trend. In practice, 50 slides are probably the most we can get. Thus, we used $n = 50$ throughout this chapter for our simulation studies of model properties.

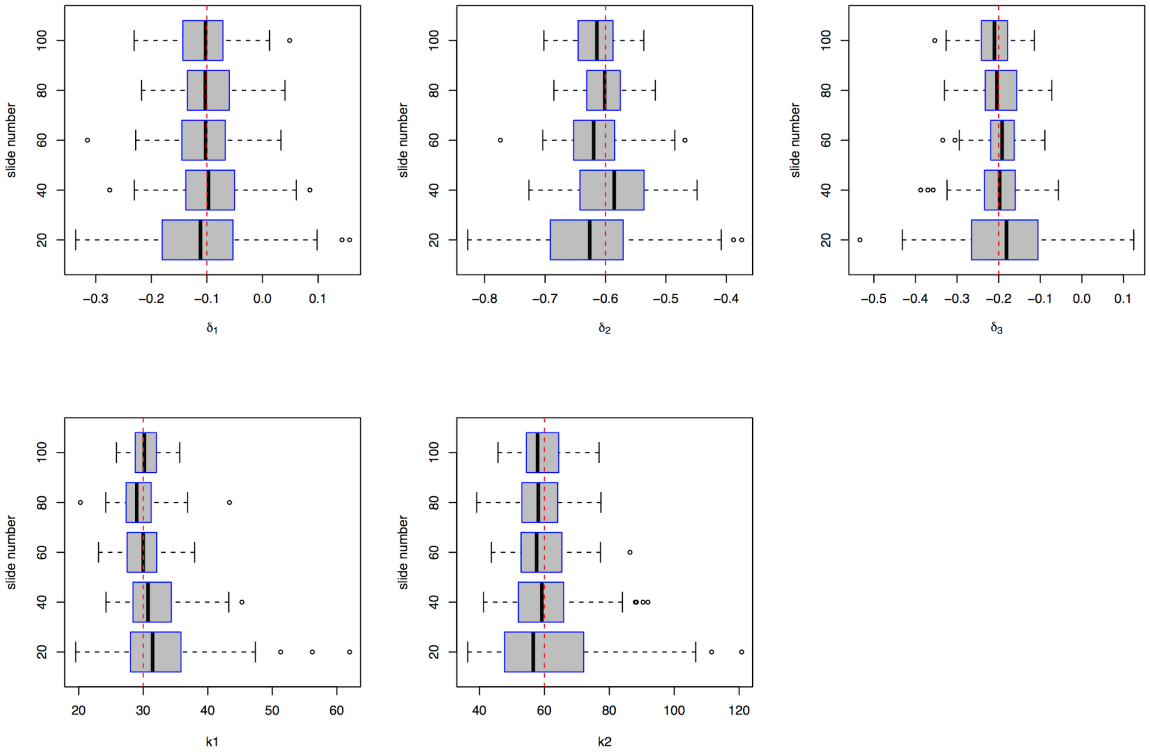


Fig. 3.17. Distribution of posterior means for case 2

## 3.11 Robustness of the Proposed Model

To investigate how our model performs when the observed data do not follow a Dirichlet distribution, we simulated observed data using logistic Normal distributions with positive correlations. The procedure of simulating logistic Normal data is as follows:

(i) Simulate 50 reference means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{50}$ from a Dirichlet prior $\mathcal{D}(\boldsymbol{\mu}_p, k_p)$ with $\boldsymbol{\mu}_p = (0.2, 0.3, 0.3, 0.2)$ and $k_p = 10$.

(ii) Based on the relationship between Dirichlet and logistic Normal (Equation 1.1, page 12), the corresponding logistic Normal means $\mathbf{m}_i$ can be calculated. This would ensure the logistic Normal slide means are comparable to the Dirichlet slide means. To allow for positive covariances, we simulated the $\Sigma_i$ for each slide using the Inverse-Wishart ($\Psi$, $v$). Set $\Psi = \psi_{ij}$, where $\psi_{ii} = 1$ and $\psi_{ij} = 0.3\,(i \neq j)$, and $v = 5$, such that the simulated logistic Normal data have moderate strength of positive correlation among components. Then generate Rater A's observations as $\mathbf{x}_i \sim \mathcal{L}(\mathbf{m}_i, \Sigma_i)$.

(iii) Given the 50 reference means and a fixed set of shifts $(\delta_1, \delta_2, \delta_3)$, calculate the 50 Dirichlet means $\boldsymbol{\mu}_{y1}, \ldots, \boldsymbol{\mu}_{y50}$ for Rater B. Again the comparable logistic Normal distribution means $\mathbf{m}_{yi}$ can be obtained and $\Sigma_{yi} \sim$ Inverse-Wishart$(\Psi, v)$ for each slide. Rater B's observations $\mathbf{y}_i$ are generated from this logistic Normal.

(iv) In order to compare the bias and the coverage, 50 pairs of observations from Dirichlet distributions are also generated based on $\mathcal{D}(\boldsymbol{\mu}_i, k)$ and $\mathcal{D}(\boldsymbol{\mu}_{yi}, k)$.

For each scenario, the above procedure is repeated 100 times and 100 posterior means and 95% credible intervals are stored. The numbers in Table 3.7 represent the average of the 100 posterior means and how many times the 95% credible interval contained the true parameters. Two intra-rater variabilities of Dirichlet distributions and three sets of shifts are considered here. The intra-rater parameter $k$ doesn't impact the covariance structure $\Sigma$ but it does impact how close the logistic Normal means approximate the corresponding Dirichlet means. Recall that the bigger $k$ is (thus the bigger $\boldsymbol{\alpha}$ since $\boldsymbol{\alpha} = \boldsymbol{\mu}k$), the closer the logistic Normal distribution approximates the Dirichlet distribution. Each of the shift sets represents a typical agreement case: two raters generally agree, two raters vary in terms of distinguishing the middle categories, and two raters have relatively poor agreement.

Table 3.7.
Logistic Normal data versus Dirichlet data[a]

| $(\delta_1, \delta_2, \delta_3)$ | $k$ | Dirichlet | | | Logistic Normal | | |
|---|---|---|---|---|---|---|---|
| $(-0.1, +0.2, +0.1)$ | 10 | -0.14(0.90) | 0.20(0.94) | 0.18(0.90) | 0.01(0.80) | 0.24(0.88) | 0.33(0.74) |
| | 50 | -0.12(0.96) | 0.20(0.94) | 0.11(0.94) | -0.06(0.78) | 0.21(0.85) | 0.18(0.83) |
| $(-0.1, -0.6, -0.2)$ | 10 | -0.13(0.80) | -0.60(0.88) | -0.15(0.92) | -0.13(0.73) | -0.41(0.70) | -0.01(0.77) |
| | 50 | -0.12(0.87) | -0.61(0.92) | -0.18(0.92) | -0.12(0.71) | -0.58(0.84) | -0.14(0.75) |
| $(+0.8, +0.5, +0.2)$ | 10 | 0.78(0.84) | 0.54(0.82) | 0.19(0.88) | 0.70(0.84) | 0.53(0.71) | 0.36(0.63) |
| | 50 | 0.80(0.95) | 0.51(0.93) | 0.20(0.86) | 0.81(0.88) | 0.52(0.85) | 0.10(0.70) |

[a] The average of 100 posterior means (coverage of 100 95% credible intervals)

When we compare the average posterior means, we do observe some bias from the logistic Normal data when $k = 10$. The bias tends to get very small when $k$ increases to 50. Generally, our model with Dirichlet data provides greater than 80% chance ($> 90\%$ in most cases) that the 95% credible intervals cover the true parameters. When $k = 10$, our approach using logistic Normal data can still provide $\geq 70\%$ chance to cover the true parameters (except one in the poor agreement case). When $k$ increases to 50, the coverages given logistic Normal data slightly improve. This shows our model is generally robust to logistic Normal data though a moderate positive covariance structure is assumed. Further studies may still be needed to investigate how our model performs given other covariance structures.

The average coverages from our model do not obtain 95% and this is particularly true for small $k$'s. We used the highest posterior density (HPD) (the narrowest intervals) to calculate the credible intervals, and this might be one of the reasons for the under-coverage. In addition, the 95% credible intervals represent the Bayesian coverages which don't necessarily have the 95% coverage properties in a frequentist setting (Wasserman, 2004). Adding sample size and increasing $k$ would increase our Bayesian coverages.

## 3.12 Discussion

Motivated by the inter-rater agreement problem in IHC scoring, we propose a Bayesian method to assess inter-rater agreement for compositional data. Due to the sum-to-one constraint of the compositional data, the Dirichlet distribution serves as a reasonable distribution to describe these compositional scores and we consider a logistic link between pairs of rater means to describe pattern differences in response.

Our proposed model is generalized to handle not only continuous but also discrete or rounded compositional data across multiple raters with or without replication. Simulation studies show that our model can provide unbiased estimates of parameters with decent robustness. The interpretation of this model is fairly easy, as the shift parameters serve as an indication of the pattern of rater agreement and the BC serves as an overall agreement index.

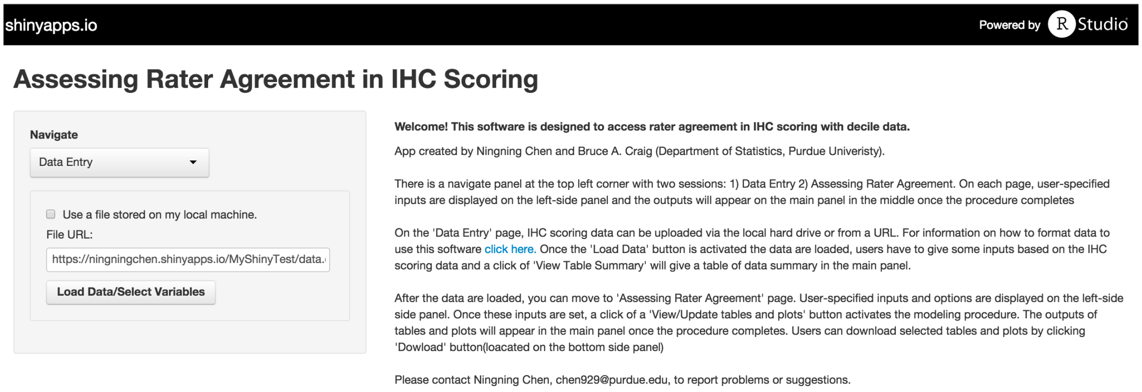
Besides the proportional odds assumption inherent to our model through shifts of the logit link, we have also considered another view of scoring. That is that the cutpoints for both raters on the latent intensity scale are fixed and the distribution of cell intensities on each slide varies. The difference between these two approaches was discussed in Section 3.1.3.

