CHAPTER 3

Intraclass Correlation: A Measure of Rater Agreement

OBJECTIVE
This chapter presents a general overview of the use of Intraclass Correlation Coefficients for quantifying the extent of agreement among raters when the ratings are in the form of quantitative measurements. A high-level description of the underlying statistical models is provided as well as a discussion on the limitations associated with their use. After reading this chapter the practitioner will be able to decide which model is appropriate for the study that was conducted, and will know the related challenges that must be overcome. This chapter also describes the Bland-Altman plot, a popular graphical method for analyzing agreement between two raters. The reader will find an introduction to sample size calculations in this chapter, and a more detailed treatment of the sample size problem in subsequent chapters. Figure 3.2 represents a flowchart showing how to find the correct intraclass correlation coefficients based on the way the ratings were gathered and the type of analysis to be done.

Contents

3.1 Introduction

In the past few chapters of parts I and II, I presented many techniques for quantifying the extent of agreement among raters. Although some of these techniques were extended to interval and ratio data, the primary focus has been on nominal and ordinal data. This chapter as well as the other chapters of Part II, are devoted to the study of inter-rater reliability for quantitative outcomes whose possible values are defined on a continuum, as opposed to being a predetermined set of specific values.

Why do we need to care about intraclass correlation when weighted versions of the chance-corrected measures can be used to handle quantitative outcomes? It is because the notion of "perfect" agreement associated with two raters assigning the exact same score to the same subject, does not translate well to quantitative measurements. Consider for example two electronic devices used to measure the knee joint laxity on 15 human subjects. Even if both devices are equally reliable, you would not expect them to produce the exact same quantitative measurement on the same subjects, since these values belong to a continuum. Likewise, two very competent raters that measure the height or the weight of the same human subject will likely produce slightly different numbers regardless of their proficiency level in the use of the measuring instrument. With agreement no longer referring to an exact match, the notions of chance agreement and percent agreement evaporate.

The solution to this problem is to use the portion of variation in the data that is due to subjects, and to compare it to the other portion of that variation due to the raters. If the rater-induced variation exceeds that of the subjects by a wide margin then the raters are said to have low inter-rater reliability. Otherwise, the raters are said to have high inter-rater reliability. This comparison of variance components is achieved by calculating the ratio of the subject variance component to total variance[1], which is known as the Intraclass Correlation Coefficient (ICC). This approach to inter-rater reliability will work only if the reliability experiment is designed in such a way that the different variation components can be separated. You will see in the next few sections how this task can be accomplished. Several approaches can be used to design an inter-rater reliability study, depending on the goal aimed at for the study. In the next section, I will describe a few designs commonly used in the context of inter-rater reliability analysis.

The ICC approach to inter-rater reliability is based on a sound statistical theory. Nevertheless, the statistical foundation of this method requires the subjects taking

[1]Total variance includes the variance components due to raters and that due to experimental errors. Consequently, large experimental errors or large variations among raters will inevitably lead to low inter-rater reliability. Therefore, minimizing measurement or experimental errors is crucial to a successful inter-rater reliability experiment.

part of the experiment to have been randomly selected from a larger and heterogeneous population of subjects. At times researchers want to quantify inter-rater reliability when the subject population is homogeneous. With a small subject variance component to begin with, even a small to moderate rater variance component will lead to a small ICC, making it impossible to obtain a high ICC even when the raters agree. To deal with this issue, Finn (1970) proposed a method known by researchers as the within-group inter-rater reliability, and which is discussed in chapter 7.

In most inter-rater reliability projects involving quantitative measurements, the extent of agreement among raters is evaluated with respect to a single characteristic of interest. For example, the characteristic could be a person's peak expiratory flow indicating his ability to breathe out air. The inter-rater reliability experiment in this case consists of getting several raters to assign a single measurement to each human subject. However, the field of medical imaging for example requires that the same subject be measured on multiple characteristics. Likewise, the performance of a computer system will likely be evaluated on many aspects. The use of multiple characteristics often requires a multivariate approach to inter-rater reliability in order to provide a global measure of the extent to which the raters agree on all characteristics of interest.

To address the multivariate problem describe in the previous paragraph, various authors have proposed multivariate versions of the classical ICC. Shou et al. (2013) proposed the I2C2 coefficient, while Yue et al. (2015) recommended a more generalized version of I2C2 known as GICC. These 2 proposals present the advantage of producing a single global coefficient that adequately summarizes the extent of agreement among raters in a multivariate setting. The main disadvantage of these methods is their complexity, and the lack of software packages that implement them. I propose a simpler multivariate approach to the ICC in section 3.5 that can be implemented using most standard software packages.

3.2 Statistical Models

Consider the reliability data shown in Table 3.1. That data represents scores that 4 raters assigned to 6 subjects, and could be interpreted in various ways depending on how it was collected. Here are 4 possible study designs (or data models) that could have produced Table 1 data:

- **Model 1A**: *Each subject is rated by a different group of raters*

  According to this model, each row of Table 3.1 is not necessarily associated with the same set of 4 raters. Although the 4 raters are consistently labeled as

1, 2, 3, and 4, they could represent different individuals, or different measuring instruments. One may average this data row-wise to study the subject effect, but will not be able to average column-wise to obtain the rater effect. It is why this is often known as a one-factor (or one-way) model, the single factor here being the subject.

