

**Extending the Linear Mixed
Modelling Framework for Method
Comparison Studies involving
Functional Responses with
Applications in Elite Sports**

PHD THESIS

by

Kishor Das

Supervisors:

Prof. John Newell

and

Dr. Thiago Oliveira

SCHOOL OF MATHEMATICAL & STATISTICAL SCIENCES

UNIVERSITY OF GALWAY



OLLSCOIL NA GAILLIMHÉ
UNIVERSITY OF GALWAY

June 2022

Acknowledgement

First, I would like to express my gratitude and love towards my mother Dipali Rani Halder whose continuous love and support made all the work possible. I also like to express my gratitude towards my father Rajendra Nath Das who always supported my interest in science from my early childhood. I would like to thank my elder brother Rathindra Nath Das who has always been an inspirational figure to me in all aspects of life.

I am very thankful to my friend Dr. Jaynal Abedin, who introduced University of Galway to me, for his guidance during the admission process of my PhD program and his support during my early days here in Galway. I am also extremely thankful to the Hardiman PhD scholarships program for supporting my PhD.

I would also like to express my appreciation to Dr. Lara Coyne and Professor Charlie Pedlar for providing the interesting datasets which were essential for this thesis.

I also like to express my sincere appreciation to all members of the Statistics Discipline and the administrative staff in the School of Mathematical and Statistical Sciences in University of Galway for their friendship and support.

I would like to express special thanks to my co-supervisor Dr. Thiago de Paula Oliveira for his friendship and expertise. Finally, I would like to express my deepest gratitude to my friend and supervisor Professor John Newell for providing insightful guidance throughout the journey and introducing key people along the way. Without his dedication, friendship, motivation, and immense knowledge this PhD thesis would not have been achievable.

Abstract

A method comparison study compares two or more different methods measuring some quantity of interest and the aim is to determine the level of agreement between the methods. Most of the methodological developments in this research area have focused on studies where the response variable of interest is a univariate continuous response. More recently there has been interest on method comparison studies where the response is functional in nature (i.e. a discrete realisations of an underlying functional form). This is still a new and growing field of research in Statistics with many open questions.

The specific aim of this thesis is to extend the Linear Mixed Modelling (LMM) framework to method agreement studies i) involving a continuous response when the observed bias between the methods of measurement is non-linear and ii) where the response variable is functional in nature.

Initially the focus is on method comparison studies involving a univariate continuous response where the LMM is used to extend the “classical” Bland and Altman approach to account for non-linear bias between two methods of measurement.

Following this, a further extension of the LMM framework is proposed for method comparison studies (of increasing design complexity) involving functional responses. A natural alternative analytical approach to consider is the use of Functional Data Analysis (FDA), given the nature of the response. A detailed description of the use of FDA in method comparison studies is given including a new functional equivalent to the Bland-Altman plot. An approach to adapt the LMM framework to accommodate functional responses is then given and the benefits of this approach for study designs with increasing complexity are discussed. The computational issues that arise when fitting a nonparametric LMM are highlighted and a new elegant solution to circumvent

this problem is proposed using an eigenbasis for the random-effects regressor matrix.

A simulation study is presented to investigate the performance of the FDA and LMM when used to generate functional limits of agreement in studies with no replicates. The performance of the eigenbasis approach to a full B-spline basis implementation is compared in terms of coverage and computational time.

All the graphical and analytical approaches proposed are demonstrated using data from two case studies in elite sports: one relating to blood biomarkers with a univariate continuous responses and the other a comparison of two motion capture systems with functional responses.

Notation

This thesis uses the following notation unless specified otherwise.

Concept	Notation	Example
Random variable	Capital Latin letter	X, Y, Z
Value of a random variable	Small Latin letter	x, y, z
Vectors (random variable or its value)	Bold face small Latin letter	$\mathbf{x}, \mathbf{y}, \mathbf{z}$
Matrices	Bold face capital Latin letter	$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$
Parameters	Small Greek letters	α, β, γ
Parameter vector	Bold face small Greek letters	$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$
Parameter matrices	Bold face capital Greek letters	$\boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\Gamma}$

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Datasets	4
1.2.1	C-Reactive Protein (CRP) Analyser Study	4
1.2.2	Motion Capture Study	9
1.3	Structure of the Thesis	18
2	Method Agreement Studies involving a Univariate Response	22
2.1	Introduction	22
2.2	Statistical Models for Measurements	23
2.3	Introduction to Linear Models	26
2.3.1	Linear Regression Models	26
2.3.2	Introduction to Linear Mixed Models (LMM)	28
2.3.3	Introduction to Smoothing	37
2.4	Reliability and Validity of a Measurement Method	38
2.4.1	Indices of Reliability	39
2.4.2	Application: CRP Analyser Study	41
2.5	Agreement between two Methods	43
2.5.1	Indices of Agreement	43
2.5.2	Application: CRP Analyser Study	48
2.5.3	Adjusting for Non-linear Bias	50
2.6	Summary	55

3	Method Agreement Studies involving a Functional Response	57
3.1	Introduction	57
3.2	Introduction to Functional Responses	58
3.3	Functional Responses in Method Comparison Studies	58
3.3.1	Introduction to Functional Data Analysis (FDA)	60
3.4	Reliability Indices for Functional Responses	69
3.5	Measuring Agreement between Functions using FDA	72
3.5.1	Introduction to Nonparametric LMM	77
3.6	Measuring Agreement between Functions using LMM	79
3.6.1	95% fLoA using a LMM for Studies with no Replicates	80
3.6.2	95% fLoA using a LMM for Studies with Replicates	94
3.7	Faster Computational Approach for Nonparametric LMM	99
3.7.1	The Karhunen–Loève Theorem	99
3.7.2	Eigenvalue Decomposition of a Variance Covariance Matrix	100
3.7.3	Application: Motion Capture Study	102
3.8	Chapter Summary	104
4	Simulation Study and Results	106
4.1	Introduction	106
4.2	Simulation Study Outline	107
4.3	Assessment Criteria	109
4.4	Comparing LMM and FDA in Studies with No Replicates	110
4.4.1	Interpreting a 95% fLoA?	114
4.4.2	Comparing 95% fLoA generated by FDA and LMM	115
4.5	Coverage and Computational Performance of a LMM in Studies with Replicates	117
4.5.1	Comparing Coverage When Using a B-spline and Eigenbasis for the Random-effects Regressor Matrices	119
4.6	Application: Motion Capture Study	123
4.6.1	R Code to Fit a LMM to Calculate 95% fLoA	128

4.7	Summary	130
5	Conclusions and Further Work	132
5.1	Introduction	132
5.2	CRP Analyser Study Summary and Conclusion	132
5.3	Motion Capture Study Summary and Conclusion	135
5.4	Summary of the Thesis	136
5.5	Open Questions and Further Work	139

List of Figures

1.1	Profile plot of c-reactive protein (CRP) measurements over seven different measurement occasions by two different methods of measurement from 35 athletes. These two measurement methods are a laboratory test (lab test) and a point of care test (POC test). A separate colour is used to represent each athlete. Different athletes provided a different number of measurements as not all were available at each testing time point. The minimum, median, and maximum number of measurements per subject per method are one, three and seven, respectively.	5
1.2	Gaussian kernel density plot for c-reactive protein (CRP) and log CRP measurements. Here measurements from both the laboratory test and a point of care test are combined to create the density plot.	6
1.3	Mean log c-reactive protein (CRP) profile over calendar months and seasonal months. One mean profile is from the measurements by a laboratory test (lab test, in red) and the other is from a point of care (POC) test (in blue).	7
1.4	Scatter plot of log c-reactive protein (CRP) measurements by the laboratory test (lab test) and a point of care (POC) test with the line of equality (black line) and a locally estimated scatterplot smoothing (LOESS) line (blue line).	8
1.5	Six degrees of freedom for the knee joint, which include three rotational and three translational motions (Komdeur et al., 2002).	10
1.6	The XYZ axis for the kinematic model along with corresponding degrees of freedom (McLean et al., 2005).	11

1.7	Right leg lunge exercise used to concurrently compare the markerless and marker-based systems.	12
1.8	Hierarchical study design for the motion capture study.	14
1.9	Transformation of subtracting the initial angle from the whole curve. This transformation was done to all the functional responses. For the purpose of demonstration only right hip flexion ange is presented here.	15
1.10	Multiplying by -1 to the measurements made by the markerless motion capture system to align with the direction of positive/negative angle of the marker-based system for the low spine angle during a lunge.	15
1.11	Shape of different functional responses for different directions at different body segments measured by two different methods of measurement during a right leg lunge.	17
1.12	Functional responses measured by a markerless and a marker-based motion capture system for right hip flexion during a lunge.	18
1.13	Functional responses by a markerless and a marker-based motion capture system for right hip abduction during a lunge.	19
1.14	Functional responses by a markerless and a marker-based motion capture system for right hip rotation during a lunge.	20
2.1	Properties of a measurement method.	39
2.2	Dot plot for C-reactive protein (CRP) measurements by both a point of care (POC) and a laboratory (lab) method. Each vertical line on the X-axis represents an athlete and points on a line represents replicate measurements from the athlete on different measurement occasions.	41
2.3	Indices of reliability and agreement.	44
2.4	Scatter plot with the line of equality for c-reactive protein (CRP) measurements made by a laboratory method (lab test) and a point of care method (POC test).	49

2.5	The Bland-Altman plot for the two methods measuring c-reactive protein (CRP). Here the two measurement methods are a point of care (POC) test and a laboratory (lab) test. The blue line represent a locally estimated scatterplot smoothing (LOESS) line.	50
2.6	A correlation plot of the differences between the measurements made by a point of care method and a laboratory method. Each subplot is a scatter plot of the differences where X and Y-axis represents two different measurement occasions.	51
2.7	95% limits of agreement to compare two methods of measurement: a laboratory test (lab) and a point of care test (POC). Four different variance functions were considered: A. no variance function; B. a fixed variance function; C. a power variance function; D. an exponential variance function.	53
2.8	95% limits of agreement using log transformed c-reactive protein (CRP) measurements.	55
3.1	Right hip abduction angle curve of nine different athletes from first measurement session. Only first replicate by each method of measurement are displayed here. A dot on a curve represents the maximum value of the angle curve. One way to reduce functional responses to univariate responses would be by only keeping the maximum angle for the analysis.	59
3.2	Right hip abduction angle curve measured by a marker-based and a markerless method of measurement during a lunge from an athlete in a given session.	61
3.3	B-spline basis functions for order 1 (A), order 2 (B), and order 3 (C) splines with knot sequence $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$	65
3.4	Observed difference (MB - ML) curves and smoothed functions from nine different athletes. Here MB is the measurement by the marker-based and ML is the measurement made by the markerless method. Blue dots are the observed differences and red solid curves are the smoothed functions obtained from it using smoothing splines.	68

3.5	Intraclass correlation coefficient (ICC) curve for two methods of measurement: markerless and marker-based motion capture system. Only two replicates from the first measurement session were used to calculate the ICC.	70
3.6	Standard error of measurement (SEM) curve for two methods of measurement: markerless and marker-based motion capture system. Only two replicates from the first measurement session were used to calculate the SEM.	71
3.7	Generalised cross validation (GCV) score for different values of log lambda.	73
3.8	95% functional limits of agreement estimated from one replicate measurement per method from a given session using the functional data analysis approach. Here, MB is the marker-based and ML is the markerless method of measurement.	74
3.9	Three dimensional plot for the functional limits of agreement. Here time is added as the third dimension to the Bland-Altman plot.	76
3.10	The Bland-Altman plot at different selected time points (frame) of the time domain. Here, MB is the marker-based and ML is the markerless method of measurement. Here, at each time point Bland-Altman plot was created considering the measurements as univariate responses. . . .	77
3.11	Cubic B-spline basis functions with a different number of equally spaced inner knots.	82
3.12	Adjusted R-squared values after fitting a linear regression model to each individual difference curve using B-spline basis functions with different numbers of inner knots as the covariates. Here different points on x-axis are different difference curve. In total 54 difference curves (9 athletes, 2 sessions, 3 replicates) were used for the right hip abduction angle measured by two different measurement methods: markerless and marker-based.	83

3.13 Akaike information criterion (AIC) and Bayesian information criterion (BIC) value for different mixed-effects models fitted with different knot sequences for the random-effects structure for the right hip abduction angle data.	85
3.14 Auto-correlation plots of residuals after fitting a mixed-effects model with different temporal correlation structures for the right hip abduction angle data. Here, AR1 is an autoregressive model with order 1; CAR1 is a continuous autoregressive model with order 1; ARMA(p, q) is an autoregressive-moving-average model with order p for the autoregressive model and order q for the moving-average model.	86
3.15 Akaike information criterion (AIC) of the mixed-effects models fitted with different correlation structure for the right hip abduction angle data.	87
3.16 Standardised residual over time frames from different mixed-effects models fitted with different variance functions.	88
3.17 Comparison of the estimate of the mean curve using the functional data analysis (FDA) and different linear mixed-effects models (LMM). The sample mean curve is also shown with the label “raw data”.	89
3.18 Comparison of the sample standard deviation curve (raw data) and estimate of the standard deviation curve by functional data analysis (FDA) with different specifications of the mixed-effects model (LMM).	90
3.19 95% functional limits of agreement for a study with no replicates by functional data analysis approach (FDA) and mixed-effects modelling (LMM) frameworks. Here, MB is the marker-based and ML is the markerless method of measurement.	93
3.20 A: differences against averages ignoring the time frame for the right hip abduction angle curves. Only one replicate measurement from the first measurement session was considered here. B: residuals against averages for the same data after fitting a linear mixed-effects model. Here, MB is the marker-based and ML is the markerless method of measurement.	93

3.21	95% functional limits of agreement for the right hip abduction angle from the motion capture study using a linear mixed-effects modelling framework. Here all the replicates from two different measurement sessions were used. MB is the marker-based and ML is the markerless method of measurement.	97
3.22	A: differences against averages ignoring the time frame for the right hip abduction angle curves. All the replicate measurements from two measurement sessions were considered here. B: residuals against averages for the same data after fitting a linear mixed-effects model. Here, MB is the marker-based and ML is the markerless method of measurement.	98
3.23	95% functional limits of agreement estimated by a linear mixed-effects modelling framework using two different basis systems for the random-effects regressor matrices. B-spline and eigenbasis systems were used for the random-effects regressor matrices. Here, MB is the marker-based and ML is the markerless method of measurement.	102
3.24	A: differences against averages ignoring the time frame for the right hip abduction angle curves. All the replicate measurements from two measurement sessions were considered here. B: residuals against averages for the same data after fitting a linear mixed-effects model with eigenbasis for the random-effects regressor matrices. Here, MB is the marker-based and ML is the markerless method of measurement.	104
4.1	Simulation strategy to compare the performance of both the functional data analysis (FDA) and the linear mixed-effects modelling (LMM) framework to calculate 95% functional limits of agreement (fLoA).	110
4.2	Observed and simulated difference curves. Here MB is the marker-based and ML is the markerless method of measurement.	111

4.3	95% pointwise envelope, 95% functional limits of agreement (fLoA) and 99.4% pointwise envelope that gives a 95% global envelope for the simulated population. Here MB is the marker-based and ML is the markerless method of measurement.	115
4.4	Simulation results to compare functional data analysis (FDA) and linear mixed-effects modelling (LMM) framework to calculate 95% fLoA. Panel-A, B, C show 95% fLoA calculated from a sample with the sample size 10, 50 and 100, respectively. Panel-D, E, F show coverage curves corresponding to samples of size 10, 50 and 100 respectively. 100 different random samples were used to calculate the coverage curve for any given sample size. Overall coverage (average over the domain) corresponding to each framework are shown in panel-D, E and F. Here MB is the marker-based and ML is the markerless methods of measurement. .	116
4.5	Observed and simulated data for studies with replicate measurements. Here MB is the marker-based and ML is the markerless method of measurement.	117
4.6	Simulation results to compare B-spline basis and eigenbasis for the random-effects regressor matrices using a linear mixed-effects modelling (LMM) framework to calculate 95% functional limits of agreement. Here, 100 different random samples were used to calculate the coverage curve for any given sample size. Overall coverage (average over the domain) corresponding to each framework are shown in panel-A, B and C.	118
4.7	Adjusted R-squared values after fitting a linear regression model to each individual difference curve using B-spline basis functions with different numbers of equally spaced inner knots for different angles measured in the motion capture study. Here different points on x-axis are different difference curve. For each angle, 54 difference curves (9 athletes, 2 sessions, 3 replicates) were used measured by two different measurement methods: markerless and marker-based.	122

4.8	Autocorrelation plot of the residuals after fitting a LMM of the difference curves with autoregressive moving average ARMA(2,1) error structure for different angles during a lunge from the motion capture study. . . .	124
4.9	Residuals against average plot after fitting a linear mixed model of the difference curves for different angles during a lunge from the motion capture study.	125
4.10	95% functional limits of agreement using a linear mixed-effects model for all angles measured during a lunge in the motion capture study. For each angle, all the replicate measurements in the two measurement sessions were used to calculate the 95% functional limits of agreement. Here MB is the marker-based and ML is the markerless method of measurement.	127

List of Tables

2.1	Different variance functions in <code>nlme</code> package in R	32
2.2	Name of the correlation functions available in the <code>nlme</code> package in R. .	36
2.3	SEM and ICC for the two measurement methods using two measurements per method.	42
3.1	Different candidate knot sequences for the random-effects structure. . .	84
4.1	Time needed to fit LMMs for the two basis systems used for the random- effects regressor matrix.	120
4.2	LMM specification for the difference curves by angles with the time taken to fit the model.	126

Chapter 1

Introduction

1.1 Introduction

An essential ingredient of all scientific studies is that the measurements of all variables, especially the response variable, are reliable in the sense that if they were measured again on the same subjects under identical conditions the measurements would agree (Newell et al., 2014).

The potential error for a measurement can be categorised into two types: systematic bias and a random component (Watson and Petrie, 2010). Systematic bias refers to an error which occurs in the same direction for all observations in a study. This error can happen, for example, due to a poorly calibrated measurement instrument or a flaw in the design of the study. Random error does not occur in a particular direction and it refers to natural variability within subjects as well as variability across observers, measuring apparatus etc. When the random error, possibly from multiple sources, is too large then the usefulness of the measurement method is in question and hence any conclusions from the study must be in doubt.

Studies where methods of measurement are compared are often referred to as method comparison studies (Altman and Bland, 1983). The main focus of this thesis is to extend the linear mixed effects modelling framework to analyse method comparison studies where a variable of interest is measured by two or more methods in order to determine the level of agreement between the methods. The challenge is to determine if these

methods of measurement agree, which ultimately leads to a decision on which method to use. The most common method comparison study is one where a new method is compared to a gold standard method, i.e. a reliable method already in use, to find out if the new method is an adequate alternative to the gold standard (Altman and Bland, 1983). Given their importance, studies assessing the level of agreement between two or more measurement methods are very common and are a well-studied research area in statistics.

One of the earliest references to method agreement using indices of reliability was in a review paper where Cochran (1968) showed the complexity of estimating measurement error in method agreement studies. In their paper, Cochran attempted to investigate the problem by proposing different statistical models for different situations to estimate the different variance components of the model using the analysis of variance (ANOVA) method. Bartko (1966) emphasised the usefulness of the intra-class correlation coefficient (ICC) as a measure of the reliability of a measurement method in the social sciences. There, the author, strongly recommended calculating the ICC based on a statistical model with the help of ANOVA to estimate the different variance components for the model. Shrout and Fleiss (1979) approached the use of ICC more systematically and proposed six different forms of the ICC with corresponding statistical models to assess the reliability of a measurement method. The authors also used ANOVA to estimate different variance components, using those estimates to estimate a set of possible ICCs. This work was so influential that still, to date, many research articles are being published to clarify what form of ICC should be used and in what situations. According to Google Scholar, the Shrout and Fleiss (1979) paper has received more than 23,000 citations and the ICC is still considered a valuable measure of reliability for a method of measurement. The reliability of a measurement method can be quantified using the ICC, but appropriate indices were needed in order to assess the reliability of a method in the same unit of measurement as the response, namely the standard error of measurement (SEM) (Weir, 2005).

For a valid measurement method to be usable it must be reliable. However, a different question is whether an existing method is replaceable or interchangeable with

a new method. To assess the agreement between methods of measurement, Altman and Bland (1983) proposed the use of 95% limits of agreement (95% LoA). Bland and Altman criticised the use of the mean comparison and the Pearson correlation coefficient as analytical tools in method comparison studies (Altman and Bland, 1983). Lee et al. (1989) suggested using the ICC to measure the agreement between two methods which was later criticised by Bland and Altman (1990). However, they acknowledged the appropriateness of ICC to measure the reliability of a method of measurement. Carstensen et al. (2008) proposed estimation of variance components to generate limits of agreement using a linear mixed-effects model (LMM) framework and argued that this provides an elegant estimation of the 95% LoA when dealing with replicate measurements.

To measure agreement among the observations from the same measurement method or measurements from two different methods of measurement, Lin (1989) proposed the concordance correlation coefficient (CCC), a correlation based index to assess the agreement based on how close the pairs of measurements are from the line of equality. Carrasco and Jover (2003) proposed a LMM approach to estimate CCC which allows for more than two methods of measurement and the adjustment of confounding variables. Oliveira et al. (2018) extended the estimation of CCC for longitudinal study design using a LMM.

Rather than estimating any quantity to assess agreement, Roy (2009) proposed a statistical test based approach to formally assess agreement for two methods of measurement. Roy described that there could be three source of disagreement: bias between the two methods of measurement, differences in the between-subject variability, and differences in the within-subject variability estimated by the two different methods. A LMM approach was proposed to conduct statistical tests to provide evidence of the existence of significant bias, a difference between between-subject variability and a difference between within-subject variability between two methods of measurement.

Most of the methodological developments in this research area described above have focused on studies where the response variable of interest is a continuous measure. More recently there has been interest on method agreement studies where the response

is functional in nature (i.e. discrete realisations of an underlying functional form) (Røislien et al., 2012; Olsen et al., 2013). This is still a new and growing field of research in statistics with many open questions.

The specific focus of this thesis is on how best analyse method agreement studies when the

- observed bias is non-linear between the two methods of measurement;
- response variable is functional in nature;
- study design is hierarchical involving replicates.

The work presented in this thesis addresses these open questions and new analytical approaches are given for each.

1.2 Datasets

There are two case studies used throughout this thesis. The first study involves the comparison of a laboratory test and a point of care (POC) test measuring c-reactive protein (CRP), a blood-based biomarker, longitudinally over time on a sample of elite soccer players. The response variable, CRP measurement, for this study is a continuous univariate response. The second study involves the comparison of two motion capture systems for measuring human movement involving a functional response.

The background, question of interest and numerical and graphical summaries for each case study are as follows.

1.2.1 C-Reactive Protein (CRP) Analyser Study

Soccer match play is characterised by performance of prolonged and high intensity intermittent movements (Souglis et al., 2015). These strenuous and multi-directional movements require repeated eccentric muscle contractions, inducing skeletal muscle fibre damage, and include numerous alterations in the immune system leading to inflammation similar to the acute phase inflammatory response to infection or injury

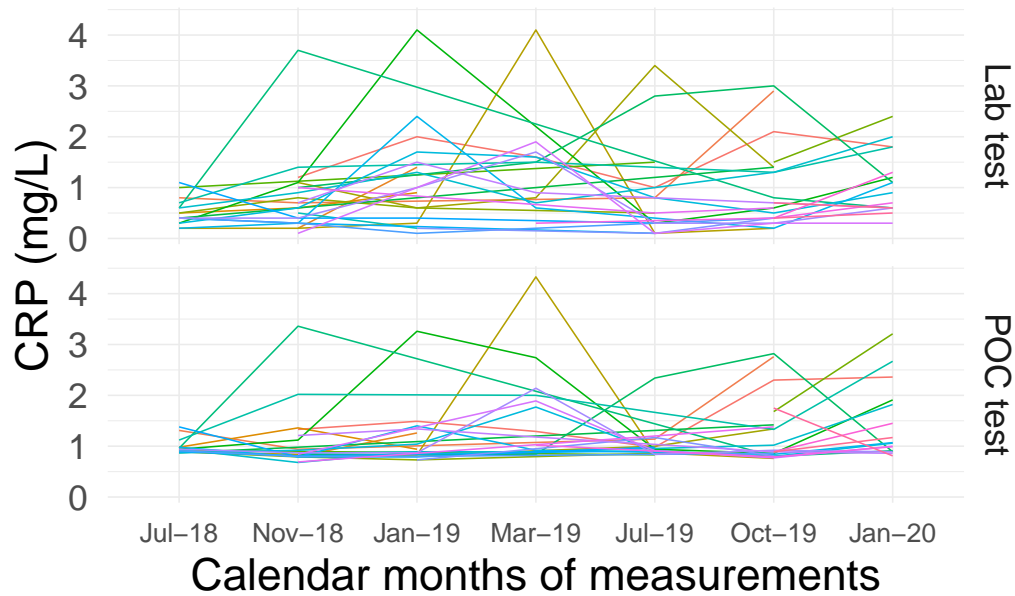


Figure 1.1: Profile plot of c-reactive protein (CRP) measurements over seven different measurement occasions by two different methods of measurement from 35 athletes. These two measurement methods are a laboratory test (lab test) and a point of care test (POC test). A separate colour is used to represent each athlete. Different athletes provided a different number of measurements as not all were available at each testing time point. The minimum, median, and maximum number of measurements per subject per method are one, three and seven, respectively.

(Fehrenbach and Schneider, 2006). As a result, a soccer match produces significant increase in inflammatory response markers, producing an acute performance deterioration in elite soccer players (Ispirlidis et al., 2008).

Repeated exposure to acute inflammatory episodes and limited periods of rest throughout a season places athletes at high risk of under recovery. The monitoring of inflammatory response in elite athletes may therefore be a medical and performance objective for protecting players' health, recovery and adaptation.

The measurement of c-reactive protein (CRP) is widely used to monitor inflammatory states of athletes (Du Clos, 2000). Longitudinal studies examining biomarkers of inflammation in athletes are rare as frequent collection and transportation of blood samples to a laboratory is challenging.

Using a point of care (POC) test of a capillary blood sample may present a practical solution to such problems for medical and performance staff. The simple application,

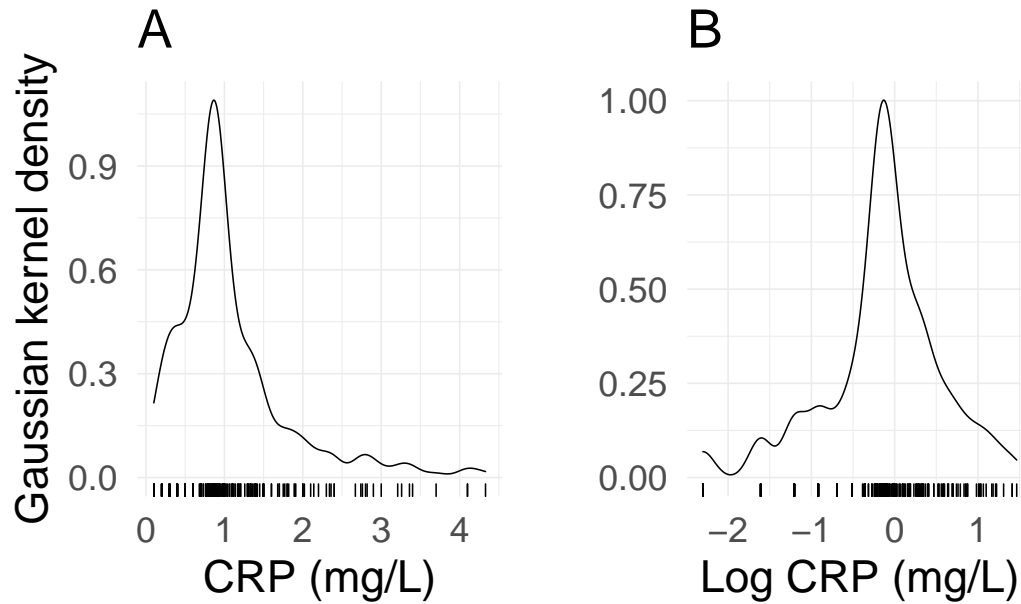


Figure 1.2: Gaussian kernel density plot for c-reactive protein (CRP) and log CRP measurements. Here measurements from both the laboratory test and a point of care test are combined to create the density plot.

small sample volume, and rapid result reporting indicate that there are a number of practical advantages to POC testing over a laboratory test in the elite high performance environment. To use a POC test to reliably identify biomarkers of inflammation, the medical and performance staff will require evidence that the laboratory test and a POC test are comparable for the measurement of CRP.

The aim of this study was to investigate the level of agreement between the two methods of measurement, the standard laboratory method and a POC test, for CRP obtained from professional soccer players with physiological concentrations below 5 mg/L, over the course of three English Premier League (EPL) seasons.

The testing protocol was as follows:

POC Test: Whole blood capillary samples ($20\mu\text{L}$) were taken from each participant's ear lobe, and immediately analysed at room temperature according to the manufacturer's instructions. Inflammation was measured using an immunoturbidimetric high sensitivity CRP assay on the Cube-S POC analyser (Eurolyser Diagnostica GmbH, Salzburg, Austria) using a $20\mu\text{L}$ capillary sample.

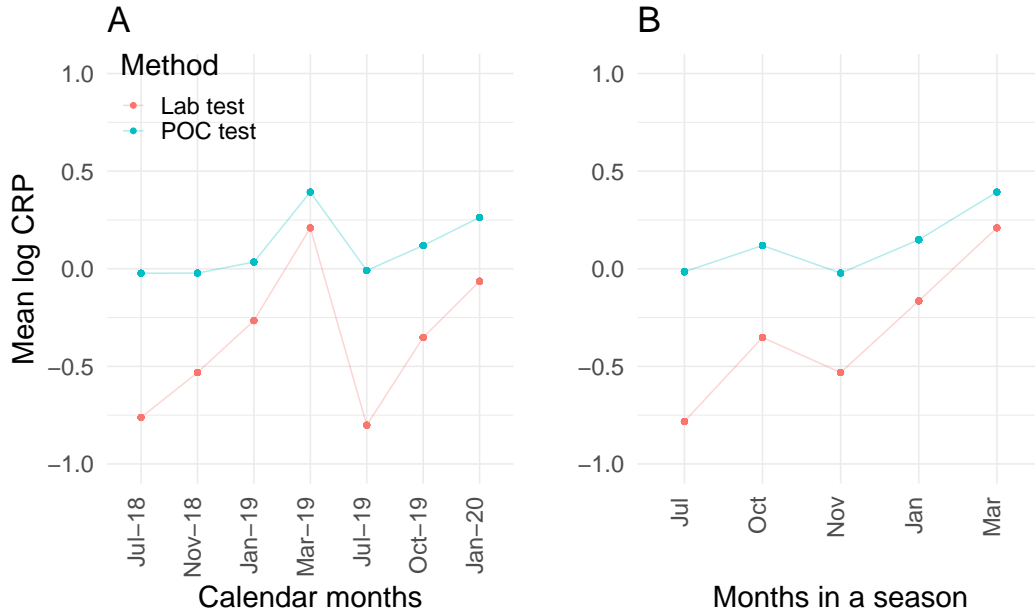


Figure 1.3: Mean log c-reactive protein (CRP) profile over calendar months and seasonal months. One mean profile is from the measurements by a laboratory test (lab test, in red) and the other is from a point of care (POC) test (in blue).

Laboratory Test: 1×5 mL venous blood samples were drawn via the antecubital vein in a serum separator tube for the analysis of CRP. Samples were immediately capped and transported to the laboratory on the morning of testing, for analysis on the same day. Serum CRP was measured using the CRP Vario test (ultra-sensitive method), a latex immunoassay (Architect System, Abbott laboratories, USA).

Numerical and Graphical Summaries

Thirty five professional soccer players were recruited for the study with age mean (SD): 25.3 (3.1) years; weight 75.2 (7.2) kg; height 182 (7.2) cm, body mass index (BMI) 22.7 (1.6), who were participating in a biomarker monitoring program over the course of three EPL seasons. Measurements were collected at seven different measurement occasions (Figure 1.1). Not all the 35 players provided measurements at each measurement occasion. Five players provided measurements in one measurement occasion, eight players provided measurements in two different measurement occasions, five players provided measurements in three different measurement occasions, five players provided measure-

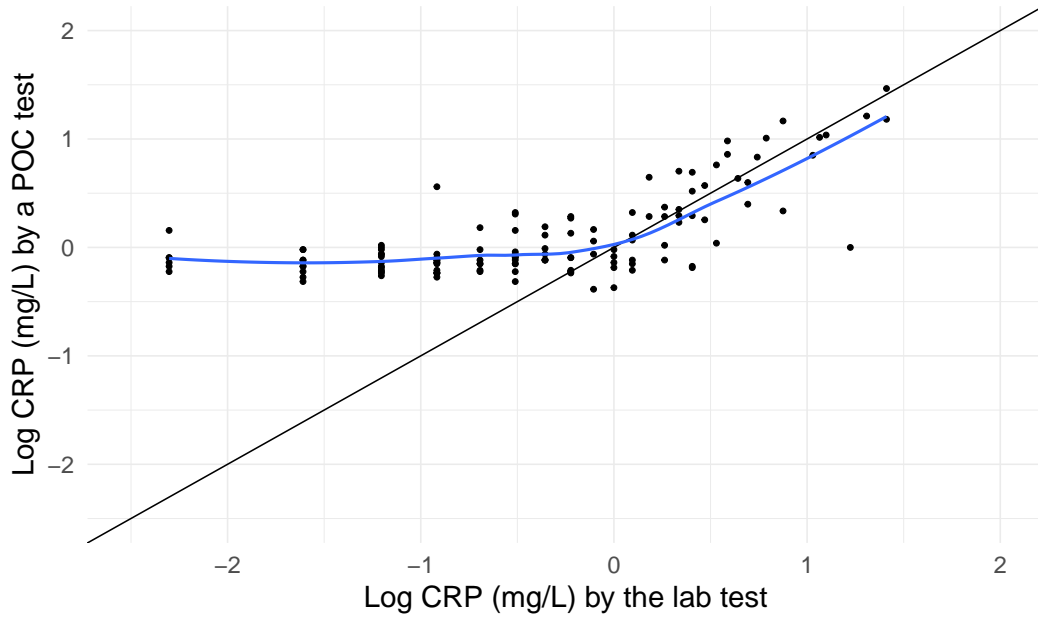


Figure 1.4: Scatter plot of log c-reactive protein (CRP) measurements by the laboratory test (lab test) and a point of care (POC) test with the line of equality (black line) and a locally estimated scatterplot smoothing (LOESS) line (blue line).

ments in four different measurement occasions, five players provided measurements in five different measurement occasions, four players provided measurements in six different measurement occasions, and only three players provided measurements in seven different measurement occasions. When a player provided samples, measurements of CRP using both the tests were obtained. For a given measurement occasion a minimum of 13 and maximum of 25 players provided samples. In total 256 CRP measurements were obtained using both the tests. The maximum value of CRP was 4.33 mg/L and the minimum was 0.10 mg/L. The mean and median CRP measurements were 1.07 mg/L and 0.89 mg/L, respectively.

A Gaussian kernel density plot of the CRP values combining the measurements from both the measurement methods suggests that the distribution is slightly skewed towards right (Figure 1.2 A). Since CRP values are non-negative this phenomenon of having a skewed right tail in the distribution is not surprising. Although the distribution is slightly right skewed, the normality assumption does not seem too unrealistic. However, a log transformed CRP values seems to provide a more symmetrical shaped distribution

(Figure 1.2 B).

The mean log CRP profile over calendar months displays that the mean profile is not constant over the different months (Figure 1.3 A). The mean value is the lowest around July and highest around March (Figure 1.3 B). This is quite natural as the EPL season runs from August to May. It is noticeable that the mean values of CRP by these two methods do not agree. There might be some relative bias between these two methods of measurement. Another interesting fact, also noticeable from the figure, is that the bias seems large when the mean CRP values are smaller, but the bias decreases as the mean value increases (Figure 1.3 B).

If there was perfect agreement then all the points of the scatter plot of the individual measurements by these two methods of measurement would fall on the line of equality. It seems that the agreement is good when the CRP values are larger and the agreement seems poor when the values are smaller (Figure 1.4). A locally estimated scatterplot smoothing (LOESS) line was drawn (blue solid line) to highlight the fact that there seems to be a non-linear bias between these two methods of measurement (Figure 1.4). Here a non-linear bias means that the magnitude of the bias between the two methods of measurement varies non-linearly across different magnitude of the CRP measurements.