# CHAPTER 4. A REAL DATA APPLICATION IN IHC ASSAYS

## 4.1 Introduction

In Section 3.1, we provided an overview of IHC assays and the common scoring method that results in a compositional vector. We also emphasized the importance of inter-rater agreement of these vector scores for consistency in prognosis. In this chapter, we apply our proposed methodology to some real IHC agreement study data and provide a comparison of our assessment with those based on the H-score. Moreover, we discuss power calculations based on our model to assist in determining sample size for future studies.

To facilitate the use of our methodology, an online software platform was constructed using R Shiny (Version 0.12.1). We demonstrate the features of this program throughout the chapter. It can be accessed at:
https://ningningchen.shinyapps.io/MyShinyTest. The program opens to a welcome main page (Figure 4.1) which provides users with basic navigational instructions on how to use the software, contact information if there are any questions, and a link to a more detailed instruction document.



shinyapps.io

Powered by R Studio

Assessing Rater Agreement in IHC Scoring

Navigate

Data Entry

□ Use a file stored on my local machine.

File URL: https://ningningchen.shinyapps.io/MyShinyTest/data.

Load Data/Select Variables

**Welcome! This software is designed to access rater agreement in IHC scoring with decile data.**

App created by Ningning Chen and Bruce A. Craig (Department of Statistics, Purdue Univeristy).

There is a navigate panel at the top left corner with two sessions: 1) Data Entry 2) Assessing Rater Agreement. On each page, user-specified inputs are displayed on the left-side panel and the outputs will appear on the main panel in the middle once the procedure completes

On the 'Data Entry' page, IHC scoring data can be uploaded via the local hard drive or from a URL. For information on how to format data to use this software click here. Once the 'Load Data' button is activated the data are loaded, users have to give some inputs based on the IHC scoring data and a click of 'View Table Summary' will give a table of data summary in the main panel.

After the data are loaded, you can move to 'Assessing Rater Agreement' page. User-specified inputs and options are displayed on the left-side side panel. Once these inputs are set, a click of a 'View/Update tables and plots' button activates the modeling procedure. The outputs of tables and plots will appear in the main panel once the procedure completes. Users can download selected tables and plots by clicking 'Dowload' button(loacated on the bottom side panel)

Please contact Ningning Chen, chen929@purdue.edu, to report problems or suggestions.

Fig. 4.1. Screenshot: welcome page

## 4.2 Study Description

This real study was provided by the Oncology Lab, Eli Lilly & Company and it follows the layout of a typical multivariate IHC study (Table 3.1), specifically to compare three raters to a “Gold Standard” (a very experienced rater). Here is a summary of the details:

**Number of unique IHC slides:** $n = 30$.

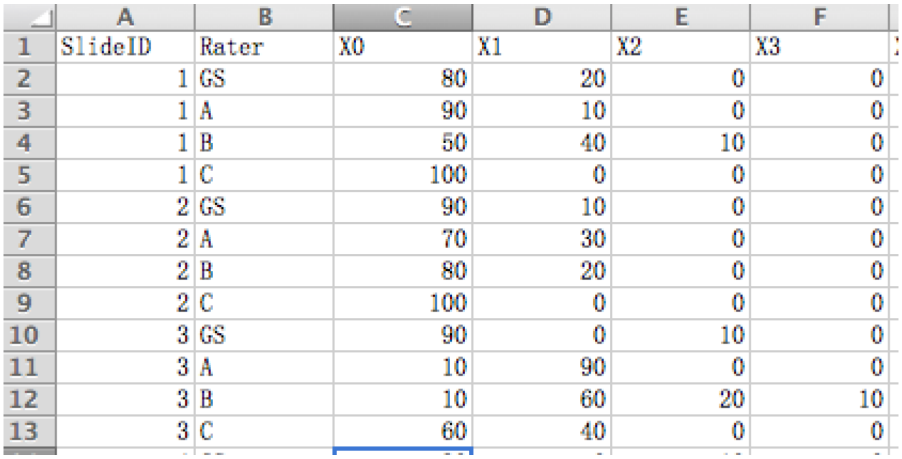
**Rater labels:** Rater GS (experienced rater), Rater A, Rater B, Rater C.

**Replication:** GS scored all 30 IHC slides once; Raters A, B and C scored 15 of these IHC slides twice and the other 15 IHC slides once.

**Missing data:** Rater A has a missing score on one of the IHC slides that is scored once.

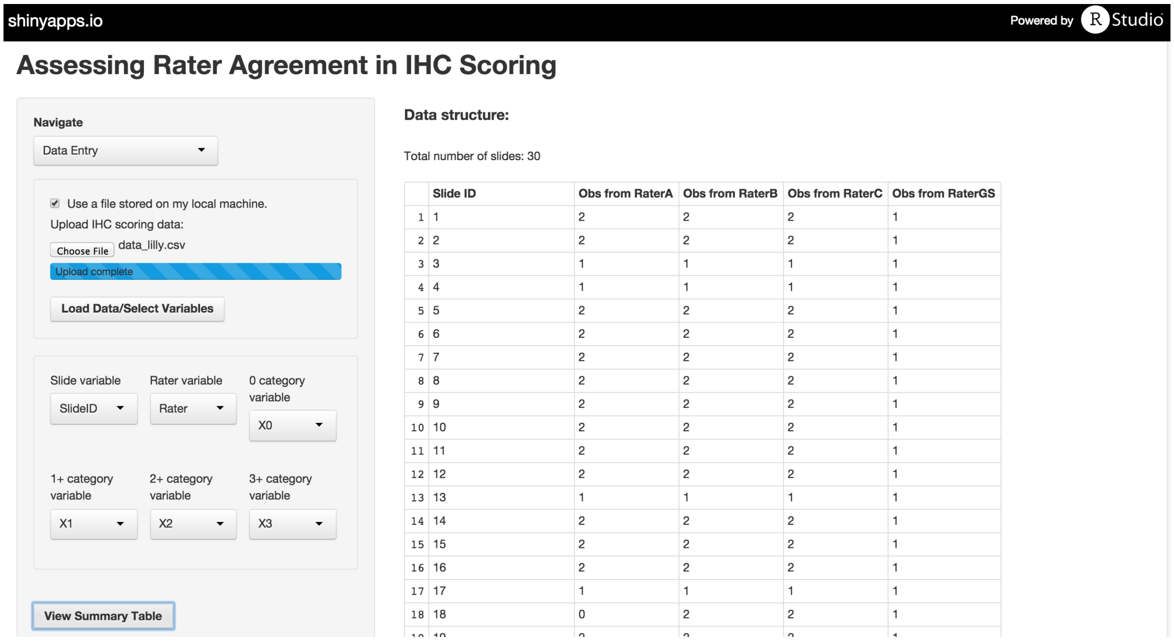
To upload a data set into our platform it needs to be in a particular form. The data set should include one column that identifies the rater, one column that identifies the slide, and four columns that identify the four staining intensity categories. Figure 4.2 is an example of a data set ready for our software. The scores for each staining category can be either in percent format as shown in the figure or in decimal format. Users use the “Data Entry” option from the Navigate panel to upload data sets from either a local machine, or provide a link where the data set is stored. The default data set provided with the program is the real IHC data we use throughout this chapter.

After uploading, users specify the corresponding variable names by choosing them from the drop-down lists on the left panel (Figure 4.3). The “View Summary Table” option provides a more detailed summary of the data once the data set is uploaded and the variable names are specified. The right side of the window lists each slide as a row and each rater as a column. The number in each cell represents the number of replicate scores. For our real data set (Figure 4.3), we can see Rater A has a missing score for Slide 18.



|  | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | SlideID | Rater | X0 | X1 | X2 | X3 |
| 2 | 1 | GS | 80 | 20 | 0 | 0 |
| 3 | 1 | A | 90 | 10 | 0 | 0 |
| 4 | 1 | B | 50 | 40 | 10 | 0 |
| 5 | 1 | C | 100 | 0 | 0 | 0 |
| 6 | 2 | GS | 90 | 10 | 0 | 0 |
| 7 | 2 | A | 70 | 30 | 0 | 0 |
| 8 | 2 | B | 80 | 20 | 0 | 0 |
| 9 | 2 | C | 100 | 0 | 0 | 0 |
| 10 | 3 | GS | 90 | 0 | 10 | 0 |
| 11 | 3 | A | 10 | 90 | 0 | 0 |
| 12 | 3 | B | 10 | 60 | 20 | 10 |
| 13 | 3 | C | 60 | 40 | 0 | 0 |

Fig. 4.2. Screenshot: Data format



Assessing Rater Agreement in IHC Scoring

Navigate

Data Entry

☑ Use a file stored on my local machine.

Upload IHC scoring data:

data_lilly.csv

Upload complete

Slide variable: SlideID

Rater variable: Rater

0 category variable: X0

1+ category variable: X1

2+ category variable: X2

3+ category variable: X3

Data structure:

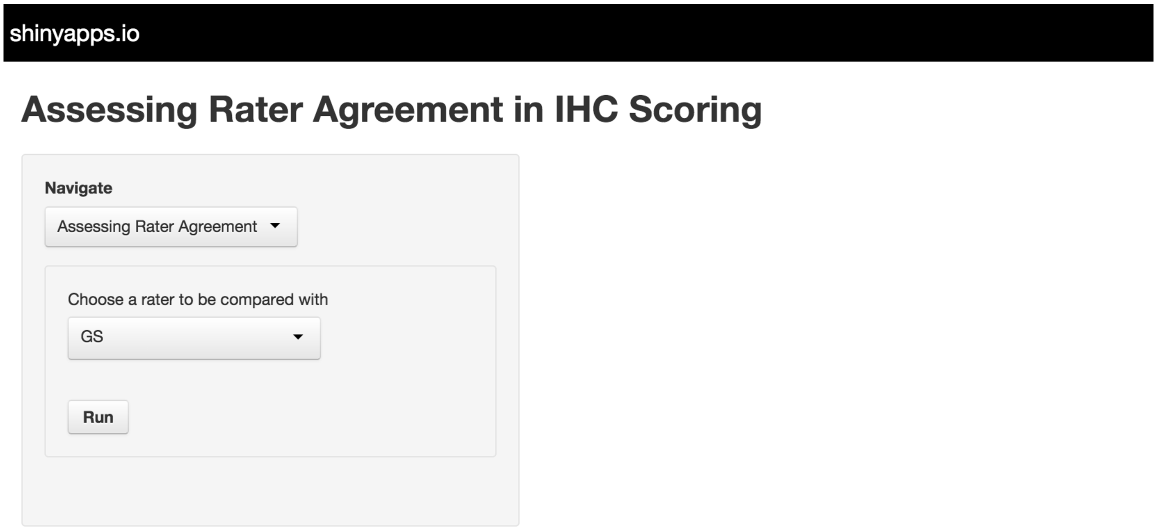
Total number of slides: 30

|  | Slide ID | Obs from RaterA | Obs from RaterB | Obs from RaterC | Obs from RaterGS |
|---|---|---|---|---|---|
| 1 | 1 | 2 | 2 | 2 | 1 |
| 2 | 2 | 2 | 2 | 2 | 1 |
| 3 | 3 | 1 | 1 | 1 | 1 |
| 4 | 4 | 1 | 1 | 1 | 1 |
| 5 | 5 | 2 | 2 | 2 | 1 |
| 6 | 6 | 2 | 2 | 2 | 1 |
| 7 | 7 | 2 | 2 | 2 | 1 |
| 8 | 8 | 2 | 2 | 2 | 1 |
| 9 | 9 | 2 | 2 | 2 | 1 |
| 10 | 10 | 2 | 2 | 2 | 1 |
| 11 | 11 | 2 | 2 | 2 | 1 |
| 12 | 12 | 2 | 2 | 2 | 1 |
| 13 | 13 | 1 | 1 | 1 | 1 |
| 14 | 14 | 2 | 2 | 2 | 1 |
| 15 | 15 | 2 | 2 | 2 | 1 |
| 16 | 16 | 2 | 2 | 2 | 1 |
| 17 | 17 | 1 | 1 | 1 | 1 |
| 18 | 18 | 0 | 2 | 2 | 1 |

Fig. 4.3. Screenshot: data summary

## 4.3 Modeling Procedure and Results

### 4.3.1 Estimation using Rater GS as the reference



Assessing Rater Agreement in IHC Scoring

Navigate
Assessing Rater Agreement

Choose a rater to be compared with
GS

Fig. 4.4. Screenshot: preparation for model fit

To apply our model to the real IHC data set, users need to choose "Assessing Rater Agreement" from the drop-down list in the Navigate panel and then choose the reference rater (Figure 4.4). The Bayesian MCMC procedure will initiate once a user clicks "Run". A progress bar will show up on the right corner to indicate the progress and results will be displayed after the procedure is completed. This software can automatically detect different types of the data uploaded and use the appropriate model to fit the data (Chapter 3, Section 3.7-3.9). For example, the software can detect if there is any replication (i.e., multiple scores by the same rater on the same slide). If no replication is detected, a single intra-rater variability parameter is estimated. Moreover, if there are missing scores from the chosen reference rater, the software will generate an error message indicating the missingness. Thus, for our case, Rater A cannot be chosen as the reference rater.

For this study, there are four raters assessing 30 IHC assays, with three raters having partial duplicates. Due to the inclusion of replication and multiple raters, we utilize our extended model summarized in Section 3.8 for the analysis.

Table 4.1.
Summary for model estimates[a]

| Parameter | Rater GS | Rater A | Rater B | Rater C |
|---|---|---|---|---|
| $\delta_1$ | — | -0.74 (-1.1, -0.45) | -0.82 (-1.14, -0.52) | 0.49 (0.18, 0.81) |
| $\delta_2$ | — | 0.54 (0.18, 0.86) | -0.96 (-1.25, -0.66) | 0.54 (0.18, 0.84) |
| $\delta_3$ | — | 0.40 (0.04, 0.78) | -0.90 (-1.23, -0.48) | 0.25 (-0.17, 0.65) |
| $k$ | 9.09 (5.92, 12.16) | 13.74 (8.3, 18.76) | 14.80 (10.64, 19.71) | 8.77 (5.78, 11.11) |

[a] Posterior means (95% credible interval)

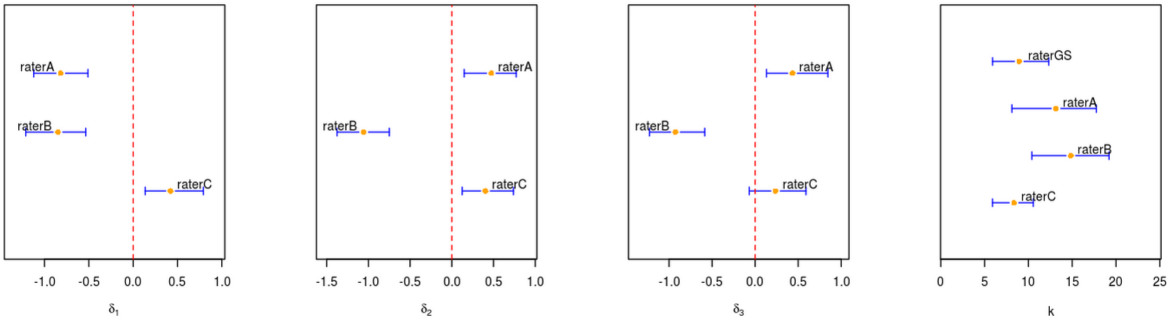


Fig. 4.5. Posterior means and 95% credible intervals of model estimates

Since we are interested in comparing Rater A, B and C to Rater GS, there are three sets of shifts and intra-rater variabilities corresponding to these three raters that need to be estimated. Table 4.1 and Figure 4.5 summarize the model estimates. To give readers a better idea of how these three raters differ from Rater GS in terms of the percents rather than the shifts on the logistic scale, the presumed mean scores for three different IHC assays are provided in Table 4.2.

Table 4.2.
Expected scores for Rater A, B, and C given three reference assays

| **Rater GS (reference)(%)** | Rater A | Rater B | Rater C |
|---|---|---|---|
| **(25, 25, 25, 25)** | (14, 49, 19, 18) | (13, 15, 27, 45) | (35, 28, 16, 21) |
| **(10, 40, 40, 10)** | (5, 57, 31, 7) | (5, 21, 52, 22) | (15, 45, 31, 9) |
| **(5, 15, 20, 60)** | (3, 28, 19, 50) | (2, 6, 12, 80) | (8, 21, 16, 55) |

The overall population mean $\boldsymbol{\mu_p}$ is estimated to be [0.25, 0.29, 0.19, 0.18] and the population dispersion $k_p$ is estimated to be 2 based on the observed data from Rater GS. This indicates our IHC slides have very different mixture means. All four raters show a very large amount of intra-rater variability. Surprisingly, Rater GS has slightly bigger intra-rater variability than Rater A and Rater B, but not significantly different from any rater. Given these small value of $\hat{k}$'s, we might still see some overestimation of $k$ because of the poor estimation of $k_p$. In addition, this small value of $k$ definitely introduces large uncertainty in estimating the shifts. This is why we see relatively wide credible intervals for $\delta_1$, $\delta_2$, and $\delta_3$ similar to our simulation studies in Section 3.7, Chapter 3 (Table 3.5). We also expect that raters vary more when they score the middle two staining levels than the lower and upper two levels. However, the shift estimates and Table 4.2 doesn't necessarily show the expected pattern except for Rater C.

Overall, Rater A can be defined as a "moderate" rater because of the negative shift (-0.74) between "negative" and "weak" staining categories and the positive shift (0.40) between "moderate" and "positive" staining categories. All shift estimates of Rater A are significantly different from zero based on the 95% credible intervals. This implies Rater A tends to assign most of the cells to the middle two categories. Rater B and Rater C have the opposite behaviors in terms of scoring IHC assays. Rater B is a "bold" rater as the three shifts are consistently negative and significantly different from zero, implying the cells are assigned to the higher categories. Rater C on the other hand is more "conservative" in assigning cells to the "positive" category and tends to assign more cells to "negative" category. Only Rater C's shift between "moderate" and "positive" categories is not significantly different from zero.

### 4.3.2 Estimation using Rater C as the reference

Our software allows users to specify the reference rater. Instead of GS as the reference, this section shows the output using Rater C as the reference rater. We switch the reference rater from Rater GS to Rater C using the drop-down list option shown in Figure 4.4 and re-run the analysis. In Table 4.3, we match Rater C's mean scores in Table 4.2 and calculate the expected mean scores from Rater GS, Rater A, and Rater B based on the new estimates of model parameters.

Comparing Table 4.2 and Table 4.3, the shift estimates are consistent no matter which rater is chosen as the reference since the estimated expected means are similar across these two tables. The deviance of the expected mean percent in each category is within 5% except the third reference row of Rater GS. This, again, may be due to no replication from Rater GS. The $k$ estimates for all raters are almost the same with the values shown in Table 4.1, thus we omit listing a redundant table here.

Moreover, we notice that in Table 4.1 and Figure 4.5, Rater A and Rater C have the same $\hat{\delta}_2$ and overlapped credible intervals for $\delta_2$. After we switch the reference rater to Rater C, we get $\hat{\hat{\delta}}_2 = 0.02$ with the credible interval $(-0.21, 0.29)$. This confirms again our previous result using Rater GS as the reference rater.

Table 4.3.
Expected scores for Rater GS, A, and B given three reference assays

| Rater GS | Rater A | Rater B | **Rater C (reference)(%)** |
|---|---|---|---|
| (20, 33, 28, 19) | (13, 53, 19, 15) | (12, 17, 31, 41) | **(35, 28, 16, 21)** |
| (8, 42, 42, 8) | (5, 58, 31, 6) | (4, 22, 54, 20) | **(15, 45, 31, 9)** |
| (4, 18, 25, 53) | (2, 29, 23, 46) | (2, 7, 15, 76) | **(8, 21, 16, 55)** |

### 4.3.3 Overall agreement using the BC index

Our purpose for this analysis is to compare all other three raters to GS, thus we calculate the posterior mean estimates of the BC and the corresponding 95% credible interval for the three pairwise comparisons. The reference slide used here is the overall observed mean calculated from Rater GS. To also investigate and compare other agreement indices, the CCC and the $\psi$ (1-CIV) indices based on the H-score method, as well as the iota coefficient using the raw compositional vectors are calculated (Table 4.4).

