CHAPTER 9

Intraclass Correlation & Multivariate Analysis

OBJECTIVE

This chapter goes beyond the description of abstract concepts and the associated computational methods. The objective is to explore and understand the story that is told by the intraclass correlation coefficient (ICC) and their precision measures. I will first discuss benchmarking techniques that would help you qualify the magnitude of the ICC based on a benchmark scale that determine the ICC's degree of acceptability. I will also discuss about some influence analysis techniques that can be used to identify problem raters responsible for lowering the ICC. Such techniques could be useful for indetifying raters who may need further training to improve the overall ICC. Finally, this chapter will address the important problem of multivariate intraclass correlation. This problem results from the need for a global agreement coefficient when several measurements are taken on the same subject. Existing multivariate Intraclass Correlation Coefficients will be discussed as well as a new approach based on principal component analysis.

Contents

9.1 *Overview* 265
9.2 *Benchmarking Intraclass Correlation Coefficients* 265
9.2.1 *Benchmarking under Model 1A* 267
9.2.2 *Benchmarking under Alternative ANOVA Models* 269
9.3 *Influence Analysis* 270
9.4 *A Multivariate Approach to Inter-Rater Reliability* 272
9.4.1 *Review of Existing Multivariate Approaches* 274
9.4.2 *Introduction to Principal Components Analysis (PCA)* 277
9.4.3 *Multivariate Coefficients with Quantitative Ratings* 280
9.4.4 *Some R- Scripts for Calculating the Φ Coefficient* 290

9.1 Overview

From chapter 3 through chapter 7, we have explored various methods for evaluating the extent of agreement among raters that produced quantitative measurements. Agreement was quantified using either the intraclass correlation coefficient under different ANOVA models, or the Finn's coefficient. I also discussed the way to calculate precision measures in the form of confidence intervals and p-values. This chapter focuses on interpreting these agreement coefficients and on extending the univariate methods of the previous chapters so that multivariate rating data can be analyzed as well.

For the sake of interpreting ICC coefficients, I will confine myself the 2 tasks of benchmarking and influence analysis. Benchmarking an ICC estimate consists of matching it against a benchmark scale so as to determine its degree of acceptability. The benchmark scale is an exhaustive list of ICC ranges and the associated description of the strength of agreement. The specific benchmark scale that is used in the chapter is that proposed by Koo and Li (2016). It is one of the very few benchmark scales in the literature of quantitative inter-rater reliability that I know of, and which is discussed in section 9.2. Influence analysis on the other hand, aims at identifying problem raters in cases where the ICC is deemed low or not sufficiently high. A measure of influence of each individual rater is calculated and those raters with the highest negative impact on the overall ICC are identified and corrective actions can be taken. Influence analysis is discussed in section 9.3.

Section 9.3 is devoted to multivariate analysis of inter-rater reliability. When dealing with a single variable, the ICC can be calculated using one of the many methods discussed in Part-II chapters. However, there are situations where each rater takes several often correlated measures on the same subject. Although separate independent univariate analyses can always be conducted on each of the measurements, providing a global inter-rater reliability assessment for all these variables may be difficult due to the possibly high correlation among these variables. In section 9.3, I will briefly review the literature on this topic before proposing a new approach based on principal component analysis.

9.2 Benchmarking Intraclass Correlation Coefficients

In chapters 4, 5 and 6, I presented several intraclass correlation coefficients (ICC) without discussing any criterion or standard of acceptability. In other words, what magnitude of the ICC should represent "poor," or "good" agreement among raters? The set of ICC ranges that would define such criteria or standards is generally referred to as the benchmark scale, and the task of evaluating any specific ICC against it is known as *Benchmarking*. If this issue was not discussed, it is primarily because

there is no known criterion or standard of acceptability in the inter-rater reliability literature. However, Koo and Li (2016) briefly discussed this problem and proposed a benchmark scale that will be discussed in this chapter.

Table 9.1: The Koo-Li ICC Benchmark Scale[a]

ICC Range	Strength of Agreement
$0 \leq \texttt{ICC} < 0.50$	Poor
$0.50 \leq \texttt{ICC} < 0.75$	Moderate
$0.75 \leq \texttt{ICC} < 0.90$	Good
$0.90 \leq \texttt{ICC} < 1$	Excellent

[a]This benchmark scale is from Koo and Li (2016)

Benchmarking the ICC is a task that must be conducted carefully. Several factors could have a substantial impact on the magnitude of the ICC, and should be accounted for when interpreting your analysis. Here are 4 of the most important of these factors:

- `Subject Variance Magnitude`
  The size of the subject variance is a key factor impacting the magnitude of the ICC. Therefore, ICC-based inter-rater reliability experiments should be based on subject samples that were randomly selected from a relatively heterogenous subject population. Otherwise, a homogeneous subject sample will inevitably leads to a low ICC.