1.2.2 Motion Capture Study

Functional competence is one of the core elements of the athletic development pathway used by elite soccer academies (Ryan et al., 2018). Functional competence refers to muscle flexibility, strength imbalances, and general movement pattern proficiency of athletes. Improvement in functional competence has been suggested for long-term development in athleticism for youth athletes to enhance fitness and prevent injuries (Lloyd et al., 2016). Assessment of functional competence provides a foundation for the progression of athletic development programs, injury-free performance in the game and career longevity of athletes (Lloyd et al., 2015; Chorba et al., 2010; Lloyd et al., 2016).

From a clinical and sports medicine standpoint, human movement analysis serves to quantify baseline movement quality which can be used to develop an athletic devel-

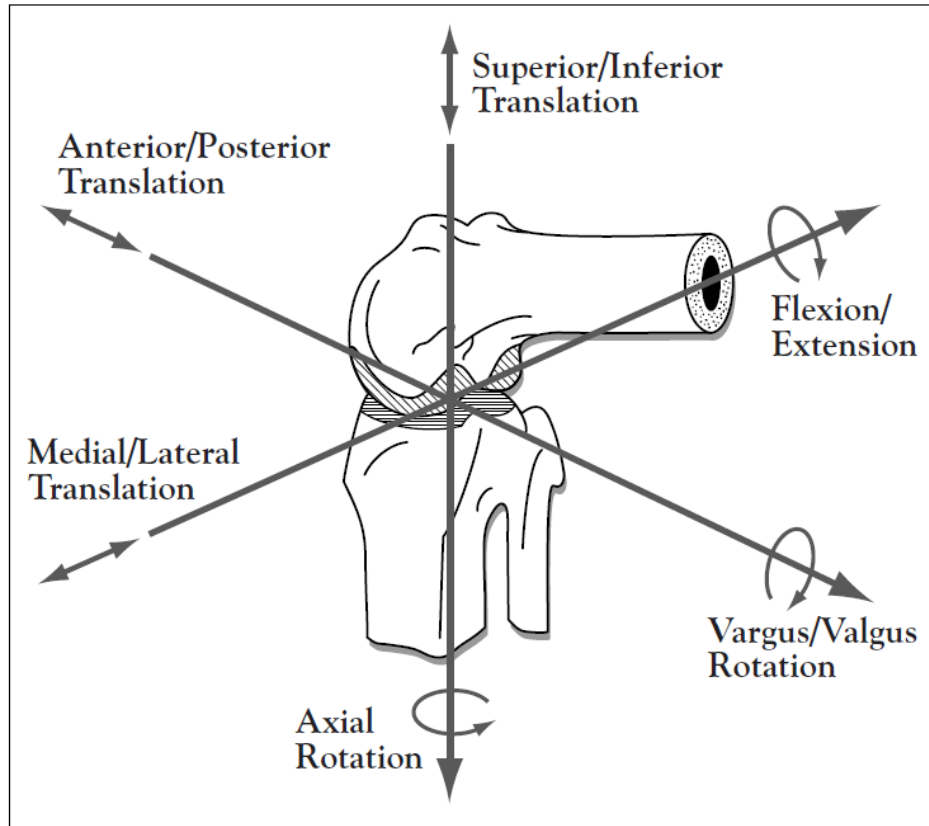


Figure 1.5: Six degrees of freedom for the knee joint, which include three rotational and three translational motions (Komdeur et al., 2002).

opment program suitable for a particular athlete to improve upon individual's weaknesses, ensuring a sustained improvement in movement capabilities (Hunter Bennett et al., 2020). Fundamental movement skills and athletic movement efficiency provide the foundation for progression of athletic development programs, tailored to develop strength and power essential for elite level, as part of a long-term athletic development approach.

Human movement analysis involves measurement of the position and orientation of the body segments during the execution of locomotion and physical exercises (Cappozzo et al., 1995). Kinematics, a branch of biomechanics, deals with the study of human motion that describes the position and orientation of a body segment relative to another with respect to time thus enabling quantification of motion without reference to the forces causing the motion (Cappozzo et al., 1995; Cereatti et al., 2017; Colyer et al., 2018; Robertson et al., 2013). The motion of the body segments may be represented

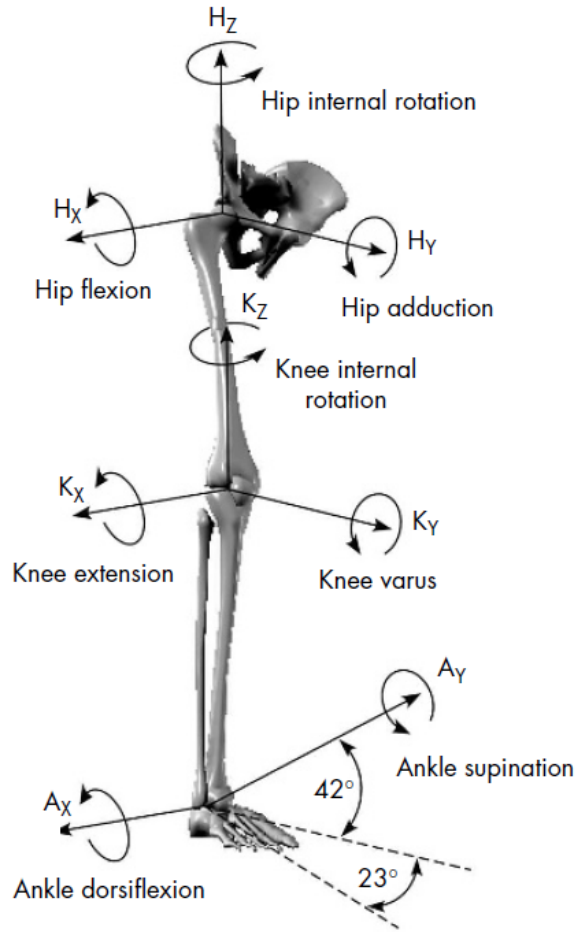


Figure 1.6: The XYZ axis for the kinematic model along with corresponding degrees of freedom (McLean et al., 2005).

with up to six degrees of freedom (6DOF): three translations and three rotations (Žuk and Pezowicz, 2015; Robertson et al., 2013) .

These three rotational angles are often measured with Cardan (Euler) sequence of angles (Glasoe et al., 2014). Cardan sequence of angles are defined as ordered rotations about 3 axes and thus represent the body segments' motion in the anatomical body planes: frontal plane, sagittal plane, and transverse plane. These three axes are denoted by X, Y, and Z. The X-axis is the medial/lateral axis parallel to the floor and perpendicular to the frontal plane, the Y-axis is the anterior/posterior axis perpendicular to the sagittal plane, and the Z-axis is the vertical axis perpendicular to the transverse plane (Figure 1.5). Flexion/extension motion occurs about the X-axis,

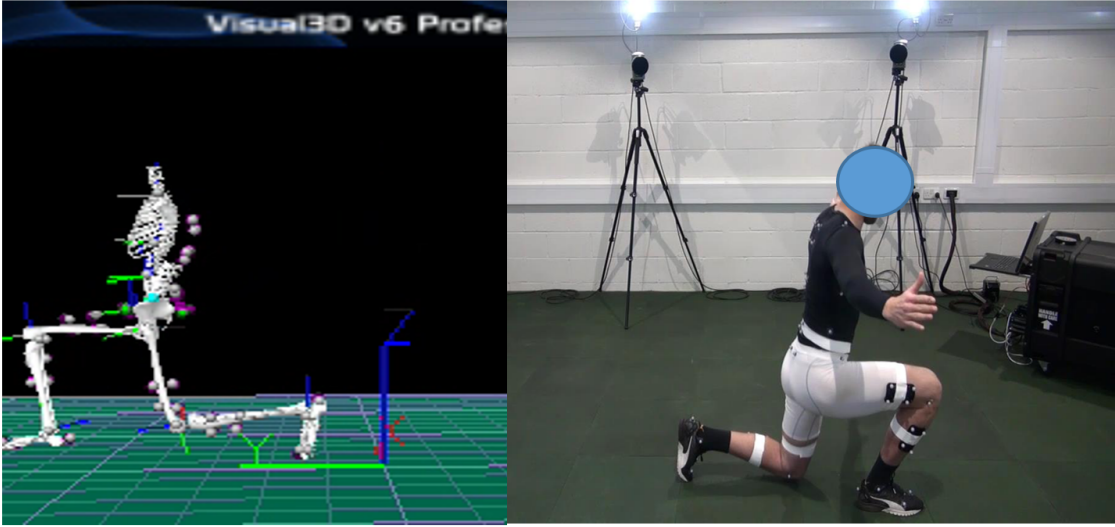


Figure 1.7: Right leg lunge exercise used to concurrently compare the markerless and marker-based systems.

abduction/adduction occurs about the Y-axis, and medial/internal rotation and lateral/external rotation occur about the Z-axis (Figure 1.6). Among the six degrees of freedom, only the flexion, abduction, and rotation angle were collected in the original study and are the data under consideration in this thesis (Coyne, 2021). The response represent the angles that were measured during a lunge exercise and hence represent discrete realisations of an underlying continuous curve which is a function of time (i.e. camera capture frame) (Figure 1.7).

A marker-based motion capture system is considered a gold standard for kinematic measurements such as flexion, abduction, and rotation angle (Ceseracciu et al., 2014). This system, using reflective markers on bony landmarks to define anatomical joint centres, collects data on the trajectory of the reflective markers in space and with respect to the global co-ordinate system (Cappozzo et al., 2005). The trajectory of these markers are simultaneously captured by multiple cameras to determine the motion of the 3D coordinates of a particular body segment.

Markers are placed on anatomical landmarks according to the recommendations made by the International Society of Biomechanics (ISB). Marker-based systems are considered the optimal laboratory based method to measure kinematics data with accuracy between 3-5 degrees for most lower extremity segments (McGinley et al., 2009;

Miller et al., 2016). However, the application of markers to anatomical bony points is a time consuming process and constrained by inter-, intra-tester and session reliability issues (McGinley et al., 2009; Rácz et al., 2018; Borhani et al., 2013).

The emergence of novel markerless motion capture systems without the use of reflective markers is a rapidly growing field with an attractive future advancement in motion analysis (Colyer et al., 2018; Yang et al., 2014). A markerless motion capture system offers a fully automatic, non-invasive, markerless approach, which would ultimately provide a major breakthrough for research and practice within sports biomechanics and rehabilitation. This new technology uses multiple synchronised video cameras surrounding a capture space. If it is deemed reliable it could remove the difficulty of quantitatively assessing movement quality in elite academy footballers and ultimately have wider scope in scientific research.

The purpose of this motion capture pilot study was to explore the level of agreement of a commercially available “DARI” markerless system with an established gold standard marker-based system, for the concurrent evaluation of lower quadrant kinematics during a lunge motion in an elite youth football population.

Nine elite soccer players (n=9) with mean age (SD) 18.5 (1.3) years, height 1.83 (0.04) metres and weight 79.2 (6.2) kg participated in the study. All nine players reported their right leg as the dominant kicking leg. All participants wore tight fitting clothing during the measurement with the same footwear recommended by the standardised testing protocols. To reduce extraneous marker movement, tight lycra clothing was essential during the measurement.

All the participants were full time academy football players for a EPL club for over a minimum of two years and involved in supervised strength training. Participants had achieved full maturation status or 100% of peak adult height at the time of testing and were able to complete the lunge exercise without any physical restriction.

An experienced musculoskeletal physiotherapist with over 20 years of experience placed all the markers on the participants under the guidance of experienced biomechanists with over 15 years of motion capture experience. The position and orientation of body segments was defined according to the recommendations of the IBS (Wu et al.,

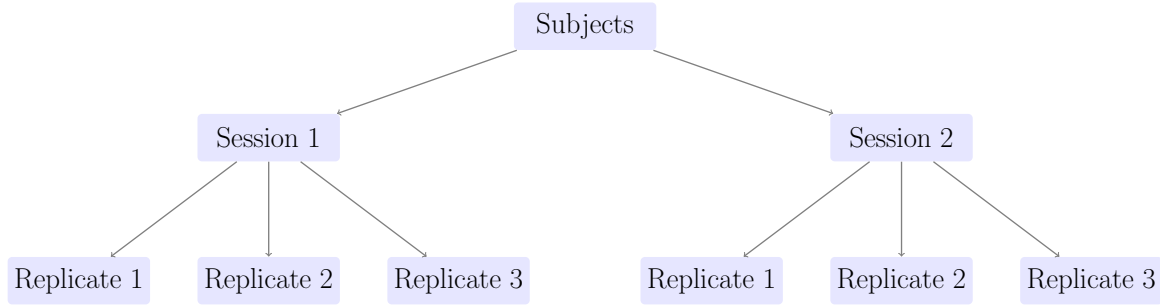


Figure 1.8: Hierarchical study design for the motion capture study.

2002).

Kinematics data were collected using both the marker-based and markerless motion capture system simultaneously while all the participants performed the lunge exercise (Figure 1.7). Three repetitions of lunge on two separate testing occasions, one week apart, were collected (Figure 1.8). All participants familiarised themselves with both movements prior to data collection under the supervision of a chartered physiotherapist and the same verbal and visual demonstration was used to instruct the participants in both the occasions. The details of the study and study design can be found elsewhere (Coyne, 2021).

The time domain of the measurements for both systems were time normalised to a ‘frame 0’ starting point and ‘frame 100’ end point. This time domain is referred to as ‘normalised’ time or ‘time frame’ throughout the thesis. The lunge datasets were time normalised to 101 data points including the start and end of the exercise. The start of the lunge was defined as the first peak of right knee flexion as participants initially lifted their right leg up, and the end of the lunge was defined as the final peak in knee flexion as participants bend their knee prior returning to the start position.

For each performance of a lunge, three different body segments were targeted for measurement: low spine, right hip, right knee. For each of the body segment two motion capture systems were used simultaneously to measure three functional responses: flexion, abduction, and rotation angle curve.

Although the data for this motion capture study are obtained directly from the markerless system feed a considerable amount of data processing was required before any analyses could be carried out. As the start of the lunge was defined on the first

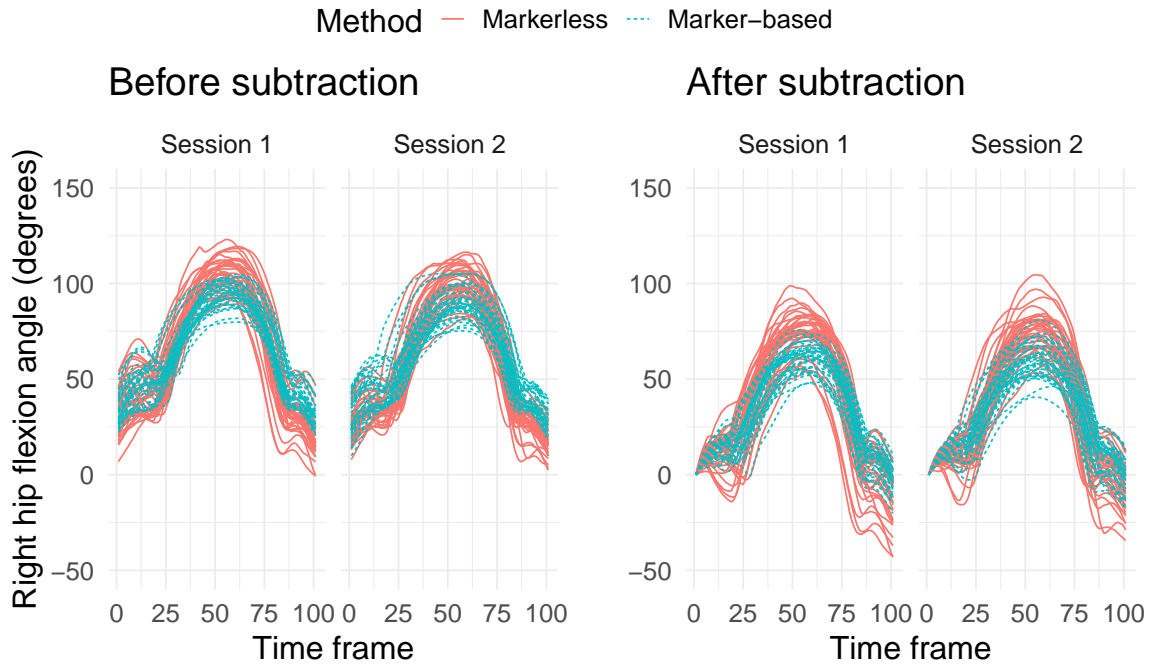


Figure 1.9: Transformation of subtracting the initial angle from the whole curve. This transformation was done to all the functional responses. For the purpose of demonstration only right hip flexion angle is presented here.

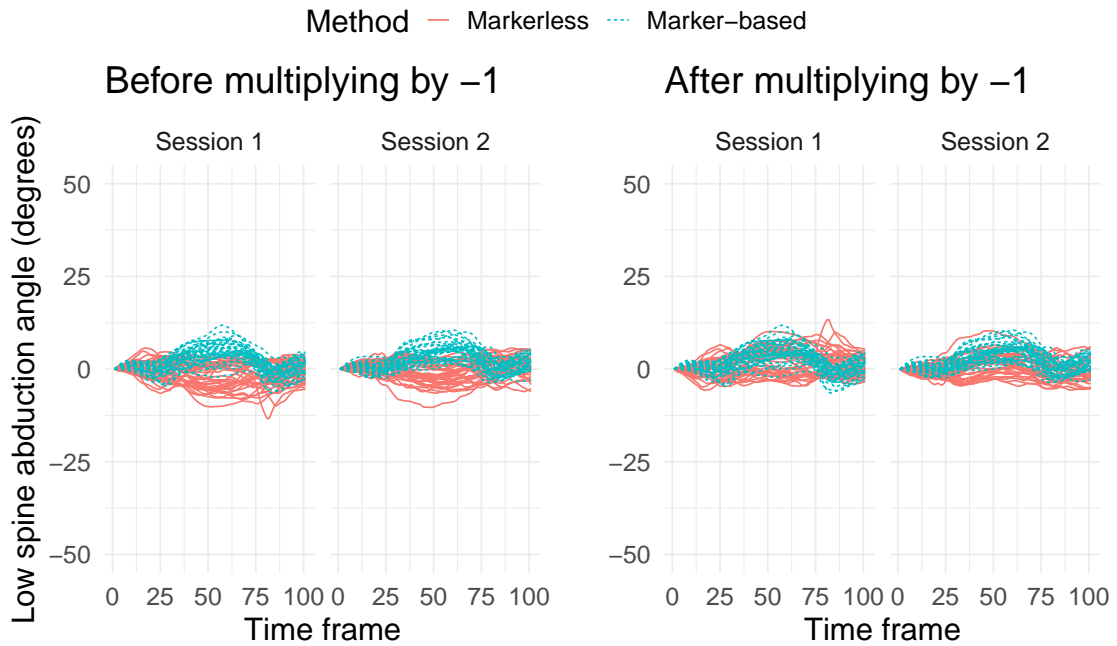


Figure 1.10: Multiplying by -1 to the measurements made by the markerless motion capture system to align with the direction of positive/negative angle of the marker-based system for the low spine angle during a lunge.

peak of right knee flexion, the initial angle for different athletes were not the same. Based on the recommendation of a biomechanist involved in the original study, the initial angle was subtracted from the whole curve for each of the responses. This new angle is known as range of motion (ROM) in the biomechanics literature. This subtraction of the initial angle was done for every angle recorded in the study, however for the purpose of demonstration only the right hip flexion angle during a lunge is demonstrated (Figure 1.9).

As the measurements of angles were taken using two different methods of measurement, the direction of the positive angle and negative angle were not similar for low spine body segments. For this reason all measurements of low spine angles by the markerless method were multiplied by -1 to align them and make them comparable to the direction of the positive angle of the marker-based method (Figure 1.10).

Graphical summaries provide a useful description of the data. For the right leg lunge exercise three body segments were considered, i.e. low spine, right hip and right knee. For each body segment, responses were measured in three different directions: flexion, abduction and rotation angle curve. These three directions and three body segments constitute nine different scenarios. The shape of the functional responses and their variability over the domain are different for different scenarios (Figure 1.11). It appears that agreement between the two method of measurements is different for the different directions at different body segments. For example, agreement between the method of measurement for the flexion angle at right hip (Figure 1.12) seems better compared to the agreement for abduction and rotation (Figure 1.13 and 1.14). For this reason assessment of agreement between the two methods of measurement will be investigated separately for each of the different scenarios.

For each athlete, 108 functional responses were measured for a lunge exercise alone. For a given scenario, e.g. flexion at right hip abduction, 12 functional response were measured. These 12 functional responses are three replicates from two different sessions by two different methods of measurement. There were 9 different scenarios considered for a lunge (Figure 1.11).

Now that the two case studies have been introduced and the questions of interest

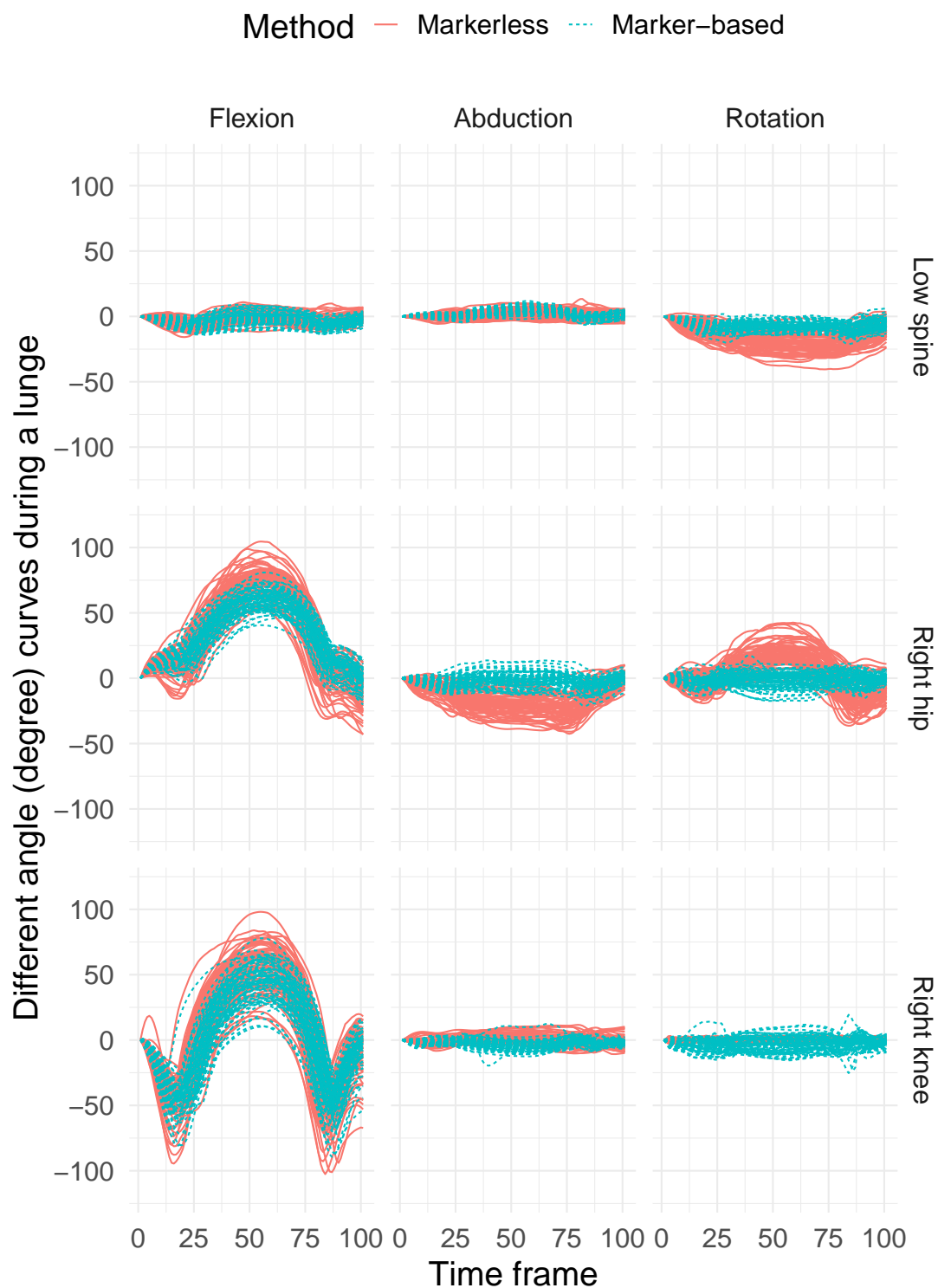


Figure 1.11: Shape of different functional responses for different directions at different body segments measured by two different methods of measurement during a right leg lunge.

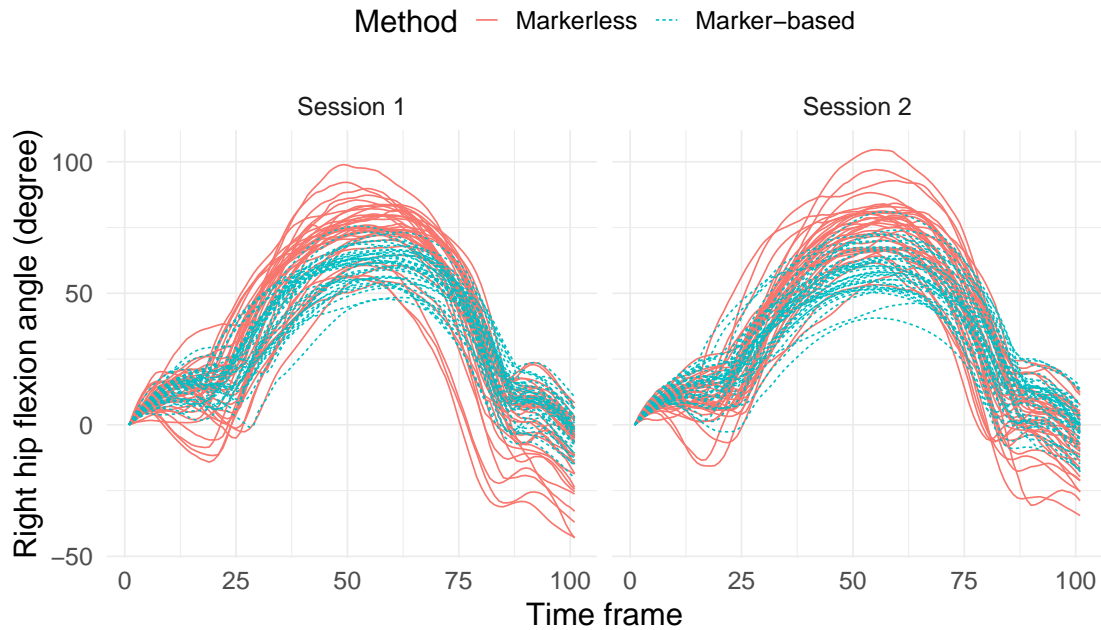


Figure 1.12: Functional responses measured by a markerless and a marker-based motion capture system for right hip flexion during a lunge.

described, a summary of how these problems will be addressed is given by way of an overall summary of the structure of the thesis.

1.3 Structure of the Thesis

In this chapter a brief introduction was presented to identify important open research questions in method comparison studies. A method comparison study can involve a univariate or a functional response depending on the intrinsic nature of the context and the measurement methods involved. Two datasets from two different method comparison studies have been described. One of these studies collected a univariate response longitudinally and the other collected a functional response. These case studies will be used throughout the thesis to demonstrate the statistical approaches proposed.

Chapter 2 focuses entirely on univariate continuous responses in method comparison studies. Different statistical models for a univariate continuous response in method comparison studies are discussed. Methods to calculate indices of reliability for a measurement method using the statistical models are presented. The concordance corre-

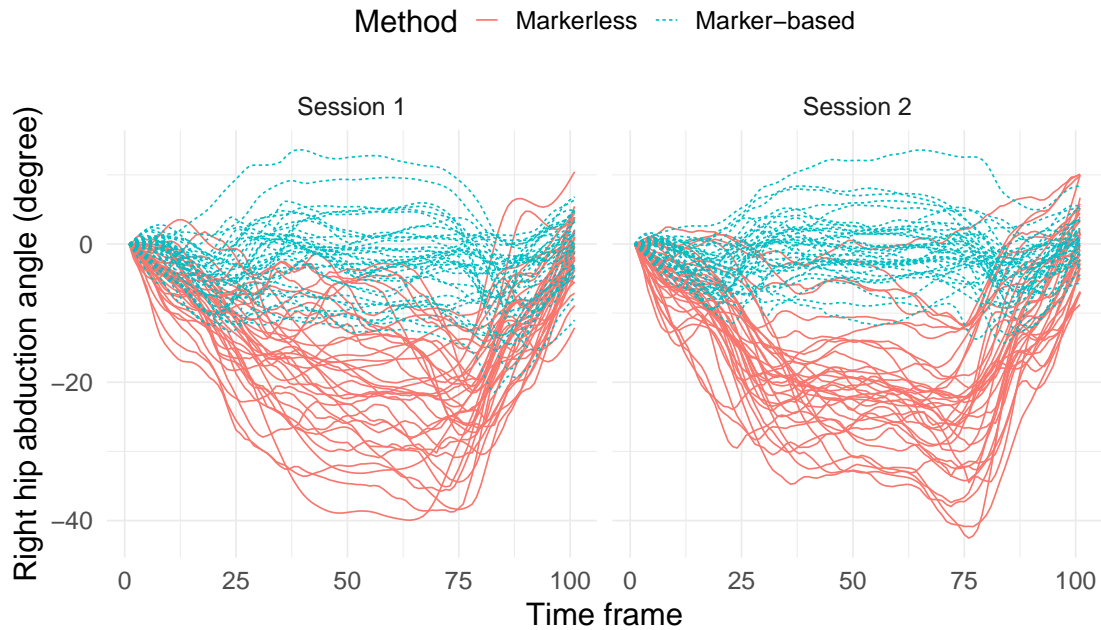


Figure 1.13: Functional responses by a markerless and a marker-based motion capture system for right hip abduction during a lunge.

lation coefficient and 95% limits of agreement to assess the agreement between two or more methods of measurement are introduced and discussed. A novel graphical approach to identify non-linear bias between two methods of measurement in method comparison studies is proposed. A novel statistical approach is presented which involves an extension of the linear mixed-effects modelling framework to calculate 95% limits of agreement in the presence of non-linear bias between two methods of measurement. The proposed graphical approach and statistical approach for analysis are demonstrated using the CRP analyser study.

Chapter 3 deals entirely with functional responses in method comparison studies. A description of a functional response is given including a literature review of the use of functional responses in method comparison studies. An introduction to functional data analysis is then given with the focus on method comparison studies. Following this, indices of reliability for a method measuring a functional response is discussed. Estimation methods for 95% functional limits of agreement for simple method comparison studies using functional data analysis are discussed, followed by a discussion of nonparametric linear mixed-effects models as an alternative approach. This chap-

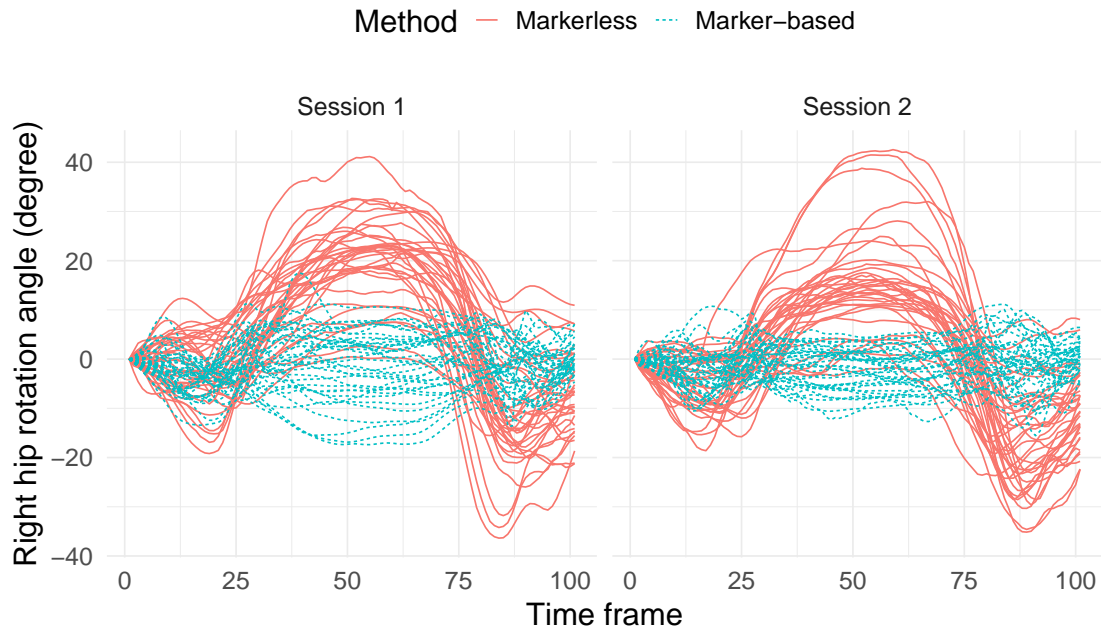


Figure 1.14: Functional responses by a markerless and a marker-based motion capture system for right hip rotation during a lunge.

ter proposes a novel statistical approach by extending linear mixed-effects modelling framework to calculate 95% functional limits of agreement in simple method comparison studies and more complicated studies with replicates and hierarchical study designs. A novel graphical approach is presented as a ‘functional’ equivalent of the Bland-Altman plot for univariate responses. The computational cost while using a nonparametric linear mixed-effects model is discussed and a novel and time efficient computational approach to calculate 95% functional limits of agreement is proposed.

A simulation study is presented in Chapter 4 in order to guide the appropriate interpretation of 95% functional limits of agreement in practice. The performance of the proposed LMM extension and the use of Functional Data Analysis, in terms of coverage, when calculating 95% functional limits is evaluated under different scenarios. Initially data were simulated from simple method comparison studies followed by more complicated studies with replicates. In this Chapter, a comparison is made of the performance and computational costs when calculating 95% functional limits of agreement using two basis systems for the random-effects regressor matrix in nonparametric linear mixed-effects models. This chapter concludes with an application of the proposed linear

1.3. Structure of the Thesis

mixed-effects modelling framework to analyse the motion capture case study.

The thesis concludes with Chapter 5 where an overall summary of the work presented in the thesis, a discussion of the results of the two case studies and areas of further work.

Chapter 2

Method Agreement Studies involving a Univariate Response

2.1 Introduction

In this chapter statistical approaches for analysing method comparison studies with a univariate continuous response are described. The chapter begins with a brief discussion of different statistical models where the aim is to identify the different components in a measurement. An introduction to linear models is presented as an approach to estimate the model parameters of the simple statistical models followed by a discussion of how the linear mixed modelling framework is a more attractive choice for more complicated statistical models. Indices for reliability are then introduced and approaches to estimate these using the statistical models are outlined. The role of the concordance correlation coefficient and the 95% limits of agreement to assess the agreement between two or more measurement methods is discussed. The chapter concludes with an example of how the linear mixed-effects modelling framework can be used to calculate limits of agreement in the presence of non-linear bias between two methods of measurement.

2.2 Statistical Models for Measurements

A wide variety of measurement methods are used in scientific studies. For example, wright peak flow meter measures peak expiratory flow rate while a turbidometric assay measures the concentration of Vancomycin, a tricyclic glycopeptide antibiotic, in premixed parenteral solution (Bland and Altman, 1986; Lawrence and Lin, 1992). Due to the intrinsic nature of a quantity being measured, the measurement that a method produces can be categorised into four different scales: nominal, ordinal, interval, and ratio scale. The nominal and ordinal scales are often referred to as a categorical scale and a continuous scale as an interval or ratio scale. Statistical techniques that assess the properties of a measurement method depend on these measurement scales. This thesis will focus only on those measurement methods that produce measurements on a continuous scale.

All measurement methods that produce values on a continuous scale are not necessarily similar in nature. For example, a measurement method could be a medical device or it could be a human. When humans are measurement methods, they are often referred to as raters, judges, or observers. When a measurement method is a human or a device, the statistical techniques for assessing the quality of measurements do not vary, but careful attention should be given when estimating assessment indices as the sources of variability might be different.