Before we make any conclusions based on these agreement indices, we investigate the sensitivity of the BC, the CCC, the $\psi$, and the iota coefficient to changes in $\delta_2$. We focus on changes just in $\delta_2$ because this was the breakpoint considered to be the most variable. Recall the BC and the $\psi$ ranges from 0 to 1 with a value of 1 indicating perfect overlap/agreement between two raters while the CCC and the iota coefficient both range from -1 to 1 for two-rater case. Though they have different ranges, most of the values of the CCC and the iota coefficient range from 0 to 1 and a value between 0 and 1 has the same interpretation among these four agreement indices.

Based on the data simulation procedure (Section 3.5.1, page 75), we simulated 100 slides for each $\delta_2$ from the Dirichlet population with $\boldsymbol{\mu_p} = [0.25, 0.25, 0.25, 0.25]$ and $k_p = 5$. The intra-rater variability parameter $k$ is set to be 30 for both raters. This setup introduces a large amount of between-slide variability. The BC was calculated using $\boldsymbol{\mu_p}$ as the reference mean, which is denoted as BC$_o$. Since only one data set is simulated for each $\delta_2$, we do expect some noise in the results. To assess the trend, we plot loess curves.

Figure 4.6 displays the index scores when $\delta_1$ and $\delta_3$ are fixed at zero and $\delta_2$ varies from -1.0 to 1.0 by 0.1 increment. Figure 4.6 clearly shows that the BC is very sensitive to changes in $\delta_2$ and ranges from 0.99 ($\delta_2 = 0$) to nearly 0.07 ($\delta_2 = 1$). The second most sensitive index is $\psi$ because the between-slide variability isn't included and the between-rater component dominates over the intra-rater variability. Due to the loss of information when the H-score is used, the $\psi$ is not as sensitive as the BC. The iota coefficient shows some sensitivity, ranging from 0.85 to 0.57, but it factors in the between-slide variability which is very large for this example. In fact, because of the loss of information using the H-score and its dependence on the between-slide variability, the CCC doesn't really vary at all over the range of $\delta_2$.

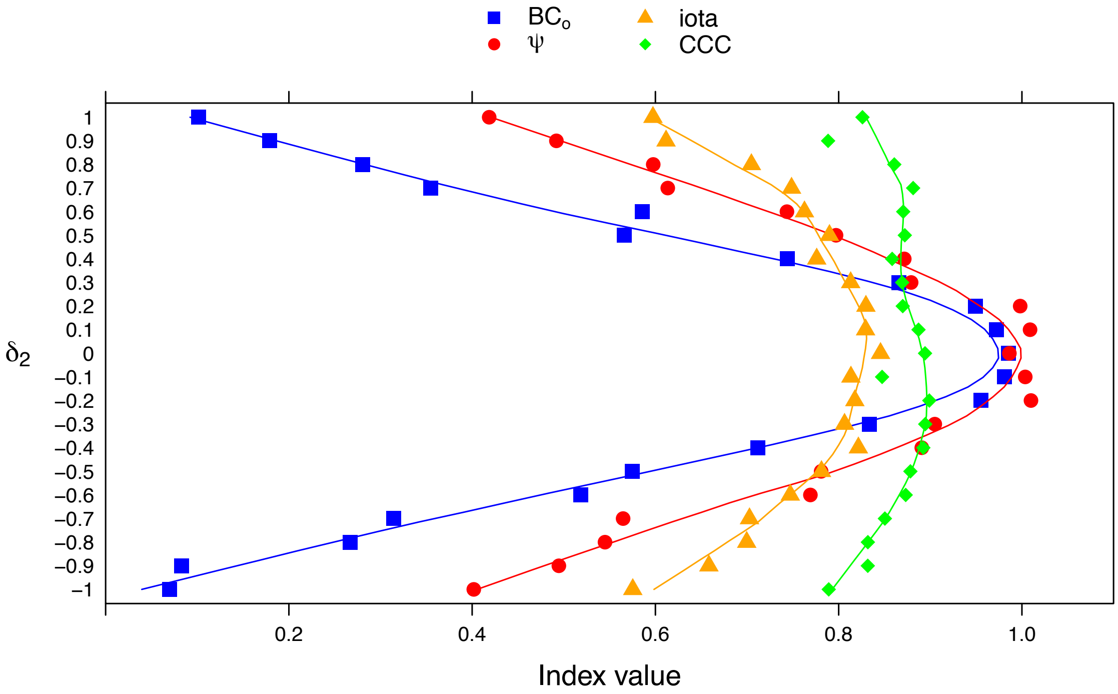


Fig. 4.6. The values of BC, iota coefficient, CCC, $\psi$ based on uniform population

Figure 4.7 shows a similar experiment but here the slides come from a non-central distribution $\mathcal{D}([0.1, 0.1, 0.2, 0.6], 80)$. The most noticeable difference here is that the CCC and iota coefficient now range between 0.1 to 0.25, much lower than observed previously. This is due to the dramatic decrease of the between-slide variability. The BC$_o$ and $\psi$ are more sensitive to positive changes of $\delta_2$ due to the skewness of the slide population. Similar trends and degrees of sensitivity, however, are observed here for all four indices.

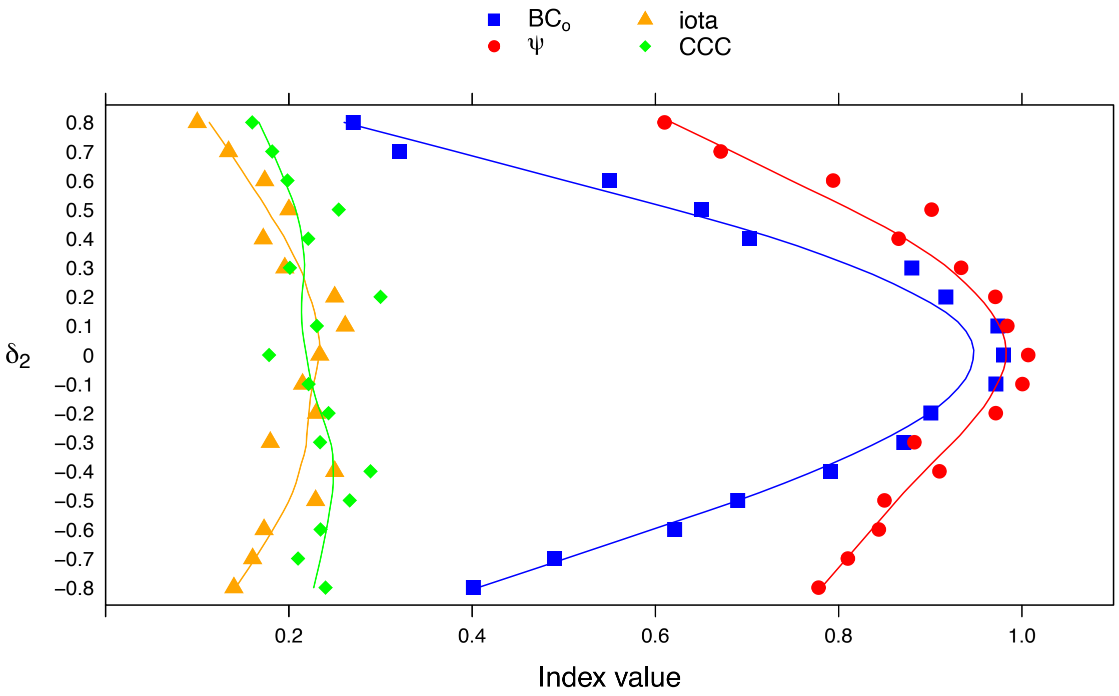


Fig. 4.7. The values of BC, iota coefficient, CCC, $\psi$ based on skewed population

These two simulations assumed $k1 = k2 = 30$. Increasing $k$ will reduce the intra-rater variability, thereby increasing the sensitivity of BC$_o$ and $\psi$ while simply increasing the average value for the other two indices. Different values of $k_1$ and $k_2$ will not impact iota coefficient and the CCC much but will generally lower the index value for BC$_o$.

We can understand the diminished sensitivity of the $\psi$ and the insensitivity of the CCC better by examining the distributions of the differences of the H-score between the two raters. In Figure 4.8, when $\delta_2$ is small (0.2), the differences of the H-score between Rater A and Rater B are almost centered at 0. When $\delta_2$ increases to 1.0, we want to see this distribution shift far away from the center. However, we only observe a small shift to the right which indicates that the H-score distribution is insensitive to big changes of $\delta_2$. This example demonstrates the biggest drawback of using the H-score for agreement analysis.

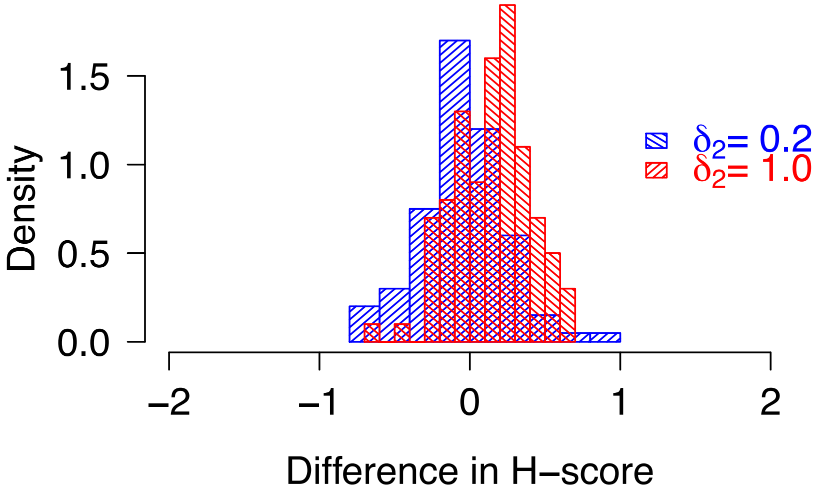


Fig. 4.8. Distributions of H-score

Now that we have an idea of the sensitivity of all the agreement indices, we calculate them based on our real data set. The results in Table 4.4 confirm our findings from Figure 4.6 and Figure 4.7. Generally, the CCC does not vary as much as the other three agreement indices. Given this 30-slide population, the BC index shows Rater C has the best agreement with Rater GS, followed by Rater A, and then Rater B, which is consistent with the iota coefficient. In contrast, the $\psi$ and the CCC show that the agreement of Rater A > Rater C > Rater B based on the corresponding H-score distribution. We notice that the value of $\psi$ for Rater B is really low. This can be explained by the three large negative shifts of Rater B resulting in more percent assignments to the positive staining level. Since the H-score gives the biggest weight to the percent in the positive staining level, the low value of $\psi$ is expected. Similarly, the value of $\psi$ for Rater A is close to 1 even though there is a relatively big negative value of $\delta_1$ and moderate positive values of $\delta_2$ and $\delta_3$, because the H-score doesn't not change much in this scenario.