  In general, the ICC is quantified as the ratio of the subject variance to the total variance. Total variance includes the rater variance among other variance components. This rater variance that appears in the denominator may be measurable or not depending on the underlying ANOVA model used. Either way, this rater variance component does exist. When the ICC is high, then you know the subject variance exceeds the rater variance by a significant margin, and the extent of agreement among raters can rightfully be considered to be good. However, if the ICC is low then you can claim there is no evidence suggesting a high agreement among raters. It is a conclusive proof of the absence of agreement. Instead, it proves that the dominant component in total variance is either the rater variance (suggesting low agreement) or the error variance (suggesting poor experiment planning).

- `Underlying ANOVA Model`
  The study of ICCs in the past few chapters has revealed that the ANOVA model that underlies the ICC also impacts its magnitude and the way it should

be interpreted. Therefore, the ANOVA model used should be an integral part of the reported results as suggested by Koo and Li (2016).

- `Subject Sample Size`
  A small number of subjects will yield an ICC with a large variance. This lack of precision will make the ICC an unreliable estimate of the extent of agreement among raters. That is, two subject samples of the same small size could possibly yield 2 substantially different ICC estimates. Therefore, the precision associated with the ICC estimate must be an integral part of the benchmarking process.

- `Raters' Background`
  The raters participating in an inter-rater reliability experiment need to receive basic training in the scoring of a pilot subject sample. This will minimize the risk of a systematic bias in the scoring, producing a constant gap between 2 series of ratings. The ICC may not interpret such a systematic bias as a disagreement.

An intraclass correlation coefficient that is calculated based on experimental data is a rough approximation (i.e the estimate) of the actual intraclass correlation (the estimand). The benchmark scale of Table 9.1 is normally meant to be used with the estimand. Since the estimate is subject to statistical error, benchmarking it against the Koo-Li benchmark scale requires a more elaborate procedure.

9.2.1 *Benchmarking under Model 1A*

Consider the children lung function data of Table 9.2. The ICC and associated 95% confidence interval for this dataset under the Model 1A ANOVA model are given by,

$$\text{ICC} = 0.752 \textit{ and the 95\% confidence interval is } (0.557 \text{ to } 0.894). \tag{9.2.1}$$

By simply using the ICC and the Koo-Li benchmark table, you would conclude that the strength of agreement is "Good." However, the 95% confidence interval suggests that "true" ICC value could be between 0.557 and 0.894, which qualifies the strength of agreement as "Moderate" to "Good." The confidence interval is a measure of precision, which accounts for the underlying Model 1A, the number subjects and raters in the experiment. Its use in the benchmarking process is essential to account for the many factors that impact the ICC magnitude. The confidence interval has a disadvantage. It does not tell us where the extent of agreement is more likely to Moderate than Good or vice-versa. To resolve this problem, I recommend assigning a membership probability to each ICC range of Table 9.1 as shown in Table 9.3.

Table 9.3 indicates that the true ICC could be good with a 51.82% probability whereas it could Moderate with a 45.55% probability. Moreover, it is on the "Mod-

erate" interval that the cumulative membership probability exceeds the threshold of 95%. Therefore, I recommend qualifying the extent of agreement as Moderate.

Note that the benchmarking process described in this section is based on Model 1A. Consequently, for any given interval (l, u) the associated membership probability based on Model 1A is given by,

$$\theta(l,u) = \begin{cases} P\left(0 \le F \le \dfrac{F_0}{1 + Ml/\left[n(1-l)\right]}\right), & \text{if } u = 1, \\ P\left(\dfrac{F_0}{1 + \dfrac{M}{n}\dfrac{u}{1-u}} \le F \le \dfrac{F_0}{1 + \dfrac{M}{n}\dfrac{l}{1-l}}\right), & \text{otherwise.} \end{cases} \tag{9.2.2}$$

where F is a random variable following the F distribution with $n-1$ and $M-n$ degrees of freedom, $F_0 = \text{MSS}/\text{MSE}$, n is the number of subjects that were rated and M the total number of ratings produced by all raters. For Table 9.2 data, $n = 15$ and $M = 60$.