When a measurement method involves a continuous response, the response can be decomposed into two parts: the true value being measured and some random error. For a specific observation on a given subject, a measurement by a measurement method can be represented as follows:

$$Y = \tau + \epsilon \tag{2.1}$$

where τ is the true value being measured, and ϵ is the random error, generally, assumed as $\epsilon \sim N(0, \sigma_\epsilon^2)$. If the true value changes from time to time for a given subject, then it can be considered a random variable. If the true value does not change, it is a fixed quantity. For example, the blood pressure of an individual changes over

2.2. Statistical Models for Measurements

time, but the height of an adult remains constant for a long period. Assuming the quantity being measured does not change over time, the τ in model (2.1) is a fixed quantity. If a method measures the true value exactly, then the random error will be zero. However, it is rare to find such a method. The measurement error ϵ in model (2.1) is due to the measurement method. This error may not be entirely due to the measurement method in some situations. For example, how a technician administers a medical device to measure a quantity may introduce some error in the measurement, which is not due to the measurement method (Cochran, 1968). Assuming there are no external influences on the measurement in model (2.1), the random error ϵ is entirely due to the measurement method.

Model (2.1) may contain other terms. If there is a systematic bias due to the measurement method, then the model (2.1) should contain one more term for the systematic bias. In a situation where the true value being measured changes over time, a random variable should be added to model (2.1) to incorporate this. If the measurement method is changed, the random error ϵ may change.

In a situation where the measurement method is being used to measure a quantity on different subjects, not on a specific subject, then the model (2.1) becomes

$$Y_i = \tau_i + \epsilon_i \tag{2.2}$$

where, Y_i is the measurement made on the i^{th} subject by a measurement method, τ_i is the true value of the quantity being measured for the i^{th} subject. Now, τ_i is no longer a fixed value, it is a random variable. It is also useful to assume that $\tau_i \sim N(\mu, \sigma_\tau^2)$, where μ is the true average value in the population. For the model (2.2) to be identifiable, it is assumed that τ_i and ϵ_i are uncorrelated random variables, i.e. $\text{cov}(\tau_i, \epsilon_i) = 0$. Model (2.2) can be re-written as follows:

$$Y_i = \mu + \alpha_i + \epsilon_i \tag{2.3}$$

where, α_i is the deviation from the average value μ for the i^{th} subject assumed to be sampled from $\alpha_i \sim N(0, \sigma_\alpha^2)$ with $\text{cov}(\alpha_i, \epsilon_i) = 0$.

2.2. Statistical Models for Measurements

Let Y_{mi} be the measurement made by method m on the i^{th} subject. Y_{mi} can be written as follows:

$$Y_{mi} = \mu + \alpha_i + \beta_m + \epsilon_{mi}. \quad (2.4)$$

In addition to the components in the model (2.3), β_m in this case represents the systematic bias due to the m^{th} method, and ϵ_{mi} the method specific random error term with $\epsilon_{mi} \sim N(0, \sigma_{\epsilon_m}^2)$. The random error for each of the methods could have different standard deviations. Notice that in model (2.4) the true value of the i^{th} subject remains the same, but the measurements produced by each of the methods may or may not be the same due to the systematic term β_m and random error term ϵ_{mi} .

When replicates measurement are taken then model (2.4) becomes:

$$Y_{mij} = \mu + \alpha_i + \beta_m + \epsilon_{mij}. \quad (2.5)$$

where Y_{mij} is the j^{th} measurement for the i^{th} subject by the method m and ϵ_{mij} is the random error of j^{th} measurement for the i^{th} subject by the method m .

The measurement variability for Y_{mij} in model (2.5) comes from two different sources. The first source is σ_{α}^2 and the second source is $\sigma_{\epsilon_m}^2$. For a given subject, α_i is fixed, so the measurement varies only due to $\sigma_{\epsilon_m}^2$. Since ϵ_{mij} varies within a given subject, $\sigma_{\epsilon_m}^2$ is called the *within-subject variability*. Note that in this situation, the within-subject variability only includes ϵ_{mij} . In other situations where the true value being measured changes, the within-subject variability may contain other sources. For example, when measuring blood pressure the true value of the quantity changes at different points of time. In this situation the within subject variability constitutes of the variability of the true value and the measurement error. The α_i is fixed for a given subject, but it varies from subject to subject, and hence σ_{α}^2 is called the *between-subject variability*.

If a subject-method interaction term is added in model (2.5), then the model becomes:

$$Y_{mij} = \mu + \alpha_i + \beta_m + \gamma_{mi} + \epsilon_{mij}. \quad (2.6)$$

2.3. Introduction to Linear Models

where, γ_{mi} is the interaction term between the i^{th} subject and m^{th} method with $\gamma_{mi} \sim N(0, \sigma_{\gamma_m}^2)$. This will allow the interaction between different magnitude of the true value with a measurement method to be modelled.

Model (2.6) can be expressed in the following way when separate estimate of α_i and γ_{mi} are not possible or not necessary as follows:

$$Y_{mij} = \mu + \alpha_{mi} + \beta_m + \epsilon_{mij}. \quad (2.7)$$

where, α_{mi} is the combined term of the subject-specific random deviation from the average (α_i) and subject-method interaction term (γ_{mi}) with $\alpha_{mi} \sim N(0, \sigma_{\alpha_m}^2)$.

2.3 Introduction to Linear Models

In section 2.2, different statistical models were described, all of which are variations of linear models. The term ‘linear model’ contains a broad category of model, however, only linear regression models and linear mixed-effects models will be considered in this thesis as they are best suited for the type of measurement data under consideration.

2.3.1 Linear Regression Models

Assume \mathbf{y} is a vector of responses from all n observations in a given sample, and p covariates form the columns of the matrix \mathbf{X} , in addition to the first column which contains all 1s. The linear regression model can be written (in matrix notation) as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \mathbf{I}\sigma^2) \end{aligned} \quad (2.8)$$

where $\boldsymbol{\beta}$ is the vector of coefficients, \mathbf{I} is the identity matrix and $\boldsymbol{\epsilon}$ is the vector of errors.

The first aim is to find $\hat{\boldsymbol{\beta}}$, which is an estimate of the unknown parameter $\boldsymbol{\beta}$, typically using *ordinary least squares*(OLS) or *maximum likelihood* (ML) estimation procedures (Freedman, 2009).

2.3. Introduction to Linear Models

The OLS estimate of $\boldsymbol{\beta}$ is obtained by minimising the following quantity (Ruppert et al., 2003):

$$\|\boldsymbol{\epsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad (2.9)$$

where $\|\mathbf{a}\|$ is the standard norm of any vector \mathbf{a} . It can be shown that the estimate of $\boldsymbol{\beta}$ that minimises the above quantity is (Ruppert et al., 2003)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (2.10)$$

Under the condition that errors are uncorrelated to each other and have equal variance, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimate (BLUE) of $\boldsymbol{\beta}$ (Freedman, 2009). Although an estimate of σ^2 is not a part of the OLS estimation procedure, the following estimate is used (Ruppert et al., 2003):

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n - p} \quad (2.11)$$

For ML estimation it is assumed that all the observations are independently and identically distributed normal variates. ML estimation estimates both the $\boldsymbol{\beta}$ and σ^2 parameters by maximising the following likelihood function (Seber and Lee, 2012):

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) = \prod_{i=1}^n p(Y_i | \boldsymbol{\beta}, \sigma^2) \quad (2.12)$$

where $p(\cdot)$ is the probability density function, \mathbf{y} is a n -dimensional response vector, Y_i is the response for the i^{th} observation, and $L(\cdot)$ is the likelihood function. It can be shown that the estimate of $\boldsymbol{\beta}$ and σ^2 that maximises the function (2.12) is as follows (Seber and Lee, 2012):

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y} \\ \hat{\sigma}^2 &= \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n}. \end{aligned} \quad (2.13)$$

Note that in equation (2.13) for the estimator of σ^2 , the denominator is n not $n - p$.

It can be shown that the above ML estimate is a biased estimate of σ^2 (Verbeke and Molenberghs, 2000). Restricted maximum likelihood (REML) estimation remove this bias (Verbeke and Molenberghs, 2000).

The maximum likelihood estimator of σ^2 is biased for ML estimation since it does not account for the degrees of freedom lost to estimate the β in the model. In REML the loss of degrees of freedom is accounted using a set of error contrasts $\mathbf{u} = \mathbf{A}'\mathbf{y}$, where \mathbf{A} is any $n \times (n - p)$ matrix with $(n - p)$ linearly independent columns orthogonal to the space spanned by the columns of the design matrix \mathbf{X} . Now \mathbf{u} follows a Normal distribution with mean vector $\mathbf{0}$ and variance covariance matrix $\sigma^2\mathbf{A}'\mathbf{A}$. This \mathbf{u} is now independent of β . It can be shown that maximising the likelihood corresponding to the \mathbf{u} yields the following estimate of σ^2 (Verbeke and Molenberghs, 2000).

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}{n - p} \quad (2.14)$$

The estimator $\hat{\sigma}^2$ in equation (2.14) is called the REML estimator of σ^2 since it is restricted to $(n - p)$ error contrasts (Verbeke and Molenberghs, 2000). Note that estimation of β is not a part of the REML estimation (Verbeke and Molenberghs, 2000).

Linear regression is only appropriate when there is one measurement per method per subject. If multiple measurements are made by a given measurement method on the same subject, then the assumption that observations are independent of each other is violated and linear regression should not be used. LMM offers a modelling framework that can handle correlated responses. In the next section, an introduction to LMM is given.

2.3.2 Introduction to Linear Mixed Models (LMM)

A linear statistical model can be expressed as a combination of *fixed-effects* and *random-effects* (Pinheiro and Bates, 2000). Fixed-effects are related to the population parameters which remain fixed for the entire population (Pinheiro and Bates, 2000). On the other hand, random-effects are related to individuals drawn at random from the

population of interest (Pinheiro and Bates, 2000). For example, the relative bias between two measurement methods remains constant for a given population, and it can be considered a fixed effect. However, if there is any interaction between a measurement method with subjects, this interaction effect varies from subject to subject and can be considered a random effect. A linear model that incorporates both fixed-effects and random-effects is called a *mixed-effects* model (LMM) (Pinheiro and Bates, 2000).

Linear mixed-effects models are usually used to model the relationship between a continuous response variable and a set of explanatory variables where the observations are grouped according to one or more categorical variables (Pinheiro and Bates, 2000). *Longitudinal data, repeated measures data, clustered data, multilevel data, and hierarchical data* can be considered as *grouped data* (Pinheiro and Bates, 2000). In grouped data, experimental units under the same group are correlated. In a mixed-effects model, the correlation from the random effects corresponding to each factor is modelled flexibly through the variance-covariance matrix of the random-effects (Pinheiro and Bates, 2000). In addition, heteroscedastic and correlated errors can also be modelled under the same modelling framework through the variance-covariance matrix for the errors.

An introduction and computational details for fitting linear and non-linear mixed-effects models can be found in the book by Pinheiro and Bates (2000). A detailed description of applications of LMM for longitudinal data can be found in the book by Verbeke and Molenberghs (2000). Diggle et al. (2002) discussed LMM in the context of generalised linear models for longitudinal data.

LMM for Single Level Grouping

Consider \mathbf{y}_i as an n_i dimensional response vector of observations for the i^{th} group/subject. A mixed-effect model for this response can be written as:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, n, \\ \mathbf{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{2.15}$$

where n is the total number of groups, $\boldsymbol{\beta}$ is a p -dimensional vector of fixed-effects, \mathbf{b}_i is a q -dimensional vector of random-effects, \mathbf{X}_i (of size $n_i \times p$) is the fixed-effects

regressor matrix, \mathbf{Z}_i (of size $n_i \times q$) is the random-effects regressor matrix, and $\boldsymbol{\epsilon}_i$ is the n_i -dimensional error vector. Both the matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$ are positive-definite symmetric. $\boldsymbol{\Psi}$ is the variance-covariance matrix for the random effects that incorporates the correlation between the observations in the same group. $\boldsymbol{\Sigma}$ is the variance-covariance matrix that incorporates the correlated and heteroscedastic residuals. The random-effects \mathbf{b}_i are assumed to be independent for different groups, the \mathbf{b}_i and the error $\boldsymbol{\epsilon}_i$ are independent within-group and between-groups.

Different Random-effects Structures

Different random-effects structures in the model (2.15) can be specified using a patterned variance-covariance matrix ($\boldsymbol{\Psi}$) for the random-effects. If it can be assumed that random-effects are independent of each other, then the $\boldsymbol{\Psi}$ matrix would be a diagonal matrix. When the random-effects are independent and have the same variance, then the $\boldsymbol{\Psi}$ matrix would be an identity matrix multiplied by a constant. But if there is no structure for the random-effects then the $\boldsymbol{\Psi}$ matrix would be unstructured and for a $q \times q$ variance-covariance matrix for the random-effects $q \times (q + 1)/2$ parameters will be estimated in the $\boldsymbol{\Psi}$ matrix.

Different Error Structures

A linear mixed-effects model often assumes that the within-group error is white noise. However, it may not be a suitable assumption to consider in many situations. In those situations, the within-group error could be *correlated* or *heteroscedastic* or both correlated and heteroscedastic. This correlated and heteroscedastic error is modelled using the covariance matrix $\boldsymbol{\Sigma}$. This is done using the decomposition of the matrix $\boldsymbol{\Sigma}$ into a product of matrices (Pinheiro and Bates, 2000)

$$\boldsymbol{\Sigma} = \mathbf{V}\mathbf{C}\mathbf{V} \tag{2.16}$$

where, \mathbf{V} is a diagonal matrix and \mathbf{C} is the correlation matrix. To uniquely identify $\boldsymbol{\Sigma}$ all the diagonal elements of \mathbf{V} must be positive (Pinheiro and Bates, 2000).

For a single level LMM, it can be shown that

$$\text{Var}(\epsilon_{ij}) = [V]_{jj}^2, \quad \text{cor}(\epsilon_{ij}, \epsilon_{ik}) = [C]_{jk} \quad (2.17)$$

where ϵ_{ij} is the error of the j^{th} measurement from the i^{th} subject and ϵ_{ik} is the error for the k^{th} measurement from the same subject. This decomposition of the covariance structure Σ into a *variance structure* \mathbf{V} and a *correlation structure* \mathbf{C} is useful both theoretically and computationally. It allows one to model the two structures separately and then combine them in a more flexible framework. The variance functions for variance component \mathbf{V} and correlation structure for the correlation component \mathbf{C} will now be described.

Variance functions are used to model different variance structures which can address the issue of heteroscedasticity. The general variance structure for the single level LMM can be defined using variance functions as follows (Pinheiro and Bates, 2000):

$$\text{Var}(\epsilon_{ij}|\mathbf{b}_i) = \sigma^2 g^2(\mu_{ij}, \mathbf{v}_{ij}, \boldsymbol{\delta}) \quad (2.18)$$

where, $\mu_{ij} = E(y_{ij}|\mathbf{b}_i)$, \mathbf{v}_{ij} is the vector of variance covariates, $\boldsymbol{\delta}$ is the vector of variance parameters, and $g(\cdot)$ is the variance function assumed to be continuous over $\boldsymbol{\delta}$. The choice of the function $g(\cdot)$ depends on the context. If it is believed that the variance of the within-group error increases linearly with `time` then the variance model would be

$$\text{Var}(\epsilon_{ij}) = \sigma^2 \text{time}_{ij} \quad (2.19)$$

and the corresponding variance function is

$$g(\text{time}_{ij}) = \sqrt{\text{time}_{ij}}. \quad (2.20)$$

There are different variance functions and correlation structures available in the `nlme` package in R (R Core Team, 2021; Pinheiro and Bates, 2000). This makes variance functions and correlation structures easily accessible when fitting a mixed-effects model in R. Table 2.1 lists the different variance functions available in the `nlme` package in R.

Table 2.1: Different variance functions in `nlme` package in R

Function name in <code>nlme</code>	description	$\text{Var}(\epsilon_{ij})$	$g(\cdot)$
<code>varFixed</code>	fixed variance	$\sigma^2 \text{time}_{ij}$	$\sqrt{\text{time}_{ij}}$
<code>varIdent</code>	variances per stratum	$\sigma^2 \delta_{s_{ij}}^2$	$\delta_{s_{ij}}$
<code>varPower</code>	power of covariate	$\sigma^2 \text{time}_{ij} ^{2\delta}$	$ \text{time}_{ij} ^\delta$
<code>varExp</code>	exponential of covariate	$\sigma^2 \exp(2\delta \text{time}_{ij})$	$\exp(\delta \text{time}_{ij})$
<code>varConstPower</code>	constant plus <code>varPower</code>	$\sigma^2 (\delta_1 + \text{time}_{ij} ^{\delta_2})^2$	$\delta_1 + \text{time}_{ij} ^{\delta_2}$

In the mixed-effects modelling framework, the correlation structure is used to model the dependency of the within-group error. Historically, these dependency structures have been developed in two areas of statistics: time series analysis and spatial data analysis. The difference between these two types of data is that response variable in time series is usually indexed by one variable (i.e. time) whereas in spatial data, the response is usually indexed by two coordinates of a spatial plane. As the CRP and the motion capture study involves a time series type response, dependency structures needed for time series analyses will be considered.

To develop a general structure similar to the variance function, consider the case where the dependency of the within-group errors only depend on some position vector \mathbf{p}_{ij} . In the situation where a univariate response is considered, this position vector is just a scalar. In the case of spatial data analysis for example, this position vector may contain a multidimensional vector. It is also assumed that the correlation structure is *isotropic*. This means the correlation between two errors depends only through some distance, say $d(\mathbf{p}_{ij}, \mathbf{p}_{ik})$, between these two positional vectors relating to those errors. A general correlation structure for errors is therefore as follows (Pinheiro and Bates, 2000):

$$\text{cor}(\epsilon_{ij}, \epsilon_{ik}) = h[d(\mathbf{p}_{ij}, \mathbf{p}_{ik}), \boldsymbol{\rho}], \quad (2.21)$$

where, $\boldsymbol{\rho}$ is the vector of correlation parameters and $h(\cdot)$ is the correlation function.

This is a very general structure for modelling the dependency in the within-group errors both for the spatial data and time series data. Since the nature of the data in

this thesis is similar to time series data, from now on the focus will be on correlation structures relevant for time series data. The correlation in time series data is known as *serial correlation*. Since only time series data are considered, \mathbf{p}_{ij} will only contain a scalar position index which will be denoted as p_{ij} since it is no longer a vector. The isotropic assumption will further be simplified to a situation where the correlation only depends on the absolute value of the difference between two position indexes. The general serial correlation structure can now be modelled as (Pinheiro and Bates, 2000)

$$\text{cor}(\epsilon_{ij}, \epsilon_{ik}) = h[|p_{ij} - p_{ik}|, \boldsymbol{\rho}] \quad (2.22)$$

In time series data, the correlation function $h(\cdot)$ is referred to as *autocorrelation* function. A nonparametric estimate of the autocorrelation function is known as the *empirical autocorrelation* function and can be used to examine the serial correlation in the data. Let $r_{ij} = (y_{ij} - \hat{y}_{ij})/\hat{\sigma}_{ij}$ denote the standardised residual from a fitted mixed-effect model where $\hat{\sigma}_{ij}^2$ is the estimate of $\text{Var}(\epsilon_{ij}) = \sigma_{ij}^2$, then the empirical autocorrelation function at lag l is defined as (Pinheiro and Bates, 2000)

$$\hat{\rho}(l) = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i-l} r_{ij} r_{i(j+l)} / N(l)}{\sum_{i=1}^n \sum_{j=1}^{n_i} r_{ij}^2 / N(0)} \quad (2.23)$$

where, n is the total number of subjects, n_i is the number of observations for the i^{th} subject, $N(l)$ is the number of residual pairs used in the summation to define the numerator of $\hat{\rho}(l)$, and $N(0)$ is the total number of residuals.

The simplest serial correlation structure is *compound symmetry*, which can be defined as follows:

$$\text{cor}(\epsilon_{ij}, \epsilon_{ik}) = \rho, \quad \forall j \neq k \quad (2.24)$$

In this situation the autocorrelation function is $h(l, \rho) = \rho$, where $l = |j - k|$. This is a very simplistic correlation structure and might not be very useful in general. The other extreme is the general correlation structure, where the autocorrelation function is

$$h(l, \boldsymbol{\rho}) = \rho_l.$$

2.3. Introduction to Linear Models

This may also not be a useful correlation structure as it requires many correlation parameters to be estimated. Therefore, this may be useful only to find a more parsimonious correlation structure for exploratory purposes.

The correlation structure that is most relevant for this thesis comes from a different class of linear stationary models: *autoregressive* models and *moving average* models. These models assume that the measurements were taken at discrete time points. Consider ϵ_t as the measurement at time point t . The distance, or *lag*, between two measurements ϵ_t and ϵ_s is $|t - s|$ where lag-1 means the measurements are one unit apart. The autoregressive model assumes that the measurement at the current time is linearly dependent upon the previous measurements plus homoscedastic white noise, a_t , centred at zero, $E(a_t) = 0$ (Pinheiro and Bates, 2000).

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \dots + \phi_p \epsilon_{t-p} + a_t \quad (2.25)$$

The number of previous observations on which the current measurement depend upon is called the *order* of the autoregressive model. The order of the autoregressive model here is p and the model is denoted as an AR(p) model. Note that p is used previously to denote the number of fixed-effects in a LMM. For this section, p will be used as the order of an AR model. The model also contains the same number of correlation parameters, $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)'$ as the order of the model. The correlation function for an AR(1) model is follows (Pinheiro and Bates, 2000):

$$h(k, \boldsymbol{\phi}) = \phi^k, \quad k = 0, 1, 2, \dots \quad (2.26)$$

where k is the distance between two time points. The correlation function beyond the AR(1) model does not have any simple representation. It is defined recursively through the difference equation (Pinheiro and Bates, 2000)

$$h(k, \boldsymbol{\phi}) = \phi_1 h(|k - 1|, \boldsymbol{\phi}) + \dots + \phi_p h(|k - p|, \boldsymbol{\phi}), \quad k = 1, 2, 3, \dots \quad (2.27)$$

Moving average models assume that the current observations are a linear combination

2.3. Introduction to Linear Models

of independent and identically distributed white noise terms (Pinheiro and Bates, 2000)

$$\epsilon_t = \theta_1 a_{t-1} + \dots + \theta_q a_{t-q} + a_t. \quad (2.28)$$

The number of white noise terms with lag, q , is the order of the moving average model which is denoted by MA(q) model. Note that q is used previously to denote the number of random-effects in a LMM. For this section, q will be used as the order of a MA model. There are q correlation parameters in this model $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$.

The correlation function for an MA(q) model for the observations with k distance apart is as follows (Pinheiro and Bates, 2000):

$$h(k, \boldsymbol{\theta}) = \begin{cases} \frac{\theta_k + \theta_1 \theta_{k-1} + \dots + \theta_{k-q} \theta_q}{1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2}, & k = 1, 2, \dots, q, \\ 0, & k = q + 1, q + 2, \dots \end{cases} \quad (2.29)$$

A combination of an autoregressive and a moving average model is called an autoregressive-moving average model and denoted by ARMA (p, q) with the order p for the autoregressive model and order q for the moving average model. This model can be written as follows (Pinheiro and Bates, 2000):

$$\epsilon_t = \sum_{i=1}^p \phi_i \epsilon_{t-i} + \sum_{j=1}^q \theta_j a_{t-j} + a_t \quad (2.30)$$

The correlation function for this model is defined recursively as follows (Pinheiro and Bates, 2000):

$$h(k, \boldsymbol{\rho}) = \begin{cases} \phi_1 h(|k-1|, \boldsymbol{\rho}) + \dots + \phi_p h(|k-p|, \boldsymbol{\rho}) + \\ \theta_1 \psi(k-1, \boldsymbol{\rho}) + \dots + \theta_q \psi(k-q, \boldsymbol{\rho}), & k = 1, 2, \dots, q \\ \phi_1 h(|k-1|, \boldsymbol{\rho}) + \dots + \phi_p h(|k-p|, \boldsymbol{\rho}), & k = q + 1, q + 2, \dots, \end{cases} \quad (2.31)$$

where $\psi(k, \boldsymbol{\rho}) = E(\epsilon_{t-k} a_t) / \text{Var}(a_t)$. Table 2.2 lists the available correlation functions in the `nlme` package. Note that this table includes correlation structures for time series data and correlation structures for spatial data. Only the correlation structure for time

Table 2.2: Name of the correlation functions available in the `nlme` package in R.

Name of correlation function in <code>nlme</code>	description
<code>corCompSymm</code>	compound symmetry
<code>corSymm</code>	general
<code>corAR1</code>	autoregressive of order 1
<code>corARMA</code>	autoregressive-moving average
<code>corExp</code>	exponential
<code>corGaus</code>	Gaussian
<code>corLin</code>	linear
<code>corRatio</code>	rational quadratic
<code>corSpher</code>	spherical

series data has been discussed in this thesis and the details for the spatial correlation structure can be found in Pinheiro and Bates (2000).

LMM for Multi-level grouping

A single level LMM can be easily extended to a multi-level LMM. In this section, the single level LMM is extended into a two level LMM. For a two nested level LMM the response vector for the inner-most level can be written as $\mathbf{y}_{ij}, i = 1, \dots, n, j = 1, \dots, n_i$, where n is the number of groups in the first level grouping, outermost, and n_i is the number of second level groups within the first level i^{th} group. The size of \mathbf{y}_{ij} is n_{ij} and the LME model is (Pinheiro and Bates, 2000)

$$\begin{aligned} \mathbf{y}_{ij} &= \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{i,j}\mathbf{b}_i + \mathbf{Z}_{ij}\mathbf{b}_{ij} + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, n_i, \\ \mathbf{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Psi}_1), \quad \mathbf{b}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Psi}_2), \quad \boldsymbol{\epsilon}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{2.32}$$

where $\boldsymbol{\beta}$ is the p -dimensional vector of fixed effects, \mathbf{b}_i is the q_1 -dimensional vector of random effects for first level grouping, \mathbf{b}_{ij} is the q_2 -dimensional vector of random effects for the second level grouping within the first level grouping, \mathbf{X}_{ij} (of size $n_{ij} \times p$) is the fixed effects regressor matrix, $\mathbf{Z}_{i,j}$ (of size $n_{ij} \times q_1$) is the random effect regressor matrix corresponding to the \mathbf{b}_i , \mathbf{Z}_{ij} (of size $n_{ij} \times q_2$) is the random effect regressor matrix corresponding to the \mathbf{b}_{ij} , $\boldsymbol{\epsilon}_{ij}$ is the error vector. $\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \boldsymbol{\Sigma}$ are positive-definite symmetric matrices. The level-1 random-effects \mathbf{b}_i are assumed to be independent of

different i , the level-2 random effect are assumed to be independent of different i or j and of \mathbf{b}_i , the within-group errors ϵ_{ij} are assumed to be independent of different i or j and of random effects.

Estimation for LMM

Maximum likelihood and restricted maximum likelihood estimation can be used to estimate the parameters in models (2.15) and (2.32). The details of the estimation procedure for both the ML and REML method can be found in Pinheiro and Bates (2000), and the discussion between the difference in the ML and REML method can be found in Verbeke and Molenberghs (2000). For this thesis, only REML method will be used to estimate model parameters for a LMM.

2.3.3 Introduction to Smoothing

The limitation of a linear regression model to account for multiple observations from the same measurement method on the same subject, can be accommodated using a LMM. However, if the relative bias is non-linear, that cannot be readily modelled by a LMM. In this case a nonparametric regression approach is needed to model the non-linear bias under the LMM framework.

The relationship between two continuous variables can be written as follows:

$$Y = f(X) + \epsilon \tag{2.33}$$

where Y is a random variable, and X is a covariate. $f(X)$ can be considered the mean value of Y for a given value of X , and ϵ is the error. If the function $f(\cdot)$ is linear, only two parameters needed to be estimated. As the relationship has a parametric form, it is called a parametric regression. If the relationship is non-linear but the mathematical form of the relationship is known then it is still a parametric regression. However, a non-linear parametric form often requires a great deal of effort to find (Ruppert et al., 2003). An easier way is to estimate the function $f(\cdot)$ from the data using a smoothing technique.

The aim of smoothing in statistics is to find an appropriate and robust underlying smooth function. There are many smoothing techniques available. One of the popular techniques in regression analysis is *regression spline smoothing* (Ramsay and Silverman, 2005). This technique first creates a system of basis functions for any smooth functions; then it represents the given function in that space with the help of a regression analysis. There are many basis systems available for regression spline smoothing, but due to the computational advantages, a *B-spline* basis system is widely used. The next chapter will give details of smoothing techniques and the B-spline basis system in particular. The number of parameters for these smoothing technique is not fixed and both the number of parameters and the value of the parameters need to be estimated from the data. Since there is no fixed parametric form for the relationship, this approach of finding the $f(\cdot)$ using a regression model is called nonparametric regression. If a regression model involves two covariates where the relationship of one covariate with the response is parametric and the relationship of the other covariate is nonparametric, the model is called a semi-parametric regression model (Ruppert et al., 2003). This semi-parametric regression model, under the LMM framework, provides much flexibility when fitting models.

In the next section two important properties of a measurement method, namely *validity* and *reliability*, will be considered. Both can be estimated using the models introduced so far.

2.4 Reliability and Validity of a Measurement Method

The validity of a measurement method means how accurately it can measure the true value of a quantity (Kimberlin and Winterstein, 2008). If in model (2.5), the bias term β_m is not equal to zero, then the method of measurement would not be a valid one. The reliability of a measurement method indicates whether a measurement method produces the same measurement when used to measure the same quantity on the same subject (Kimberlin and Winterstein, 2008). In model (2.5), ϵ_{mij} is related to the reliability of a measurement method. If $\sigma_{\epsilon_m}^2$ is very small, then the measurement method would

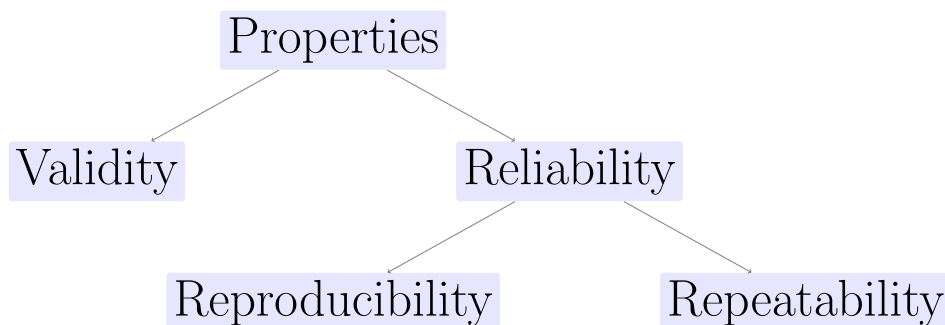


Figure 2.1: Properties of a measurement method.

be considered reliable, but if it is large, then the measurement method would not be considered a reliable one.

The properties of reliability can further be broken down into *repeatability* and *reproducibility* (Figure 2.1) (Newell et al., 2014).

The closeness of the measurements from one person using the method is called repeatability (Newell et al., 2014). On the other hand, reproducibility means the closeness of the measurements produced by the same method but operated by different people (Newell et al., 2014). In model (2.5) the value $\sigma_{\epsilon_m}^2$ can be used to assess both the repeatability and reproducibility depending upon the study design.

2.4.1 Indices of Reliability

The ICC and SEM are the two widely used indices of reliability (Weir, 2005) will now be discussed in more detail.

Standard Error of Measurement (SEM)

In model (2.5), $\sigma_{\epsilon_m}^2$ is directly related to the measurement method. For this reason, $\sigma_{\epsilon_m}^2$ is used to assess the reliability of a measurement method. The SEM is defined as follows (Hopkins, 2000):

$$\text{SEM} = \sqrt{\sigma_{\epsilon_m}^2}. \quad (2.34)$$

This index of reliability is used to assess the reliability in the same units of the measurement and is hence called an *absolute measure of reliability* (Weir, 2005).

Intra-class Correlation Coefficient (ICC)

The magnitude of $\sigma_{\epsilon_m}^2$ can be assessed in comparison with σ_α^2 . If σ_α^2 is large in comparison with $\sigma_{\epsilon_m}^2$ then the method is deemed reliable, and if σ_α^2 is similar to $\sigma_{\epsilon_m}^2$ or even less than $\sigma_{\epsilon_m}^2$, then one may say that the measurement method is not reliable. A different measure of reliability can be used, which is called *intra-class correlation coefficient* (ICC). Using model (2.5) for a given measurement method, the ICC can be defined as follows (Weir, 2005):

$$\text{ICC} = \text{Cor}(Y_{mij}, Y_{mik}) = \frac{\text{Cov}(Y_{mij}, Y_{mik})}{\sqrt{\text{Var}(Y_{mij}) \text{Var}(Y_{mik})}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_{\epsilon_m}^2}, \quad (2.35)$$

where Y_{mij} be the j^{th} measurement made on the i^{th} subject and Y_{mik} be the k^{th} measurement made on the same subject by the method m , with $j \neq k$. Here, the correlation is restricted to be non-negative $0 \leq \text{cor}(Y_{ij}, Y_{ik}) \leq 1$. Although a value of zero is possible for ICC, most of R packages, such as `nlme` and `lme4`, handle the fact that variance estimates must be non-negative by fitting variances on the log scale, which means $\sigma_\alpha^2 = 0$ never happens in practice while using these packages.

This approach is unitless and hence called *relative measure of reliability* (Weir, 2005). The ICC can be interpreted in different ways. One way to interpret it is the proportion of between-subject variability compared to the total variability of the measurements. If the between subject variability is high compared to the total variability then in the measurements the information content is high compared to the noise or error, hence the method is more reliable.

Since the ICC depends on σ_α^2 , it also depends on the range of the true value being measured (Bland and Altman, 1990). For example, assume the same method is applied to a population where the range of the true value being measured is large and to a population where the range of the true value being measured is small. The method may appear more reliable for the population with the more extensive range of true values, while the same method is considered less reliable for a smaller range of true values.

The different measures of reliability introduced will now be calculated and inter-

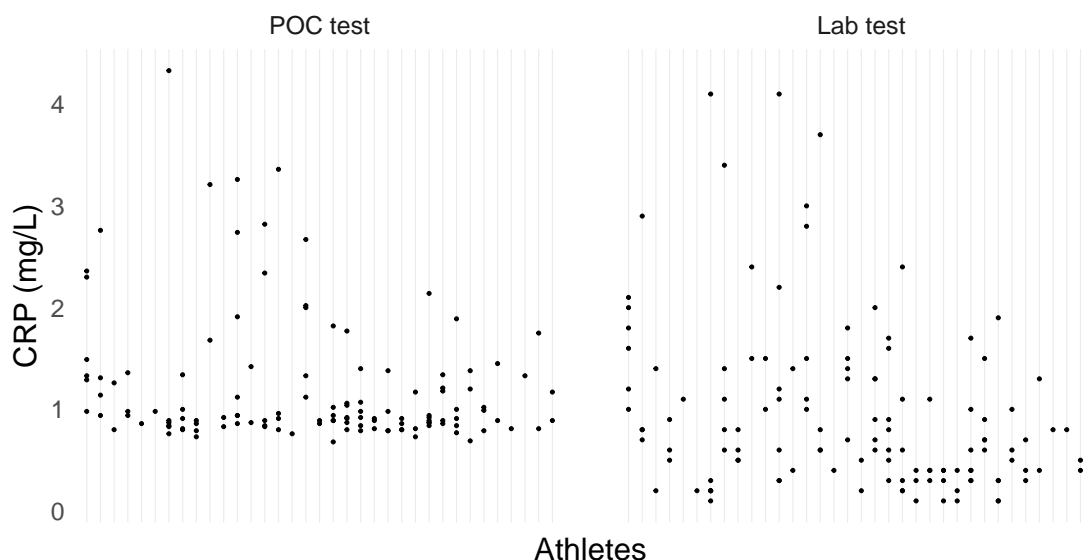


Figure 2.2: Dot plot for C-reactive protein (CRP) measurements by both a point of care (POC) and a laboratory (lab) method. Each vertical line on the X-axis represents an athlete and points on a line represents replicate measurements from the athlete on different measurement occasions.

preted for the blood biomarker case study.