The main implication of this model is that one rater may not have the opportunity to score more than one subject. Consequently, this model makes it impossible to evaluate *Intra-rater Reliability*, which is a measure of the rater's self-consistency. However, the raters under this model still score the same subjects, making it possible to compute *Inter-rater reliability*.

The main advantage for using this model is that the raters could be located in different geographic areas, and rate local subjects. There is no need to move subjects around to allow different groups of raters to rate the same subjects. This model may also be suitable in situations where subjects are hard to recruit and the availability of the same group of raters cannot be guaranteed when a subject is able to participate in the experiment.

Table 3.1: Scores assigned by 4 raters to 6 subjects

Subject	Rater[a]				Average
	1	2	3	4	
1	9	2	5	8	6
2	6	1	3	2	3
3	8	4	6	8	6.5
4	7	1	2	6	4
5	10	5	6	9	7.5
6	6	2	4	7	4.75
Average	7.67	2.5	4.33	6.67	5.29

[a]This data is taken from Shrout and Fleiss (1979), although I replaced the terms Target and Judge with Subject and Rater respectively, and added row and column marginal averages.

► **Model 1B**: *Each rater rates a different group of subjects*

If Table 3.1 data were collected according to this design, then the 6 subjects may differ from rater to rater. That is, each rater scored his own set of subjects, even though I may have decided to consistently labeled them as 1, 2, 3, 4, 5, and 6. One may evaluate the rater effect by averaging Table 3.1's columns. Any row-wise averaging would be meaningless as such an operation would involve different subjects as well as different raters. Therefore, the only factor that can be studied is the rater factor, and this model will later be referred to as a one-factor or one-way model.

The main implication of this model is that it allows for the evaluation of intra-rater reliability, and not that of inter-rater reliability. Evaluating inter-rater reliability always requires different raters to score the same subjects.

► **Model 2**: *The Random Factorial Design*

According to this model, each subject is scored by the same group of raters. Both the subjects and the raters are random samples selected from the respective populations they represent, hence the naming "random" design. Moreover, the column and row marginal averages are meaningful, and the effects of subject and rater factors can be evaluated. It is because both factors (rater and subject) can be studied that this design is known as a "factorial design". The experimental design that produces Table 3.1 data is called a two-way factorial design.

► **Model 3**: *The Mixed Factorial Design*

According to this design, each subject is scored by the same group of raters, and is also in this regard a factorial design. Unlike Model 2, here only the group of subjects represents a random sample selected from a larger subject population, while the group of raters does not represent a random sample. Because the group of raters that participate in the reliability experiment is not randomly selected from a larger rater population, these raters only represent themselves. The resulting inter-rater reliability coefficient can therefore not be applied to raters beyond those in the experiment. Therefore, the subject effect is random, while the rater effect is fixed. This combination of random and fixed effects gave this design the name "Mixed Factorial Design." When the number of factors considered is limited to two as is the case for Table 3.1, it is renamed the "Two-Way Mixed Factorial Design."

Each of these models requires a different method for calculating the intraclass correlation coefficient. Shrout and Fleiss (1979) discussed models 1A (although it was referred to as model 1), 2, and 3. The same models were also discussed by McGraw and Wong (1996), who presented methods for computing the intraclass correlation for each of them. However, these authors did not deal with the important problem of missing ratings, which is very common in inter-rater reliability experiments. The missing-rating issue extensively will be extensively discussed in chapters 4, 5 and 6.

When all the conditions required to ensure the validity of ANOVA models are not met, the nonparametric methods of chapter 7 should be used. These nonparametric methods were advocated by several authors including Finn (1970), James et al. (1984) and O'Neill (2017)

3.3 The Bland-Altman Plot

A mainly graphical method often used as alternative to the intraclass correlation for analyzing inter-rater reliability data was proposed by Bland and Altman (1986). It combines a graphical approach and a quantitative analysis of the magnitude of the rating differences. This method can only analyze two raters at a time, and has become popular over time among researchers, although many of its users are often unaware of its limitations. In this section, I will present an overview of this method, and will discuss its merits as well as its limitations.

Suppose that we want to study the extent of agreement between the two raters labeled as 3 and 4 using Table 3.1's ratings. The Bland-Altman method is implemented as follows:

- The first step consists of creating a scatterplot that depicts the differences in ratings between raters 4 and 3 as a function of their averages. Table 3.2 shows the ratings being analyzed as well as the two series of averages and differences used to create the scatterplot of Figure 3.1.

- The next step is to display on the scatterplot created in the previous step, the two "limits of agreement". The dotted line at the bottom is the lower limit of agreement and the one at the top represents the upper limit of agreement. The lower limit of agreement is -1.169 while the upper limit of agreement is 5.836. This indicates that you can expect the difference between raters 4 and 3 to be as high as 3.763 and as low as 0.904. Depending on the application at hand, such a gap may be acceptable or may be too wide. Ultimately, this gap will help the researcher decide whether the extent of agreement between the two raters 4 and 3 is acceptable or not. If $\overline{d}$ is the average difference and s the standard deviation of the differences, then the lower limit of agreement is $\overline{d} - 2s/n$ and the upper limit of agreement $\overline{d} + 2s/n$.