To see if this pattern consistently occurs, we simulated 100 data sets using the posterior means as our parameters, fit each data set, and calculated all the agreement indices. We always observe the agreement pattern of Rater C > Rater A > Rater B based on the estimated BCs. However, the CCC, the $\psi$ and the iota coefficients show Rater B always has the lowest agreement with Rater GS but Rater A and Rater C are very close since roughly half of the time we observe Rater A having the best agreement. Except for Rater C, the BC scores are not particularly large. This is due to the large estimated shifts and large amount on intra-rater variability. In fact, even though Rater C has a BC value close to one, we may observe very different slide scores between GS and C because of this intra-rater variability.

Table 4.4.
Comparisons of overall agreement indices: BC, CCC and the iota coefficient

| Agreement Index | Rater A | Rater B | Rater C |
|---|---|---|---|
| BC[a] | 0.64 (0.51, 0.77)[b] | 0.58 (0.44, 0.75) | 0.91 (0.78, 0.99) |
| $\psi$[c] | 0.98 | 0.38 | 0.91 |
| CCC | 0.88 (0.79, 0.94)[d] | 0.79 (0.66, 0.88) | 0.84 (0.71, 0.91) |
| iota | 0.63[e] | 0.55 | 0.66 |

[a] Reference assay mean is the observed overall mean $(25, 28, 29, 18)\%$

[b] Posterior means (95% credible interval)

[c] Point estimate using ANOVA based method

[d] Point estimate (95% confidence interval)

[e] Point estimate (no confidence interval available)

The BC calculated in Table 4.4 is conditional on the observed overall mean, but it varies if we specify different reference means (page 73-74). The BC index has its advantage that it provides more information regarding the assay population. For

Table 4.5.
The BCs based on three assay populations

| Poppulation | Rater A | Rater B | Rater C |
|---|---|---|---|
| $(25, 25, 25, 25)\%$ | 0.65 (0.52, 0.81) | 0.57 (0.42, 0.69) | 0.87 (0.74, 0.99) |
| $(10, 40, 40, 10)\%$ | 0.80 (0.67, 0.88) | 0.60 (0.47, 0.75) | 0.89 (0.77, 0.98) |
| $(5, 15, 20, 60)\%$ | 0.75 (0.58, 0.88) | 0.71 (0.59, 0.86) | 0.87 (0.72, 0.99) |

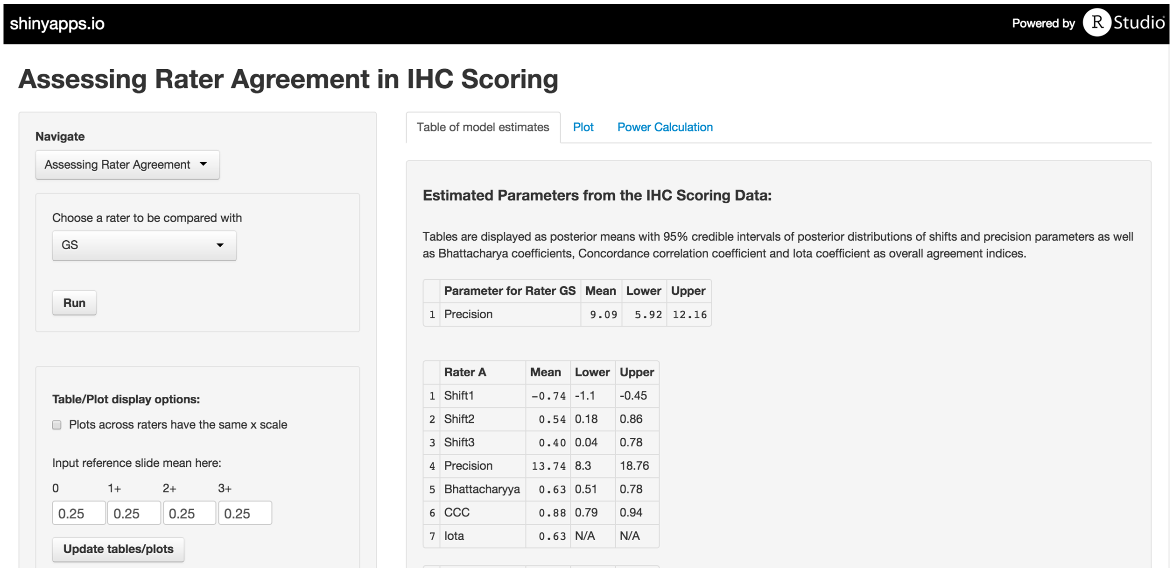
example, people may care more about agreement in the subpopulation having high percents in the "positive" category. Table 4.5 lists three typical assay populations with the corresponding posterior means and credible intervals of the BC. Generally, we observe larger values of BC when one or two categories dominate.

### 4.3.4 Software instructions

All the tables and figures displayed and interpreted can be obtained using our web application except the value of $\psi$. We do not present $\psi$ because it is not widely used yet and no existing standard software for calculating the CIV/CIA. In Figure 4.8, on the left panel of "Assessing Rater Agreement" Navigation, users can specify the reference mean for the BC calculation and expected scores below "Table/plot display option". There are table and plot tabs on the main panel where users can obtain all the desired results in forms of tables and plots. A download option is also provided if users want to save the table/plot results on their local machines. To better compare estimates across raters, a customize option is provided to plot each parameter estimates on the same scale (Figure 4.10).

## 4.4 Power and Sample Size

Our model can also be used to help answer design questions, specifically how many slides to include and how many replicates per slide. This is done through simulation



Assessing Rater Agreement in IHC Scoring

Choose a rater to be compared with: GS

Table/Plot display options:

□ Plots across raters have the same x scale

Input reference slide mean here:

| 0 | 1+ | 2+ | 3+ |
|---|---|---|---|
| 0.25 | 0.25 | 0.25 | 0.25 |

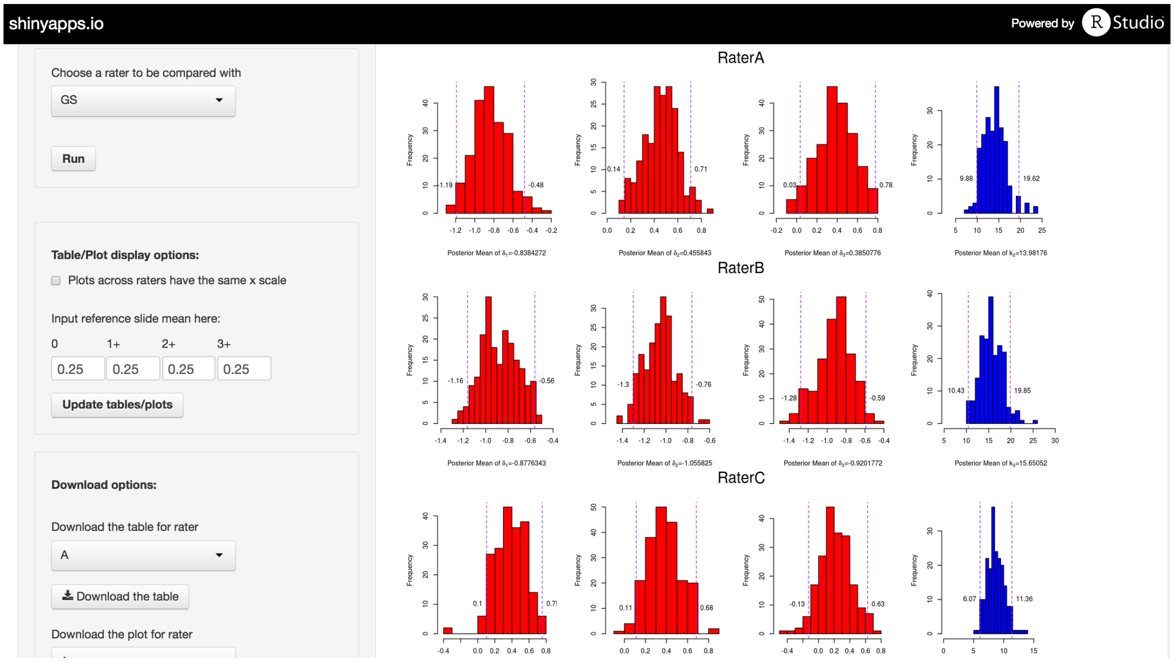
Estimated Parameters from the IHC Scoring Data:

Tables are displayed as posterior means with 95% credible intervals of posterior distributions of shifts and precision parameters as well as Bhattacharya coefficients, Concordance correlation coefficient and Iota coefficient as overall agreement indices.

| | Parameter for Rater GS | Mean | Lower | Upper |
|---|---|---|---|---|
| 1 | Precision | 9.09 | 5.92 | 12.16 |

| | Rater A | Mean | Lower | Upper |
|---|---|---|---|---|
| 1 | Shift1 | −0.74 | -1.1 | -0.45 |
| 2 | Shift2 | 0.54 | 0.18 | 0.86 |
| 3 | Shift3 | 0.40 | 0.04 | 0.78 |
| 4 | Precision | 13.74 | 8.3 | 18.76 |
| 5 | Bhattacharyya | 0.63 | 0.51 | 0.78 |
| 6 | CCC | 0.88 | 0.79 | 0.94 |
| 7 | Iota | 0.63 | N/A | N/A |

Fig. 4.9. Screenshot: tables model estimates


shinyapps.io
Powered by R Studio
Run
Update tables/plots
Download the table
RaterA
Frequency
-1.19
-0.48
Posterior Mean of δ1=-0.8384272
0.14
0.71
Posterior Mean of δ2=0.455843
0.03
0.78
Posterior Mean of δ3=0.3850776
9.88
19.62
Posterior Mean of k2=13.98176
RaterB
-1.16
-0.56
Posterior Mean of δ1=-0.8776343
-1.3
-0.76
Posterior Mean of δ2=-1.055825
-1.28
-0.59
Posterior Mean of δ3=-0.9201772
10.43
19.85
Posterior Mean of k3=15.65052
RaterC
0.1
0.75
0.11
0.68
-0.13
0.63
6.07
11.36

Choose a rater to be compared with: GS

Table/Plot display options:

□ Plots across raters have the same x scale

Input reference slide mean here:

| 0 | 1+ | 2+ | 3+ |
|---|---|---|---|
| 0.25 | 0.25 | 0.25 | 0.25 |

Download options:

Download the table for rater: A

Download the plot for rater

Fig. 4.10. Screenshot: plots of model estimates

under different scenarios. Recall in Section 3.10, we simulated decile data sets with different sample sizes and investigated the posterior means of model estimates (Figure 3.15 and Figure 3.16, page 89-91). In particular, when slide numbers increases from 20 to 40, the variation of posterior means goes down quickly. However, as the slides number continue to increase, there is only a slight variance reduction benefit.