2.4.2 Application: CRP Analyser Study

CRP is a biomarker commonly used to measure systematic inflammation in elite athletes. A gold standard laboratory test (lab test) measures the CRP level through a blood biomarker. However, this measurement method is invasive as a full venous draw is needed and therefore not suitable if a quick measurement is needed. An alternative POC method is available, and a method comparison study was needed to compare this to the gold standard laboratory test.

Before measuring the agreement between the two measurement methods, the reliability of the methods will be assessed using the SEM and ICC indices.

A dot plot of the measurements for different athletes is useful for visualising the between-subject and within-subject variability (Figure 2.2). The POC method does not measure any CRP values below 0.68 mg/L which keeps the measurements for a given subject very close to each other (Figure 2.2). This may artificially makes the

2.4. Reliability and Validity of a Measurement Method

Table 2.3: SEM and ICC for the two measurement methods using two measurements per method.

	SEM	ICC
POC method	0.60	0.06
Lab method	0.67	0.11

SEM for this method smaller.

To calculate the SEM and ICC in order to assess reliability, *analysis of variance* (ANOVA) is usually suggested in the literature (Li et al., 2015). However, caution should be exercised when using ANOVA to estimate the ICC since the calculations in ANOVA depend on whether the study design was balanced or not (Searle et al., 2006). If the same number of measurements are available for each method per subject, then the study design is balanced, otherwise it is not. This may not be the case for many method comparison studies, hence a robust and general approach would be to use a LMM to estimate the ICC and SEM as the LMM estimation procedures do not depend on a study being balanced or not. Model (2.7) will be used to estimate the ICC and SEM using a LMM which will provide an estimate of the within- and between-subject variability for each measurement method separately.

One may notice that the SEM of the POC method is lower than the lab method. This may mean that the POC method produces more precise estimates than the lab method. However, looking at the Figure 2.2 may clarify this apparent confusion. The figure shows that the CRP measurements by the POC method never go below a measurement value of 0.68 mg/L. On the other hand, the lab test result produces measurements below the CRP value 0.68 mg/L for the same subject on the same occasion. As the POC method produces a value of 0.68 mg/L for all values below the threshold of 0.68 mg/L, those measurements cluster together around the value 0.68 mg/L. This makes the POC method appear artificially more precise.

The ICC for both the methods shows that the lab method is more reliable than the POC method. Given the fact that the maximum value of ICC is 1, the ICC for both the methods is not as high as one may expect. Note that each athlete's measurements were

taken in different months. The true value of an athlete is likely to differ from month to month. For this reason, there is an additional source of within-subject variability other than the measurement error in the measurements. This additional source of variability was included in the error variance, and hence it reduced the estimate of ICC. However, for a study with one measurement per method at each time point, this additional variability cannot be estimated separately from the error variance.

2.5 Agreement between two Methods

For a valid measurement method to be useful it must be reliable, i.e. repeatable and reproducible. However, a different question is whether an existing method is replaceable or interchangeable with a new method. The investigation of such a question is called a method comparison study (Altman and Bland, 1983). In a method comparison study, the main aim is to determine the level of agreement between two measurement methods, i.e. can a new method produce the same results (or acceptable results) as the current method in use.

2.5.1 Indices of Agreement

To measure agreements between methods of measurement, Altman and Bland (1983) proposed 95% limits of agreement (95% LoA) while Lin (1989) proposed the concordance correlation coefficient (CCC) (Figure 2.3). These two indices of agreement will be discussed next.

Concordance Correlation Coefficient (CCC)

When two methods measure the same quantity on a continuous scale, both measurements should fall on the line of equality (i.e. the line which goes through the origin with slope 1) if there is perfect agreement between them. The Pearson correlation coefficient measures the degree of linear relationship between two variables measured on a continuous scale. However, the Pearson correlation coefficient cannot be used to measure agreement since it quantifies any linear relationship, not just the one relevant

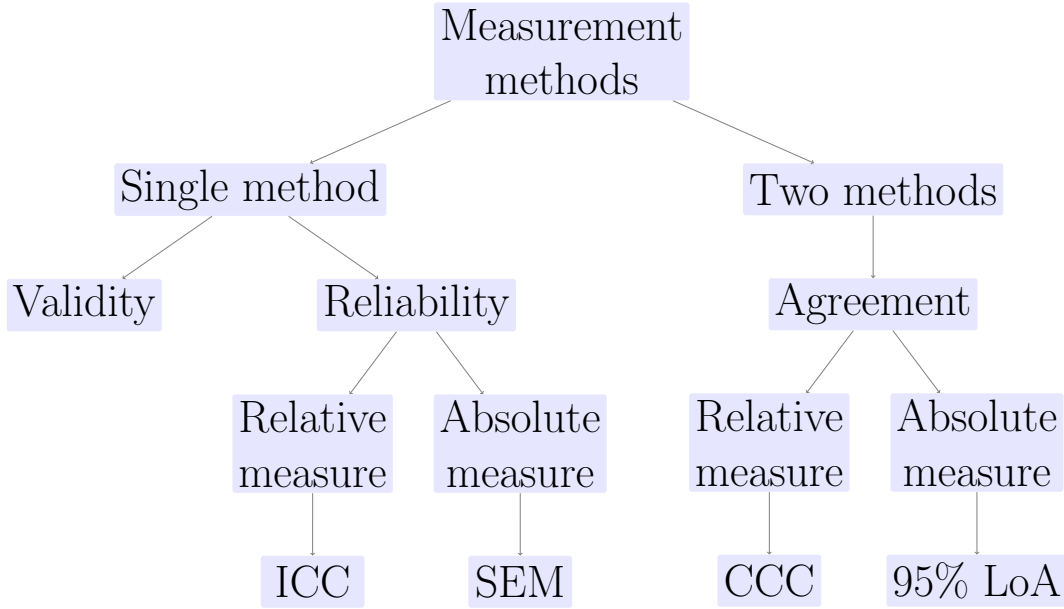


Figure 2.3: Indices of reliability and agreement.

for method agreement (Lin, 1989). For example, if a linear relationship between two methods of measurement can be described by a straight line intersecting the Y-axis at 10 unit, then the Pearson correlation coefficient may give a high value for the co-efficient while there is still relative bias between the two methods. In this situation, an index measuring agreement should indicate that the agreement is poor. Lin (1989) proposed the concordance correlation coefficient (CCC) to measure the agreement between two methods of measurement by removing the issue with the Pearson correlation coefficient in method agreement studies. The CCC measures how close each pair of measurements are to the line of equality.

If Y_1 is the measurement by the first measurement method and Y_2 is the measurement made by the second measurement method with $Y_1 \sim N(\mu_1, \sigma_1^2)$, $Y_2 \sim N(\mu_2, \sigma_2^2)$, and $\text{cov}(Y_1, Y_2) = \sigma_{12}$, then the CCC can be defined as follows (Lin, 1989):

$$\text{CCC} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \frac{2\rho\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \rho \times C_b. \quad (2.36)$$

where ρ , the Pearson correlation coefficient ($|\rho| \leq 1$), measures how far each observation deviated from the line of best fit (a precision measure) and C_b , the bias correction factor ($0 \leq C_b \leq 1$), measures how far the line of best fit deviates from the line of equality

(Lin, 1989).

Equation (2.36) can be used when there is one measurement per method per subject. When replicate measurements are present, under the specification of the model (2.5), the CCC can be defined as follows (Carrasco and Jover, 2003):

$$\begin{aligned} \text{CCC} &= \frac{2\sigma_{\alpha}^2}{(\sigma_{\alpha}^2 + \sigma_{\epsilon_{m1}}^2) + (\sigma_{\alpha}^2 + \sigma_{\epsilon_{m2}}^2) + (\beta_{m1} - \beta_{m2})^2} \\ &= \frac{\sigma_{\alpha}^2}{\sigma_{\alpha}^2 + \frac{1}{2} [\sigma_{\epsilon_{m1}}^2 + \sigma_{\epsilon_{m2}}^2 + (\beta_{m1} - \beta_{m2})^2]} \end{aligned} \quad (2.37)$$

The CCC in equation (2.37) ranges from 0 to 1. If the value of the CCC is 0, then there is no agreement between the two methods; the closer it gets to 1, the better the agreement, and when the value is 1, there is perfect agreement between the two methods.

Like the ICC, the CCC is also unitless, hence it is a relative measure of agreement. For example, the distinction between a point estimate of 0.9 and 0.8 for the CCC is not possible to interpret.

The CCC assumes that the relationship between the measurements made by the two methods of measurement is linear. If the relationship is not linear then in this case it is not possible to interpret the CCC. The relationship may not be linear in a situation when there is non-linear bias between the measurements made by the two methods of measurement.

95% Limits of Agreement (LoA)

Bland and Altman criticised the use of the Pearson correlation coefficient, two-sample t-test and ICC as a measure of agreement (Altman and Bland, 1983; Bland and Altman, 1990), and proposed the use of the 95% limits of agreement (LoA) as an appropriate approach (Altman and Bland, 1983; Bland and Altman, 1986, 1999). Their approach is quite intuitive; the difference between the two methods' pair of measurements is calculated, the average and standard deviation (SD) of these differences are then used to calculate two quantities: mean \pm 2 SD. These two quantities are called the lower and upper limits of agreement where the average of the differences is the relative bias

2.5. Agreement between two Methods

between the two measurement methods quantifying whether these two methods agree on average and the limits of agreement provide a reference as to the level of agreement within each pair of measurements. The LoA are an absolute measure of agreement as they measure the level of agreement in the same unit of measurement. A more detailed description of 95% LoA is now given.

To calculate the 95% LoA, the difference between two measurements by the two methods on the same subject needs to be calculated. Let D_{ij} represents the difference for the j^{th} measurement on the i^{th} subject by the two methods of measurement which can be calculated under the model (2.6) as follows.

$$\begin{aligned}
 D_{ij} &= Y_{m_1ij} - Y_{m_2ij} \\
 &= (\mu + \alpha_i + \beta_{m_1} + \gamma_{m_1i} + \epsilon_{m_1ij}) - (\mu + \alpha_i + \beta_{m_2} + \gamma_{m_2i} + \epsilon_{m_2ij}) \\
 &= (\beta_{m_1} - \beta_{m_2}) + (\gamma_{m_1i} - \gamma_{m_2i}) + (\epsilon_{m_1ij} - \epsilon_{m_2ij}) \\
 &= \delta + \gamma_i + \epsilon_{ij} \\
 &= \text{Bias} + \text{Error}.
 \end{aligned} \tag{2.38}$$

where, $\delta = (\beta_{m_1} - \beta_{m_2})$, $\gamma_i = (\gamma_{m_1i} - \gamma_{m_2i})$, $\epsilon_{ij} = (\epsilon_{m_1ij} - \epsilon_{m_2ij})$, and $\text{Bias} = \delta$, $\text{Error} = \gamma_i + \epsilon_{ij}$. In equation (2.38), one can notice that the true value from the same subject ($\mu + \alpha_i$) was cancelled out and only two parts remain: a systematic difference (δ) between the two methods and random error components ($\gamma_i + \epsilon_{ij}$). The part δ is considered as the relative bias between the two methods. The $(\gamma_i + \epsilon_{ij})$ part is responsible for the variability of the individual value of the differences (D_{ij}). Let $\gamma_i \sim N(0, \sigma_\gamma^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, and $\text{Error} \sim N(0, \sigma_{\text{Error}}^2)$. To measure the level of agreement Altman and Bland (1983) proposed the following 95% LoA:

$$\begin{aligned}
 \text{Lower limit} &= \text{Bias} - 2 \sigma_{\text{Error}} \\
 \text{Upper limit} &= \text{Bias} + 2 \sigma_{\text{Error}}
 \end{aligned} \tag{2.39}$$

From equation (2.38), it is clear that the lack of agreement may come from two different sources. However, if the relative bias is zero and the value of σ_{Error}^2 is very

2.5. Agreement between two Methods

small, the two methods are deemed in agreement and can be used interchangeably.

If the term $\delta = 0$ then it is concluded that these two methods agree on average. However, if two methods agree on average, there is no guarantee that individual measurements also agree. Given the relative bias is zero if σ_{Error}^2 is large, then it means that the level of agreement is poor, while if σ_{Error}^2 is small, then the level of agreement is good. The value of σ_{Error}^2 being large or small depends on the context it is being used for. For example, the CRP biomarker is an indicator of inflammation. In elite sports a measurement method with high accuracy might be desirable due to the competitive nature of the game. For this reason a deviation of 0.5 mg/L unit from the true value may be of importance. However, in a clinical setting 0.5 mg/L deviation from the true value might not be important since this error may not affect any health related decisions.

When σ_{Error}^2 is small but δ is large, then there is also a lack of agreement; given these estimates the new method can be adjusted to remove the bias by adjusting for δ accordingly .

The term 95% LoA was coined by Bland and Altman (1986). If the difference in the pair of measurements is normally distributed, then approximately 95% of differences would fall within the range $(\bar{d} - 2s, \bar{d} + 2s)$, where \bar{d} is the sample mean, and s is the sample standard deviation of the differences which are the estimate of the \bar{D} and σ_{Error}^2 , respectively. They named this range as the 95% limits of agreement. The 95% LoA should contain 95% of the differences in the sample not in the population. If the 95% LoA were calculated from a different sample, the limits of agreement would be different.

One question of interest is what would be a suitable approach to calculate the 95% LoA when the bias increases as the magnitude of the measurement increases. Bland and Altman answered this question by modelling the relationship between the differences and the average value of the measurements using a linear regression model (Bland and Altman, 1999).

Bland and Altman (1999) provided the necessary calculation to compute the 95% LoA when replicates were taken without using any modelling framework; however, it was not clear what would be the appropriate approach to calculate the 95% LoA when the data were obtained using a complex study design, and when there is a need to

2.5. Agreement between two Methods

adjust for covariates. These questions were addressed by Carstensen et al. (2008) by proposing a LMM for such designs and then calculating the 95% LoA from the model parameters. This approach offers an elegant way of calculating the 95% LoA where replicates were taken.

At this point one might think that the problem of method comparison is solved. However, some open questions remain unanswered. For example, what is the most suitable method to estimate the 95% LoA if there is a non-linear bias between methods of measurement remains unanswered. This question will be considered in the next section of the thesis by way of example.

2.5.2 Application: CRP Analyser Study

After assessing the reliability of the two measurement methods, the agreement between these two methods can be investigated using the CCC and 95% LoA. A scatter plot of the measurements by the two methods with a line of equality is a useful visual way to assess agreement. Most of the points are not close or on the line of equality, indicating that the agreement between these two methods of measurement might be poor (Figure 2.4).

Using the measurement model (2.5) and the formula (2.37) the CCC between the two methods of measurement can be estimated. The estimated value of the CCC between these two methods of measurement is 0.14. This indicates that the agreement between these two methods of measurement is poor since the CCC can range from 0 to 1. However, the value of the CCC is interpretable based on the assumption that the relationship between the two method of measurement is linear. It is clear that the relationship between the measurement by these two method of measurement is not linear (Figure 2.4). For this reason, the value of the CCC in this context is not possible to interpret.

The agreement between these two methods can also be investigated by the 95% LoA. The first step in calculating the 95% LoA is to calculate the difference between each pair of measurements by the two methods. Using model (2.38) and equation (2.39), the bias between the two methods of measurement was estimated as 0.27 mg/L,

2.5. Agreement between two Methods

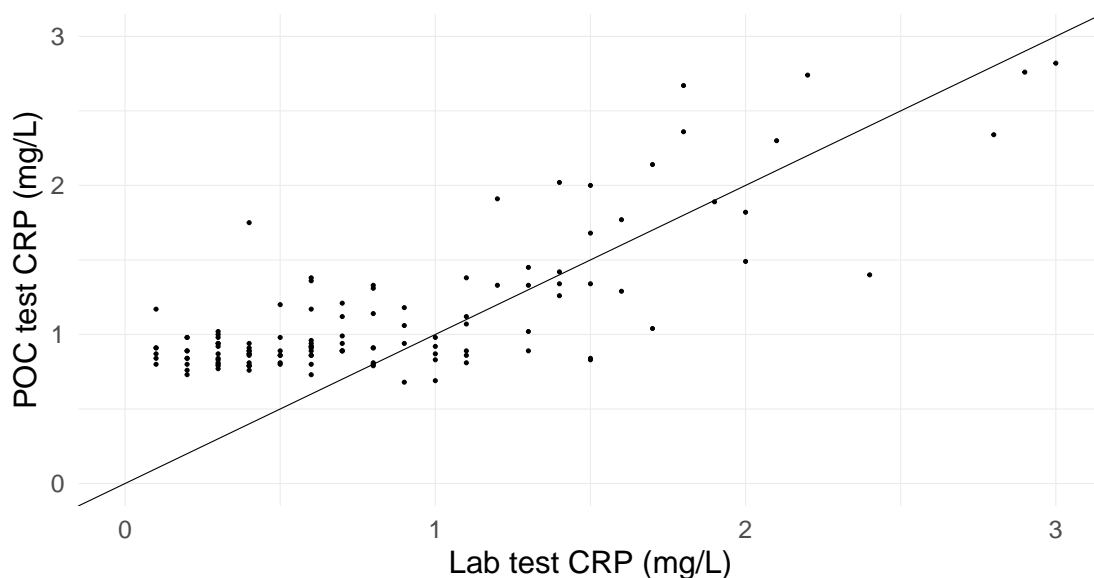


Figure 2.4: Scatter plot with the line of equality for c-reactive protein (CRP) measurements made by a laboratory method (lab test) and a point of care method (POC test).

and the corresponding 95% LoA are $(-0.67, 1.21)$. Before making any conclusion, the assumptions underlying this calculation need to be checked. One assumption for the 95% LoA is that both the average and the standard deviation of the differences are constant for the different magnitude of CRP values.

Figure 2.5 shows the scatter plot of the average value of each pair of measurements on the x-axis and their difference on the y-axis. The three horizontal line shows the relative bias, upper and lower limits of agreement. This plot is known as the Bland-Altman plot. The reason for drawing the bias line is to assess visually whether the raw data are consistent with the average difference being the same across the range of the CRP values. The reason to draw the upper and lower limits is to check whether the standard deviation of the difference seems constant for the different magnitudes of CRP values. By including a locally estimated scatterplot smoothing (LOESS) line on the Bland-Altman plot, it is clear that the relative bias is not constant over the different magnitude of the CRP values, and the assumption of constant relative bias is violated (Figure 2.5).

The bias between these two methods of measurement is not linear. Due to this

2.5. Agreement between two Methods

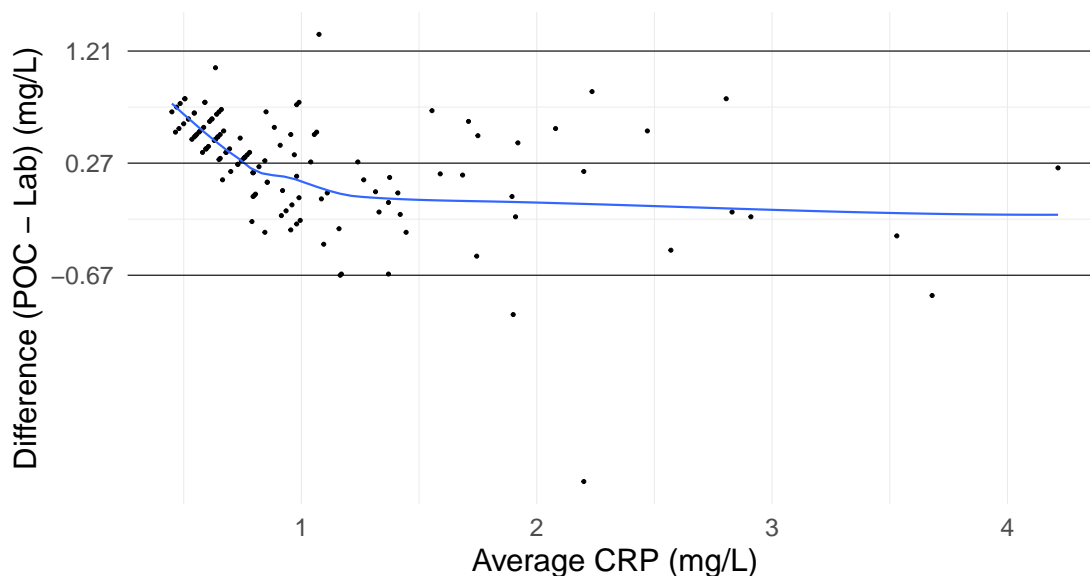


Figure 2.5: The Bland-Altman plot for the two methods measuring c-reactive protein (CRP). Here the two measurement methods are a point of care (POC) test and a laboratory (lab) test. The blue line represent a locally estimated scatterplot smoothing (LOESS) line.

non-linear bias, model (2.38) cannot be used. This non-linear bias needs to be handled appropriately to produce appropriate 95% LoA for this study.

2.5.3 Adjusting for Non-linear Bias

In the CRP analyser study, both the replicate measurements and the non-linear bias needed to be taken into account when calculating the 95% LoA to compare the two measurement methods. In the presence of replicates, the application of the LMM modelling framework to properly estimate the variance components to calculate the 95% LoA has been suggested by Carstensen et al. (2008). To adjust for linear bias while calculating the 95% LoA, the use of linear regression is also suggested by Bland and Altman (1999). However, according to the authors' knowledge, no statistical technique has been suggested to calculate the 95% LoA in the presence of non-linear bias. In this thesis, a nonparametric regression approach is proposed to calculate the 95% LoA in the presence of non-linear bias between the two methods of measurement. It can be argued that a nonparametric LMM modelling framework should be used in general to

2.5. Agreement between two Methods

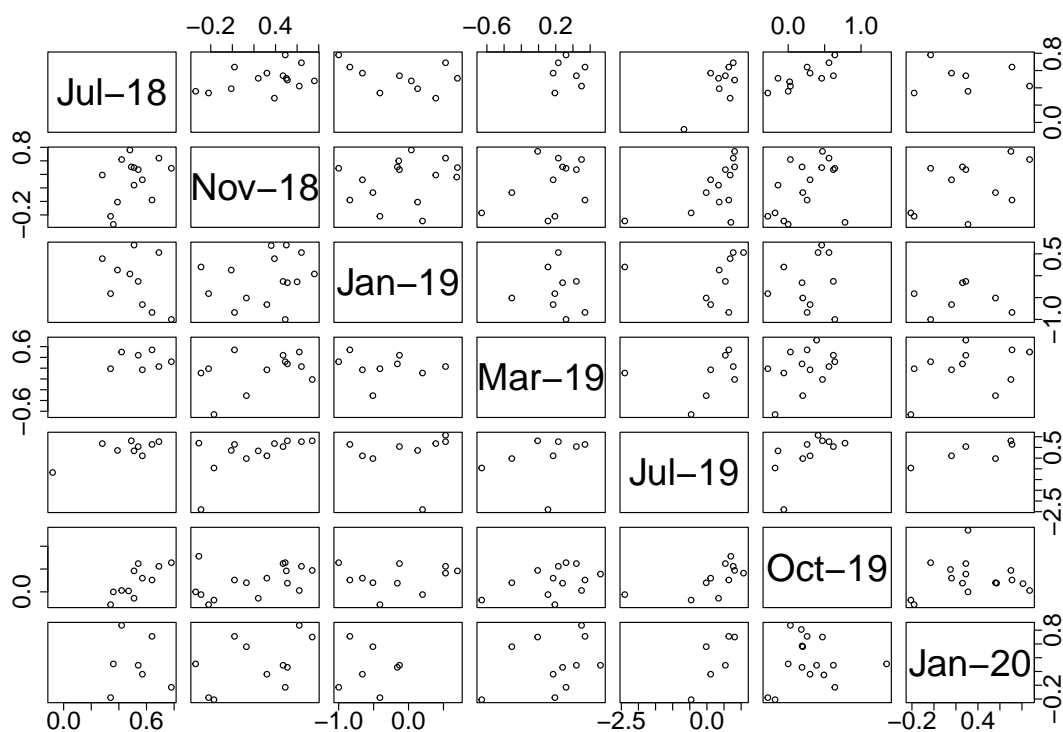


Figure 2.6: A correlation plot of the differences between the measurements made by a point of care method and a laboratory method. Each subplot is a scatter plot of the differences where X and Y-axis represents two different measurement occasions.

calculate 95% LoA in particular in the presence of replicates and non-linear bias.

A modification in model (2.38) is needed to adjust for non-linear bias. Consider for the i^{th} subject n_i pairs of replicates were taken. Here replicates are measured longitudinally and a pair of measurements are from the two different measurement methods. Note that n_i could be different for different subjects. Taking the difference of these pairs of measurements will give n_i differences for the i^{th} subject. Consider \mathbf{d}_i is a n_i -dimensional vector containing those differences for the i^{th} subject.

When there is a constant bias between the two methods of measurement the bias is constant over the range of the value being measured. When there is linear bias, this bias increases linearly over the range of the value being measured. In the case of non-linear bias, the bias varies non-linearly over the range. Measurements from neither of the methods of measurement can be considered as true values (Bland and Altman, 1995). The best estimate of the true value being measured is the average of the pair of

2.5. Agreement between two Methods

measurements taken by the two methods of measurement. As there are n_i differences for the i^{th} subject, there are also n_i averages from n_i pairs of measurement. Let \mathbf{a}_i be a n_i -dimensional vector containing those averages.

Since the bias varies over the range of the value being measured, \mathbf{d}_i varies over \mathbf{a}_i non-linearly. To model this non-linearity, a set of B-spline basis functions evaluated at different value of the averages will be considered. Consider \mathbf{X}_i is a regressor matrix containing a set of B-spline basis functions evaluated at \mathbf{a}_i to model this non-linearity.

Now, \mathbf{d}_i can be modelled as follows:

$$\begin{aligned} \mathbf{d}_i &= \mathbf{X}_i \boldsymbol{\beta} + b_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, n \\ b_i &\sim N(0, \sigma_b^2), \quad \boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma}_i) \end{aligned} \tag{2.40}$$

where, \mathbf{X}_i is a $n_i \times p$ regressor matrix containing B-spline basis functions evaluated at \mathbf{a}_i , p be the number of B-spline basis functions needed to model the non-linear pattern, $\boldsymbol{\beta}$ is a p -dimensional vector of coefficient for each column of the \mathbf{X}_i matrix, b_i is the random intercept term for the i^{th} athlete, $\boldsymbol{\epsilon}_i$ is the n_i dimensional error vector for the i^{th} athlete, σ_b^2 is the variance for random intercept terms and $\boldsymbol{\Sigma}_i$ the variance-covariance matrix for $\boldsymbol{\epsilon}_i$. Here, $\boldsymbol{\Sigma}_i$ is an $n_i \times n_i$ positive-definite variance-covariance matrix which depends on i through its dimension n_i , but the set of parameters in $\boldsymbol{\Sigma}_i$ does not depends on i (Laird and Ware, 1982).

Using the model (2.40) the relative bias can be calculated as follows:

$$\text{Bias} = \text{E}(\mathbf{d}) = \mathbf{X} \boldsymbol{\beta} \tag{2.41}$$

where, Bias is the vector of relative bias along the range of CRP values, \mathbf{X} contains all the B-spline basis functions, and $\boldsymbol{\beta}$ is the coefficient vector for the columns of the \mathbf{X} matrix.

The variance-covariance matrix of the differences can be calculated as follows:

$$\mathbf{V} = \text{var}(\mathbf{d}) = \sigma_b^2 \mathbf{I} + \boldsymbol{\Sigma} \tag{2.42}$$

2.5. Agreement between two Methods

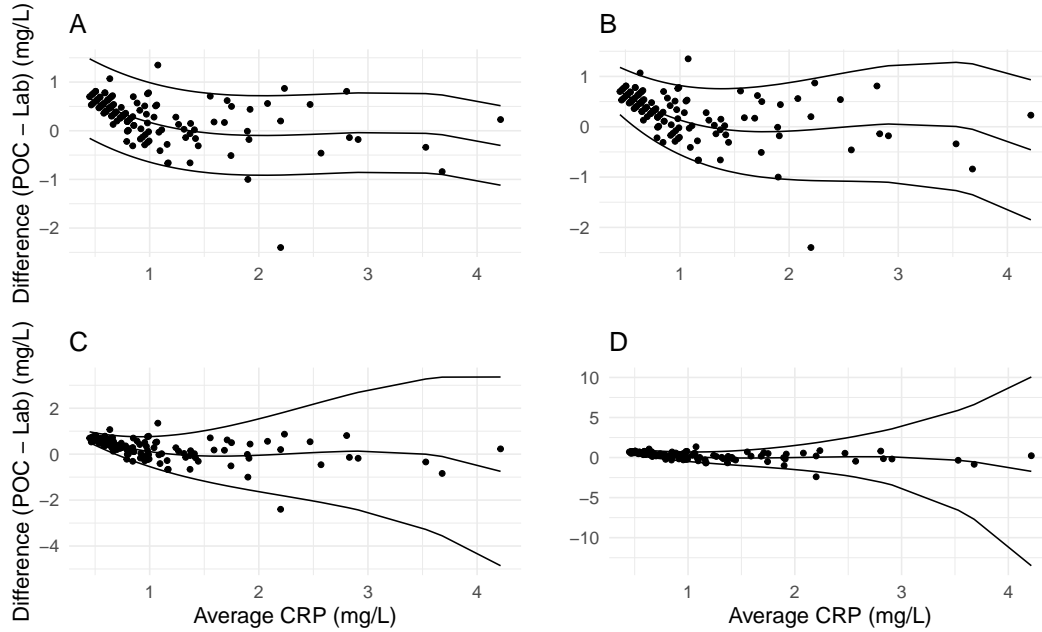


Figure 2.7: 95% limits of agreement to compare two methods of measurement: a laboratory test (lab) and a point of care test (POC). Four different variance functions were considered: A. no variance function; B. a fixed variance function; C. a power variance function; D. an exponential variance function.

Using equation (2.41) and (2.42), the 95% LoA for the CRP data can be calculated as follows:

$$\text{LoA} = \mathbf{X}\boldsymbol{\beta} \pm 2\sqrt{\text{diag}(\mathbf{V})} \quad (2.43)$$

For the CRP analyser data, using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) three cubic B-spline basis functions were found sufficient to model the non-linearity of relative bias over the range of CRP values. For the specification of the $\boldsymbol{\Sigma}$ matrix in the model (2.40), both the correlation structure and the variance function need to be specified. For model (2.40), differences between the pair of measurements by the two methods are the responses. If these differences are correlated among different measurement occasions (i.e. different months), then a correlation structure is needed to include in the model (2.40). Note whether the original measurements are correlated among different measurement occasions or not that does not affect the decision of including a correlation term in model (2.40). Only the correlation of the differences among the different measurement occasions matters.

It seems that there is little correlation for the differences among different measurement occasions (Figure 2.6). For this reason, no correlation structure will be considered for model (2.40)

If the variability of the differences increases as the magnitude of CRP increases, then a variance function is needed to include in model (2.40). From the list of available variance functions in the Table 2.1, four variance functions were considered: no variance function, a fixed variance function, a power variance function, and an exponential variance function. None of the variance functions can model the error variance appropriately (Figure 2.7). However, the three cubic B-spline functions seem to model the non-linear bias sensibly (Figure 2.7).

95% LoA using Log CRP

In a situation where the variance of differences increases over the magnitude of the measurement, Bland and Altman (1999) suggested using a log transformation to remove this issue. The 95% LoA using the log-transformed CRP values was estimated without any variance function (Figure 2.8). After taking the log transformation of the CRP values, the variance of the differences became similar across the range of CRP measurements. A log transformation did not remove the issue of non-linear relative bias. For this reason, B-spline basis functions are still needed to model the non-linear relative bias. In this case, using the AIC and BIC, a set of four B-spline functions was found to be sufficient to model the non-linearity in the bias.

The interpretation of the 95% LoA for the log transformed CRP is not the same as the interpretation of the 95% LoA without the log transformation. The bias for the 95% LoA with the log transformed CRP is no longer additive. The bias is now multiplicative. For example, when the average log CRP value is -1.1, the relative bias is of 1.9 with the 95% LoA (1.3, 2.5). This should be interpreted as when the geometric mean of the measurements by the two methods of measurement is 0.33 mg/L (taking the anti-log of -1.1) for example, on average the POC method produces measurements that are 6.7 (taking the anti-log of 1.9) times higher than the measurement produced by the laboratory method. The 95% LoA (1.3, 2.5) should be interpreted as the POC

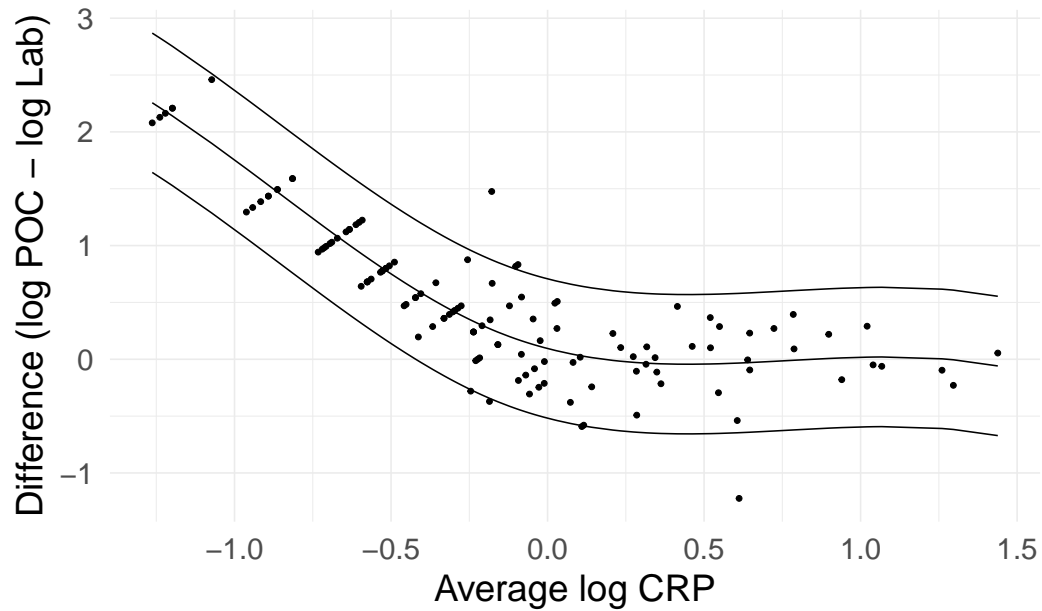


Figure 2.8: 95% limits of agreement using log transformed c-reactive protein (CRP) measurements.

method could produce measurements 3.7 to 12.2 times higher than the measurement produced by the laboratory method when the true value of the measurement is 0.33 mg/L (geometric mean). Similarly, for different average log CRP values, different 95% LoAs are obtained and interpreted accordingly.

Using the 95% LoA for log CRP, it is now clear that there is a substantial multiplicative relative bias between the methods of measurement approximately below the average log CRP value 0.0. This is the primary sources of disagreement between the two measurement methods.

2.6 Summary

In this chapter, the focus was on statistical approaches to accommodate statistical models for measurements of increasing complexity involving univariate continuous responses.

A description of how these models can be used to calculate indices of validity and reliability and the concept of a limits of agreement in method agreement studies was

given.

The limits of agreement provide a directly interpretable measure of agreement in the same units of the response. An essential assumption for using limits of agreement is that the relative bias between the two measurement methods should be constant over the different values of the quantity being measured.

A novel graphical approach was proposed in this chapter for assessing this assumption by plotting a smoothing line on the Bland-Altman plot. This proposed graphical approach is a valuable one to determine if there is a non-linear bias between the two measurement methods. Following this a new statistical approach was presented to calculate the 95% LoA when there is a non-linear bias between the two measurement methods by fitting a LMM with B-spline basis functions to model the non-linear bias between the two measurement methods. The proposed method was demonstrated with the data obtained from the CRP analyser study. In addition to the proposed new graphical and analytical method for method comparison studies, it was suggested that every method comparison study should be analysed using an appropriate linear mixed-effects model as this modelling framework provides the most flexible tools for such analyses. Ignoring the linear mixed-effects model and using earlier calculation methods, such as ANOVA and the approach proposed by Bland-Altman, to analyse method comparison study may provide a misleading conclusion as these methods can accommodate relatively simple study designs only.

In the next chapter the focus will change to method agreement studies involving a response variable that is a curve, a so called functional response, where a new framework to generate functional limits of agreement will be presented.