The two steps described above summarize what is known as the Bland-Altman method. It is intuitive and fairly straightforward to apply. Bland and Altman (1986) indicated that their plot can help study the relationship between the rating pairwise differences and the associated pairwise means, which by the way are used as surrogates for the true rating associated with the subject. The study of this relationship is one way of verifying whether the differences are independent or not. These differences must be approximately independent for the interpretation of the lower and upper limits of agreement to be valid. If these differences have for example a tendency to decrease as the averages increase, or if this relationship shows any other specific trend, this may an indication of a lack of independence. Transforming the initial ratings using the logarithm function for example may be the remedy for obtaining the independence needed.

Some researchers believe that the Bland-Altman method is the only realistic way of dealing with inter-rater agreement. That is not true. We will see in the next few chapters why the intraclass correlation is not only appropriate, but is often the better approach.

Table 3.2: Scores assigned by Raters 3 & 4 to 6 subjects

Subject	Rater #3	Rater #4	Mean Rating	Difference[a]
1	5	8	6.5	3
2	3	2	2.5	-1
3	6	8	7	2
4	2	6	4	4
5	6	9	7.5	3
6	4	7	5.5	3

[a]Difference = (Rater 4) - (Rater 3)

ISSUES WITH THE BLAND-ALTMAN METHOD

Part of the popularity of the Bland-Altman method stems from its graphical nature. You can look at the graph and see right in front of you the differences between the ratings obtained from the two raters you are analyzing. A simple visual exploration may even allow you to form an opinion about the extent to which they agree. Using the two limits of agreement helps you figure out how large the difference should be before it can be considered too large. Here are a few assumptions the Bland-Altman method is based upon, and which are often not satisfied:

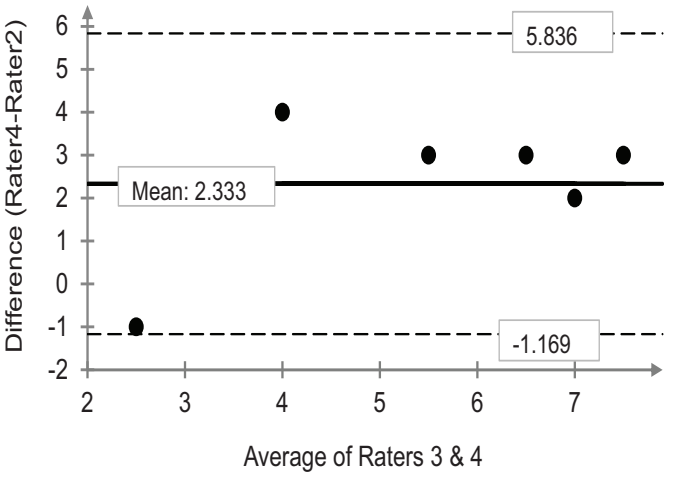

Figure 3.1: Rating Differences as a function of rating means

- Bland and Altman (1986, page 4), indicate that the "... *differences are likely to follow a Normal distribution because we have removed a lot of the variation between subjects and are left with the measurement error.*" The real problem with this assumption is that it is untrue if there is a subject-rater interaction. This is often the case when the rating is affected by the magnitude of the "true" score associated with the subjects. The subject-rater interaction does not preclude the differences from following the Normal distribution. However, the differences will be correlated and their actual standard deviation would be higher than the estimate s recommended by Bland and Altman (1986).

  If the standard deviation of the differences is underestimated then the Bland-Altman method may produce a false sense of agreement. When subjects and raters interact, inter-rater reliability is better analyzed with the intraclass correlation that relies on a formal modeling of the interaction effect.

- Another benefit of the Bland-Altman plot lies in the analysis of the the relationship between the differences and the average ratings. This relationship is important primarily because it allows you to see whether raters and subjects interact provided the average is a good surrogate for the subject's true score. The problem here is that the average is known to be close to the true value only if there is little variation in the ratings. That is if the raters are known to be in agreement, an assumption we cannot make since that very agreement is precisely what we are studying.

- The Bland-Altman method is meant for pairwise analyses only. It may not allow you to obtain a global picture of the extent of agreement among multiple

raters. When the number of raters is moderately large such as 8, the number of pairwise analyses becomes as large as 28, which can be problematic.

I would recommend using the Bland-Altman plot mainly as an exploratory technique. It allows the researcher to have a first glimpse into the inter-rater reliability results. Ultimately, an intraclass correlation based on the appropriate statistical model should be calculated.

3.4 Sample Size Calculations

When designing an inter-rater (or intra-rater) reliability study, the researcher first needs to determine how many subjects and how many raters must be part of the experiment. Sometimes, there is also a need to determine the number of trials also known as the number of replicates if replication is desired. Note that replication is about a rater taking more than one measurement from the same subject. Each of the next three chapters has a section on sample size calculation. These sections provide detailed procedures showing how the number of raters, the number of subjects and the number of trials can be determined depending on which data model is chosen.