In practice, 50 rater evaluations are probably the most we can get. Thus, we will consider different combinations of slides and replicates to see if they have an impact on power. We consider the slides coming from three different Dirichlet populations and fix $\delta_1 = \delta_3 = 0$. We vary $\delta_2$, $k_1$, $k_2$ and $n$. Power is approximated by simulating 100 data sets under each combination of ($\delta_2$, $k_1$, $k_2$, $n$) and then counting how many times the 95% credible interval of $\delta_2$ does not contain 0 (Table 4.6 and Table 4.7).

Since $\delta_1$ and $\delta_3$ are fixed to be zero, the mean percents in "negative" and "positive" categories between two raters are fixed to be the same while the change of $\delta_2$ contributes solely to the percent variation between "weak" and "moderate" categories. For example, for Population 1 and Population 2, the population slide mean ($\mu_p$) is symmetric, and ±0.1 of $\delta_2$ results in ±2.5% change to the "weak" category. Generally, the bigger the amount of percent change in original compositions is, the higher power we have to detect such change. Based on the $k$ estimates from the real data set, intra-rater variabilities for both raters are set to be the same with a value of 20, for simplicity. This introduces considerable intra-rater variability. The results of this power calculation show that when slides are from Population 1 and Population 2, over 75% power of detecting ±0.3 shifts can be obtained given 30 slides with no replication (i.e., sample size: 30×1). Population 3 is an example of very skewed slide population, therefore, the power decreases due to the decreased percent change given the same value of $\delta_2$. In summary, to be able to detect 5% percent change with close to or over 70% power, 30 slides with two replicates from each rater (60 observations from each rater) is recommended.

In Table 4.7, we also investigate the power when raters have less intra-rater variability ($k_1 = k_2 = 40$). As we expect, the power increases as $k_1$ and $k_2$ go up. Given only 20 slides with two replicates, the power of detecting 5% percent change increases to close to 80%.

Table 4.6.
Power based on 100 simulated datasets from different slide populations

| Population[a] | Sample Size [b] | -0.1, 0.1 | -0.2, 0.2 | -0.3, 0.3 |
|---|---|---|---|---|
| Population 1<br>$\mu_p = (0.1, 0.4, 0.4, 0.1)$<br>$k_p = 10$ | $30 \times 1$ | 0.23, 0.21 | 0.50, 0.52 | 0.78, 0.81 |
| | $20 \times 2$ | 0.21, 0.20 | 0.53, 0.55 | 0.86, 0.84 |
| | $25 \times 2$ | 0.25, 0.27 | 0.70, 0.63 | 0.92, 0.90 |
| | $30 \times 2$ | 0.27, 0.29 | 0.75, 0.71 | 0.92, 0.95 |
| Population 2<br>$\mu_p = (0.25, 0.25, 0.25, 0.25)$<br>$k_p = 10$ | $30 \times 1$ | 0.21, 0.20 | 0.47, 0.43 | 0.77, 0.74 |
| | $20 \times 2$ | 0.20, 0.22 | 0.50, 0.52 | 0.84, 0.84 |
| | $25 \times 2$ | 0.22, 0.25 | 0.62, 0.60 | 0.87, 0.88 |
| | $30 \times 2$ | 0.24, 0.24 | 0.66, 0.66 | 0.94, 0.92 |
| Percent change in “weak” category | | ±2.5% | ±5% | ±7.4% |
| Population 3<br>$\mu_p = (0.6, 0.2, 0.1, 0.1)$<br>$k_p = 10$ | $30 \times 1$ | 0.15, 0.12 | 0.30, 0.31 | 0.37, 0.40 |
| | $20 \times 2$ | 0.16, 0.15 | 0.34, 0.36 | 0.59, 0.51 |
| | $25 \times 2$ | 0.15, 0.18 | 0.35, 0.38 | 0.60, 0.54 |
| | $30 \times 2$ | 0.21, 0.23 | 0.45, 0.48 | 0.70, 0.62 |
| Percent change in “weak” category | | ±1.6% | $(-3.4\%, +3.0\%)$ | $(-5.0\%, +4.4\%)$ |

[a] Intra-rater variability parameters $k_1 = k_2 = 20$

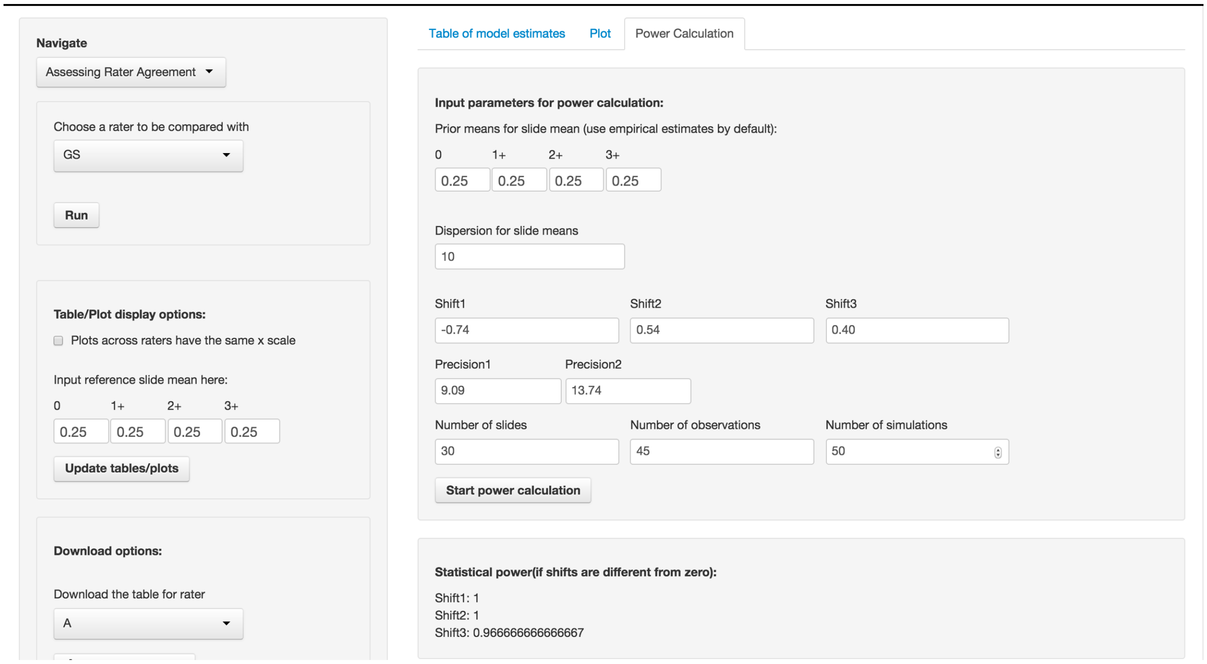
[b] Slides×Number of replicates

Table 4.7.
Power based on 100 simulated datasets from Population 3

| Intra-rater Variability | Sample Size | -0.1, 0.1 | -0.2, 0.2 | -0.3, 0.3 |
|---|---|---|---|---|
| $k_1 = k_2 = 20$ | $30 \times 1$ | 0.15, 0.12 | 0.30, 0.31 | 0.37, 0.40 |
| | $20 \times 2$ | 0.16, 0.15 | 0.34, 0.36 | 0.59, 0.51 |
| | $30 \times 2$ | 0.21, 0.23 | 0.45, 0.48 | 0.70, 0.62 |
| $k_1 = k_2 = 40$ | $30 \times 1$ | 0.18, 0.18 | 0.40, 0.39 | 0.57, 0.55 |
| | $20 \times 2$ | 0.18, 0.20 | 0.46, 0.41 | 0.78, 0.70 |
| | $30 \times 2$ | 0.20, 0.25 | 0.58, 0.45 | 0.87, 0.78 |

Next we demonstrate the power calculation for the real data set using our software. In Figure 4.11, the “Power Calculation” tab on the main panel provides the built-in power calculation function for users. Users can input the population parameters (i.e., prior means and dispersion for slide means), rater-specific parameters (i.e., shifts and precisions) as well as the sample size (i.e., number of slides and number of observations) to start the power calculation. By default, the software uses model estimates if users do not specify these options.

Since the procedures of power calculation are computationally intensive, we do not recommend trying a huge number of simulations using our software. The software uses $n = 50$ simulations to investigate the power for detecting the estimated shifts from the real IHC scoring data (Table 4.8). Overall, we obtain at least 75% power for the four different IHC populations we specify, except for Rater C’s shift 3 (=0.25).



Choose a rater to be compared with: GS

**Table/Plot display options:**

□ Plots across raters have the same x scale

Input reference slide mean here:

| 0 | 1+ | 2+ | 3+ |
|---|---|---|---|
| 0.25 | 0.25 | 0.25 | 0.25 |

Download the table for rater: A

**Input parameters for power calculation:**

Prior means for slide mean (use empirical estimates by default):

| 0 | 1+ | 2+ | 3+ |
|---|---|---|---|
| 0.25 | 0.25 | 0.25 | 0.25 |

Dispersion for slide means: 10

Shift1: -0.74
Shift2: 0.54
Shift3: 0.40

Precision1: 9.09
Precision2: 13.74

Number of slides: 30
Number of observations: 45
Number of simulations: 50

**Statistical power(if shifts are different from zero):**

Shift1: 1
Shift2: 1
Shift3: 0.966666666666667

Fig. 4.11. Screenshot: power calculation

Table 4.8.
Power based on 100 simulations

| $\boldsymbol{\mu}_p$ | Rater A<br>(-0.74, 0.54, 0.40)[a] | Rater B<br>(-0.82, -0.96, -0.90)[a] | Rater C<br>(0.49, 0.54, 0.25)[a] |
|---|---|---|---|
| (0.25, 0.28, 0.29, 0.18)[b] | (1, 0.95, 0.95) | (1, 1, 1) | (0.85, 0.85, 0.65) |
| (0.25, 0.25, 0.25, 0.25) | (1, 1, 0.97) | (1, 1, 1) | (1, 1, 0.80) |
| (0.10, 0.40, 0.40, 0.10) | (1, 1, 0.80) | (1, 1, 1) | (0.75, 1, 0.60) |
| (0.60, 0.20, 0.10, 0.10) | (1, 0.95, 1) | (1, 1, 1) | (1, 0.95, 0.95) |

[a] The estimated shifts for three raters

[b] Estimated population mean using the average of Rater GS

## 4.5 Summary

This chapter discusses the inter-rater agreement application using our proposed Bayesian method on a real IHC scoring data. Our aim here is to provide a user-friendly software that allows researchers and raters to implement our method as well as compare this new agreement index to some traditional agreement indices. The software we introduced and described in this chapter is designed to assess inter-rater agreement for decile IHC scores given multiple raters while one rater is chosen as the reference rater. It provides a summary of the data set uploaded for agreement analysis, Bayesian model estimates (i.e., posterior mean estimates and associated credible intervals) displayed in both tables and plots, and power calculation function that allow users to not only calculate the power for the current study but also the sample size for a future design.