Chapter 3

Method Agreement Studies involving a Functional Response

3.1 Introduction

In this chapter the focus changes to method comparison studies involving a functional response. The chapter begins with a brief introduction of functional responses, followed by a discussion of the typical statistical methods used in method agreement studies involving functional responses. An overview of functional data analysis is then given, with a particular emphasis on method agreement studies. Indices of reliability for assessing agreement between functional responses are then described and their limitations discussed. Estimating 95% functional limits of agreement (fLoA) using functional data analysis in simple method comparison studies is outlined, followed by a discussion on the suitability of nonparametric linear mixed-effects models as an alternative approach to consider. A proposed new and novel extension to the mixed modelling framework is then introduced to estimate 95% fLoA in simple method comparison studies and in more complicated studies involving replicates and hierarchical study designs. Issues relating to the computational cost when using a mixed model framework to generate functional limits of agreement are addressed through the provision of a new faster computational approach for fitting the nonparametric mixed-effects models in question.

3.2 Introduction to Functional Responses

Most statistical analyses involve a univariate or multivariate response variable when making inference about the general population from which the sample is drawn. With the advent of modern technology and continuous improvements in data capture, response variables can now be in the form of curves. The statistical analysis appropriate for such data is called Functional Data Analysis (FDA) (Ramsay and Silverman, 2002).

In its most general form, each response in a FDA is considered a random function instead of a random variable. For example, in biomechanical research, collection of integrated three-dimensional kinematics, kinetics, and muscle activation patterns are commonplace (Pini et al., 2019; Lencioni et al., 2019). The high intrinsic dimensionality of such data brings challenges to theory and computation. However, the data's high dimensional structure is a rich source of information, as will be seen in the marker comparison study presented in this chapter.

3.3 Functional Responses in Method Comparison Studies

Statistical techniques for analysing univariate response data in method comparison studies are not readily applicable to functional data. Although the methodology for analysing functional responses is available, researchers in biomechanics often reduce the functional response into a single value by choosing one aspect of the curve and ignoring all other information available (Donoghue et al., 2008; Richter et al., 2014). The single value chosen is usually the maximum, minimum, or the timing of a specific event (Harsted et al., 2019; Markström et al., 2018). For example, in Figure 3.1 a plot of each athlete's right hip abduction angle curve is given with each athlete's maximum value highlighted. Is it clear that ignoring the rest of the data for every individual functional response will discard a huge amount of information available in these data. This approach of only using a single value derived from the whole curve has been discouraged as a richer and more informative analysis should consider the whole functional response

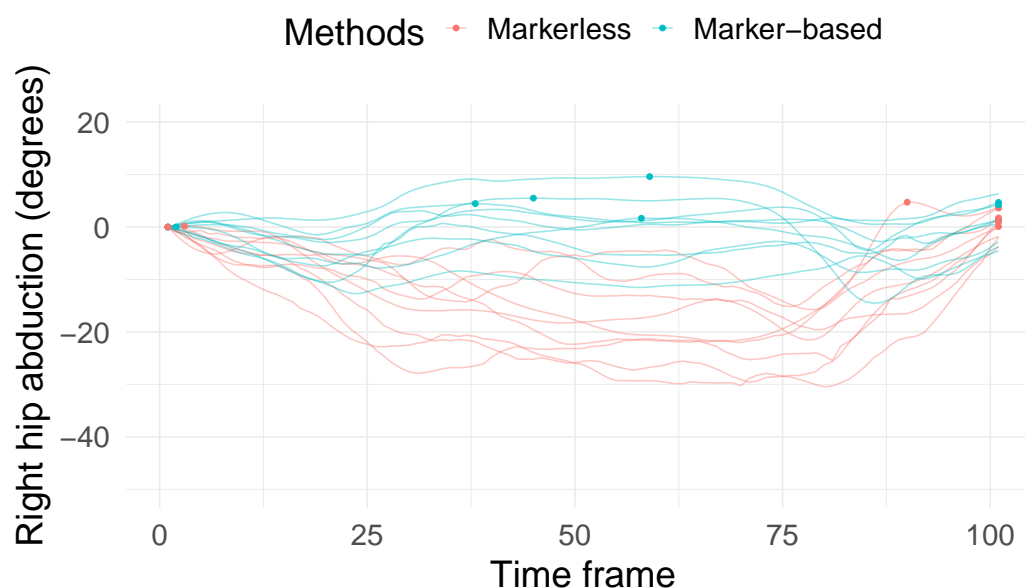


Figure 3.1: Right hip abduction angle curve of nine different athletes from first measurement session. Only first replicate by each method of measurement are displayed here. A dot on a curve represents the maximum value of the angle curve. One way to reduce functional responses to univariate responses would be by only keeping the maximum angle for the analysis.

(Donoghue et al., 2008; Richter et al., 2014; Pataky et al., 2013).

If the aim of the analysis is to make inference on the curves (using the functional responses), the investigation of the measurement methods should also be based on the functional responses rather than on derived univariate responses. Due to the lack of satisfactory statistical approaches for method comparison studies for functional responses, applied researchers still tend to use the reduced univariate responses instead of the available functional responses when assessing agreement (Perrott et al., 2017; Harsted et al., 2019).

Methods do exist however to generate 95% fLoA (Røislien et al., 2012). In this approach, each subject's set of observations first need to be converted into a smooth functional response. A difference curve is then calculated by taking the difference between the functional responses for the two different methods of measurement under investigation. The mean and standard deviation curves of these difference curves are then calculated and used to calculate the 95% fLoA. This approach is useful for

functional responses; however, it is limited in that it can only be used when functional responses measured by each measurement method are independent of each other. Thus, one can only apply this approach when one functional response per subject by a given measurement method was taken. If replicates for functional responses are available, this approach is no longer applicable.

To calculate 95% fLoA for functional responses from designs involving replicates, Olsen et al. (2013) proposed fitting separate LMMs at each time frame of the functional response. For example, if the functional responses were collected on a grid of 100 time frames, one needs to fit 100 different LMMs to account for the replicate responses and then calculate 100 separate estimates to calculate the 95% LoA for the 100 different time frames. Finally, all the 100 different 95% LoA are combined to get the overall 95% fLoA for the functional response. There are many elements in this approach that one might criticise, including that it is not an elegant or efficient way to use a statistical model. First, there is no smoothness in the estimated mean or standard deviation function, which are the essential characteristics of a functional response. Secondly, statistical tests for such situations are not well studied. Moreover, Olsen et al. (2013) only estimated the bias function (the mean of the difference curve) and its corresponding 95% confidence band. No estimate of the 95% LoA for functional responses was presented in the paper (Olsen et al., 2013).

How to estimate functional limits of agreement for studies with replicate functional responses is still an unanswered research question. The same applies for study designs that are more complex, such as hierarchical study designs and the inclusion of subject specific covariates.

A detailed introduction to FDA is now given in order to demonstrate the usefulness of this approach in method agreement studies with functional responses using the biomechanics case study by way of example.

3.3.1 Introduction to Functional Data Analysis (FDA)

Recall that in the motion capture study, the response variables of interest are functional responses, i.e. the angle over the range of movement. In this study, a sequence of

3.3. Functional Responses in Method Comparison Studies

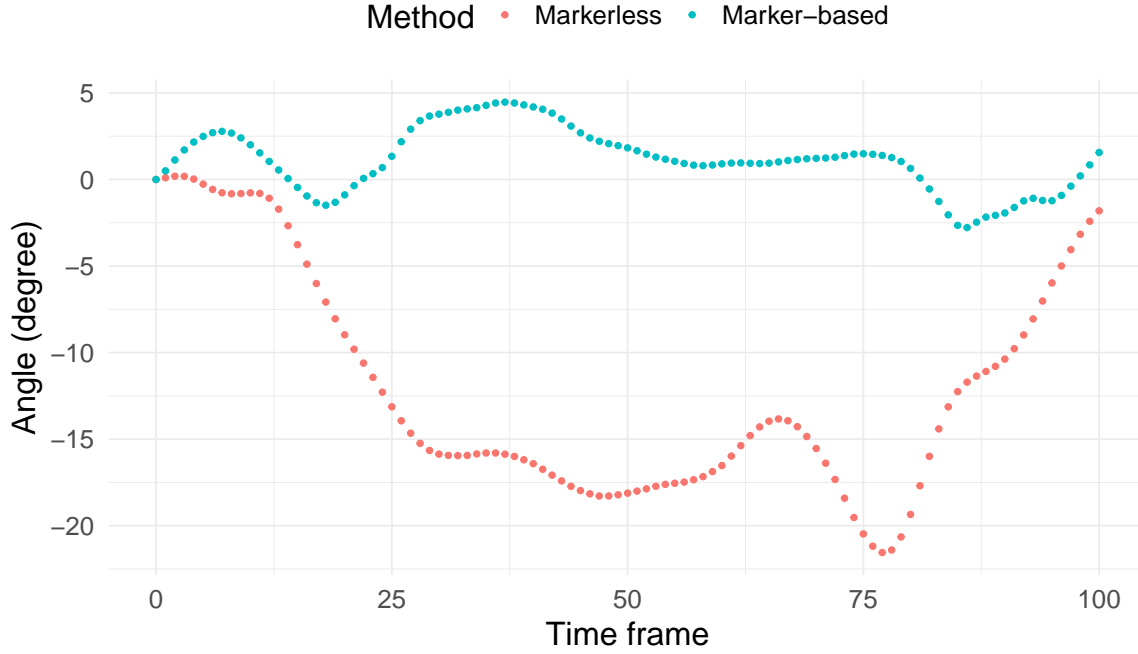


Figure 3.2: Right hip abduction angle curve measured by a marker-based and a markerless method of measurement during a lunge from an athlete in a given session.

measurements was taken using a marker and markerless system to measure the right hip angle while athletes were performing a lunge, typically taking less than a minute to complete. Figure 3.2 displays the sequence of measurements taken by the two methods of measurement for a particular athlete. The right hip abduction angle was measured at 101 equally spaced time frames for each measurement method. This 101 sequence of measurements constitute a single observation (i.e. a functional response), not 101 different univariate responses.

Consider, for a given measurement method, that these 101 discrete measurements are denoted by $Y_0, Y_1, Y_2, \dots, Y_{100}$ at time frame $t_0, t_1, t_2, \dots, t_{100}$. The measurements are discrete in nature but represent an underlying continuous function. The discrete sequence of data $\{Y_0, Y_1, Y_2, \dots, Y_{100}\}$ is called a functional response as they represent discrete realisations of an underlying function. If the underlying function can be written as $f(\cdot)$, then the following can be stated:

$$Y_j = f(t_j) + \epsilon_j, \quad j = 0, 1, 2, \dots, 100 \quad (3.1)$$

where, Y_j is the measurement at frame t_j , $f(t_j)$ is the underlying value of the function, and ϵ_j is the measurement error. Note that the underlying function $f(\cdot)$ is different for different athletes and could be different within a given athlete if replicates were taken.

One might argue that this could be considered longitudinal data or multivariate data. However, the main reason these are called functional data is that the underlying function is considered *smooth* (Ramsay and Silverman, 2002). The smoothness of a function ensures that the two adjacent values are very similar, i.e. Y_j and Y_{j+1} are very similar. Mathematically, a smooth function is a differentiable function. This smoothness requirement is why functional data are different from longitudinal data or multivariate data.

The function $f(\cdot)$ in model (3.1) can be represented as a linear combination of basis functions. Two basis systems are widely used in the literature, namely the Fourier basis and B-spline basis. The Fourier basis system considers $f(\cdot)$ as a Fourier series and is useful if it can be assumed that the function is periodic. A periodic function always has the same value for the initial and end values of the function. On the other hand, the B-spline basis system considers $f(\cdot)$ as a spline and is useful if the function is non-periodic.

For example, the right hip abduction angle can be considered periodic. However, due to the process used when collecting the data in the motion capture study, it can be seen in Figure 3.2 that the value of the function in the boundaries is not the same. For this reason, in this thesis, the B-spline basis system will be considered as the same issue arises in all angle measurements considered here.

Splines and B-splines

Mathematically, splines are piecewise polynomial functions with certain discontinuities of derivatives permitted at the joining points (Ahlberg et al., 1967). In other words, splines are a collection of polynomial functions defined on piecewise intervals on the domain of interest with certain conditions regarding differentiability must be met at the adjoining points.

3.3. Functional Responses in Method Comparison Studies

Let S be a piece-wise function defined on the interval $[a, b]$ as follows:

$$S : [a, b] \rightarrow \mathbb{R}. \quad (3.2)$$

Consider the partitioning of the interval $[a, b]$ into k pieces as follows:

$$[a, b] = [t_0, t_1) \cup [t_1, t_2) \cup [t_2, t_3) \cup \dots \cup [t_{k-1}, t_k] \quad (3.3)$$

where $a = t_0 \leq t_1 \leq \dots \leq t_k = b$. Here i^{th} interval is $[t_i, t_{i+1}]$. At each of the i^{th} interval a polynomial P_i of highest order n or degree $n - 1$ is defined such that

$$P_i : [t_i, t_{i+1}] \rightarrow \mathbb{R}. \quad (3.4)$$

where, $i = 0, 1, 2, 3, \dots, k - 1$. Now S can be defined on the interval $[a, b]$ through P_i as follows:

$$\begin{aligned} S(t) &= P_0(t), & t_0 \leq t < t_1, \\ S(t) &= P_1(t), & t_1 \leq t < t_2, \\ S(t) &= P_2(t), & t_2 \leq t < t_3, \\ &\vdots \\ S(t) &= P_{k-1}(t), & t_{k-1} \leq t \leq t_k. \end{aligned} \quad (3.5)$$

Consider at nodes $t_1, t_2, t_3, \dots, t_{k-1}$, the function S has multiplicity $m_1, m_2, m_3, \dots, m_{k-1}$, respectively and $D^j P_{i-1}(t) = D^j P_i(t)$, $j = 0, 1, \dots, n - 1 - m_i$; $i = 1, 2, \dots, k$. Here $D^j P_{i-1}(t)$ is the j^{th} derivative of $P_{i-1}(t)$. The function S is called a spline function of degree $n - 1$ or order n with nodes $t_1, t_2, t_3, \dots, t_{k-1}$ (Micula and Micula, 1999). The different t_i ; $i = 0, 1, \dots, k$ are called *knots* and t_i ; $i = 1, \dots, k - 1$ are called inner knots (Ramsay and Silverman, 2005).

A spline function can be represented by a B-spline basis system (de Boor, 2001). There are many ways such system can be constructed, however, the B-spline basis system developed by de Boor (2001) is the most popular one (Ramsay and Silverman,

2005). A B-spline basis system for splines with different order and sequences of knots can be constructed using the Cox-de Boor recursion algorithm (de Boor, 2001).

Consider a given knot sequence with p representing the number of inner knots. If the order of the B-spline is n , then there exists $n + p$ non-trivial basis functions. Figure 3.3 displays one set of B-spline basis functions for different orders of B-spline for a knot sequence of $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

A spline of order n with p interior knots can be defined as follows:

$$S(t) = \sum_l^{n+p} c_l B_{l,n}(t) \quad (3.6)$$

where c_l is the coefficient for the l^{th} B-spline basis function of order n with the given knot sequence. Equation (3.6) can be expressed with $f(t)$ instead of $S(t)$ to emphasise that the B-spline basis system can be used to represent any function $f(\cdot)$.

$$f(t) = \sum_l^{n+p} c_l B_{l,n}(t). \quad (3.7)$$

Smoothing Functional Data from Noisy Data

Equation (3.7) shows that any spline function can be represented with a suitable B-spline basis system. The same equation can be re-expressed again as follows:

$$f(t) = \sum_{l=1}^L c_l \phi_l(t) \quad (3.8)$$

where L is the number of B-spline basis functions induced by order of the spline and a given knot sequence, $\phi_l(t)$ is the l^{th} basis functions and c_l is the coefficient for the l^{th} basis function. Equation (3.8) can be simplified further using matrix-vector notation as follows:

$$\mathbf{f} = \mathbf{\Phi} \mathbf{c} \quad (3.9)$$

where, in the context considered, \mathbf{f} is an n -dimensional vector containing the value of the function evaluated at n different time frames, \mathbf{c} is the vector of length L with

3.3. Functional Responses in Method Comparison Studies

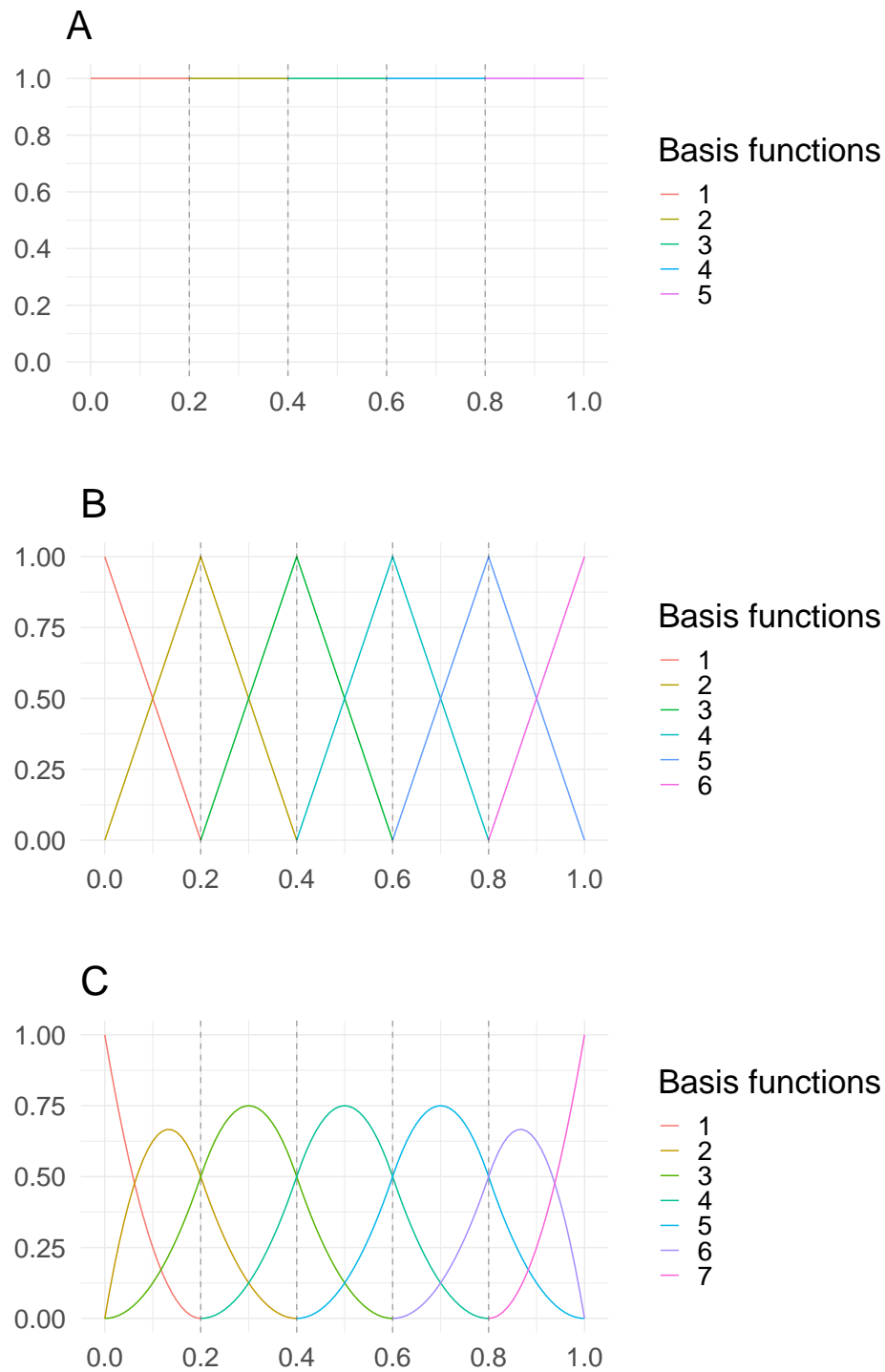


Figure 3.3: B-spline basis functions for order 1 (A), order 2 (B), and order 3 (C) splines with knot sequence $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

3.3. Functional Responses in Method Comparison Studies

l^{th} element the coefficient of the l^{th} basis function, Φ is a matrix of $n \times L$ dimension containing the l^{th} basis function at the l^{th} column. These basis functions in the Φ are also evaluated at the n different time frames.

Model (3.1) can now be written in concise vector-matrix notation as follows:

$$\mathbf{y} = \Phi \mathbf{c} + \boldsymbol{\epsilon} \quad (3.10)$$

where \mathbf{y} is a vector of length n containing the sequence of observed value of the function at points $\{t_0, \dots, t_n\}$, Φ is the $n \times L$ matrix with l^{th} column containing the value of the l^{th} basis function evaluated at points $\{t_0, \dots, t_n\}$, \mathbf{c} is a vector of length L containing the coefficient of the basis functions, and $\boldsymbol{\epsilon}$ is the n dimensional vector of measurement error at time frames $\{t_0, \dots, t_n\}$.

The least-squares criterion can be used to estimate the coefficient vector \mathbf{c} by minimising the following quantity (Ramsay and Silverman, 2005):

$$\text{SMSSE}(\mathbf{c}|\mathbf{y}) = (\mathbf{y} - \Phi \mathbf{c})'(\mathbf{y} - \Phi \mathbf{c}) \quad (3.11)$$

It can be shown that the estimate of \mathbf{c} that minimises the above quantity is as follows (Ramsay and Silverman, 2005):

$$\hat{\mathbf{c}} = (\Phi' \Phi)^{-1} \Phi' \mathbf{y}. \quad (3.12)$$

This approach is useful when it can be assumed that $\text{Var}(\mathbf{y}) = \Sigma_{\boldsymbol{\epsilon}} = \sigma^2 I$. When this is too simplistic to assume and the variance-covariance matrix $\Sigma_{\boldsymbol{\epsilon}}$ has a patterned structure then weighted-least squares can be used with the following objective function (Ramsay and Silverman, 2005) :

$$\text{SMSSE}(\mathbf{c}|\mathbf{y}) = (\mathbf{y} - \Phi \mathbf{c})' \mathbf{W} (\mathbf{y} - \Phi \mathbf{c}), \quad (3.13)$$

where $\mathbf{W} = \Sigma_{\boldsymbol{\epsilon}}^{-1}$. The corresponding estimate of the coefficient vector \mathbf{c} is (Ramsay

and Silverman, 2005):

$$\hat{\mathbf{c}} = (\Phi' \mathbf{W} \Phi)^{-1} \Phi' \mathbf{W} \mathbf{y}. \quad (3.14)$$

The choice of the number of basis functions influences the curve fitting. If one chooses a small number of basis functions, the fit might not be flexible enough. On the other hand, using too many basis functions may overfit the data. Ordinary cross-validation (OCV) error can be used to choose the optimal number of basis functions. The process of calculating such an error is as follows:

- Leave out one observation (t_i, y_i) when estimating the curve. Here y_i is the observed value of the curve at time point t_i
- From the remaining data estimate $\hat{f}_{-i}(t)$, the estimated function without the i^{th} observation being used in the estimation process
- Measure $y_i - \hat{f}_{-i}(t_i)$

The OCV score can then be calculated as $OCV(K) = \sum [y_i - \hat{f}_{-i}(t_i)]^2$. One can then choose the L , the number of basis functions, that minimises the OCV score. However, this procedure relies totally on minimising the OCV error and does not consider the smoothness of the fitted function, and this method may therefore produce a fitted function with too much wiggle. In addition, the choice of the number of basis functions depends on two factors for the B-spline basis system namely the degree of the spline and the position of the knots. If one fixes the degree, then the placement of the knots may still influence the fit.

To overcome the above issues, a penalised approach has been proposed to take control of the smoothness of the function and the SSE criterion. Using this approach the objective function is as follows:

$$\text{PENSSE}_\lambda(\mathbf{f}|\mathbf{y}) = [\mathbf{y} - f(\mathbf{t})] \mathbf{W} [\mathbf{y} - f(\mathbf{t})] + \lambda \int [D^2 f(s)]^2 ds. \quad (3.15)$$

The function $f(t)$ that minimises the $\text{PENSSE}_\lambda(\mathbf{f}|\mathbf{y})$ is a cubic spline with the knot position at the sampling point t_j (Ramsay and Silverman, 2005). This simplifies the

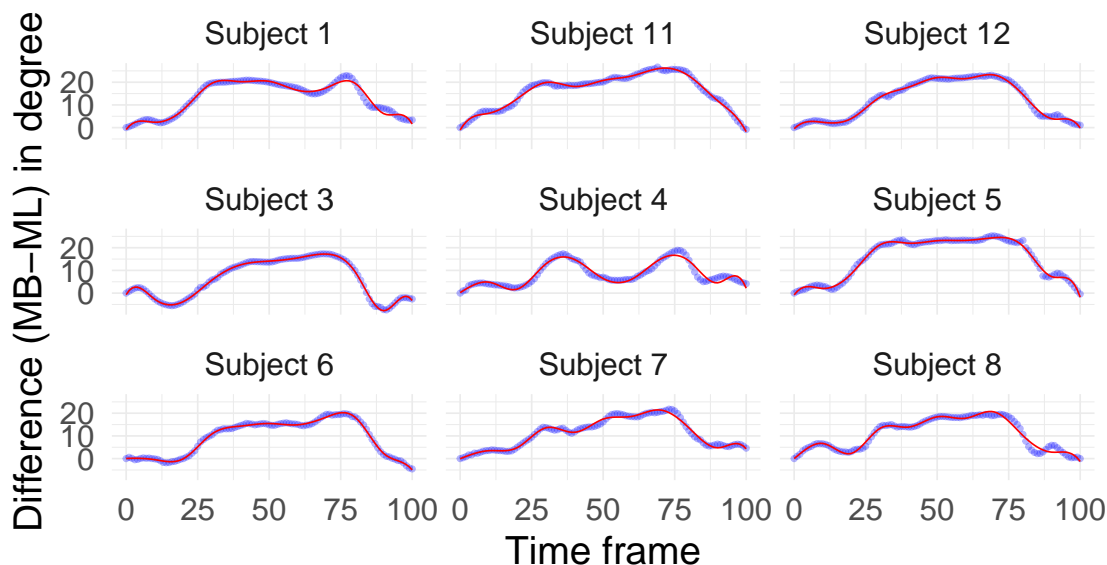


Figure 3.4: Observed difference (MB - ML) curves and smoothed functions from nine different athletes. Here MB is the measurement by the marker-based and ML is the measurement made by the markerless method. Blue dots are the observed differences and red solid curves are the smoothed functions obtained from it using smoothing splines.

estimation procedure by providing a guideline about the choice of the position of the knots and the degree of the spline allowing the focus to be on the λ , the smoothing parameter. One can choose λ by generalised cross validation (GCV) (Ramsay and Silverman, 2005). In practice, GCV provides more smooth function estimates than the OCV criterion.

All the formulas shown here are for estimating a single function from a single functional datum from a single individual. However, estimating the functional responses from all subjects can be done quickly by adding the SSE criterion from all the individual functional responses.

The first step of any functional data analysis is to obtain smooth functions from noisy data using the techniques described here. For example, observed difference curves, the difference between the measurements obtained by the two methods of measurement, were smoothed first (Figure 3.4). These smoothed difference curves are the data for any further analysis using FDA. In later section of this chapter these smoothed difference

functions will be used to calculate a 95% fLoA using FDA.

In this section techniques for obtaining smooth functions from noisy observed data have been outlined. In the next section the focus will be on how to calculate reliability and agreement indices for methods of measurement using these smoothed functions.

3.4 Reliability Indices for Functional Responses

The most straight forward approach for calculating reliability indices for functional data is to calculate the reliability indices for a single response at all time frames as suggested by Pini et al. (2019).

Consider $R(t)$ as the reliability index for a single response at time t . The reliability index could be the ICC or SEM for example. Although, Pini et al. (2019) suggested that the Pearson correlation coefficient can also be used, it can be argued that the Pearson correlation coefficient is not a suitable measure of reliability. The average value of the reliability index over the entire time domain $[a, b]$ can be calculated as follows:

$$R = \frac{\int_a^b R(t)dt}{b - a}. \quad (3.16)$$

The index in equation (3.16) is called an *integrated pointwise index* (Pini et al., 2019). For equally spaced time frames the value of R in equation (3.16) can be approximated as follows:

$$R = \frac{\sum_{t_0}^{t_n} R(t)}{t_n - t_0}. \quad (3.17)$$

The *Coefficient of Multiple Correlations* (CMC) has also been suggested to measure the reliability of methods measuring functional responses (Milner et al., 2011; Ford et al., 2007; Pini et al., 2019).

The CMC for the i^{th} individual is calculated as follows:

$$CMC_i = \sqrt{1 - \frac{\sum_{s=1}^S \sum_{t=t_0}^{t_n} (Y_{si}(t) - \bar{Y}_i(t))^2 / T(S - 1)}{\sum_{s=1}^S \sum_{t=t_0}^{t_n} (Y_{si}(t) - \bar{Y}_i)^2 / (TS - 1)}}, \quad (3.18)$$

where, $Y_{si}(t)$ is the s^{th} replicate of the functional response for the i^{th} subject, $\bar{Y}_i(t)$ is

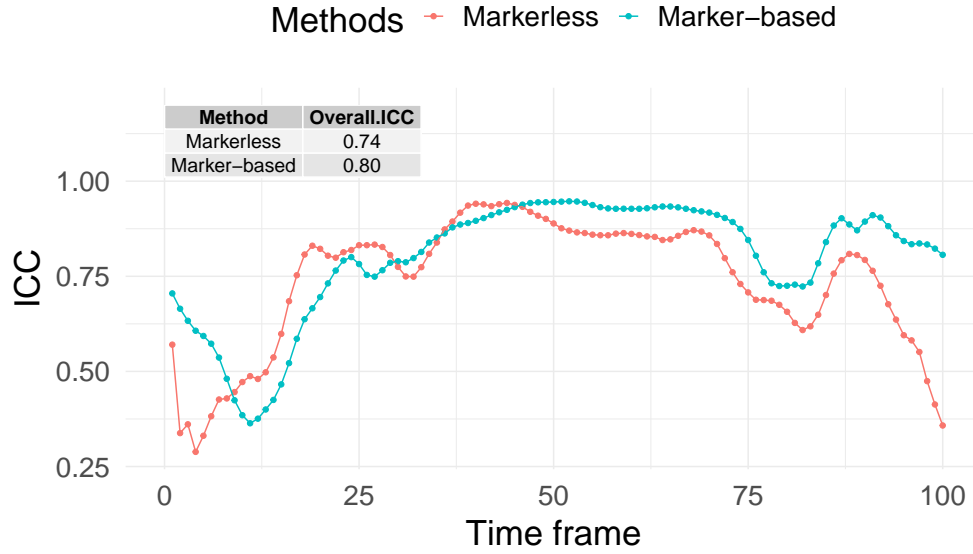


Figure 3.5: Intraclass correlation coefficient (ICC) curve for two methods of measurement: markerless and marker-based motion capture system. Only two replicates from the first measurement session were used to calculate the ICC.

the mean curve for the i^{th} subject, \bar{Y}_i is the mean of the mean curve across the time domain, T is the number of time frames, and S is the number of replicates measured for the i^{th} subject. Overall the measure of reliability is obtained by taking the average of the CMC from all subjects.

Pini et al. (2019) also suggested a distance measure and a similarity measure for assessing reliability. Both of these indices only apply when exactly two replicates were taken for a given subject. Consider, $Y_{i1}(t)$ and $Y_{i2}(t)$ as two replicates measured at time t , where $t \in [a, b]$. A distance measure can be defined as:

$$d(i) = \sqrt{\int_a^b (Y_{i1}(t) - Y_{i2}(t))^2 dt}. \quad (3.19)$$

If the measurement for the functional responses are taken at equally spaced times t_0, \dots, t_n , then the discrete approximation of the distance measure can be obtained as:

$$d(i) = \sqrt{\Delta t \sum_{t_0}^{t_n} (Y_{i1}(t) - Y_{i2}(t))^2} \quad (3.20)$$

3.4. Reliability Indices for Functional Responses

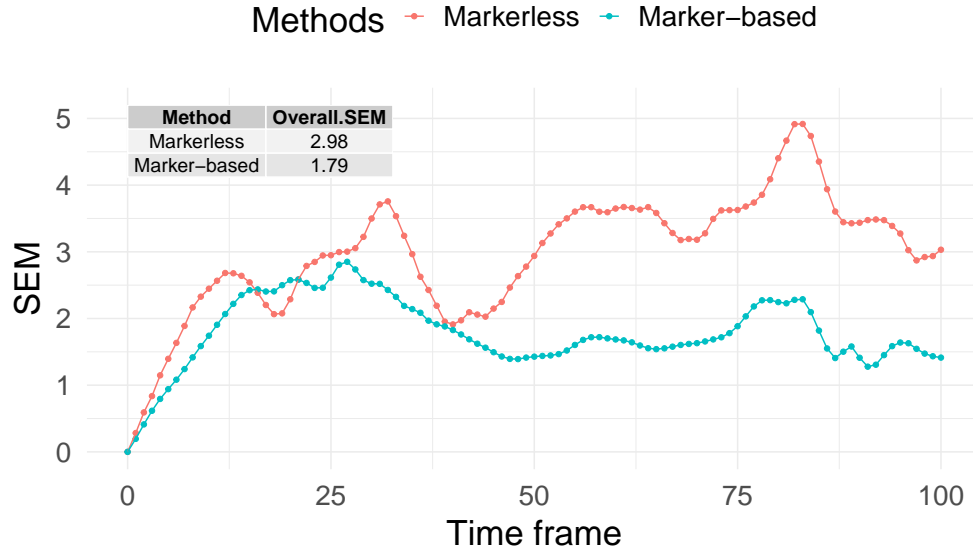


Figure 3.6: Standard error of measurement (SEM) curve for two methods of measurement: markerless and marker-based motion capture system. Only two replicates from the first measurement session were used to calculate the SEM.

where, $\Delta t = (t_n - t_0)/(n + 1)$. An overall measure of reliability can be obtained by computing the average value of the $d(i)$ for all subjects.

A similarity measure when two replicates were taken for a given subject can be obtained as follows:

$$S(i) = \frac{\int_a^b Y_{i1}(t)Y_{i2}(t)dt}{\sqrt{\int_a^b (Y_{i1}(t))^2 dt}\sqrt{\int_a^b (Y_{i2}(t))^2 dt}} \quad (3.21)$$

For equally spaced time points, the discrete approximation of the similarity is

$$S(i) = \frac{\sum_{t_0}^{t_n} Y_{i1}(t)Y_{i2}(t)}{\sqrt{\sum_{t_0}^{t_n} (Y_{i1}(t))^2}\sqrt{\sum_{t_0}^{t_n} (Y_{i2}(t))^2}} \quad (3.22)$$

Based on a simulation study, Pini et al. (2019) recommended the use of the integrated pointwise ICC and SEM as measures of reliability for functional data. For the motion capture study, pointwise ICC and SEM were calculated for both the methods of measurement using the two replicates from the first measurement session as an example. The ICC and SEM for both the methods of measurement were calculated

using the equation (3.17). In particular, the ICC for the markerless and marker-based method of measurement are 0.74 and 0.80, respectively. From the overall ICC it can be concluded that the reliability of both the measurement methods are similar. The same interpretation can be concluded from the pointwise ICC curve (Figure 3.5). However, from the pointwise SEM curve, it is clear that the marker-based method produces less measurement error compared to the markerless method of measurement throughout the measurement domain (Figure 3.6). The SEM for both measuring methods suggest that the markerless method produces about 66% more measurement error compared to the marker-based method of measurement (Figure 3.6). From these results, it appears that the marker-based method is more reliable than the markerless method of measurement for measuring the right hip abduction angle.

3.5 Measuring Agreement between Functions using FDA

An extension of the 95% LoA approach was proposed by Røislien et al. (2012). This extension is applicable when only one functional response per method per subject is taken. The details of this extension are as follows.

Let $Y_{i1}(t)$ and $Y_{i2}(t)$ be the measurements taken by method 1 and 2 at time t , where $t \in [a, b]$. The difference curve can be calculated as follows:

$$D_i(t) = Y_{i1}(t) - Y_{i2}(t). \quad (3.23)$$

Let $D_i(t)$ be represented as follows:

$$D_i(t) = f_i(t) + \epsilon_i(t), \quad (3.24)$$

where $f_i(t)$ is the smooth difference function for the i^{th} subject at time t , and $\epsilon_i(t)$ is the measurement error for the i^{th} subject at measurement time t .