Table 9.2: 15 children lung function measurements representing the peak expiratory flow rates (PEFR)[a]

Subject	Rater (j)			
(i)	1	2	3	4
1	190	220	200	200
2	220	200	240	230
3	260	260	240	280
4	210	300	280	265
5	270	265	280	270
6	280	280	270	275
7	260	280	280	300
8	275	275	275	305
9	280	290	300	290
10	320	290	300	290
11	300	300	310	300
12	270	250	330	370
13	320	330	330	330
14	335	320	335	375
15	350	320	340	365

[a]Source: Bland MJ, Altman DG. Statistics Notes: Measurement error. British Medical Journal 1996;312:1654 (extract)

Table 9.3: Implementing the Koo-Li ICC Benchmark Scale

ICC Range	Strength of Agreement	Interval probability	Cumulative Probability
$0.90 \leq \texttt{ICC} < 1$	Excellent	0.0189	0.0189
$0.75 \leq \texttt{ICC} < 0.90$	Good	0.5182	0.5371
$0.50 \leq \texttt{ICC} < 0.75$	Moderate	0.4555	0.9927
$0 \leq \texttt{ICC} < 0.50$	Poor	0.0073	1

9.2.2 *Benchmarking under Alternative ANOVA Models*

In section 9.2.1, benchmarking was presented under the assumption that the ICC is calculated based on model 1A. If you base your ICC calculations on an alternative model (e.g. Model 1B, Model 2 or Model 3) then equation 9.2.2 needs to be revised. For any interval (l, u) the following expressions must be used:

- **Model 1B**
  Under Model 1B, the interval membership probabilities are calculated as follows:

$$\theta(l,u) = \begin{cases} P\left(0 \leq F \leq \dfrac{F_0}{1 + Ml/\big[r(1-l)\big]}\right), & \text{if } u = 1, \\ P\left(\dfrac{F_0}{1 + \dfrac{M}{r}\dfrac{u}{1-u}} \leq F \leq \dfrac{F_0}{1 + \dfrac{M}{r}\dfrac{l}{1-l}}\right), & \text{otherwise,} \end{cases} \tag{9.2.3}$$

  where F is a random variable that follows the F with $r-1$ and $M-r$ degrees of freedom. $F_0 = \text{MSR}/\text{MSE}$, MSE and MSR are respectively given by equations 4.3.7 and 4.3.8 of chapter 4, and r is the number of raters.

- **Model 2**
  Under Model 2, the interval membership probabilities for ICC(2,1) of equation 5.2.3 (inter-rater reliability) are calculated as follows:

$$\theta(l,u) = \begin{cases} P\Big(0 \leq F \leq F_l\Big), & \text{if } u = 1, \\ P\big(F_u \leq F \leq F_l\big), & \text{otherwise,} \end{cases} \tag{9.2.4}$$

where F is a random variable that follows the F distribution with $n-1$ and v degrees of freedom[1]. Moreover, F_l and F_u are defined as follows: $F_l = \text{MSS}/(a_l\text{MSR} + b_l\text{MSI} + c_l\text{MSE})$ and $F_u = \text{MSS}/(a_u\text{MSR} + b_u\text{MSI} + c_u\text{MSE})$. Note that $a_l = rl/\big[n(1-l)\big]$, $b_l = 1 + r(n-1)l/\big[n(1-l)\big]$, and $c_l = (M/n - r)l/(1-l)$. a_u, b_u and c_u are calculated in a similar manner, by replacing the interval lower bound l with its upper bound u throughout.

For the intra-rater reliability coefficient estimated by $\text{ICC}_a(2,1)$ of equation 5.2.16, you need to use equations 5.3.13, 5.3.14 and 5.3.15.

- **Model 3**
  Under Model 3, the interval membership probabilities for ICC(3,1) of equation 6.2.9 (inter-rater reliability) are calculated as follows:

$$\theta(l,u) = \begin{cases} P\Big(0 \le F \le F_l\Big), & \text{if } u = 1, \\ P\big(F_u \le F \le F_l\big), & \text{otherwise,} \end{cases} \tag{9.2.5}$$

  where F is a random variable that follows the F distribution with $n-1$ and v degrees of freedom[2]. Moreover, F_l and F_u are defined as follows: $F_l = \text{MSS}/(a_l\text{MSI} + b_l\text{MSE})$ and $F_u = \text{MSS}/(a_u\text{MSR} + b_u\text{MSI} + c_u\text{MSE})$. Note that $a_l = rl/\big[n(1-l)\big]$, $b_l = 1 + r(n-1)l/\big[n(1-l)\big]$, and $c_l = (M/n - r)l/(1-l)$. a_u, b_u and c_u are calculated in a similar manner, by replacing the interval lower bound l with its upper bound u throughout.