Traditionally power calculation as done in statistics is based on the test of hypothesis involving population means, and consists of finding the optimal sample size that yields the desired power[2] for the statistical test. This procedure generally requires the researcher to specify the effect size (or the detectable difference)[3], the statistical significance (also known as α or alpha), and the desired power. The approach proposed here for the ICC is slightly different. It requires the researcher to specify the desired confidence interval length (this is equivalent to specifying the effect size), the confidence level associated with the confidence interval (this often takes the values 90%, 95%, or 99%), and the anticipated ICC value. The anticipated ICC value may be known from prior studies or from a pilot experiment. If such a value is unknown then I will recommend a conservative approach based on the anticipated ICC value that will yield the largest confidence interval length.

Our investigation has revealed that you need about 5 raters to optimize your inter-rater reliability coefficient for a given total number of ratings. The total number of ratings is the product of the number of raters by the number of subjects (assuming one trial per rater and per subject). Therefore, if your experiment is going to generate 140 ratings for example, then it would be more efficient to have 5 raters

[2]The power of a statistical test represents the probability for that test to reject the "null" hypothesis when it is false. This "null" hypothesis could be the equality of two population means, or the equality of a population mean to a hypothetical value.

[3]The detectable difference is the smallest difference between the two population means under comparison, which will cause the null hypothesis to be rejected.

and 28 subjects instead of having 10 raters and 14 subjects. A design is said to be more efficient in this context when it yields the smaller confidence interval length. Consequently increasing the number of subjects is more rewarding than increasing the number of raters beyond 5. However, if recruiting raters is cheaper than recruiting subjects then you may have to increase the number of raters beyond 5 and reduce the number of subjects.

3.5 Multivariate Analysis

The multivariate ICC proposed in the literature is based essentially on a more general version of what I previously referred to as Model 1A in section 3.2. These broad and complex statistical models attempt to describe the relationship between a series of tables such as Table 3.1 associated with the different characteristics being studied. These methods can be used if the researcher is willing to put in the efforts necessary to successfully implement them. A downside of these methods other than their complexity and the unavailability of software packages, is the difficulty extending them to the other alternative statistical models that may be more appropriate for our data. What else can we do?

Table 3.3 displays 4 measurements generated for each of 10 patients by a software that analyzed 4 of the main coronary vessel areas. It is often of interest to compare different software products and to assess the extent to which they agree on these measurements. Another dataset similar to Table 3.3 could be produced for a second software product so that an inter-rater reliability can be calculated between these 2 software products. A key question to be asked is how can one produce a single global intraclass correlation coefficient that summarizes the extent of agrement between 2 software products with respect to all 10 patients and 4 characteristics? There 3 natural options you can think of:

(i) You can compute a separate ICC for each of the 4 variables GLOBAL, LAD, LCx, and RCA before averaging them to obtain a single general agreement coefficient. However, this average is not and cannot be interpreted in anyway as an intraclass correlation coefficient. Note that the ICC is the proportion of total variance in the data that is due the subjects. Since the ICC is obtained as the ratio of subject variance to total variance, high-variance characteristics are expected to play a dominant role in this assessment. Therefore, averaging 4 ICC coefficients will put too much emphasis on low-variance characteristics by giving them the same weight as the high-variance characteristics. Consequently, the mean ICC could potentially lead to a dramatic understatement of the global extent of agreement among raters. A practical example of this fact is discussed in chapter 4.

(*ii*) A second option would be to compute for each individual patient a summary score by averaging the 4 associated measurements. This summary score could then be used for computing the global ICC coefficient. In my opinion this is likely one of the worse options to consider, and I strongly encourage researchers to avoid it at all costs. If there is a tendency for some variables take high values than others, averaging all of the variables at the subject level may result in all subjects having similar mean values. Any subject variation observed on each variable will vanish. This will inevitable lead to a near zero ICC, given the importance of subject variation in computing the intraclass correlation. Concrete examples that illustrate this issue are shown in chapter 4.

(*iii*) You may use one of the multivariate approaches advocated by Shou et al. (2013) or Yue et al. (2015). These options are available to researchers. As previously mentioned, implementing then requires some efforts.

(*iv*) My preferred option is to compute a composite score optimally computed so as to capture most of the variation in the data. Unlike the mean score of option (*ii*), which assign equal weight to all variate, I recommend computing what is known in the statistical literature as the "Principal Component." In a nutshell, the principal component is a linear combination of the variables of interest, where the variables with the most variation receive the largest weights. The statistical techniques require for obtaining the principal component are implemented in almost all known statistical packages. I will nevertheless describe later in this section the mechanics for obtaining this principal component.