To estimate each difference function $f_i(t)$, first the difference between the measure-

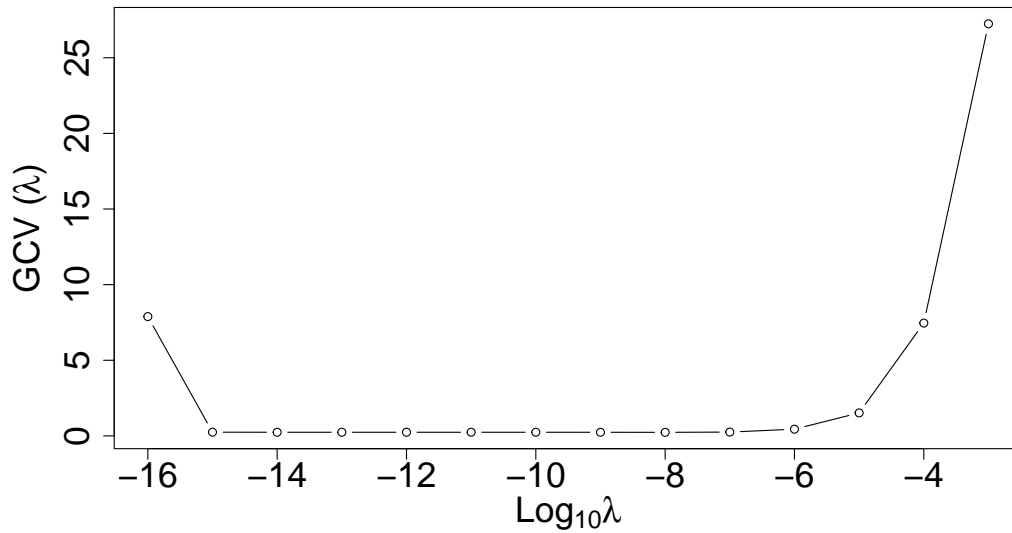


Figure 3.7: Generalised cross validation (GCV) score for different values of log lambda.

ments by the two methods for each subject at each time point is calculated. These difference curves are functional responses and hence a smoothing procedure described in Section 3.3.1 will be used to generate suitable smooth functions $\hat{f}_i(t)$.

As described in Section 3.3.1 a cubic spline with a knot position at each data point will be considered for the B-spline basis system. The difference function for each subject will be estimated by penalising the second derivative of the function to be estimated.

Figure 3.7 shows the GCV score for different values of λ , the smoothing parameter. The plot shows that the curve is flat for the value of λ from 10^{-15} to 10^{-5} . This means the choice of λ will not affect the estimation of the function much if a value is chosen between 10^{-15} to 10^{-5} . A value $\lambda = 10^{-5}$ was chosen which corresponds to about 21 degrees of freedom for the basis system.

The relative bias curve between the two measurement methods can be estimated as:

$$\begin{aligned}
 \hat{\text{Bias}}(t) &= \bar{f}(t) \\
 &= \text{E}(f_i(t)) \\
 &= \frac{\sum_i^n f_i(t)}{n},
 \end{aligned} \tag{3.25}$$

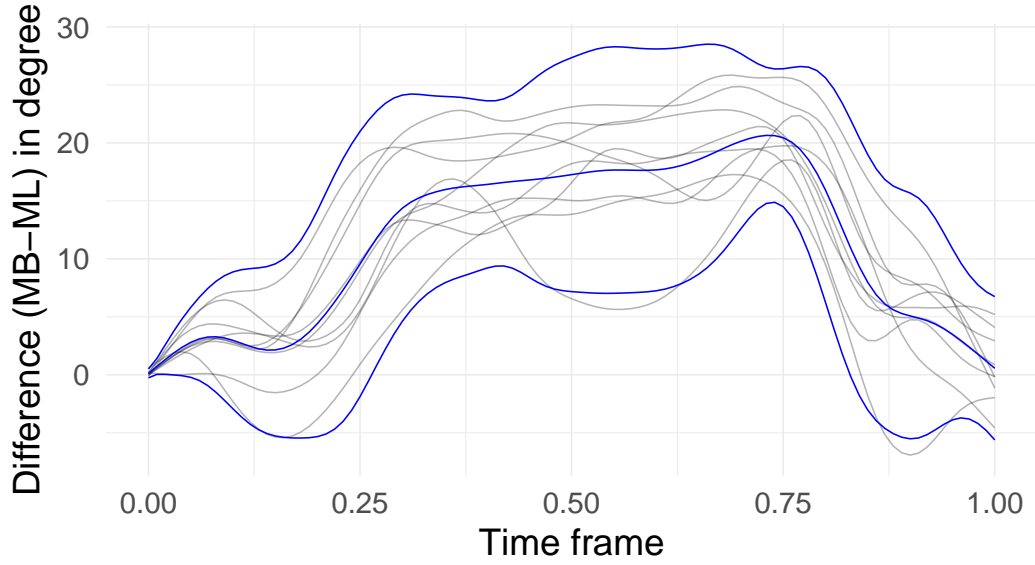


Figure 3.8: 95% functional limits of agreement estimated from one replicate measurement per method from a given session using the functional data analysis approach. Here, MB is the marker-based and ML is the markerless method of measurement.

where n is the total number of subjects. The standard deviation of the difference curve is

$$\begin{aligned}\hat{SD}(t) &= \sqrt{\text{Var}(f_i(t))} \\ &= \sqrt{\frac{\sum_i^n [f_i(t) - \bar{f}(t)]^2}{n-1}}.\end{aligned}\tag{3.26}$$

Assuming that at each time point the value of the difference function is normally distributed, the 95% fLoA can be estimated as

$$\begin{aligned}UL(t) &= \hat{\text{Bias}}(t) + 2\hat{SD}(t) \\ LL(t) &= \hat{\text{Bias}}(t) - 2\hat{SD}(t)\end{aligned}\tag{3.27}$$

where $UL(t)$ and $LL(t)$ are the upper limit curve and lower limit curve respectively.

Using the formula in equation (3.27) the 95% fLoA were calculated. The corresponding plot for the functional LoA are shown in Figure 3.8.

According to Røislien et al. (2012) fLoA should be interpreted pointwise. This means

that for a given time point the upper and lower limits of agreement should contain at least 95% of the values.

There is substantial bias between the two methods of measurement as indicated by the bias curve (Figure 3.8). On average, the markerless system overestimated the measurement by about 15 degrees or more for the right hip abduction angle during a lunge. For individual measurements, the difference between two measurements are typically between 5 to 25 degrees. This suggests that the two systems do not agree when measuring the right hip abduction angle curve. Here only one replicate measurement from one measurement session was used per method. A more detailed interpretation will be given in a later part of this chapter when the 95% fLoA are presented using all replicate measurements available from two different measurement sessions.

For the 95% LoA for a univariate response, the assumption that the mean and variance of the difference should remain constant for different magnitudes of the measured value was checked using a Bland-Altman plot. However, the same plot cannot be used here as the x-axis is already used to represent the time frame. Røislien et al. (2012) suggested concurrent functional regression between the average curve and the difference curve and a functional permutation test to verify the assumption. It can be argued that instead of using a statistical test to verify this assumption, a graphical tool is preferred over a statistical test, which is common practice in model fitting. This is particularly relevant when the number of subjects is low as statistical tests in such scenarios tend to have low power in such scenarios.

A natural choice would be to use the third dimension for the plot to verify the assumption. In the Bland-Altman plot the first two dimension were used to represent the averages and the differences. When considering functional data, the third dimension could be the time domain. Such a plot is displayed in the Figure 3.9. This figure contains all the necessary information to verify the assumptions. However, this figure is not practical as it is hard to interpret.

Figure 3.9 could be sliced at each time frame to interpret the plot more clearly. Such an attempt is shown in Figure 3.10. The figure shows the Bland-Altman plot at each time frame and suggests that the assumptions look sensible for the selected

3.5. Measuring Agreement between Functions using FDA

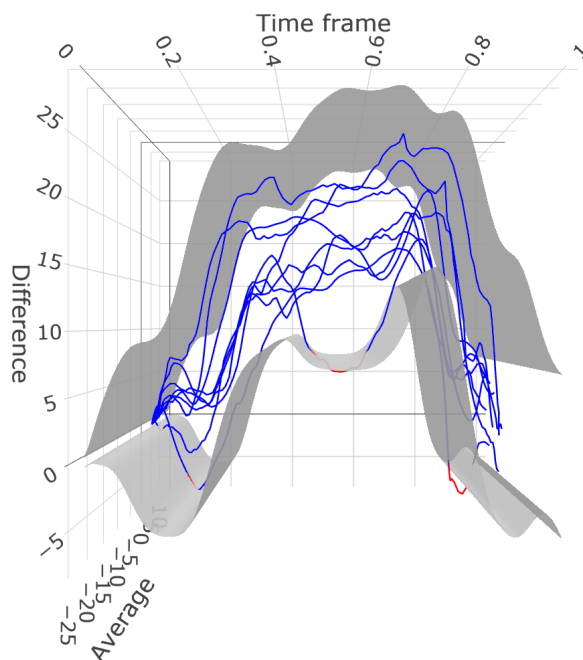


Figure 3.9: Three dimensional plot for the functional limits of agreement. Here time is added as the third dimension to the Bland-Altman plot.

frames. It is clearly not an elegant option to check the assumption as over 100 different Bland-Altman plots would need to be created and checked for this purpose. This topic will be explored further in a later part of this chapter.

In the situation where multiple replicates per method per subject are taken, the method proposed by Røislien et al. (2012) cannot be used as the approach proposed will produce incorrect estimates of the standard deviation curve. In the motion capture study, three replicate measurements per method per subject were taken. In addition, the data were collected over two different sessions. Due to the hierarchical nature of the study design, this approach for producing 95% functional limits of the agreement will not be applicable. Analysis of data of this type from complex hierarchical studies can be accommodated using a mixed-effects modelling framework. The mixed-effects model has already been used to calculate 95% LoA for single responses when multiple replicates were available. This approach will now be revisited and the suitability of

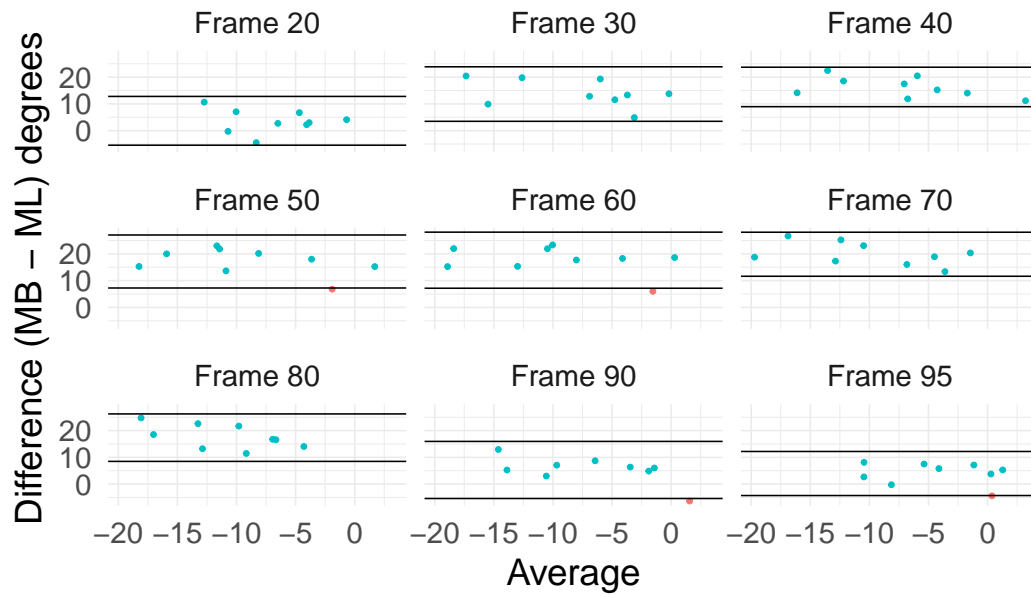


Figure 3.10: The Bland-Altman plot at different selected time points (frame) of the time domain. Here, MB is the marker-based and ML is the markerless method of measurement. Here, at each time point Bland-Altman plot was created considering the measurements as univariate responses.

nonparametric linear mixed models will be demonstrated to provide a framework for calculating 95% functional limits of agreement for functional responses in general.

Although both the FDA and nonparametric LMM can be used to analyse functional responses, their approach for estimating smooth functions is different. In FDA, first each functional response is converted into smooth functions and then any following analyses are based on these smoothed functions. As will be demonstrated in the next section, in LMM the average smooth function is first estimated and then the individual smooth functions can be obtained using the estimates of random-effects. Adjusting for replicates, covariates and hierarchical designs can then be accommodated quite easily.

3.5.1 Introduction to Nonparametric LMM

In this section the nonparametric mixed-effects regression modelling framework is introduced, initially for scenarios involving functional data where only one curve per subject per method is available.

3.5. Measuring Agreement between Functions using FDA

The idea for the method presented in this section came from Rice and Wu (2001) where they proposed a nonparametric mixed-effects model for functional data combining B-spline basis functions within the existing mixed-effects modelling framework. Their method proposed using B-spline basis functions for the fixed effects design matrix to model the nonlinear pattern in the mean curve and possibly a separate set of B-spline basis functions for the random-effects design matrix to model the covariance structure of the random-effects. The best linear unbiased prediction (BLUP) of the random effects can then be used to estimate the individual subject-specific curves.

Let n be the number of subject with n_i measurements taken from the i^{th} subject. Let Y_{ij} be the measurement on the i^{th} subject at time $t_{ij}; 0 \leq t_{ij} \leq T$. The time series of individual curves can be considered as the sum of the population mean function, a random function, and white noise. The population mean function will be approximated as a spline using a B-spline basis functions as follows (Rice and Wu, 2001):

$$E(Y_{ij}) = \mu(t) = \sum_k^K \beta_k \bar{B}_k(t), \quad (3.28)$$

where, $\mu(t)$ is the population mean function, $\bar{B}_k(t)$ is the k^{th} cubic B-spline basis function with a given sequence of knots, β_k s are the corresponding coefficients for the basis functions. Similarly, subject specific deviation curves can be approximated using a spline function $\sum_l^L b_{il} B_l(t)$. Here, $B_l(t)$ is possibly a different set of B-spline basis functions from a different space of spline functions. The coefficients b_{il} follow a Normal distribution with mean zero and covariance structure Ψ . If it can be considered that the error ϵ_{ij} follows a Normal distribution with mean zero and variance σ^2 , then the curve can be modelled using a nonparametric mixed-effects regression model as follows (Rice and Wu, 2001):

$$Y_{ij} = \sum_k^K \beta_k \bar{B}_k(t_{ij}) + \sum_l^L b_{il} B_l(t_{ij}) + \epsilon_{ij}, \quad (3.29)$$

where, ϵ_{ij} are independent of the random effects b_{il} .

Consider \mathbf{y}_i as a vector containing the time series measurements for a single curve

of the i^{th} subject, thus the mixed-effects model can then be represented as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad (3.30)$$

where, \mathbf{X}_i is a $n_i \times p$ fixed-effects design matrix containing B-spline basis functions as different columns, $\boldsymbol{\beta}$ is a p -dimensional vector containing the corresponding fixed-effects, \mathbf{Z}_i is a $n_i \times q$ random-effects design matrix containing possibly a different B-spline basis functions as different columns, \mathbf{b}_i is a q -dimensional vector containing the corresponding random-effects, $\boldsymbol{\epsilon}_i$ is the n_i -dimensional error vector. \mathbf{b}_i is normally distributed with mean zero and covariance structure $\boldsymbol{\Psi}$ and $\boldsymbol{\epsilon}_i$ is also normally distributed with mean zero and covariance structure $\sigma^2 I$. The random-effects \mathbf{b}_i are independent of the error $\boldsymbol{\epsilon}_i$. This model is only applicable if the errors can be considered white noise (i.e. unstructured). However, the situation may arise that this assumption regarding the error may not be the case. For example, if the errors are serially correlated, a simple white noise assumption is not suitable. If $\boldsymbol{\Sigma}$ can be considered as the patterned variance-covariance matrix instead of σI , the model could handle both heteroscedastic and correlated errors. The REML criterion will be used to estimate all model parameters.

In the next section this nonparametric mixed-effects modelling framework will be used to calculate the 95% fLoA for a method comparison study.

3.6 Measuring Agreement between Functions using LMM

In this section, a new approach is proposed to calculate 95% fLoA using a nonparametric LMM. First the approach will be demonstrated for a situation without any replicates. The approach will then be demonstrated by way of example using the hierarchical study design used in the motion capture study. In this case, all the functional responses for the right hip abduction angle for the motion capture study will be used to calculate the 95% fLoA to assess the agreement between the two methods of measurement.

3.6.1 95% fLoA using a LMM for Studies with no Replicates

In model (3.30), it was only considered that a single curve is measured from each subject. The model can be extended to handle a situation where two curves are measured by two different methods of measurement for each subject. Consider m as the index for measurement method, then the model can be extended to

$$\mathbf{y}_{mi} = \mathbf{X}_{mi}\boldsymbol{\beta}_m + \mathbf{Z}_{mi}\mathbf{b}_{mi} + \boldsymbol{\epsilon}_{mi} \quad (3.31)$$

where, \mathbf{y}_{mi} is the vector containing the measurement curve of the i^{th} individual by the m^{th} measurement method, $\boldsymbol{\beta}_m$ is the vector containing the fixed effects for the m^{th} method. There will be two separate mean curves for each of the methods. \mathbf{b}_{mi} is the random effects for the i^{th} subject under measurement method m . Finally, $\boldsymbol{\epsilon}_{mi}$ is the error curve. Different methods will have different sets of errors since each measurement method could have a different error structure.

To estimate the 95% fLoA, only the bias curve and variance curve of the difference curves are needed. Rather than estimating model (3.31) and then calculating the bias curve from it, the bias curve can be directly estimated by considering the difference curves as responses. For this reason, the difference curve is calculated:

$$\mathbf{d}_i = \mathbf{y}_{m_1i} - \mathbf{y}_{m_2i} \quad (3.32)$$

where, \mathbf{y}_{m_1i} and \mathbf{y}_{m_2i} are the curves measured by the measurement method m_1 and m_2 , respectively. This difference curve, \mathbf{d}_i can be modelled as follows:

$$\begin{aligned} \mathbf{d}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i ; i = 1, \dots, 9 \\ \mathbf{b}_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned} \quad (3.33)$$

where, $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$ are positive-definite matrices. \mathbf{d}_i is n_i dimensional response vector containing the difference curve for the i^{th} subject. \mathbf{X}_i is the $n_i \times p$ dimensional fixed-effect regressor matrix containing p B-spline basis functions, where p is the number of B-spline basis functions sufficient to represent the mean difference curve. A cubic

3.6. Measuring Agreement between Functions using LMM

B-spline with a suitable knot sequence is an attractive choice given its flexibility for the regressor matrix. The approach for choosing the knot sequence will be discussed later. β is a p dimensional vector containing the fixed-effects. \mathbf{Z}_i is a $n_i \times q$ dimensional random-effects regressor matrix containing q B-spline basis functions. \mathbf{b}_i is the q dimensional vector containing the random-effects. ϵ_i is the n_i dimensional error vector. The error ϵ_i are assumed to be independent of the random-effects \mathbf{b}_i .

Model 3.33 is only suitable for the situation where a single replicate has been measured using each of the measurement method. However, in the motion capture study, functional responses were taken in two different sessions, and there are three replicates in each session for each method. The model needs to be extended therefore into a hierarchical one to incorporate this. Such a hierarchical model will be considered later.

The bias between the two methods of measurement is

$$\boldsymbol{\mu}_d = E(\mathbf{d}_i) = \mathbf{X}_i\boldsymbol{\beta} \quad (3.34)$$

and the variance-covariance matrix for \mathbf{d}_i is

$$\begin{aligned} \text{Var}(\mathbf{d}_i) &= \text{Var}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i) \\ &= \mathbf{Z}_i\Psi\mathbf{Z}_i' + \Sigma \end{aligned} \quad (3.35)$$

Now the 95% fLoA are

$$\boldsymbol{\mu}_d \pm 2\sqrt{\text{diag}(\text{Var}(\mathbf{d}_i))} \quad (3.36)$$

To calculate the 95% limits of agreement for functional responses, model (3.33) needs to be fitted. There are three things that need to be chosen for the specification of the mixed-effects model. Firstly, one needs to choose the appropriate knot sequence of a B-spline basis system for the fixed-effects regressor matrix. The next step is to specify the knot sequence for a B-spline basis system for the random-effects regressor matrix and the specification of the variance-covariance matrix for the random-effects. The final step is to choose an appropriate correlation structure and a variance functions

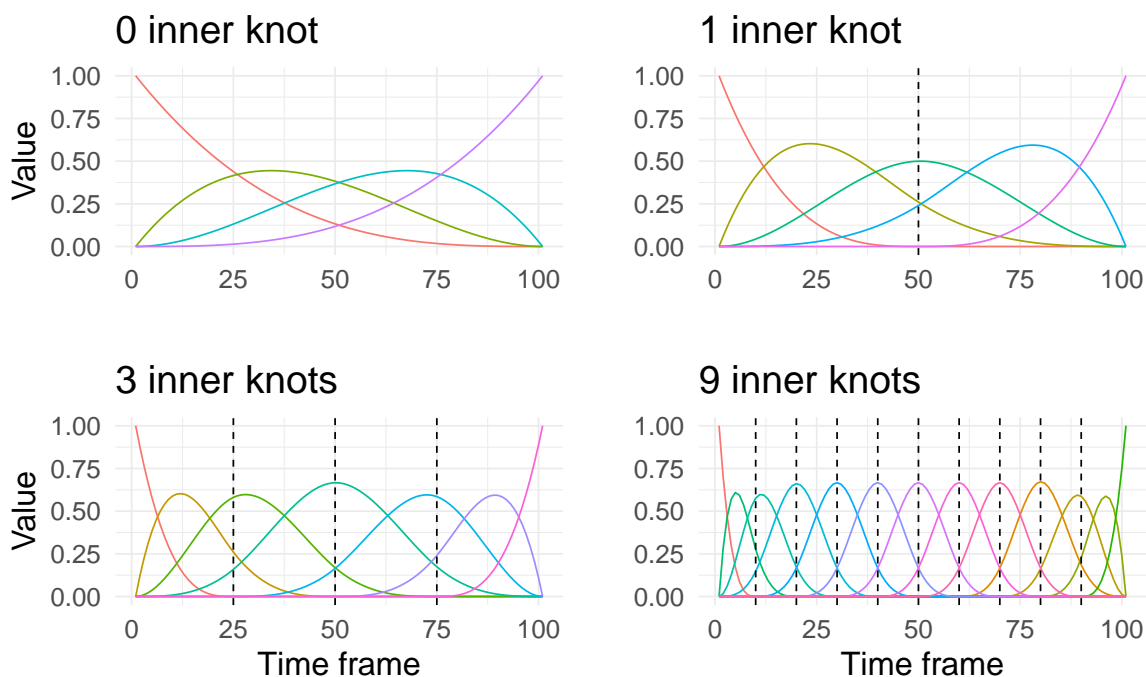


Figure 3.11: Cubic B-spline basis functions with a different number of equally spaced inner knots.

for the error.

Fixed Effect Structure

To model the mean curve, a choice of a set of B-spline basis functions for the fixed-effects regressor matrix is needed. As mentioned before, the choice of the set of the B-spline basis functions for the fixed-effects regressor matrix depends on the degree of the spline curve and the position of the knot sequence. Since derivatives of the curves are not of interest, a cubic spline will be sufficient (Ramsay and Silverman, 2005) and only a sequence of inner knots needs to be selected. There are many possible choices to consider for the knot sequence. Figure 3.11 shows four different options to choose from for the sequence of knots for the fixed effect structure. These are not the only possible choices, but when there is no reason to put knots on specific points on the domain, equally spaced knot sequence seems a reasonable choice. In Figure 3.11, there is no inner knots for the spline functions in the top left panel. There is only one inner knot for the top right panel in the middle of the domain. There are three and nine

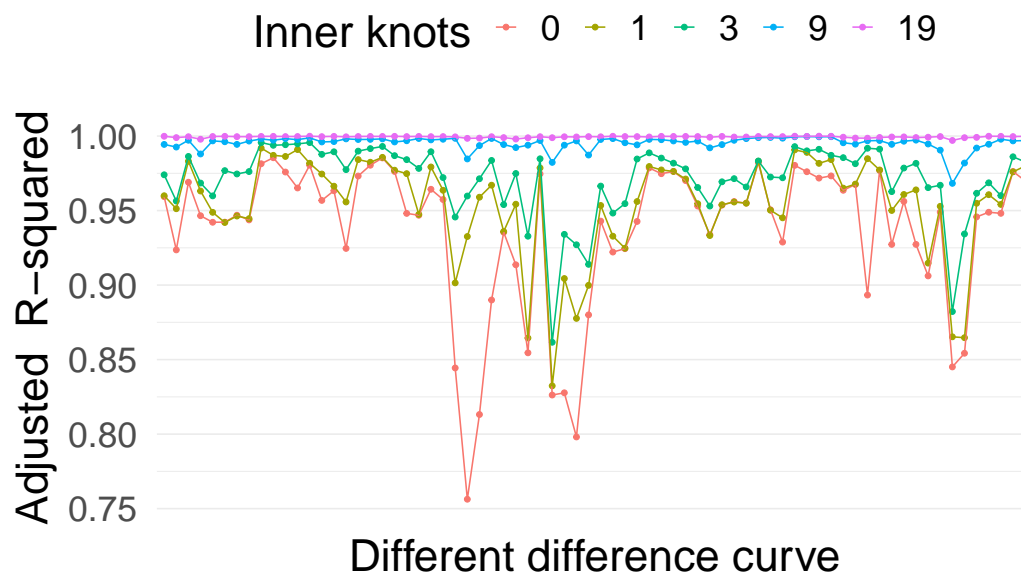


Figure 3.12: Adjusted R-squared values after fitting a linear regression model to each individual difference curve using B-spline basis functions with different numbers of inner knots as the covariates. Here different points on x-axis are different difference curve. In total 54 difference curves (9 athletes, 2 sessions, 3 replicates) were used for the right hip abduction angle measured by two different measurement methods: markerless and marker-based.

equally spaced inner knots in the bottom two panels. Depending on the number of inner knots, a different B-spline basis system will be generated with a different number of basis functions. An objective criterion is needed to pick a suitable combination of these for the context in question, in this case the motion capture study.

One way of selecting a knot sequence is to fit different LMMs with different knot sequences and then use a statistical test (e.g. likelihood ratio test) to choose one. However, this would be a cumbersome task as there are many possible choices for the sequence of knots. In this thesis, a more straightforward visual approach is proposed to guide the choice of the appropriate knot positions. The approach is based on the R-squared statistic. Separate multiple linear regressions will be fitted for each of the individual replicate measurements for different sets of B-spline basis system. Finally, a sequence of knots will be chosen where most of the adjusted R-squared values are close to 1.

This approach for choosing a sequence of knots using the R-squared criterion is

3.6. Measuring Agreement between Functions using LMM

Table 3.1: Different candidate knot sequences for the random-effects structure.

Notation	Different knot sequence	DoF ¹	AIC ²	BIC ³
ks 1	{1, 101}	3	-1293.793	-1013.212
ks 2	{1, 50, 101}	4	-1327.014	-1031.266
ks 3	{1, 25,50,75, 101}	6	-1360.924	-1034.843
ks 4	{1, 20, 40, 60, 80, 101}	7	-1377.994	-1036.747
ks 5	{1, 15, 30, 45, 60, 75, 90, 101}	9	-1387.286	-1015.705
ks 6	{1,10, 20, 30, 40, 50, 60, 70, 80, 90, 101}	12	-1176.209	-759.128

¹ DoF: degrees of freedom;

² AIC: Akaike information criterion;

³ BIC: Bayesian information criterion.

now demonstrated for the difference curves obtained for the right hip abduction angle measured during a lunge exercise.

Figure 3.12 shows the R-squared statistics calculated from the individual difference curves after fitting a multiple linear regression model to each curve using the set of basis functions as the covariates. It suggests that the sequence of knot positions with zero inner knots cannot model all the individual curves adequately. As one increases the number of inner knots, the set of basis functions can model more individual curves. For example, with 19 inner knots, the model can accommodate all the individual curves very well; however, with only nine breakpoints, the fit is almost as satisfactory as the fit with 19 inner knots. For this reason, a B-spline basis system with 9 inner knots will be considered for the fixed-effects regressor matrix.

Random Effects Structure

To choose a suitable basis system for the random-effects regressor matrix, the approach suggested by Rice and Wu (2001) will be considered here. According to this approach, different LMMs will be fitted with different B-spline basis system for the random-effects design matrix. As only cubic B-splines will be used here, the choice of basis system relies only on the knots sequence. Table 3.1 shows the list of candidate knots sequence for the random-effects regressor matrix. The *Akaike information criterion* (AIC) and *Bayesian information criterion* (BIC) can then be used to identify a suitable basis system.

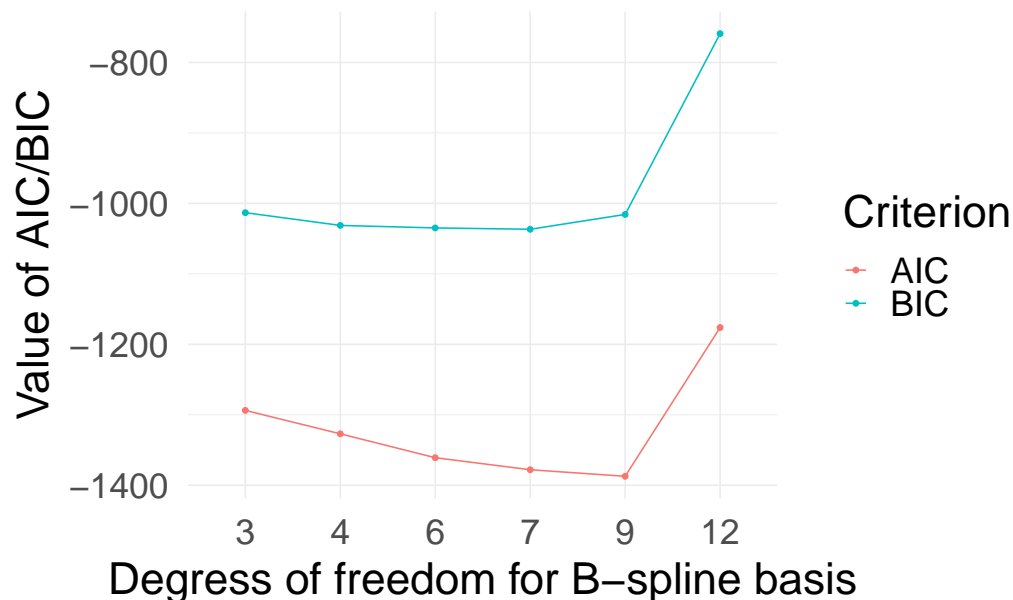


Figure 3.13: Akaike information criterion (AIC) and Bayesian information criterion (BIC) value for different mixed-effects models fitted with different knot sequences for the random-effects structure for the right hip abduction angle data.

In addition to the specification of the B-spline basis system, the structure of the variance-covariance matrix must be specified. Here a diagonal variance-covariance matrix for the random-effects will be considered.

Figure 3.13 shows the AIC and BIC value of the LMMs with different knot sequences. From this, it is clear that the AIC criterion suggests a knot sequence with 9 degrees of freedom. On the other hand, the BIC criterion suggests that any knot sequence that gives 4 to 9 degrees of freedom is suitable for the basis system for this analysis. For the random-effects regressor matrix, the knot sequence $\{1, 20, 40, 60, 80, 101\}$ for the B-spline basis system will be considered based on the value of the AIC and BIC.

Correlation Structure for Error

Choosing the most appropriate correlation structure for an LMM when used to calculate functional limits of agreement is the next consideration. The list of correlation structures and the corresponding R function in the `nlme` package has already been discussed (Table 2.2).

3.6. Measuring Agreement between Functions using LMM

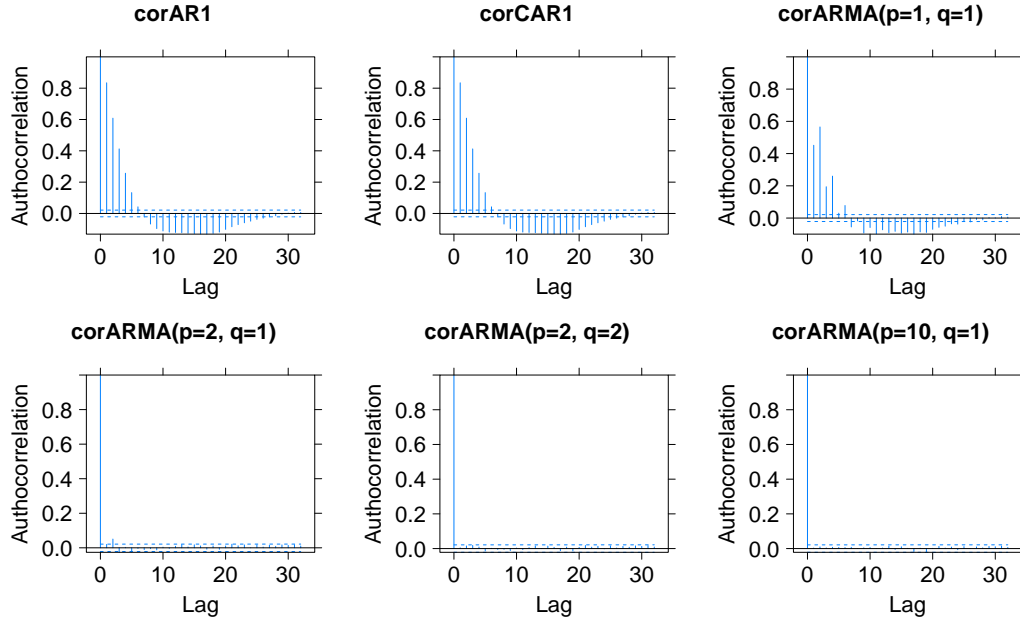


Figure 3.14: Auto-correlation plots of residuals after fitting a mixed-effects model with different temporal correlation structures for the right hip abduction angle data. Here, AR1 is an autoregressive model with order 1; CAR1 is a continuous autoregressive model with order 1; ARMA(p, q) is an autoregressive-moving-average model with order p for the autoregressive model and order q for the moving-average model.

The strategy is to fit different LMMs with different correlation structures, look at the autocorrelation plot for each model, and then decide what correlation structure best removes the serial correlation from the residuals. Figure 3.14 shows that an AR(1) model or an ARMA(1,1) model do not remove the serial correlation completely. An ARMA(2,1) and ARMA(2,2) almost removes the serial correlation from the error. An ARMA(10,1) model removes the serial correlation completely. However, specifying this correlation structure for the error will require many correlation parameters to be estimated, resulting in convergence issues. For this reason, the ARMA(2,1) model was selected as a suitable correlation structure for the right hip abduction angle curve during the lunge.