  For the intra-rater reliability coefficient estimated by $\text{ICC}_a(3,1)$ of equation 6.2.10, you need to use equations 6.3.8, 6.3.9 and 6.3.12.

9.3 Influence Analysis

Influence analysis consists of quantifying the impact (positive or negative) that each rater has on the overall extent agreement among raters. The primary objective of influence analysis is to identify problem raters that may need additional training to bring them up to speed. Therefore, you will worry about this issue only if you are dissatisfied with the low ICC value and would like explain what is causing this lack of agreement. You want to able to target specific raters who need to improve their proficiency in scoring subjects. Is low ICC the result of one rater that disagrees

[1] Note that v is defined by equation 5.3.8 of chapter 5 and can be calculated by replacing the ICC with its estimated value.

[2] Note that v is defined by equation 6.3.4 of chapter 6 and can be calculated by replacing the ICC with its estimated value.

with the rest of the group? Is it the result of 2 or 3 raters? These are the types of questions that influence analysis addresses.

The influence that a given rater k has on the ICC is defined as follows:

$$\alpha_k = \frac{\text{ICC}^{(-k)} - \text{ICC}}{\text{ICC}}, \tag{9.3.1}$$

where $\text{ICC}^{(-k)}$ is the intraclass correlation coefficient calculated based on a rater sample from which rater k had been previously removed and ICC, the intraclass correlation based on the full sample. Rater k's influence α_k represents the relative change in the ICC magnitude due to the removal of one specific rater k from the roster of rater. Each individual rater's influence must be calculated and evaluated and compare to that of other raters.

Consider once again the children lung function measurements of Table 9.2. The extent of agreement among the 4 raters was evaluated by a Model-1A-based ICC to be 0.752. The results obtained from the influence analysis are shown in Table 9.4 and depicted in Figure 9.1. It follows from this figure that removing Rater 1 from the rater sample does not affect the ICC much (it actually results in a reduction of 1% as shown in Table 9.4). However, removing Rater 2 leads to a dramatic increase of 6%. The most problematic raters are those associated with the highest (positive) influence value. In this case, it is rater 2. That is the 3 raters 1, 3 and 4 agree among themselves more than they agree with rater 2. It appears that rater 2 is an outlier in the roster of raters and must be examined.

After removing rater 3, the ICC takes a dive of more than 8%. Consequently, rater 3 appears to agree with each of the other raters more than they agree among themselves, and cannot be problematic. Similarly, rater 4' influence is only 2.76% an is relatively small. A limitation of influence analysis as a method of identifying problem raters lies in its ability to identify a single problem rater at a time (the most problematic). This is not a problem if the number of raters is limited. If the number of raters becomes large, this procedure could take time before all problem cases are identified and resolved.

Table 9.4: Influence analysis of the children lung function data of Table 9.2

k	$\text{ICC}^{(-k)}$	α_k
1	0.7429	-1.14%
2	0.7984	6.24%
3	0.6904	-8.12%
4	0.7723	2.76%
$(\star)$[a]	0.7515	0%

[a]No rater removed. Calculations based on the full roster of 4 raters.

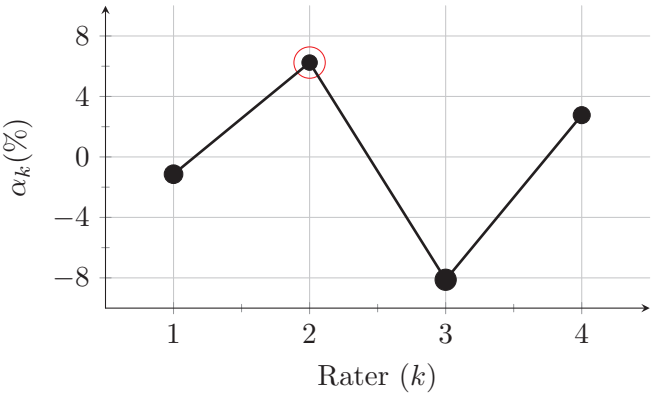

Figure 9.1: Influence analysis of the children lung function ratings of Table 9.2.

9.4 A Multivariate Approach to Inter-Rater Reliability

In section 4.1 of chapter 4, I discussed the intraclass correlation coefficient (ICC) under the one-way random subject factor model. While the focus was on the analysis of a single variable (i.e. univariate analysis), numerous applications nowadays require that several correlated variables be measured for each individual subject. Analyzing these correlated variables independently using Chapter 4 methods will ignore the extent to which the magnitude of one variable affects the other variables associated with the same subject. Consequently, your global assessment of agreement could be misleading. It is to account for the inter-dependency of several variables measured on the same subject that a number of authors have developed generalized

versions of the intraclass correlation coefficient that I am going to discuss in this section.