Table 3.3: Coronary artery disease diagnosis measures taken from the global and main coronary vessel regional levels[a]

Patient	GLOBAL	LAD	LCX	RCA
1	1.753	1.813	1.701	1.733
2	0.801	0.721	0.894	0.891
3	1.563	1.469	1.467	2.012
4	0.791	0.824	0.740	0.790
5	0.798	0.830	0.730	0.885
6	0.965	1.095	0.911	0.759
7	1.460	1.356	1.503	1.787
8	1.429	1.478	1.381	1.381
9	1.201	1.196	1.159	1.221
10	0.767	0.805	0.702	0.732

[a]LAD, LCx, and RCA are acronyms associated with the coronary vessel areas analyzed

3.5.1 *The Principal Component*

Consider Table 3.3 data and let us see step by step how to obtain the composite score using the principal component approach. The composite score is essentially a fifth variable that I will add to Table 3.3, which like the other 4 variables assign one value to each patient. The ultimate goal is for "Composite" to be used as a surrogate to the other 4 variables while capturing most of the variation among subjects shown by the original data. This technique allows us to reduce a 4-dimensional problem to a much simpler one-dimensional problem. Let COMP be the composite score and COMP_i the composite score associated with a specific patient i. The link between COMP and the other 4 variables can be described as follows:

$$\text{COMP}_i = \alpha_{\text{G}} \times \text{GLOBAL}_i + \alpha_{\text{L}} \times \text{LAD}_i + \alpha_{\text{X}} \times \text{LCX}_i + \alpha_{\text{R}} \times \text{RCA}_i, \qquad (3.5.1)$$

where GLOBAL_i for example is the value of the GLOBAL variable associated with patient i. The main question is how to determine the 4 coefficients α_{G}, α_{L}, α_{X}, and α_{R} associated with the 4 factors. These 4 coefficients are also known in the literature as the factor loadings, and are calculated in such a way that they will maximize the variance of COMP in order to capture most of the variation in the dataset. The sum of the squared loadings is also required to be 1. The sole objective of this requirement is to ensure they are unique. For the sake of simplicity, let $\mathcal{F}$ denote the list of all 4 factors under investigation. That is, $\mathcal{F} = \{G, L, X, R\}$ where the letters respectively represent the factors GLOBAL, LAD, LCX, and RCA.

Most researchers do not need to have an in-depth understanding of the complex mathematics and matrix algebra that underly the derivation of the principal component. Therefore, I present here what I consider to be the essentials. First, let me introduce a few concepts:

- *The Vector.* Let $\boldsymbol{\alpha} = (\alpha_{\text{G}}, \alpha_{\text{L}}, \alpha_{\text{X}}, \alpha_{\text{R}})$ be the list of coefficients we are attempting to calculate. Such a list is often referred to as a vector. Here, $\boldsymbol{\alpha}$ is a 4-element vector. Using Table 3.3 data, most statistical packages can produce 4 eigenvectors[4] similar to $\boldsymbol{\alpha}$ and 4 associated eigenvalues often denoted by λ_1, λ_2, λ_3 and λ_4. The eigenvector associated with the highest of the 4 eigenvalues contains the 4 coefficients or factor loadings we are looking for.

- *The Covariance Matrix.* It is common practice in statistical science to describe the statistical properties of a dataset such as that of Table 3.3 using what

[4] I will define the related notions of eigenvector and eigenvalue later in this section.

known as the covariance matrix, the variance matrix, or the variance-covariance matrix, and given by:

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{\sigma^2_G} & \sigma_{GL} & \sigma_{GX} & \sigma_{GR} \\ \sigma_{LG} & \boldsymbol{\sigma^2_L} & \sigma_{LX} & \sigma_{LR} \\ \sigma_{XG} & \sigma_{XL} & \boldsymbol{\sigma^2_X} & \sigma_{XR} \\ \sigma_{RG} & \sigma_{RL} & \sigma_{RX} & \boldsymbol{\sigma^2_R} \end{pmatrix}, \tag{3.5.2}$$

where σ^2_G is the variance of the GLOBAL variable, and σ_{LG} the covariance between the LAD and GLOBAL variables. The diagonal of the covariance matrix contains each factor's variance, and the off-diagonal elements describe how these factors are related.

To understand the importance of this covariance matrix, remember that the loadings must be calculated so as to maximize the variance of COMP. This variance can be calculated from equation 3.5.1 as follows:

$$\text{var}(COMP) = \sum_{(k,l)\in\mathcal{F}} \sum \alpha_k \alpha_l \sigma_{kl}. \tag{3.5.3}$$

It follows from equation 3.5.3 that all elements σ_{kl} of the covariance matrix $\boldsymbol{M}$ as well as all factor loadings α_{kl} of the composite score COMP are used in the calculation of the composite score variance.

Before introducing the notions of eigenvector and eigenvalue, it is essential to note that the variance of the composite score shown in equation 3.5.3 can be rewritten as follows[5]:

$$var(\text{COMP}) = \sum_{k\in\mathcal{F}} \alpha_k \left(\sum_{l\in\mathcal{F}} \alpha_l \sigma_{kl} \right). \tag{3.5.4}$$

- *Eigenvectors and eigenvalues.* The notions of eigenvalue and eigenvector are so intimately linked that one cannot mention one of them without mentioning the other. Moreover, these 2 notions are always associated with a particular matrix. In our case, it is the covariance matrix $\mathbf{M}$. *A 4-element vector* $\boldsymbol{\alpha}$ *whose squared coefficients sum 1 is an eigenvector of matrix* $\boldsymbol{M}$*, if any of its elements* α_k *is linked to the other elements as follows*:

$$\sum_{l\in\mathcal{F}} \alpha_l \sigma_{kl} = \lambda \alpha_k, \tag{3.5.5}$$

for some number λ, which is the associated eigenvalue. You want to know why we need eigenvectors and eigenvalues? Look at equation 3.5.4 and you

[5] Note that the coefficients α_k (for $k = \text{G, L, X, R}$) are still unknown to us and remain our focus.

will realize that if you have an eigenvector then term in parentheses would be replaced with $\lambda\alpha_k$, and since the squared coefficients sum to 1, the variance of the composite score will be $var(\text{COMP}) = \lambda$. The resulting COMP variable using the eigenvector's elements as coefficients is called a principal component.