The value of the AIC for these models can also be examined. Figure 3.15 shows the value of the AIC for different LMMs with different correlation structures. This graph also suggests either the ARMA(10,1) or ARMA(2,2) or ARMA(2,1) structures. Since an ARMA(2,1) uses less degrees of freedom, this correlation structure was used for the

3.6. Measuring Agreement between Functions using LMM

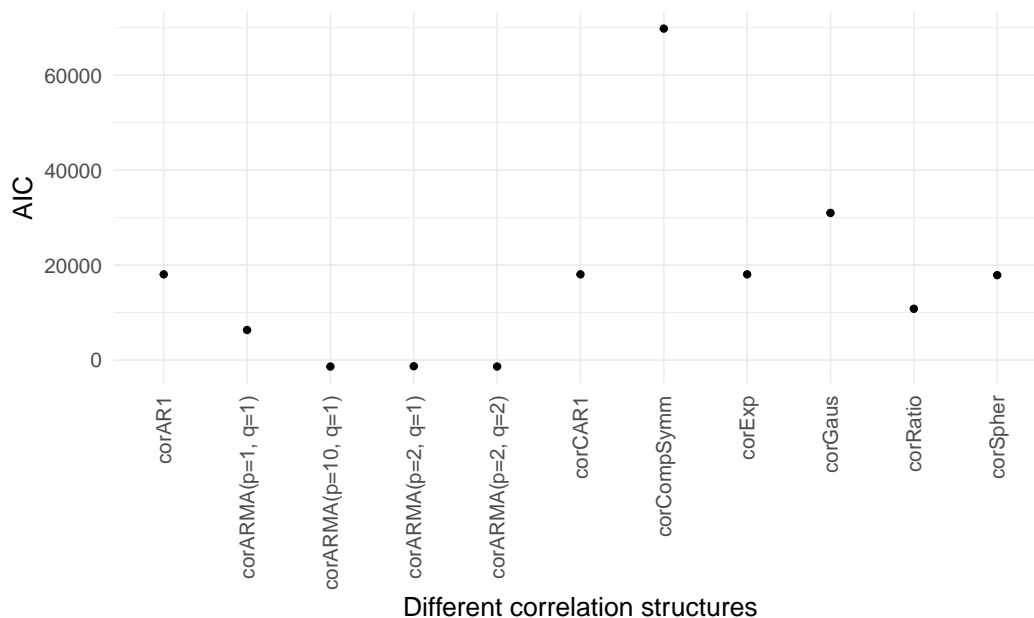


Figure 3.15: Akaike information criterion (AIC) of the mixed-effects models fitted with different correlation structure for the right hip abduction angle data.

final model.

Variance Function for Error

After selecting the correlation structure, the variance function needs to be specified to see if it improves the model fit. Selecting the correlation structure will always be necessary due to the inherent characteristics of functional responses. However, sometimes the error variance could be heteroscedastic, and the final model must account for this. It may not always be necessary, but it is always a pragmatic approach to examine this before fitting the model to calculate fLoA.

To find out what variance function is appropriate for a given situation, different LMMs with different variance functions can be fitted and compared. Figure 3.16 shows a plot of the standardised residuals over time (frame) for different LMMs with different variance functions. However, from this plot it is not obvious which variance function to choose among the no variance, `varFixed`, `varIdent`, `varPower`, and `varExp`. For this reason it can be concluded that no variance function will be sufficient for the final LMM to estimate fLoA.

3.6. Measuring Agreement between Functions using LMM

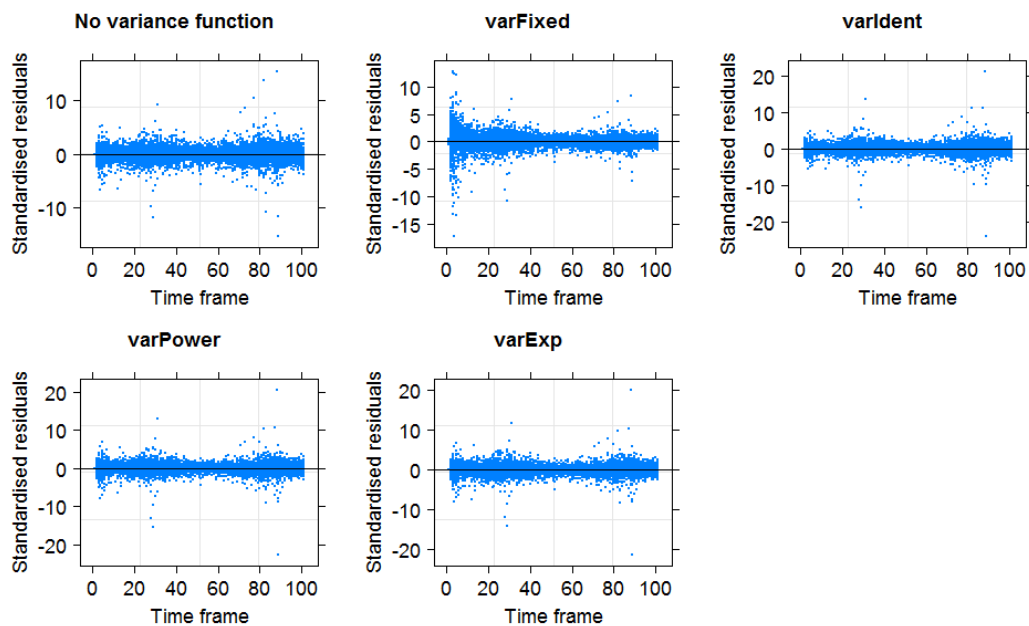


Figure 3.16: Standardised residual over time frames from different mixed-effects models fitted with different variance functions.

Model Specification of LMMs for 95% fLoA

The 95% fLoA depend on two estimates: the sample mean curve and the sample standard deviation curve. It needs to be checked whether the model specification of the LMM produces sensible estimates of the mean and standard deviation curves of the difference curves. The estimate of the mean and standard deviation curve of the difference curves should be similar compared to the estimates provided by FDA approach.

For model (3.33), guidelines for the specification of the different model components have been outlined in the previous sections. An investigation is now needed as to whether these specifications produce sensible estimates to calculate fLoA.

The standard deviation curve depends on the set of basis functions for the \mathbf{Z} matrix, the structure of the variance-covariance matrix for the random effect. It needs to be checked whether the number of B-spline basis functions and the structure of the variance-covariance matrix of the random-effects are sufficient to estimate the “true” standard deviation curve. A number of different LMMs with different specifications were chosen to identify which choices are sensitive to the estimation of the mean and standard deviation curves. To do this four different LMMs will be considered here.

3.6. Measuring Agreement between Functions using LMM

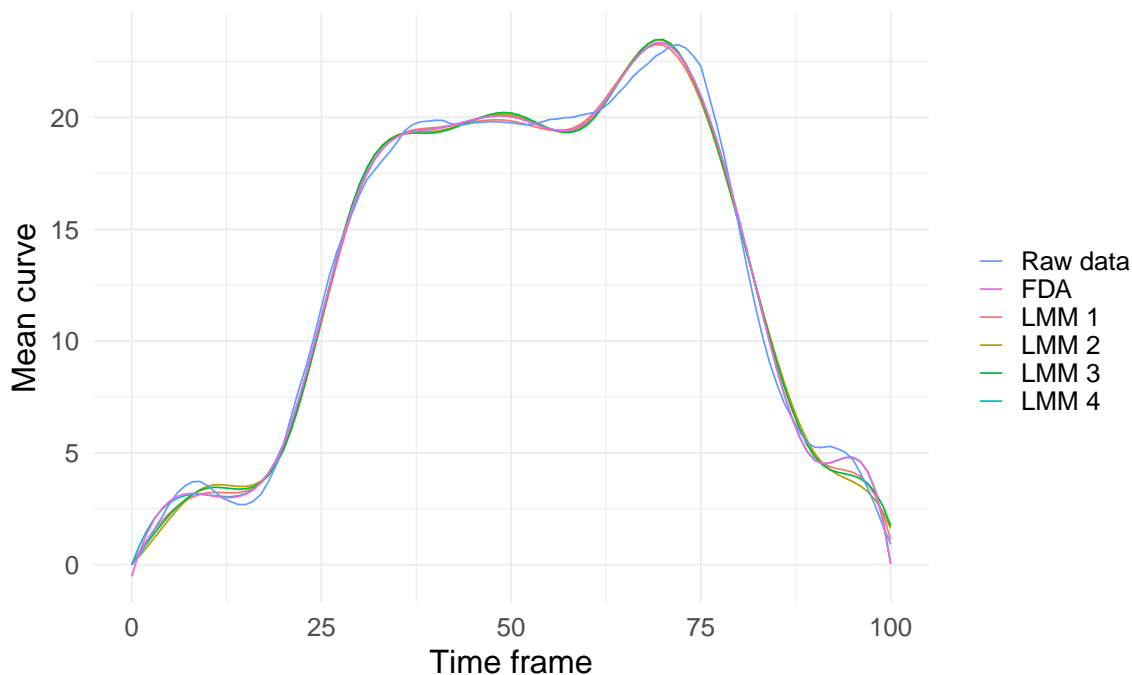


Figure 3.17: Comparison of the estimate of the mean curve using the functional data analysis (FDA) and different linear mixed-effects models (LMM). The sample mean curve is also shown with the label “raw data”.

LMM-1 has 12 basis functions for the fixed-effects regressor matrix with 7 basis functions for the random-effects regressor matrix, a diagonal variance-covariance structure for the random-effects, and exponential variance function and ARMA(1,1) correlation structure for the error. LMM-2 has 12 basis functions for the fixed-effects and random-effects regressor matrix, diagonal variance-covariance structure for the random-effects, an ARMA(1,1) correlation structure for the error, and no variance function for the error. LMM-3 has 12 basis functions for fixed-effects and random-effects regressor matrix, an unstructured variance-covariance structure for the random-effects, an ARMA(2,1) correlation structure for the error, and no variance function for the error. LMM-4 has 12 basis functions for the fixed-effects and random-effects regressor matrix, and an unstructured variance-covariance structure for the random-effects and no correlation structure or variance function for the error. Now the estimates of the mean and variance curve will be compared for these four different LMMs.

Figure 3.17 shows that the estimation of the mean is almost the same for all the

3.6. Measuring Agreement between Functions using LMM

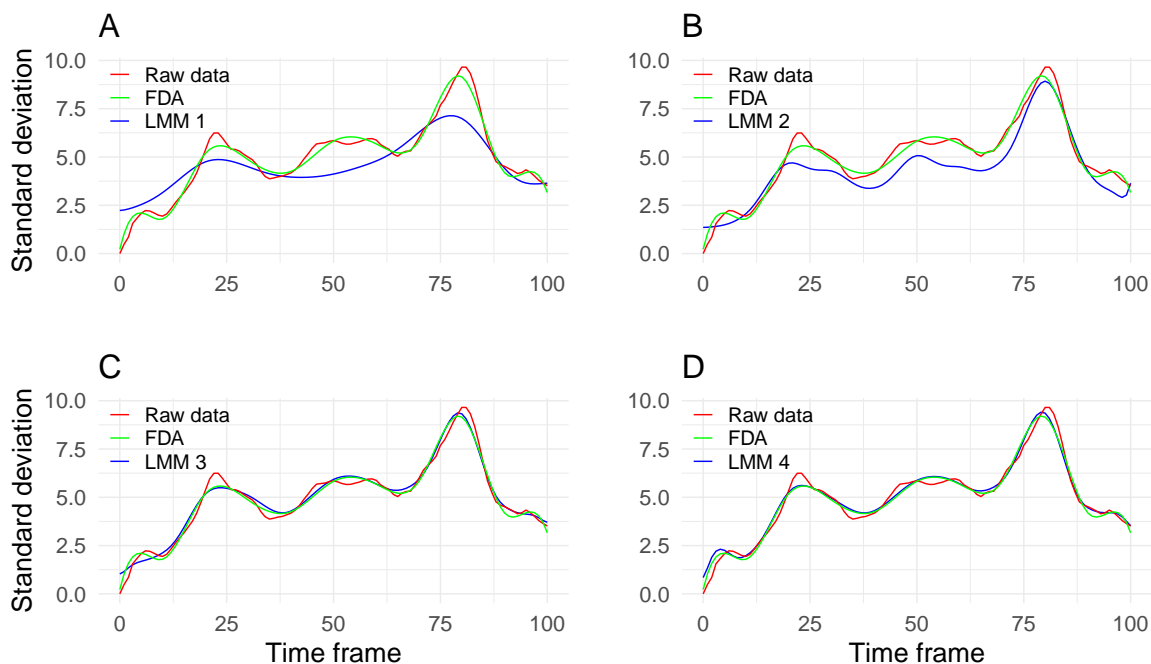


Figure 3.18: Comparison of the sample standard deviation curve (raw data) and estimate of the standard deviation curve by functional data analysis (FDA) with different specifications of the mixed-effects model (LMM).

models considered. The reason for this is that the estimation of the mean curve mainly depends on the number of basis functions used in the fixed-effects regressor matrix. In other words, the estimation of the mean curve is only influenced by the sequence of knots considered for the set of B-spline basis functions used for the fixed-effects structure. From this it can be concluded that the approach of choosing the knot sequence for the fixed-effects structure using the adjusted R-squared criterion is effective in this context.

The standard deviation curve obtained from the LMMs, the standard deviation curve calculated using the observed data and the FDA approach will be examined next. Figure 3.18-A shows a comparison of the estimate of the standard deviation curve using the FDA method and LMM-1. The sample standard deviation curve is depicted with the label “raw data”. Figure 3.18-B shows a similar comparison as in panel 3.18-A but here the LMM-2 was considered. Panel 3.18-C displays LMM-3. Panel 3.18- D displays LMM-4.

Figure 3.18-A shows that the LMM-1 was not adequate to estimate the “true” standard deviation curve. The number of basis functions for the random-effects were

not sufficient enough to capture the pattern of the standard deviation curve. Using more basis functions may produce a more flexible standard deviation curve suitable for the current context.

A re-specified LMM (LMM-2) with 12 basis functions for the \mathbf{Z} matrix, same as for the fixed-effects design matrix \mathbf{X} , with a diagonal structure for the variance-covariance matrix for the random-effects and an ARMA(2,1) correlation structure for the within group error was considered next. The estimate of the standard deviation curve is shown on Figure 3.18-B. It seems like the number of basis functions is now sufficient to capture the pattern in the standard deviation curve. However it still underestimates the standard deviation curve.

Mackenzie et al. (2005) made an observation that in a mixed-effects regression spline model, if one chooses independent random-effects while using a B-spline basis system then it underestimates the standard deviation of the response curve. In other words, a diagonal variance-covariance matrix for the random-effects underestimates the variance curve for the response curve. To address this, the LMM was further modified to make the variance-covariance matrix of the random-effects unstructured. Figure 3.18-C shows that 12 basis functions with an unstructured variance-covariance for the random-effects and ARMA(2,1) error structure provides a sensible estimate of the standard deviation curve.

It can be further investigated whether the estimation of the standard deviation curve is sensitive to the correlation structure of the within group error. Figure 3.18-D shows that the standard deviation curve is pretty robust to the specification of the correlation structure of the error in this scenario.

The various choices needed to generate 95% fLoA will clearly differ from context to context. That said, overall guidelines can be given, based on the work presented in this thesis, as to a sensible strategy to employ when using a LMM to calculate 95% fLoA. These are as follows:

- For the fixed-effects regressor matrix, choose a cubic B-spline basis system with a knot sequence suggested by the adjusted R-squared criterion.

3.6. Measuring Agreement between Functions using LMM

- Choose the same B-spline basis system for the random-effects regressor matrix.
- The variance-covariance matrix for the random-effects must be unstructured.
- Choose an appropriate correlation structure for the error based on the autocorrelation plot.

Using these guidelines and recommendations, the estimation of the 95% fLoA for the motion capture study with no replicates will be demonstrated in the next section.

Mixed-effects Model to Produce LoA for Study with no Replicates

Previously the 95% fLoA were estimated using a functional data analysis approach. This thesis proposes a new use of the mixed-effects modelling framework to calculate 95% fLoA for studies with replicates. Before demonstrating the use of this methodology to calculate fLoA for studies with replicates, the methodology will be first used to calculate fLoA for studies with no replicates.

In the motion capture study, to obtain data with no replicate measurements, only one replicate measurement per method of measurement was considered. Using these data model (3.33) was fitted. For model (3.33), a cubic B-spline basis system with 12 basis functions was considered both for the fixed-effects and the random-effects regressor matrices. An unstructured variance-covariance matrix for the random-effects was used and an ARMA(2,1) was considered for the error. No variance function was used in the model. After fitting the model, equation (3.36) was used to calculate the 95% fLoA using a LMM for studies with no replicates (Figure 3.19).

Figure 3.19 displays the 95% fLoA using both the FDA and LMM frameworks. From Figure 3.19, the first thing one can notice is that the relative bias curves are very similar using both the frameworks. The fLoA produced by the LMM framework is also very similar to the fLoA produced by the FDA framework.

When calculating 95% LoA for a univariate response using the formula proposed by Altman and Bland (1983), the assumption that differences are uncorrelated with the averages can be visually assessed using the so called Bland-Altman plot. The same plot cannot be used for a situation where a statistical model is being used to calculate

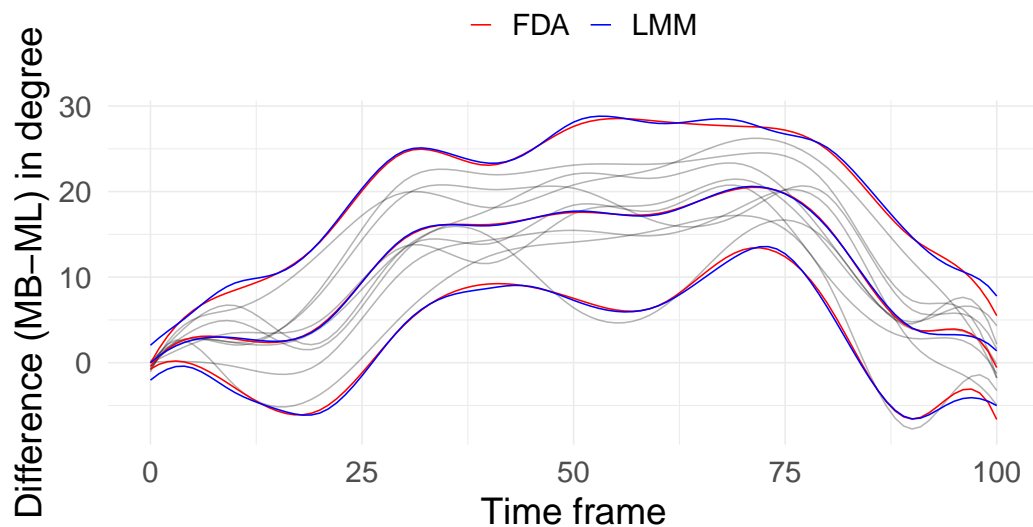


Figure 3.19: 95% functional limits of agreement for a study with no replicates by functional data analysis approach (FDA) and mixed-effects modelling (LMM) frameworks. Here, MB is the marker-based and ML is the markerless method of measurement.

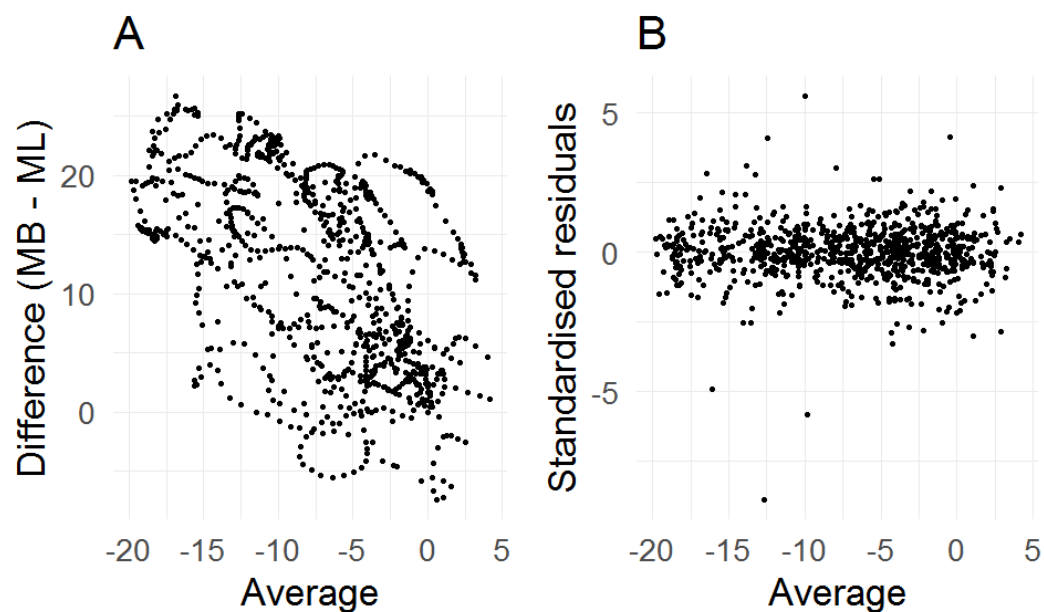


Figure 3.20: A: differences against averages ignoring the time frame for the right hip abduction angle curves. Only one replicate measurement from the first measurement session was considered here. B: residuals against averages for the same data after fitting a linear mixed-effects model. Here, MB is the marker-based and ML is the markerless method of measurement.

the 95% LoA. A statistical model can account for the scenario when the differences are correlated to the averages. When a statistical model is used, a plot of the residuals obtained after fitting the model should resemble white noise. If the modelling approach is appropriate then there should not be any pattern in the residuals when plotting them against the averages. If this is the case, it can be concluded that the modelling approach is appropriate to calculate the 95% LoA. As a modelling approach is being used here for calculating the 95% fLoA, the same strategy can be used to justify that the assumptions underlying the modelling approach were adequate. Figure 3.20 is such a plot which can be used to check that the modelling approach is adequate to calculate the 95% fLoA. Figure 3.20-A displays the differences against the averages ignoring the time dimension. Here it can be noticed that the averages are correlated with the differences. After fitting the model, the residuals were plotted against the averages in Figure 3.20-B. Here it can be noticed that the residuals are uncorrelated with the averages. This means the modelling approach appears justified here.

In the next section, the proposed LMM framework will be demonstrated to calculate the 95% fLoA for the right hip angle curves obtained through a hierarchical study design with replicates in the motion capture study.

3.6.2 95% fLoA using a LMM for Studies with Replicates

The data set described in the first chapter from the motion capture study was generated by a hierarchical study design. Therefore, a hierarchical mixed-effects model with B-spline basis functions for both fixed- and random-effects regressor matrix are needed to model these functional responses.

Let \mathbf{y}_{mijk} be the vector of angle measured by the m^{th} method ($m = m_1, m_2$), for the i^{th} subject, j^{th} session, and k^{th} replicates. The frame-wise differences can be calculated as follows:

$$\mathbf{d}_{ijk} = \mathbf{y}_{m_1ijk} - \mathbf{y}_{m_2ijk} \quad (3.37)$$

where, \mathbf{d}_{ijk} is the vector containing the difference between the measurements made by the two methods for the i^{th} subject, j^{th} session, and k^{th} replicates. This \mathbf{d}_{ijk} will be

the response vector for the LMM.

The model for this difference curve can be expressed as follows:

$$\mathbf{d}_{ijk} = \mathbf{X}_{ijk}\boldsymbol{\beta} + \mathbf{Z}_{i,jk}\mathbf{b}_i + \mathbf{Z}_{ij,k}\mathbf{b}_{ij} + \mathbf{Z}_{ijk}\mathbf{b}_{ijk} + \boldsymbol{\epsilon}_{ijk},$$

$$i = 1, \dots, 9, \quad j = 1, 2, \quad k = 1, 2, 3 \quad (3.38)$$

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_1), \quad \mathbf{b}_{ij} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_2), \quad \mathbf{b}_{ijk} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_3), \quad \boldsymbol{\epsilon}_{ijk} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

where, $\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \boldsymbol{\Psi}_3, \boldsymbol{\Sigma}$ are all positive-definite matrices. \mathbf{d}_{ijk} is a n_i -dimensional response vector containing the difference curve for a single replicate of a subject. \mathbf{X}_{ijk} is a $n_i \times p$ dimensional fixed-effects design matrix containing p B-spline basis functions, where the number p is sufficient to represent the mean difference curve. This number p depends on the number of knots and the degree of the spline. A cubic spline with a suitable knot sequence will be chosen for the B-spline basis. The approach for choosing a suitable knot sequence was discussed earlier. $\boldsymbol{\beta}$ is a p -dimensional vector containing the fixed-effects corresponding to each column of the matrix \mathbf{X}_{ijk} .

$\mathbf{Z}_{i,jk}$ is a $n_i \times q_1$ dimensional random-effects design matrix containing q_1 B-spline basis functions. This matrix models the random deviation of the individual subject from the mean curve. \mathbf{b}_i is a q_1 -dimensional vector containing the random-effects corresponding to the matrix $\mathbf{Z}_{i,jk}$.

\mathbf{b}_{ij} is a q_2 -dimensional vector containing random-effects for the j^{th} session for a given subject i . $\mathbf{Z}_{ij,k}$ is a $n_i \times q_2$ dimensional matrix containing q_2 B-spline basis functions. \mathbf{b}_{ijk} is a q_3 -dimensional vector containing the random-effects for the k^{th} replicates for a given subject i and a given session j . \mathbf{Z}_{ijk} is the corresponding random-effect design matrix which is $n_i \times q_3$ dimensional. $\boldsymbol{\epsilon}_{ijk}$ is the n_i -dimensional error vector.

The level-1 random effects \mathbf{b}_i are assumed to be independent for different i , the level-2 random effects \mathbf{b}_{ij} are assumed to be independent for different i or j and independent of level-1 random effects, the level-3 random effects \mathbf{b}_{ijk} are assumed to be independent of i, j or k and independent of level-1 and 2 random effects, the within-group errors $\boldsymbol{\epsilon}_{ijk}$ are assumed to be independent of different i, j or k and independent of all the level-1,2,3 random effects.

3.6. Measuring Agreement between Functions using LMM

The bias between two methods for a new pair of measurements for the i^{th} subject is

$$\boldsymbol{\mu}_d = E(\mathbf{d}) = \mathbf{X}_{ijk} \boldsymbol{\beta} \quad (3.39)$$

and the variance-covariance matrix of the differences between pairs of measurements is

$$\begin{aligned} \text{Var}(\mathbf{d}) &= \text{Var} \left[\mathbf{X}_{ijk} \boldsymbol{\beta} + \mathbf{Z}_{i,jk} \mathbf{b}_i + \mathbf{Z}_{ij,k} \mathbf{b}_{ij} + \mathbf{Z}_{ijk} \mathbf{b}_{ijk} + \boldsymbol{\epsilon}_{ijk} \right] \\ &= \mathbf{Z}_{i,jk} \boldsymbol{\Psi}_1 \mathbf{Z}'_{i,jk} + \mathbf{Z}_{ij,k} \boldsymbol{\Psi}_2 \mathbf{Z}'_{ij,k} + \mathbf{Z}_{ijk} \boldsymbol{\Psi}_3 \mathbf{Z}'_{ijk} + \boldsymbol{\Sigma} \\ &= \boldsymbol{\Lambda}_d \end{aligned} \quad (3.40)$$

The 95% limits of agreement for the functional response are

$$\boldsymbol{\mu}_d \pm 2 \sqrt{\text{diag}(\boldsymbol{\Lambda}_d)} \quad (3.41)$$

In the next section, all the data for the right hip abduction angle from the motion capture study will be used to calculate the 95% fLoA to assess the agreement between marker-based and markerless methods of measurement.

Application: Motion Capture Study

It was found that a knot sequence $\{1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 101\}$ for the fixed-effects regressor matrix was adequate to model the mean curve. The same set of B-spline basis functions was considered here for the fixed-effects and the random-effects regressor matrices. This means that the same B-spline basis system was used for different \mathbf{Z} matrices at different levels.

Two packages in R can be used to estimate LMMs (R Core Team, 2021). These two packages are `nlme` and `lme4`. The `nlme` package provides flexibility to include the variance function and correlation structure in the model (Pinheiro and Bates, 2000). Whereas the `lme4` package does not provide these functionalities (Bates et al., 2014). The `lme4` package can fit complex LMMs with unstructured variance-covariance matrix for the random-effects efficiently. On the other hand the estimation procedure of the `nlme` package in the presence of unstructured variance-covariance matrix is either very

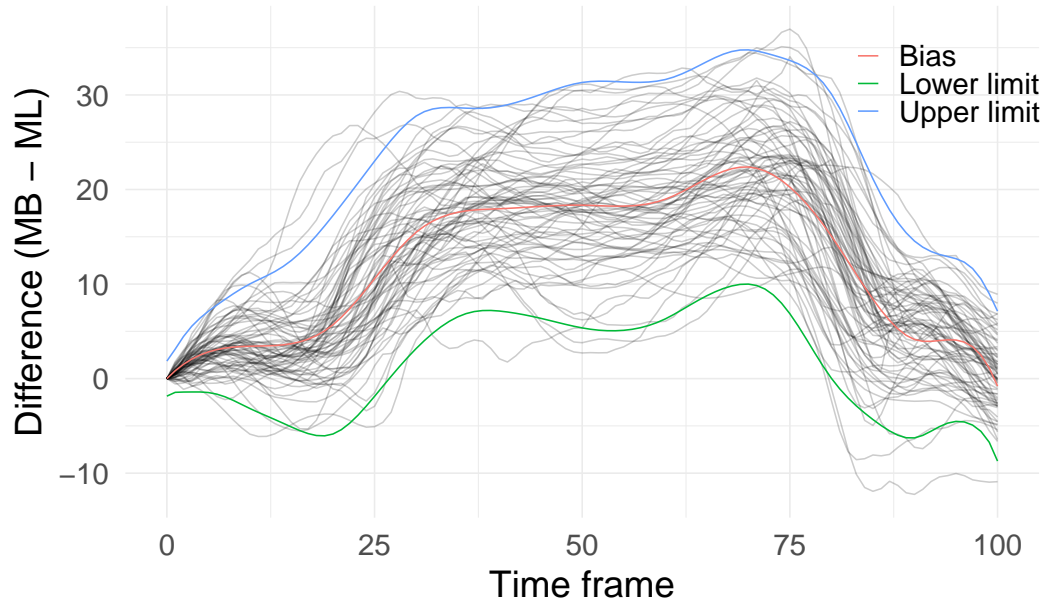


Figure 3.21: 95% functional limits of agreement for the right hip abduction angle from the motion capture study using a linear mixed-effects modelling framework. Here all the replicates from two different measurement sessions were used. MB is the marker-based and ML is the markerless method of measurement.

slow or sometimes it may happen that the optimisation procedure does not converge. In the previous section it was observed that the estimation of the variance curve does not appear to be influenced by the error structure. For this reason in this section the error structure for the LMM will be excluded in order to use the `lme4` package to estimate model (3.38) and no variance function for the error was used.

After fitting model (3.38), the formula described in equation (3.41) was used to calculate the 95% fLoA for the right hip abduction angle from the motion capture study (Figure 3.21).

From the 95% fLoA, it is clear that there is substantial bias between the two methods of measurement (Figure 3.21). It is also clear that the markerless system typically overestimates the angle compared to the marker-based system. For the middle part of the domain the relative bias is about 20 degrees on average. In theory this bias could be removed from the markerless system by subtracting this bias curve from the future measurements of the markerless system. However, the variability of the individual difference curve is substantial. It seems that the difference between the two methods of

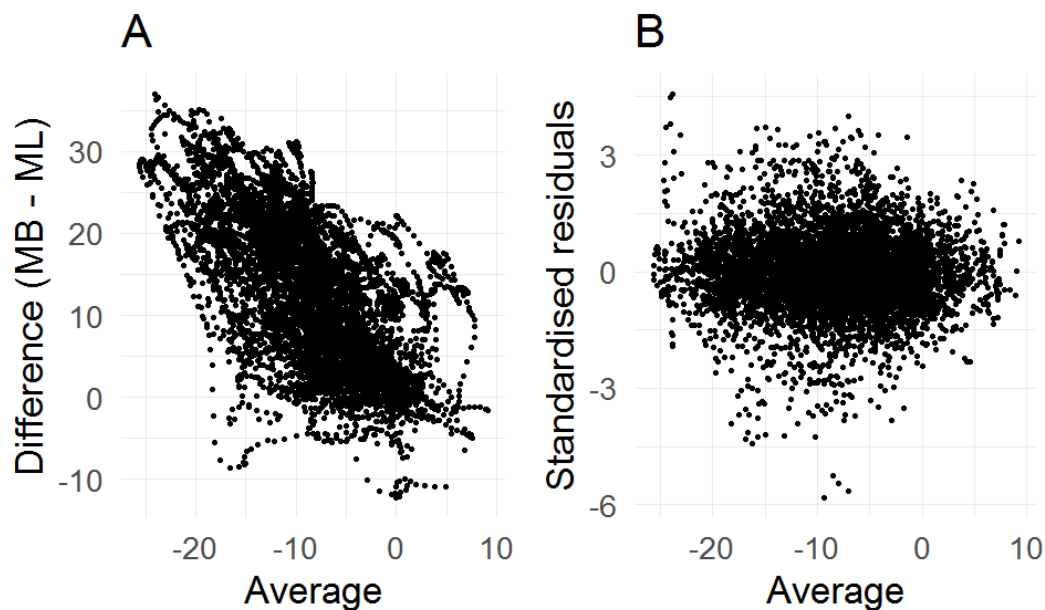


Figure 3.22: A: differences against averages ignoring the time frame for the right hip abduction angle curves. All the replicate measurements from two measurement sessions were considered here. B: residuals against averages for the same data after fitting a linear mixed-effects model. Here, MB is the marker-based and ML is the markerless method of measurement.

measurement could be 10 degrees to 30 degrees in magnitude. If this level of disagreement between the two system is considered too high for the right hip abduction angle measurement, the markerless system cannot be replaced with the marker-based system for the measurement of the right hip abduction angle during a lunge.

Figure 3.22-A shows the plot of the difference against the average, ignoring the time frame. It seems there is a pattern in the differences across the measurement range. However, when the residuals are plotted against the average in Figure 3.22-B, it shows that there is no pattern of concern regarding the residuals over the range. Figure 3.22-B can provide a visual confirmation, in a similar manner when using residual plots in model checking, that the model is adequate to remove any relation between the differences and the averages.

It was demonstrated that a large number of B-spline basis functions with an unstructured variance-covariance matrix for the random-effects are essential when using a LMM to estimate the variance curve for functional response. This necessity comes with

a cost however in terms of how long it takes to fit a LMM. From a practical point of view, the necessity of a large number of B-spline basis functions for the random-effects regressor matrix, coupled with an unstructured variance-covariance matrix requires a faster computational approach to make this approach feasible in practice in order estimate 95% functional limits of agreement. A new approach is now presented to circumvent this problem.

3.7 Faster Computational Approach for Nonparametric LMM

In this section, a new computational approach is proposed to calculate 95% fLoA using a LMM with a modified basis system for the random-effects regressor matrix which allows a diagonal variance-covariance matrix for the random-effects to be implemented. The result is based on the results of the Karhunen–Loève theorem which will now be introduced.

3.7.1 The Karhunen–Loève Theorem

Let X_t be a zero-mean and square-integrable stochastic process defined over some probability space with continuous covariance function $K_X(s, t)$. X_t is defined over a closed interval $[a, b]$. A linear operator T_{K_X} can be defined as follows:

$$T_{K_X} f = \int_b^a K_X(t, \cdot) f(t) dt \quad (3.42)$$

with k^{th} eigenvalue λ_k and corresponding eigenfunction e_k (Hsing and Eubank, 2015). According to the Karhunen–Loève theorem, the stochastic process X_t can be represented as follows (Hsing and Eubank, 2015):

$$X_t = \sum_{k=1}^{\infty} c_k e_k(t) \quad (3.43)$$

Here the coefficient c_k is a random variable with mean zero and variance λ_k with

the following definition (Hsing and Eubank, 2015):

$$c_k = \int_b^a X_t e_k(t) dt \quad (3.44)$$

Consider the functional response \mathbf{y} as a stochastic process where values were obtained at certain equally spaced time points. \mathbf{y} can be modelled as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \quad (3.45)$$

where the mean of \mathbf{y} is $\mathbf{X}\boldsymbol{\beta}$ with covariance matrix $\boldsymbol{\Sigma}$. Then $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is a zero-mean stochastic process with covariance matrix $\boldsymbol{\Sigma}$. Since

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \quad (3.46)$$

$\mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$ is also a zero-mean stochastic process with covariance matrix $\boldsymbol{\Sigma}$. Therefore, $\mathbf{Z}\mathbf{b}$ is a zero-mean stochastic process with covariance matrix different to $\boldsymbol{\Sigma}$. Consider,

$$\mathbf{u} = \mathbf{Z}\mathbf{b} \quad (3.47)$$

On comparing equation (3.43) with equation (3.47) it is clear that the column of the \mathbf{Z} can be the eigenvectors of the $\boldsymbol{\Sigma}$ and \mathbf{b} a vector of independent random coefficients of the corresponding eigenvectors. Although $\boldsymbol{\Sigma}$ is not the covariance matrix of \mathbf{u} , it does not matter since the coefficients will not be estimated by equation (3.44). Only a convenient basis system for \mathbf{u} is needed, and the REML criterion with best linear unbiased prediction (BLUP) will provide estimates of the coefficients of the random-effects \mathbf{b} .