# CHAPTER 5. SUMMARY

Compositional data are frequently encountered in a variety of research areas. This dissertation has focused on one area of inference with these data, inter-rater agreement analysis. This is an area that has not received much attention. This dissertation begins with literature reviews of compositional data analysis (Chapter 1) and methods of assessing inter-rater agreement (Chapter 2). These two chapters build the fundamentals for our proposed methodology in Chapter 3. Our novel Bayesian approach to assess inter-rater agreement is the first using the compositional vectors directly. Extensions to the approach allow applicability to different scenarios including discrete compositional scores, multiple raters, and repeated scores on partial or full slides. Based on the methodological development and simulation results in Chapter 3, we introduce a user-friendly software in Chapter 4 as a tool to implement rater agreement analysis. A real IHC data set was used to illustrate the analysis procedures and power/sample size calculations using our software.

The primary contribution of this work is towards assessment of inter-rater agreement for compositional data. There has been extensive research on methods for compositional data (e.g., Aitchison's book on compositional data analysis) and for inter-rater agreement measures. However, there has been little work connecting these two. In contrast to the other work on agreement using compositional data, our approach focuses on assessing agreement in the simplex space. Shift parameters are introduced to describe the pattern of differences between raters. We also include an agreement index using the Bhattacharyya Coefficient. This index is conditional in the sense that its value is based on particular slide mean. In many ways it is similar to the infrequently used CIV/CIA index, which assess the between-rater variability relative to the rater-related variability (between-rater + within-rater variability). We recommend using this approach instead of univariate agreement measures because often small shifts cannot be picked up by the H-score conversion. In some cases where there are big shifts between low intensity categories, the H-score distribution still won't change much because of the small weights assigned to the percents of low intensity categories when converted to the H-score. Moreover, numerous compositional vectors can be converted to the same H-score given different combinations of shifts. Simulation examples have been used to illustrate the insensitivity of different univariate agreement indices in Section 4.3.3.

Motivated by the IHC scoring example described in the beginning of Chapter 3, our work is very important to medicinal and pathological research. On one hand, when we don't have a reference rater (or method), our rater agreement index provides a guide for the consistency of raters. If there exists some inconsistency, this should prompt an investigation into where the raters are differing. Our shift parameters provide such information and could be used to help in the training of new raters. On the other hand, when we have a reference rater, our method assesses how the other raters differ systematically from the reference and thus provides the way of adjusting the other raters' scores. Last but not the least, our model can be used to provide us with the guidance of efficient experimental design for similar studies.

Even though we described the application of our work mainly in agreement analysis in IHC scoring, the application can be extended to more broader areas, such as geology, petrology, and economics, where compositional data appear very often. For example, when the interest is in assessing the agreement between ecologists when they sample plant species for the same set of sites, or the agreement between individuals' consumption behaviors across various commodity categories, our approach can be applied directly. Another example can be when they are interested in how the plant species compositions change over time/seasons within certain sites. In this case, the raters can be the two seasons.

Since this research area is new, much future work remains. Currently our approach lacks any assessment of model fit because of the limited information provided in these agreement studies. In order to assess agreement, we make some relatively strong assumptions about the distribution of scores for a given slide and the pattern of differences. When possible, we would like to assess if this assumption we impose actually fits our data. We can do this assessment only when we have replicates as they allow us to consider more general models. For example, we can allow the shifts to vary slide to slide, and do Bayesian model comparisons. In a frequentist setting, we can estimate the means of slides separately by raters and construct a likelihood ratio test to assess the proportional odds assumption. However, these all require bigger studies and more replicates from raters, and we simply don't see these data sets in practice yet. There are other potential directions we could consider to model the data. For example, if we consider standard Normal distribution as the slide distribution, then we don't have the proportional odds assumption based on the logistic distribution. Another direction regarding the shift means idea is the fixed cutpoints conceptual framework, as we briefly described on page 61-62. Again, we then need to have the appropriate experimental design and more data to validate these different models.

In the Bayesian paradigm, one often uses posterior predictive assessment to examine the model fit. However, there is no convenient distance measure to assess the deviance between the observed and predicted compositional vectors. Suppose we have a legitimate distance metric to measure the distance between two discrete compositional vectors. One way to assess our model fit would be: simulate numerous observations based on the estimated reference means, shift and intra-rater variability parameters for each rater and each slide. Calculate the distances between each simulated pairs of raters and the reference rater, e.g., distances between simulated Rater A and simulated Rater GS, simulated Rater B and simulated Rater GS, etc, for each slide. Plot the simulated distance distributions versus the observed distance. If we observe the distance distribution is very compact and has the peak at the observed distance, it provides some evidence of a good model fit to our study purpose. Figure 5.1 illustrates the model fit idea using the real IHC data set (Slide 1) and the squared Euclidean metric is used to assess the distance of two compositional vectors.

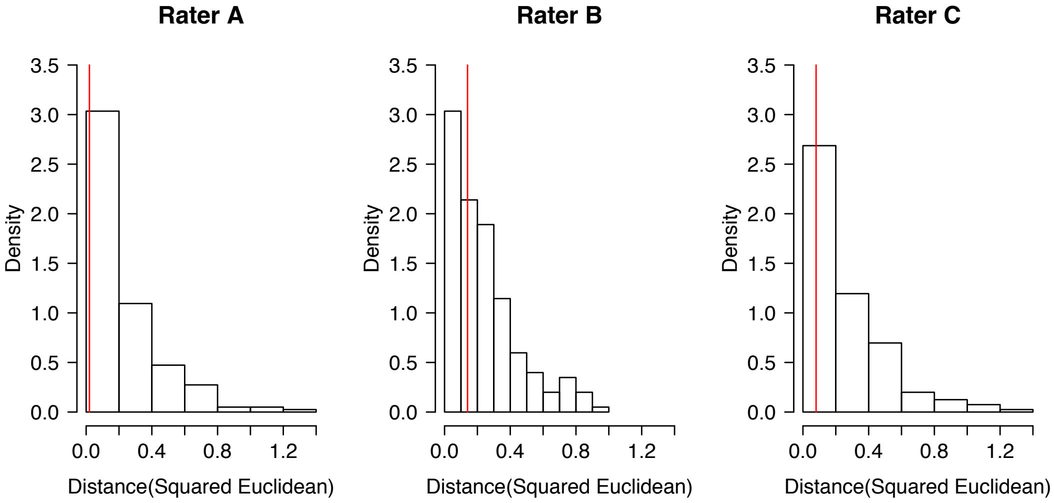


Fig. 5.1. Assessing model fit

In summary, we pioneer a methodology using Bayesian estimation to address the inter-rater agreement assessment for compositional data, especially given limited sample size and replication. This dissertation provides a comprehensive and systematic framework including conceptual and theoretical models, analysis procedure, and a easy-to-use online software for data analysis and experimental design. However, this is not the end. Some more questions have been and will be brought up and future work are needed to help answer these new questions and further expand this research.

# REFERENCES

# REFERENCES

[1] W. James and Charles Stein. Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, 1:361–379, 1961.

[2] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symp. on Math. Statist. and Prob.*, 1:197–206, 1956.

[3] Larry Wasserman. *All of statistics: a concise course in statistical inference.* Springer Texts in Statistics. Springer New York, 2004.

[4] Huiman X. Barnhart and Andrzej S. Kosinski. Assessing individual agreement. *Journal of Biopharmaceutical Statistics*, 17:697–719, 2007.

[5] Huiman X. Barnhart, Jingli Song, and Michael J. Haber. Assessing intra, inter and total agreement with replicated readings. *Statistics in Medicine*, 24:1371–1384, 2005.

[6] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[7] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, and Augusta H. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[8] Michelle Nisolle, Stéphane Gillerot, Françoise Casanas-Roux, Jean Squifflet, Martine Berliere, and Jacques Donnez. Immunohistochemical study of the proliferation index, oestrogen receptors and progesterone receptors a and b in leiomyomata and normal myometrium during the menstrual cycle and under gonadotrophin-releasing hormone agonist therapy. *Human Reproduction*, 14:2844–2850, 1999.

[9] Melina B Flanagan, David J Dabbs, Adam M Brufsky2, Sushil Beriwal, and Rohit Bhargava. Histopathologic variables predict oncotype dx$^{\mathrm{TM}}$ recurrence score. *Modern Pathology*, 21:1255–1261, 2008.

[10] Rohit Bhargava, Joan Striebel, Sushil Beriwal, John C. Flickinger, Agnieszka Onisko, Gretchen Ahrendt, and David J. Dabbs. Prevalence, morphologic features and proliferation indices of breast carcinoma molecular classes using immunohistochemical surrogate markers. *International Journal of Clinical and Experimental Pathology*, 2:444–455, 2009.

[11] K. S. McCarty, L. S. Miller, E.B Cox, and J. Konrath. Estrogen analyses: correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies. *Archives of Phthology and Laboraroty Medicine*, 109:716–721, 1985.

[12] T.W. Rauber, T. Braun, and K. Berns. Probabilistic distance measures of the dirichlet and beta distributions. *Pattern Recognition*, 41:637–645, February 2008.

[13] J. S. Uebersax and W. M. Grove. Latent class analysis of diagnostic agreement. *Statistics in Medicine*, 9:559–572, 1990.

[14] M. Aickin. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to cohen's kappa. *Biometrics*, 46:293–302, 1990.

[15] L. A. Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74:537–552, 1979.