One can prove mathematically that a total of 4 eigenvectors $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, and $\boldsymbol{\alpha}_4$ and 4 associated eigenvalues $\lambda_1 > \lambda_2, \lambda_3 > \lambda_4$ can be derived. This will lead to 4 different principal components, the variance of which equals the value of the associated eigenvalue. In reality, only the first principal component associated with the highest eigenvalue λ_1 will be retained as our composite score COMP.

- *Data Total Variation.* The sum of the 4 diagonal elements of this covariance represents the total variation in the dataset and is also known by its mathematical name of "trace of matrix $\mathbf{M}$" and denoted by $Trace(\boldsymbol{M})$ That is,

$$\boxed{\text{Total Variation} = \text{Trace}(\boldsymbol{M}) = \sigma_{\text{G}}^2 + \sigma_{\text{L}}^2 + \sigma_{\text{X}}^2 + \sigma_{\text{R}}^2.} \tag{3.5.6}$$

What is interesting here is that the total variance is also equal to the sum of all 4 eigenvalues $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$. Consequently, the proportion of the total variation that is explained by our composite score (i.e. the first principal component) is $\lambda_1/(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)$.

Example 3.1

To illustrate the principal component analysis, let us calculate the composite scores of the 10 patients of Table 3.3. My objective is to add an extra column to Table 3.3 that assigns a single summary number to each patient that will replace the previous 4 scores in further analysis.

Table 3.4 shows the 4 eigenvectors and 4 eigenvalues associated with Table 3.3 data. There a few interesting things that can be observed from this Table. It follows from the last 2 columns of this table that the first eigenvalue of 0.6118 represents about 94.8% of the total sum of all eigenvalues 0.6451. This eigenvalue is associated with the first eigenvector $\boldsymbol{\alpha}_1$ shown in the first column of the table. Consequently, the 4 numbers in the first column are precisely the 4 coefficients you need to construct the composite score (or the first principal component), shown in Table 3.5.

Table 3.4: Eigenvectors and eigenvalues associated with 3.3 data

	Eigenvectors (or factor loadings)					Eigenvalues[a]	
Factor	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_3$	$\boldsymbol{\alpha}_2$		$\boldsymbol{\lambda}$	% var
GLOBAL	0.4768	0.2225	0.0554	0.8486		0.6118	94.8%
LAD	0.4470	0.6123	0.4787	-0.4429		0.0296	4.6%
LCX	0.4692	0.1355	-0.8377	-0.2445		0.0037	0.6%
LCX	0.5939	-0.7465	0.2571	-0.1548		0.0001	0.0%
SSC[b]	1	1	1	1	Total	0.6451	100.0%

[a] $\boldsymbol{\lambda}$ is the vector of eigenvalues and "% var" is the proportion of total variation explained by each eigenvalue.
[b] Sum of Squared Coefficients

Table 3.5: Table 3.3 data along with the patients' composite scores.

Patient	GLOBAL	LAD	LCX	RCA	COMP[a]
1	1.753	1.813	1.701	1.733	3.474
2	0.801	0.721	0.894	0.891	1.653
3	1.563	1.469	1.467	2.012	3.285
4	0.791	0.824	0.740	0.790	1.562
5	0.798	0.830	0.730	0.885	1.620
6	0.965	1.095	0.911	0.759	1.828
7	1.460	1.356	1.503	1.787	3.069
8	1.429	1.478	1.381	1.381	2.810
9	1.201	1.196	1.159	1.221	2.376
10	0.767	0.805	0.702	0.732	1.490

[a] This is the first principal component, an average of the 4 factors GLOBAL, LAD, LCX, and RCA, weighted by the coefficients of the first eigenvector $\boldsymbol{\alpha}_1$

The composite score of Table 3.5 is calculated as follows:

$$\text{COMPOSITE} = 0.4768 \times GLOBAL + 0.4470 \times LAD + 0.4692 \times LCX + 0.5939 \times RCA,$$

where the coefficients come from the first column of Table 3.4. The composite score associated with patient 1 for example is calculated as follows:

$$3.474 = 0.4768 \times 1.753 + 0.4470 \times 1.813 + 0.4692 \times 1.701 + 0.5939 \times 1.733.$$

Since the first eigenvalue represents the variance of the first principal components, one can conclude that our composite score explains about 94.8% of the total variation in the data. You be quite confident in its ability to summarize your entire dataset in further analysis.