The objective in section 4.1 of chapter 4, was to quantify the extent of agreement among raters when different subjects may be rated by different groups of raters. Since intraclass correlation is not calculated for a specific group of raters under this model, ICC represents more a general measure of reproducibility than inter-rater reliability measure since the raters as well as other unspecified factors may contribute to total variation in the ratings. The statistical model used to describe the data was formulated as follows:

$$y_{ijk} = \mu + s_i + e_{ijk}, \tag{9.4.1}$$

where y_{ijk} is the rating associated with subject i, rater j and replicate k. Note that the term replicate in this context refers to an independent measurement taken on the same subject i by the same rater j under supposedly the same conditions. But all measurements are related to the same one variable labeled as y.

However, with the development of imaging technology particularly in the medical field, more and more subjects are images on which not just one variable is measured. Instead, several variables are often necessary for one rater to describe a single image. It is near impossible to describe (with any accuracy) even the simplest three-dimensional picture with a single number. A cube is the only one I can think of. Any other three-dimensional image will require 2 variables or more. Depending on what you are investigating, any image of the human heart may require a considerably large number of variables to be measured.

Table 9.5 from Wu et al. (2019) summarizes some MRI measurements of the plantar fascia thickness. It follows from this table that for each human subject (male or female), a total of $6 \times 2 = 12$ measurements must be taken. If there must be replication, then another 12 measurements will have to be taken again. Note that each individual measurement in Table 9.5 can be analyzed using model 9.4.1 and the associated intraclass correlation coefficient. However, this approach presents two problems that you need to know:

- The number of individual models to develop can become quite large. For Table 9.5 only, 12 models are needed for males, or females or for males and females combined.

- The bigger issue however, stems from the fact that the univariate approach just mentioned, does not account for the correlation or the covariance structure that is likely to exist between the variables associated with the same subject.

Consequently, several authors have developed alternative approaches for analyzing multivariate rating data, some of which are reviewed in the next section.

Table 9.5: MRI Measurements of the Average Thickness of the Plantar Fascia (in Millimeter)

	Male		Female	
Section	Left	Right	Left	Right
Origin (LS[a])	2.87	2.86	2.56	2.55
Central Part (LS)	2.02	2.01	1.82	1.81
Insertion (LS)	1.31	1.29	1.16	1.17
Origin (TS[b])	2.88	2.85	2.59	2.58
Central Part (TS)	2.05	2.01	1.82	1.83
Insertion (TS)	1.32	1.35	1.20	1.19

[a]LS = Longitudinal Section
[b]TS = Transverse Section

9.4.1 *Review of Existing Multivariate Approaches*

Shou et al. (2013) proposed a general method for quantifying the generalized Intraclass Correlation Coefficient for quantitative ratings, which they referred to as the I2C2 for Image Intra-Class Correlation (IICC or I2C2). When the ratings are binary, Yue et al. (2015) proposed the graphical intra-class correlation (GICC), which is based on a multivariate logit model. Since the logit model does not link the formulation of GICC to an explicit correlation coefficient, its validity may require more justification. Consequently, I focus on the I2C2 coefficient of Shou et al. (2013).

Let us assume that each subject is represented by T different variables labeled as $\big(y_{ijk}(1), \cdots, y_{ijk}(t), \cdots, y_{ijk}(T)\big)$, where $y_{ijk}(t)$ is the rating pertaining to the t^{th} variable, and associated with subject i, rater j and replicate k. In the field of medical imaging for example, T may be known to researchers as the number of voxels (could be the different sections of an image researchers have decided to investigate). For the general framework I am dealing with, I will refer to T as the number of variables being measured. Let $\boldsymbol{y}_{ijk}$ be the T-dimensional vector representing subject i, rater j and replicate[3] k. Therefore,

$$\boldsymbol{y}_{ijk} = \big(y_{ijk}(1), \cdots, y_{ijk}(t), \cdots, y_{ijk}(T)\big). \tag{9.4.2}$$

Shou et al. (2013) hypothesized the ANOVA model 1A for each of the T variables (see section 4.1 of chapter 4 for a discussion of univariate ANOVA model 1A) defined

[3]Note that what Shou et al. (2013) refer to as replicate j is actually rater j in my current context. Replicates here represent multiple measurements the same rater j may take on the same subject i using the same variable t under similar conditions.