[16] G. Molenberghs, T. Vangeneugden, and A. Laenen. Estimating reliability and generalizability from hierarchical biomedical data. *Journal of Biopharmaceutical Statistics*, 17:595–627, 2007.

[17] T. Vangeneugden, A. Laenen, H. Geys, D. Renard, and G. Molenberghs. Applying linear mixed models to estimate reliability in clinical trials with repeated measurements. *Controlled Clinical Trials*, 25:13–30, 2004.

[18] Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1:30–46, 1996.

[19] D.A. Bloch and H.C. Kraemer. 2x2 kappa coefficients: Measures of agreement or association. *Biometrics*, 45:269–287, 1989.

[20] T. Byrt, J. Bishop, and J. B. Carlin. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46:423–429, 1993.

[21] D. V. Cicchetti and A. R. Feinstein. High agreement but low kappa: Ii. resolving the paradoxes. *Journal of Clinical Epidemiology*, 6:551–558, 1990.

[22] A. R. Feinstein and D. V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 6:543–549, 1990.

[23] John S. Uebersax. Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101:140–146, January 1987.

[24] A. Agresti. Modeling patterns of agreement and disagreement. *Statistics in Medicine*, 1:201–218, 1992.

[25] A. Agresti. A model for agreementbetween ratingson an ordinal scale. *Biometrics*, 44:539–548, 1988.

[26] P. Graham. Modelling covariate effects in observer agreement studies: The case of nominal scale agreement. *Statistics in Medicine*, 14:299–310, 1995.

[27] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

[28] J Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220, 1968.

[29] J Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[30] Martin A. Tanner and Michael A. Young. Modeling ordinal scale agreement. *Psychological Bulletin*, 98:408–415, 1985.

[31] Martin A. Tanner and Michael A. Young. Modeling agreement among raters. *Journal of the American Statistical Association*, 80:175–180, 1985.

[32] W. A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19:321–325, 1955.

[33] Brian Smith and William S. Rayens. Conditional generalized liouville distributions on the simplex. *Journal of Theoretical and Applied Statistics*, 36:185–194, 2002.

[34] William S. Rayens and Cidambi Srinivasan. Dependence properties of generalized liouville distributions on the simplex. *Journal of the American Statistical Association*, 89:1465–1470, 1994.

[35] Rafiq H. Hijazi. An em-algorithm based method to deal with rounded zeros in compositional data unde dirichlet models. *Proceedings of the 4th International Workshop on Compositional Data Analysis*, 2011.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[37] R.F. Sandford, C.T. Pierson, and R.A. Crovelli. An objective replacement method for censored geochemical data. *Mathematical Geology*, 25(1):59–80, 1993.

[38] A.P. Dawid. Properties of diagnostic data distributions. *Biometrics*, 32:647–658, 1976.

[39] J.A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

[40] C. Philip Cox and Thomasd Roseberry. A large sample sequential test, using concomitant information, for discrimination between two composite hypotheses. *Journal of the American Statistical Association*, 61:357–367, 1966.

[41] K. Peason. 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*, pages 489–502, 1897.

[42] Felix Chayes. *Ratio correlation; a manual for students of petrology and geochemistry.* University of Chicago Press, 1971.

[43] Jane M. Fry, Tim R. L. Fry, Keith R. Mclaren, and Tanya N. Smith. Modeling zeroes in microdata. *Applied Economics*, 33:383–392, 2001.

[44] Jane M. Fry, Tim R. L. Fry, and Keith R. Mclaren. Compositional data analysis and zeros in micro data. *Applied Economics*, 32:953–959, 2000.

[45] C.W. Thomas and John Aitchison. Log-ratios and geochemical discrimination of scottish dalradian limestones: a case study. *Geological Society Special Publication*, 264:25–41, 2006.

[46] Michael J. Haber, Huiman X. Barnhart, Jingli Song, and James Gruden. Observer variability: A new approach in evaluating interobserver agreement. *Journal of Data Science*, 3:69–83, 2005.

[47] Huiman X. Barnhart, Michael J. Haber, and Lawrence I. Lin. An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics*, 17:529–569, 2007.

[48] Ruth Etzioni, Sarah Hawley, Dean Billheimer, Lawrence D. True, and Beatrice Knudsen. Analyzing patterns of staining in immunohistochemical studies: Application to a study of prostate cancer recurrence. *Cancer Epidemiol Biomarkers Prevention*, 14:1040–1046, May 2005.

[49] Alan Agresti and Joseph B. Lang. Quasi-symmetric latent class models, with application to rater agreement. *Biometrics*, 49(1):131–139, March 1993.

[50] Nicholas J. Aebischer, Peter A. Robertson, and Robert E. Kenward. Compositional anaysis of habitat use from animal radio-tracking data. *Ecology*, 74:1313–1325, July 1993.

[51] Harald Janson and Ulf Olsson. A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, 61(2):277–289, April 2001.

[52] Andrea Ongaro, Sonia Migliorati, and Gia Serafina Monti. A new distribution on the simplex containing the dirichlet family. *Conference Proceedings*, 2008.

[53] Bacon-Shone John. Modelling structural zeros in compositional data. *n Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*, 2003.

[54] John Aitchison and Jim W. Kay. Possible solution of some essential zero problems in compositional data analysis. *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*, 2003.

[55] J. F. Toucher. Pigmentation survey of school children in scotland. *Biometrika*, 5:129–235, 1908.

[56] R.N. Thompson, J. Esson, and A. C. Dunham. Major element chemical variation in the eocene laves of the isle of skye, scotland. *Journal of Petrology*, 13:219–253, 1972.

[57] H.R. Rollinson. Another look at the constant sum problem in geochemistry. *Mineralogical Magazine*, 56:469–475, 1992.

[58] N.M.S Rock. Numerical geology. *Lecture Notes in Earth Sciences*, 1988.

[59] John Aitchison. The statistical analysis of geochemical compositions. *Mathematical Geology*, 16(6), 1984.

[60] Vernon M. Chinchilli, Juliann K. Martel, Shiriki Kumanyika, and Tom Lloyd. A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, 52:341–353, March 1996.

[61] Chia-Cheng Chen and Huiman X. Barnhart. Comparison of icc and ccc for assessing agreement for data without and with replications. *Computational Statistics and Data Analysis*, 53:554–564, September 2008.

[62] Tony Vangeneugden, Annouschka Laenen, Helena Geys, Didier Renard, and Geert Molenberghs. Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, 61:295–304, March 2005.

[63] Huiman X. Barnhart and John M. Willianmson. Modeling concordance correlation via gee to evaluate reproducibility. *Biometrics*, 57:931–940, September 2001.

[64] Tonya S. King, Vernon M. Chinchilli, Kai-Ling Wang, and Josep L. Carrasco. A class of repeated measures concordance correlation coefficients. *Journal of Biopharmaceutical Statistics*, 17:653–673, 2007.

[65] Tonya S. King, Vernon M. Chinchilli, and Josep L. Carrasco. A repeated measures concordance correlation coefficient. *Statistics in Medicine*, 26:3095–3113, July 2007.

[66] Huiman X. Barnhart, Michael Haber, and Jingli Song. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, 58:1020–1027, December 2002.

[67] Muller Reinhold and Buttner Petra. A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, 13:2465–2476, 1994.

[68] Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45:255–268, March 1989.

[69] Patrick E. Shrout and Joseph L. Fleiss. Intraclss correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, March 1979.

[70] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for cateforical data. *Biometrics*, 33:159–174, March 1977.

[71] William Barlow, Mei-Ying Lai, and Stanley P. Azen. A comparison of methods for calculating a stratified kappa. *Statistics in Medicine*, 10:1465–1472, 1991.

[72] Joseph L. Fleiss, Jacob Cohen, and B.S. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72:323–327, November 1969.

[73] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications. *The American Statistical Association*, 49(268):732–764, December 1954.

[74] Josep Daunis i Estadella, Josep Antoni Martín-Fernández, and Javier Palarea-Albaladejo. Bayesian tools for zero counts in compositional data. *Conference Proceeding*, 2008.

[75] Javier Palarea-Albaladejo, Josep A. Martín-Fernández, and Juan Gómez-García. A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39:625–645, February 2007.

[76] J. A. Martin-Fernandez, C. Barcelo-Vidal, and V. Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3), April 2003.

[77] Peter Flizmoser, Karel Hron, and Clemens Reimann. Principal component analysis for compositional data with outliers. *Environmetrics*, 20:621–632, September 2009.

[78] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), April 2003.

[79] John Berhm, Scott Gates, and Brad Gomez. A monte carlo comparison of methods for compositional data analysis. *The 1998 annual meeting of the Society for Political Methodology*, July 1998.

[80] Robert J. Connor and James E. Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, March 1969.

[81] P Filzmoser, K Hron, and M Templ. Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics*, 27(4):585–604, 2012.

[82] Jane M. Fry, Tim R. L. Fry, and Keith R. McLaren. The stochastic specification of demand share equations: Restricting budget shares to the unit simplex. *Journal of Econometrics*, 73(2):377–385, 1996.

[83] A. Woodland. Stochastic specification and the estimation of share equations. *Journal of Econometrics*, 10(1):361–383, 1979.

[84] G. Ronning. Share equations in econometrics: A story of repression, frustration and dead ends. *Statistical Papers*, 33(1):307–334, 1992.

[85] Vera Pawlowsky-Glahn and Antonella Buccianti. *Compositional Data Analysis: Theory and Applications*. John Wiley and Sons, Ltd, 2011.

[86] Lajos Ó. Kovács, Gabor P. Kovács, Josep Antoni Martín-Fernández, and Carles Barceló-Vidal. Major-oxide compositional discrimination in cenozoic volcanites of hungary. *Geological Society Special Publication*, 264:11–23, 2006.

[87] Paul F. Carr. Distinction between permian and post-permian igneous rocks in the southern sydney basin, new south wales, on the basis of major-element geochemistry. *Journal of the International Association for Mathematical Geology*, 13(3):193–200, 1981.

[88] John Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall, 29 West 35th Street, New York, NY 10001, 1986.

[89] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982.

# VITA

# VITA

Ningning Chen was born in Hunan, China. She received a Bachelor of Economics with a double-major in Statistics and Finance from Guangdong University of Foreign Studies in June 2011. In August 2011, she entered the PhD program of Statistics at Purdue University, West Lafayette. She received a Doctor of Philosophy in December 2015. She was hired by Apple Inc. as a Statistician/Algorithm Analyst in Cupertino, CA after graduating from Purdue.