3.5.2 *The Multivariate ICC*

In section 3.5.1, I recommended that the first principal component be used as composite score for calculating the multivariate intraclass correlation coefficient. This composite score is expected to be used in a univariate approach along with the univariate models to be discussed in chapters 4, 5 and 6. The composite score summarizes the original variables, captures most of the variation in the dataset and incorporates its correlation structure. Although this approach will generally work well, there are a few important issues the researcher must know when using this approach.

Since the composite score is based on the first principal component, it is expected to explain the highest proportion of total variance of all principal components. However, in some instances even that highest proportion of total variance may not be sufficiently large for the composite score to have an adequate representation of total variation in the dataset. For example, if that proportion is 50% then half of the total variation will not be reflected in the composite score. Overlooking such an issue could lead to an overstatement of the extent of agreement among raters. With less variation comes an artificially higher agreement. In this case, it would be more effective to use not one, but 2 or 3 principal component scores and a weighted average as discussed in section 9.4 and chapter 9.

When a composite score is calculated as in example 3.1, can it translate any agreement among raters very well? The multivariate ICC will be an aggregate that is expected to summarize the extent of agreement among raters not on one factor, but on 2 factors or more. However, the raters may agree well on one factor while severely disagreeing on another one. In this situation, what story would we want the multivariate ICC to tell us? The ICC as a measure of agreement is far more reliable when applied to an heterogeneous population of subjects, because the subject variance is compared to the rater variance. It is when the subject variance exceeds the rater variance by a wide margin that one concludes that the raters agree. Therefore, in a multivariate setting the raters' agreement at the level of individual factors will translate well in a composite score if this score carries most of subject variation shown in the data.

If the number of factors under investigation is very large, adequately summarizing

them with a single composite score becomes a difficult task. This situation may require a more refined analysis. It may be necessary to retain the first 2 principal components, and to quantify the extent of agreement separately for these 2 composite scores. I suggest that you also compute the correlation coefficients between each of the 2 composite scores and the original variables. The highest correlation coefficients will tell you what factors are summarized by each composite score.

Composite scores must be built based on factors using the same units of measurements. If the original factors have different units of measurements, then it will be necessary to standardize all of them. This transformation is essential for obtaining an adequate composite score. If one factor is expressed in the millions while another factors is expressed in the hundreds, then the data variation will be artificially dominated by the factor measured in millions giving it an unduly high influence in the construction of the composite score. To be "fair" to all factors, they must be standardized, striping them of any unit of measurement they may have.

3.6 Concluding Remarks

Although practical constraints often dictate the statistical model to be used, researchers must at times decide which statistical best fits their analytic goals. This is generally the case when the researcher is involved in the study at the design stage. The question then becomes which model should one pick? The answer to this question depends on whether the main focus of your investigation is on the inter-rater reliability, on the intra-rater reliability or on both.

As previously discussed, models 1A and 1B are limited with respect to the type of ICC statistic they can produce. Model 1A allows for the calculation of inter-rater and not intra-rater reliability. Model 1B does the opposite. It allows for the calculation of intra-rater and not inter-rater reliability. Consequently, if computing both the inter-rater and the intra-rater reliability is among your study objectives then both models must be rejected in favor of models 2 and 3.

For the purpose of optimizing the inter-rater reliability assessment, I recommend the use of models 2 or 3 if possible as opposed to model 1A. The only reason model 1A should ever be considered is if getting the same group of raters to rate all subjects is challenging. In this case, only under model 1A will you have the luxury to assign a different group of raters to different subjects and still be able to compute inter-rater reliability. However, the use of a large number of different groups of raters will increase the experimental error, making it more difficult to obtain a high inter-rater reliability coefficient. Under models 2 and 3, the same group of raters must rate all subjects. However, if the group of raters is assumed to have been randomly selected from a larger group of raters it represents, then the rater effect is a random

variable and only model 2 can be used. However, if the raters taking part of the inter-rater reliability experiment are not seen as coming from a larger group of raters representing the primary target of the investigation, then you will need model 3. You will be penalized by using model 2 as opposed to model 3 in the latter situation. For, the rater variance calculated under model 2 is expected to have a negative impact on the inter-rater reliability coefficient.

For the purpose of optimizing intra-rater reliability, I still recommend the use of models 2 or 3 whenever possible. You may nevertheless use the simple model 1B which does not require all raters to rate the same group of subjects. A major downside of the intra-rater reliability coefficient under model 1B is its vulnerability to a diverse subject sample. The approach I recommend is the use of either model 2 or model 3 depending on whether the participating raters are seen as representing a larger subject universe or not. However, using model 2 is now more advantageous than using model 3. This stems from the fact that under model 2, the within-rater variation looks smaller when compared to the combined variation due to raters and subjects than when compared to subject variation alone. All these issues and many more are further discussed in-depth in the next few chapters.

Methods for calculating the optimal number of subjects and raters under the various models will be presented in the respective chapters. I will demonstrate among other things that for a fixed number of ratings per rater, you need no more than 4, 5, or 6 trials to obtain the most accurate intra-rater reliability coefficient under models 2 and 3. That is if a rater must produce 40 ratings, it would be more effective to use 8 subjects and 5 trials rather than 20 subjects and 2 trials. Also addressed in chapter 4, is the important issue of multivariate analysis where each rater rates the subjects on several variables and not just on a single variable as is often the case. For this situation, I recommend using the data reduction technique of principal component analysis. The first principal component associated with the original multivariate dataset of ratings is to be used as a composite score. This composite score coupled with our model of choice will lead us to the desired intraclass correlation coefficient.

You may want to know about another agreement statistic for quantitative ratings occasionally mentioned in the literature is Lin's Concordance Correlation Coefficient proposed by Lin (1989). Although this coefficient is an improvement over the classical Pearson's[6], it still cannot properly handle ratings with more complex data structures such as those described by the models studied in the next few chapters. Consequently, this coefficient is not discussed in this book.

Figure 3.2 represents a decision tree showing which equations or subsections in

[6]The classical Pearson's correlation coefficient quantifies the extent to which series of ratings from 2 raters are linked by a linear relation of any type. Lin's coefficient of concordance quantifies the extent to which the 2 series of ratings are identical.

the subsequent chapters should be used to compute the correct agreement coefficient and associated p-values and confidence intervals, depending on the model dictated by your study design. The numbering of these equations (or subsections) is descriptive, and the first digit refers to the chapter number, the second digit to the section within the chapter, and the third number to a specific equation or subsection.

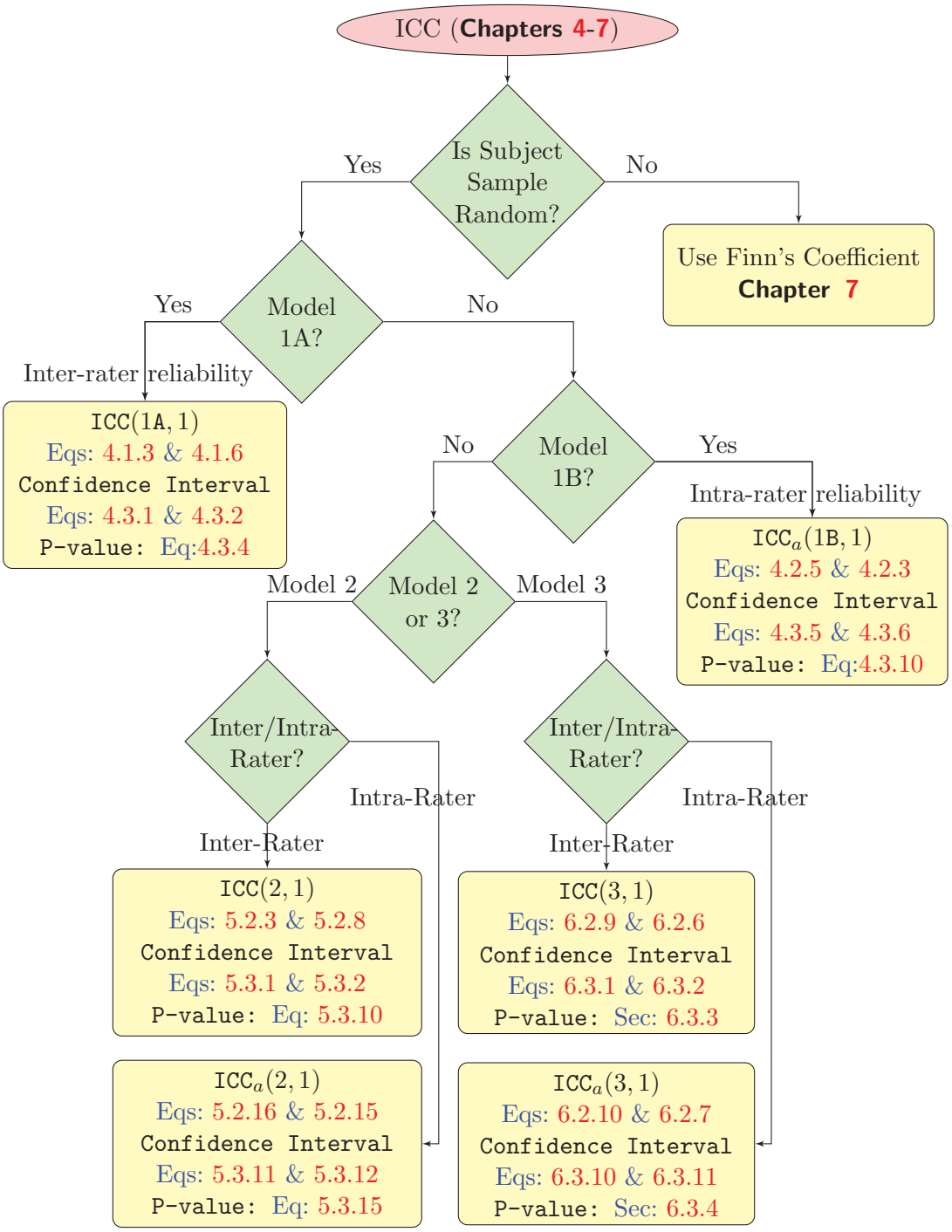

Figure 3.2: Choosing the Correct Intraclass Correlation Coefficient