3.7.2 Eigenvalue Decomposition of a Variance Covariance Matrix

Let \mathbf{S} be the sample covariance matrix for a sample of n curves. These n curves are n realisation of the random vectors \mathbf{y} . Each curve consists of n_i equally spaced observations. The sample covariance matrix is an $n_i \times n_i$ -dimensional matrix. The

eigenvalue decomposition of the matrix \mathbf{S} is as follows:

$$\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}' \quad (3.48)$$

where, $\mathbf{\Lambda}$ is a $n_i \times n_i$ diagonal matrix with eigenvalues as diagonal elements, \mathbf{Q} is a $n_i \times n_i$ orthogonal matrix where each column contains an eigenvector. The i^{th} column of the \mathbf{Q} matrix is the eigenvector associated with the $[\mathbf{\Lambda}]_{ii}$ eigenvalues. After this decomposition, one can look at the number of non-zero eigenvalues. The eigenvectors associated with those non-zero eigenvalues would then be the set of basis functions or vectors in the \mathbf{Z} matrix in the equation (3.45). In practice, not all the eigenvectors will be considered. Eigenvectors corresponding to the smallest eigenvalues can be ignored. This representation provides a basis system for the random-effects design matrix where a diagonal covariance structure for the random-effects can be assumed since the eigenvectors are orthonormal to each other.

It will be a time consuming optimisation procedure when there is a need to use a large number of B-spline basis functions for the random-effects regressor matrix. The approach of using an eigenbasis for the random-effects regressor matrix with a diagonal variance-covariance for the random-effects matrix will be very useful in such situations.

A recommendation on how to specify a LMM, when computation time is an issue, to calculate the 95% fLoA using eigenbasis can be obtained as follows:

- For the fixed-effects regressor matrix, choose a cubic B-spline basis system with a knot sequence suggested by the adjusted R-squared criterion.
- Calculate the variance-covariance matrix for the functional responses and obtain an eigenvalue-eigenvector decomposition of the matrix.
- Choose the first few eigenvectors that explain at least 99% of the variation of the responses.
- Use a diagonal variance-covariance matrix for the random-effects.
- Choose an appropriate correlation structure for the error based on the autocorrelation plot.

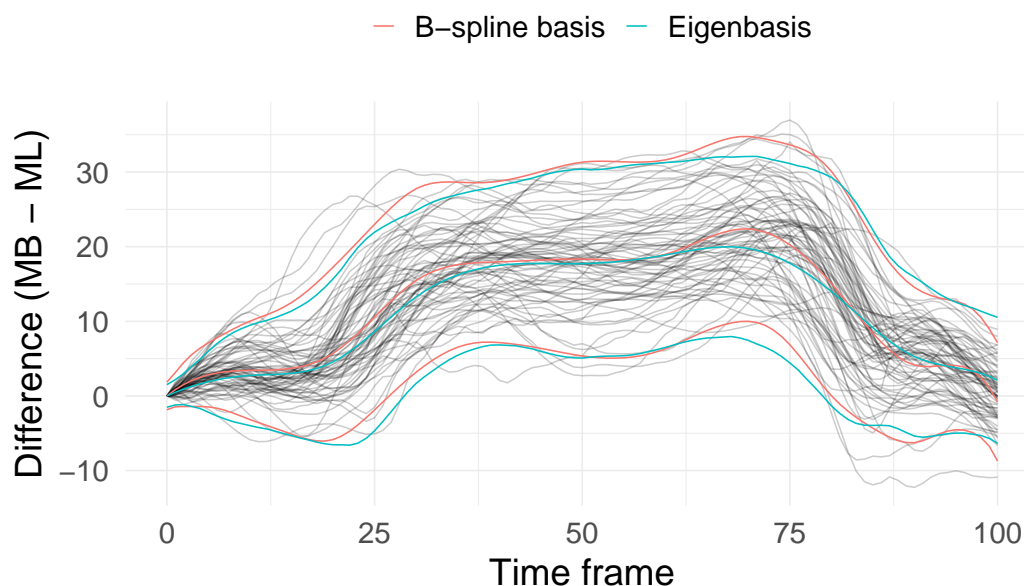


Figure 3.23: 95% functional limits of agreement estimated by a linear mixed-effects modelling framework using two different basis systems for the random-effects regressor matrices. B-spline and eigenbasis systems were used for the random-effects regressor matrices. Here, MB is the marker-based and ML is the markerless method of measurement.

An example of this approach is now given for the motion capture study and the results compared in terms of computational time and precision.

3.7.3 Application: Motion Capture Study

To estimate the 95% fLoA, model (3.38) will be used again but with a different basis system for the random-effects regressor matrices. For the fixed-effects regressor matrix, the same set of 12 B-spline basis system will be considered. To get the basis system for the random-effects regressor matrix, first a variance-covariance matrix for the functional response is obtained. As the functional responses were collected at 101 equally spaced points on the domain, the size of the variance-covariance matrix for the response will be 101×101 . After obtaining the eigenvalue decomposition of the variance-covariance matrix of the functional response, it was found that there are 21 non-zero eigenvalues. The first twelve eigenvalues had a value above or equal to 8. The next five eigenvalues ranged from 2 to 5 in value, the rest of the eigenvalues were one. The first 12 eigenvectors

3.7. Faster Computational Approach for Nonparametric LMM

were considered as they provided a space where more than 99% of the variability of functional response can be explained. The same set of 12 eigenvectors were considered for all the \mathbf{Z} matrices at different levels of random-effects. As the coefficients of the eigenvectors are uncorrelated, a diagonal variance-covariance matrix for the random-effects was considered.

As the `nlme` package can estimate diagonal variance-covariance matrices for the random-effects efficiently and it offers functionality to include different error structures, the model was estimated using the `nlme` package and an ARMA(2, 1) was used to model the error structure.

The 95% fLoA were calculated after fitting the LMM with the eigenbasis system for the random-effects (Figure 3.23). In Figure 3.23, estimates of the 95% fLoA using the B-spline basis system with an unstructured variance-covariance matrix for the random-effects and the eigenbasis system with the diagonal variance-covariance matrix for the random-effects are plotted. The estimates of the 95% fLoA using the two systems are very similar in a sense that the decision of whether these two methods of measurement agree or not will be the same. It should be noted that these two estimates are not from exactly the same model as the ARMA(2,1) structure was dropped from the model with the B-spline basis system for the random-effects regressor matrices in order to use the `lme4` package which can estimate unstructured variance-covariance for the random-effects efficiently.

The model with the B-spline basis system, an unstructured-variance covariance matrices for the random-effects, and no structure for the error took about 53 minutes to fit whereas the model with an eigenbasis system, diagonal variance-covariance matrices for the random-effects and ARMA(2, 1) structure for the error took 16 minutes to fit. In general, the estimation procedure is faster for a model without an error structure than with an error structure included in the model. Fitting the model with the eigenbasis was consistently more than three times faster than one one with the B-spline basis and the former has the added advantage that an error structure can be included in the model.

Figure 3.24 displays that there is no structure of concern remaining in the residuals

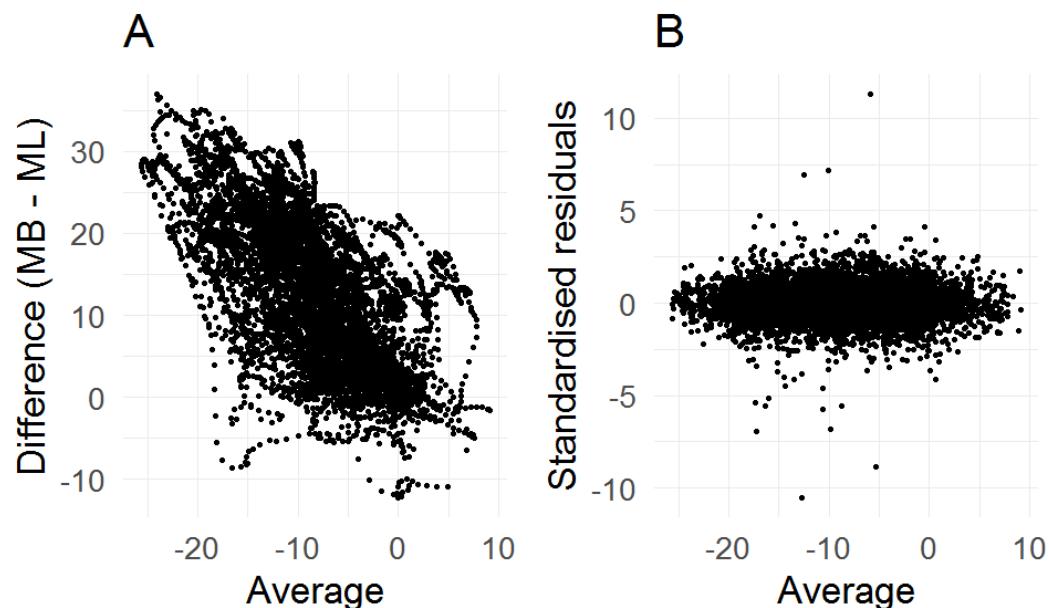


Figure 3.24: A: differences against averages ignoring the time frame for the right hip abduction angle curves. All the replicate measurements from two measurement sessions were considered here. B: residuals against averages for the same data after fitting a linear mixed-effects model with eigenbasis for the random-effects regressor matrices. Here, MB is the marker-based and ML is the markerless method of measurement.

while plotting against the different magnitude of the angle being measured.

3.8 Chapter Summary

This chapter deals entirely with functional responses in method comparison studies. Existing statistical approaches for evaluating reliability of a measurement method were presented and discussed. Methodological development for assessing agreement between two measurement methods was presented, and limitations were identified. The current statistical approach to calculating 95% limits of agreement for a functional response relies on functional data analysis.

A new approach, an extension of the mixed model framework to accommodate functional responses, was proposed which can accommodate more complex study designs with ease. This novel approach is based on a nonparametric mixed-effect model. Techniques and guidelines for selecting a suitable basis system for the fixed-effects regressor

matrix were suggested by way of a case study. It was found that the same basis system is also suitable for the random-effects regressor in the examples considered. An unstructured variance-covariance matrix is essential to estimate the standard deviation curve accurately. Guidelines for choosing an appropriate error structure were also suggested. In the case study presented it was found that specifying a variance function for the error was not necessary. A new graphical approach was presented to visually assess whether the modelling approach is adequate when calculating 95% fLoA. This new graphical tool plays a similar role for a functional response to the Bland-Altman plot for a univariate response.

The modelling framework comes with a cost in terms of the time required to fit the complex nonparametric linear mixed-effects model. A new computational approach was presented to address this problem by finding a suitable basis systems for the random-effects regressor matrix by obtaining a basis system for the random-effects regressor matrix using the eigenvectors of the sample variance-covariance matrix. This allows a diagonal variance-covariance matrix for the random effects in the model specification to be used. The proposed method provides an elegant way of computing the 95% fLoA using a nonparametric linear mixed-effects model which is time efficient.

The different techniques and model specifications needed when using the mixed model frameworks proposed in this chapter were discussed and suggestions given by way of example using the motion capture case study.

In the next chapter, a simulation study is presented to compare the FDA and the linear mixed model approaches in more detail when used to assess the level of agreement in method comparison studies with functional responses.

Chapter 4

Simulation Study and Results

4.1 Introduction

The main aim of this chapter is to evaluate the performance of the proposed linear mixed-effects modelling framework to calculate functional limits of agreement using a simulation study. The simulation study in this chapter is not a “classical” simulation study as the performance of the framework will not be assessed based on comparing the bias of the estimates and coverage (in the traditional sense) of the interval estimate of the parameter under different scenarios. The main aim is to assess the performance of the LMM framework in terms of coverage with respect to the functional limits of agreement in designs with and without replicates at the individual level.

If it can be found that the linear mixed-effects modelling framework provides a valid estimate for the specific situations considered in this simulation, it can be concluded that this may be the case in such biomechanical studies in general as the simulation reflects the type of functional response one would expect in such studies. A more extensive simulation study could have involved a larger variety of curve shapes. The main benefit of using a linear mixed-effects modelling framework is its flexibility in being able to accommodate complex designs quite easily where it is expected that the B-spline basis functions used would adapt to other functional forms given the flexibility of B-splines.

In this chapter valuable information on how best to interpret functional limits of

agreement is also provided. Then, for studies with no replicate measurements, a comparison is made between the FDA and LMM frameworks to calculate 95% fLoA. Following this, for studies with replicate measurements, a comparison is made of the performance of the B-spline and eigenbasis system for the random-effects regressor matrices to calculate 95% fLoA. In addition, an investigation into the computational time taken using the eigenbasis system compared to the B-spline basis system for the random-effects regressor matrices is made. The chapter concludes with an application of the LMM framework to calculate the 95% fLoA for all the lunge angles measured in the motion capture study and an summary of these results is given.

4.2 Simulation Study Outline

When considering a continuous univariate response, the interpretation of the 95% LoA is not generally well understood which causes confusion when interpreting the results. Some authors have interpreted 95% LoA as 95% prediction intervals (Carstensen et al., 2008). This confusion exists also when interpreting 95% fLoA as there are two different ways to interpret such limits, namely as pointwise or global bands. Therefore a clear understanding of the interpretation of a 95% fLoA is needed. One aim of this simulation study is to clarify this confusion and to provide guidance on how to interpret a 95% fLoA accurately.

In this thesis, two statistical frameworks were presented to produce 95% fLoA for studies without replicate measurements for each subject. The framework using FDA is straightforward as this only requires one to find a suitable basis system. The model specification for a LMM in a method agreement study needed to calculate 95% fLoA is more complex, as illustrated using the motion capture example in the last chapter. The model specification included the choice of the basis system for the fixed and random-effects, the structure of the variance-covariance matrix of the random-effects, and the choice of the correlation structure for the error. In this chapter a simulation study is carried out where one aim is to investigate whether these model specifications for a LMM produce the same 95% fLoA as produced by the FDA framework i.e. are the

LMM and FDA frameworks comparable in terms of coverage.

The proposed framework of calculating a 95% fLoA using a LMM framework is also applicable for hierarchical studies with replicate measurements. A simulation study is needed to verify whether the LMM achieves the nominal coverage with respect to fLoA for studies with replicates. A faster computational approach was proposed in Chapter 3 using an eigenbasis for the random-effects regressor matrix. A simulation study is also needed to verify whether this approach is comparable to using a full B-splines basis implementation and what the computational time gains are likely to be in general.

In this chapter, the following questions will be examined using simulation studies:

- What is the correct interpretation of a 95% fLoA?
- Are the FDA and LMM frameworks comparable in terms of coverage when calculating a 95% fLoA in studies with no replicates?
- Does the LMM framework achieve nominal coverage when calculating 95% fLoA in studies with replicates?
- How does the eigenbasis implementation compare to a full B-splines implementation in terms of computational time and coverage when calculating 95% fLoA in studies with replicates?

There are many scenarios that can be considered in any simulation study and this is no different in this thesis, for example different sample sizes, different population variance curves, different bias curves, and different sampling rates for a functional response.

In this chapter the right hip abduction angle data will be used to mimic the data generating mechanism of such angles in order to guide how relevant functional responses can be simulated. This means that the simulated data will be comparable to what would be expected in a biomechanics fitness battery involving elite athletes. Using the right hip abduction angle to guide the simulation process will also narrow down the different choices for the bias and variance for the difference curves. There are of course many choices for the sample sizes. Here samples of size 10, 50, and 100 will be considered

to represent small, medium and large experiments. Different sampling rates for the functional response can also be considered, however only the sampling rate (i.e. 101 frames) used in the right hip abduction angle will be used.

4.3 Assessment Criteria

The simulation strategy to assess the performance of both the FDA and LMM framework to calculate 95% fLoA is now described. First a population of 5,000 subjects were created. For studies with no replicates a functional response (i.e. a curve) for each subject was simulated. For studies with replicates, three replicate measurements for each subject were simulated.

A sample size of n means n subjects were chosen randomly from the population. For studies with no replicates each subject has one functional response but for studies with replicates each subject has three replicate functional responses.

For a given sample of size n , 100 different random samples of size n were taken from the population. For a given sample, 95% fLoA were calculated using FDA, LMM with B-spline basis for the random-effects regressor matrix, or LMM with an eigenbasis for the random-effects regressor matrix. To assess the performance of each framework, a random selection of 100 subjects from the population was taken and then the pointwise coverage of the corresponding fLoA was calculated. Here the pointwise coverage is the percentage of the 100 curves that are within the 95% fLoA at each time frame. If at a given time point 95 of these 100 curves are inside the 95% fLoA and 5 of them are outside of the 95% fLoA, then the coverage is 95% at that time point. From this, a coverage curve for a given framework can be obtained for a single sample. This procedure was repeated 100 times and the average of these coverage curves was calculated to produce the estimated coverage curve for a given framework. The simulation strategy is summarised graphically in Figure 4.1 where samples of size 10, 50, and 100 are considered.

4.4. Comparing LMM and FDA in Studies with No Replicates

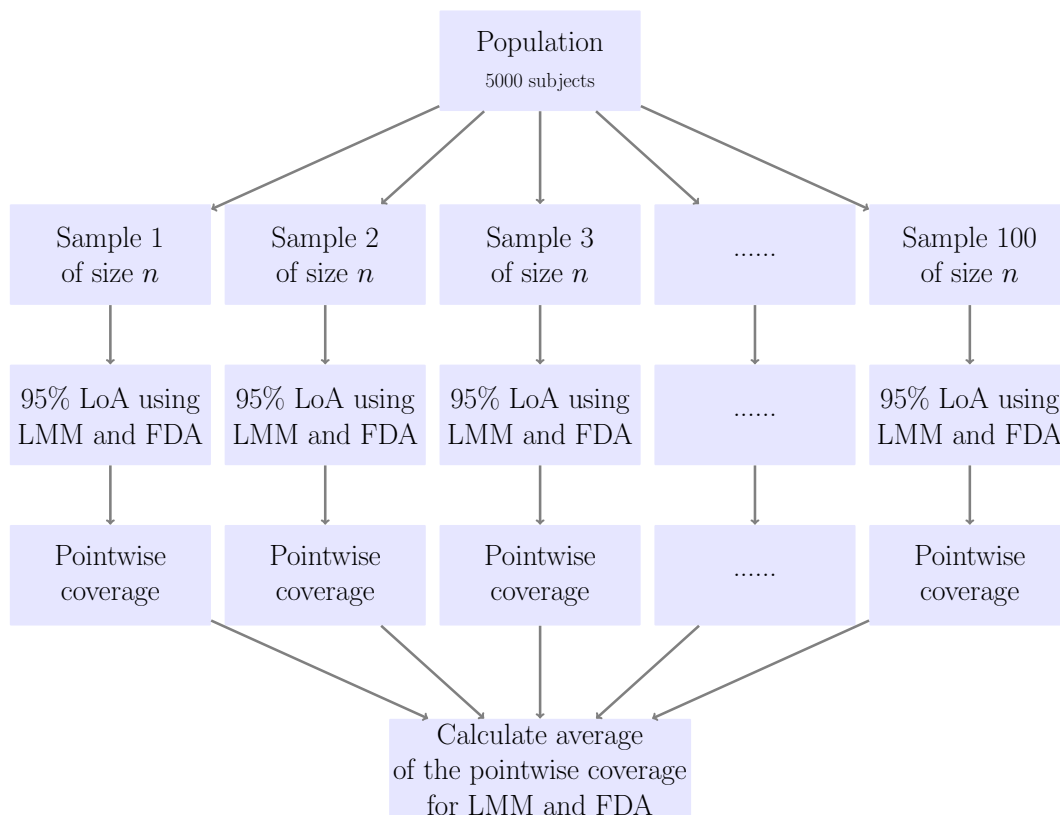


Figure 4.1: Simulation strategy to compare the performance of both the functional data analysis (FDA) and the linear mixed-effects modelling (LMM) framework to calculate 95% functional limits of agreement (fLoA).

4.4 Comparing LMM and FDA in Studies with No Replicates

As one of the objectives here is to compare the functional data analysis and mixed-effects modelling framework, it is important that neither framework is used to simulate data as the model used to generate the data will, as a consequence, be identified as the best framework to use. As both frameworks rely on a normality assumption relating to the response in the population, a sensible choice would be to use a multivariate Normal distribution to generate the individual curves. The choice of the population mean vector and variance-covariance matrix will be guided by the difference curves obtained from observed right hip abduction angle data by the two different methods of measurement.

Let \mathbf{y} be a vector of random variables with the following distribution.

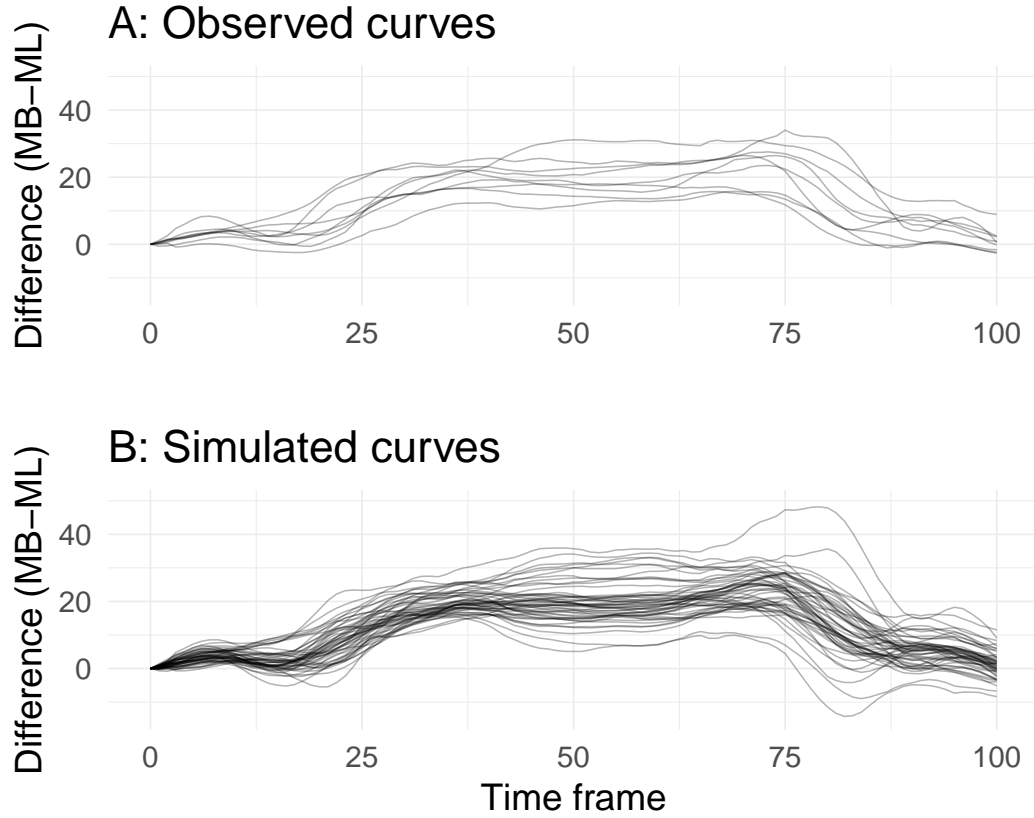


Figure 4.2: Observed and simulated difference curves. Here MB is the marker-based and ML is the markerless method of measurement.

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu}$ is the m dimensional mean vector and $\boldsymbol{\Sigma}$ is the $m \times m$ dimensional variance-covariance matrix. Each realisation of this vector valued random variable \mathbf{y} is a functional response. In the motion capture study, the length of a data vector was 101 which will also be used for this simulation where $m = 101$ will be considered. The sample mean vector and the sample variance-covariance matrix will be calculated using the difference curves obtained from the right hip abduction angle measurements from nine athletes for a given measurement session without replicates. This sample mean vector and variance-covariance matrix will be considered as the population mean vector $\boldsymbol{\mu}$ and the population variance-covariance matrix $\boldsymbol{\Sigma}$ respectively in the simulation study. Figure 4.2 shows both the observed and the simulated curves of the difference between

the measurements of the right hip abduction angle. The observed difference (marker-based minus markerless) curves from nine athletes is displayed in Figure 4.2-A. The simulated difference curves based on the same observed data are displayed in Figure 4.2-B. The simulated curves look similar to the observed curves. Only 50 curves are shown in Figure 4.2-B for visual clarity. In total, 5,000 curves were simulated using this approach and were considered as the hypothetical population of the difference curves for the right hip abduction angle measurement by the two methods.

To calculate the 95% fLoA using the FDA framework, the following model will be considered:

$$y_j = f(t_j) + \epsilon_j, \quad j = 1, \dots, m \quad (4.1)$$

where, for a given subject y_j is the j^{th} measurement measured at time t_j , $f(\cdot)$ is the true function of the angle curve, and ϵ_j is the measurement error at time t_j which follows a Normal distribution. The same model can be expressed using vector-matrix notation to compare it with a LMM which will be introduced later on. For a single subject the model (4.1) can be re-written as follows:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{c} + \boldsymbol{\epsilon} \quad (4.2)$$

where, \mathbf{y} is a m -dimensional response vector containing m measurements, $\mathbf{\Phi}$ is an $m \times k$ -dimensional matrix containing B-spline basis functions in different columns evaluated at time t_j , \mathbf{c} is the k -dimensional coefficient vector, $\boldsymbol{\epsilon}$ is the m -dimensional error vector. For the different columns of the $\mathbf{\Phi}$ matrix, a cubic B-spline basis functions with knot positions at $\{0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ will be considered. A cubic spline is of order 4 and there are 9 different inner knots. This gives $9 + 4 = 13$ different B-spline basis functions. The dimension of the $\mathbf{\Phi}$ matrix will be 101×13 .

In Chapter 3, it was observed that for the difference curves, the GCV score curve is flat for values of the smoothing parameter (λ) from 10^{-15} to $\lambda = 10^{-5}$. It can be concluded that the value of the smoothing parameter does not overly influence the curve fitting much. For this reason the value of the λ will not be estimated for each sample.

4.4. Comparing LMM and FDA in Studies with No Replicates

The value of the $\lambda = 10^{-5}$ gave a reasonable GCV score in Chapter 3 for the difference curves with less degrees of freedom for the B-spline basis system. For this reason the same value will be used as the value of the smoothing parameter here. After estimating \mathbf{c} , the methods described in Chapter 3 will be used to calculate the 95% fLoA using the FDA framework.

A mixed-effects model needs to be specified for the simulated data in order to calculate the 95% fLoA. As the response vector is the same, \mathbf{y} will be used to denote the response vector. Following LMM can be considered to estimate the 95% fLoA:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} \mathbf{b} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Psi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \end{bmatrix} \right) \quad (4.3)$$

where, \mathbf{y} is a m -dimensional response vector, \mathbf{X} is a $m \times p$ -dimensional fixed-effects design matrix, $\boldsymbol{\beta}$ is a p -dimensional fixed-effects vector, \mathbf{Z} is a $m \times q$ -dimensional random-effects design matrix, \mathbf{b} is a q -dimensional random-effects vector, and $\boldsymbol{\epsilon}$ is a m -dimensional error vector. $\boldsymbol{\Psi}$ is a $q \times q$ -dimensional diagonal variance-covariance matrix for the random-effects \mathbf{b} and $\boldsymbol{\Sigma}$ is a $m \times m$ -dimensional variance-covariance matrix for the error vector $\boldsymbol{\epsilon}$. For $\boldsymbol{\epsilon}$, an ARMA(1, 1) correlation structure will be considered. The columns of the \mathbf{X} matrix are a cubic B-spline basis functions with knots position at $\{0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ evaluated at the time points where observations were made. A order 4 spline with 9 interior knots provides a B-spline basis system with $4 + 9 = 13$ different basis functions. As the initial value of the function is zero, the first B-spline basis function will be dropped to ensure an intercept equal to zero. Here, the fixed-effects design matrix has dimension 101×12 . For the random-effects design matrix, the same B-spline basis system for the fixed-effects will be considered and the matrix \mathbf{Z} also has the dimension 101×12 . The REML criterion will be used to estimate all model parameters. After, estimating the model parameters, the procedure described in Chapter 3 will be used to calculate the 95% fLoA using a LMM.

4.4.1 Interpreting a 95% fLoA?

The first question that the simulation study aims to investigate is to provide insight into how best interpret a 95% fLoA for functional responses i.e. should one interpret 95% fLoA as a pointwise or global band? To answer this question a population with 5,000 difference curves were used. The population mean curve or vector (μ) and the population standard deviation curve or vector (σ) were calculated from the 5,000 difference curves. Following this the upper and lower functional limits of agreement were calculated using the following formula:

$$\begin{aligned}\text{Upper limits of agreement} &= \mu + 2\sigma \\ \text{Lower limits of agreement} &= \mu - 2\sigma\end{aligned}\tag{4.4}$$

The number of curves that go through the 97.5 and 2.5 percentile of the differences at each time frame were calculated. The 97.5 percentile curve and 2.5 percentile curve creates an envelope that contains the 95% of the differences pointwise for the simulated population. If the pointwise percentile curves and the 95% fLoA are the same curve then it can be concluded that one should interpret the 95% fLoA as a pointwise envelope. Figure 4.3 displays a random selection of 200 simulated curves from the 5,000 simulated curves along with a 95% pointwise envelope, 95% fLoA and 99.4% pointwise envelope that are equivalent to a 95% global envelope for the curves. Here, all the envelopes and the limits of agreement were calculated based on the 5,000 simulated curves. Only 200 curves were plotted for visual clarity.

Figure 4.3 shows that the 2.5 pointwise percentile curves and the lower limit curves of the 95% fLoA are the same in this simulated population. The observation is the same for the 97.5 pointwise percentile curve and the upper limit curve of the 95% fLoA. From this observation, it is clear that a 95% fLoA should be interpreted as a pointwise band.

Based on the simulated data it was found that the 2.5 and 97.5 percentile curves are equivalent to a 72 % global envelope for the curves, containing 72% of the curves for

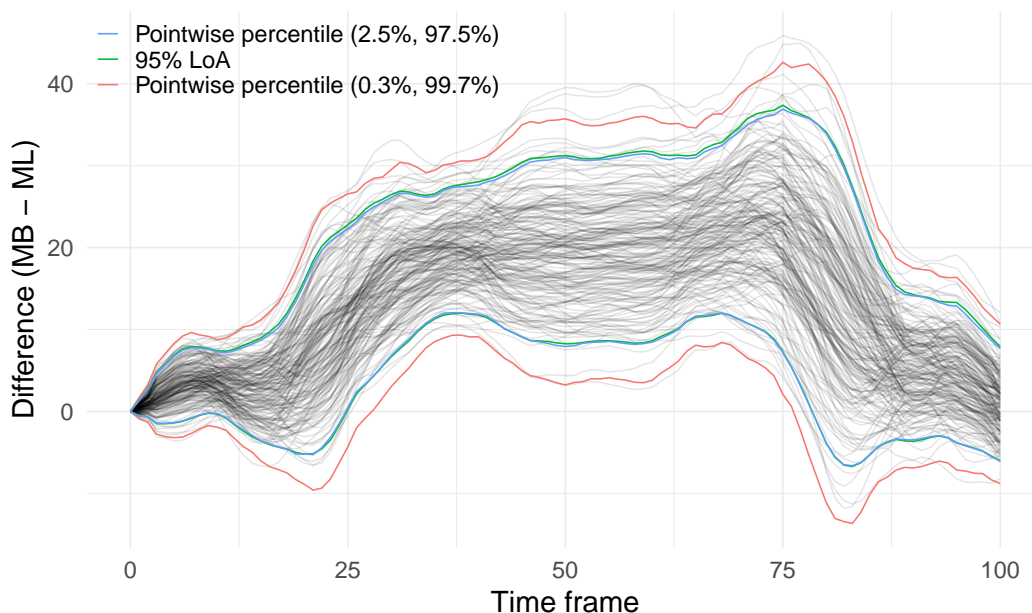


Figure 4.3: 95% pointwise envelope, 95% functional limits of agreement (fLoA) and 99.4% pointwise envelope that gives a 95% global envelope for the simulated population. Here MB is the marker-based and ML is the markerless method of measurement.

the entire domain. To generate a 95% global envelope for the curves, wider pointwise percentile bands are needed (Davison and Hinkley, 1997). For the given simulation settings, it was found that the region created by the 0.3 and 99.7 pointwise percentile curves give a 95% global envelope for the entire curves (Davison and Hinkley, 1997). Figure 4.3 shows the 95% global envelopes for the entire curves. If so desired, this approach can be used to create 95% global limits of agreement based on a sample of independent difference curves.

From the results obtained in this section, it can be concluded that 95% fLoA should be interpreted as pointwise bands.

4.4.2 Comparing 95% fLoA generated by FDA and LMM

The top three panels in Figure 4.4 show the estimated 95% fLoA using both the FDA and LMM frameworks for different sample sizes. For each sample size, 100 different samples were chosen at random from the same simulated population. One of these samples was chosen, at random, to overlay on each plot for the different sample sizes

4.4. Comparing LMM and FDA in Studies with No Replicates

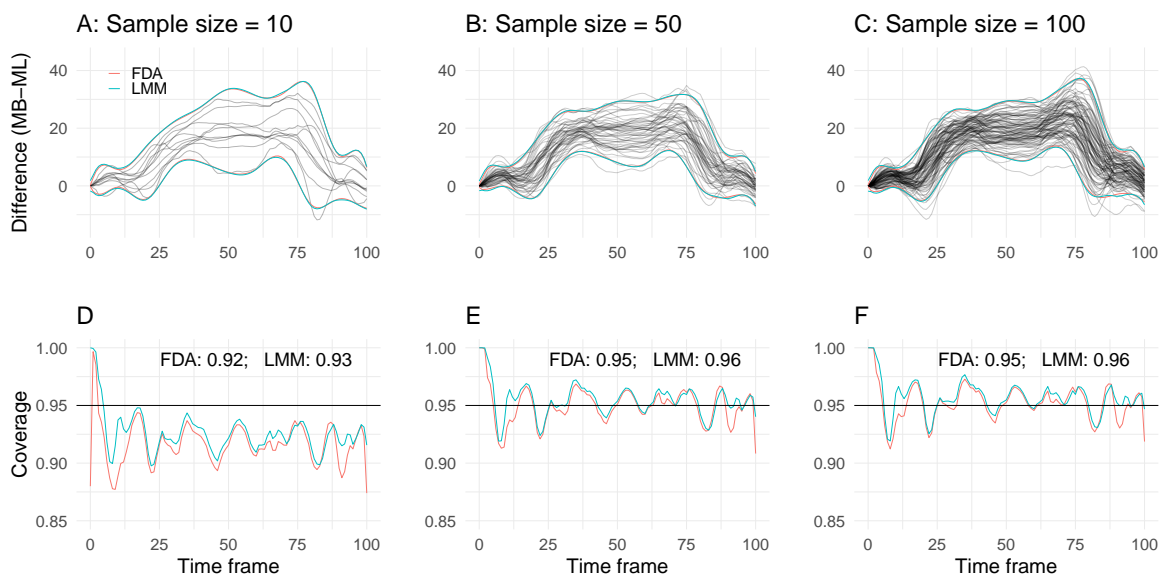


Figure 4.4: Simulation results to compare functional data analysis (FDA) and linear mixed-effects modelling (LMM) framework to calculate 95% fLoA. Panel-A, B, C show 95% fLoA calculated from a sample with the sample size 10, 50 and 100, respectively. Panel-D, E, F show coverage curves corresponding to samples of size 10, 50 and 100 respectively. 100 different random samples were used to calculate the coverage curve for any given sample size. Overall coverage (average over the domain) corresponding to each framework are shown in panel-D, E and F. Here MB is the marker-based and ML is the markerless methods of measurement.

(Figure 4.4-A, B, C). The bottom three panels show the percentage of difference values that fall inside the estimated 95% fLoA. These percentages are considered as coverage and this coverage provides a metric to assess the performance of the framework being used to estimate fLoA. In principle, 95% of the differences should be contained in the band spanned by the estimated 95% limits of agreement on average. A dark line was drawn at 0.95 as reference.

From Figure 4.4, it is clear that both the FDA and LMM frameworks are giving reasonable coverage for samples of size 50 and 100. For a sample of size 10 the coverage is not as good as for the other two sample sizes. For samples of size 50 and 100, the overall coverage when averaging over the domain, was 0.95 and 0.96 for the FDA and LMM framework, respectively. The overall coverage for the two frameworks were less than 0.93, indicating that a sample of size of greater than 10 is required to attain 95% coverage for the given context.

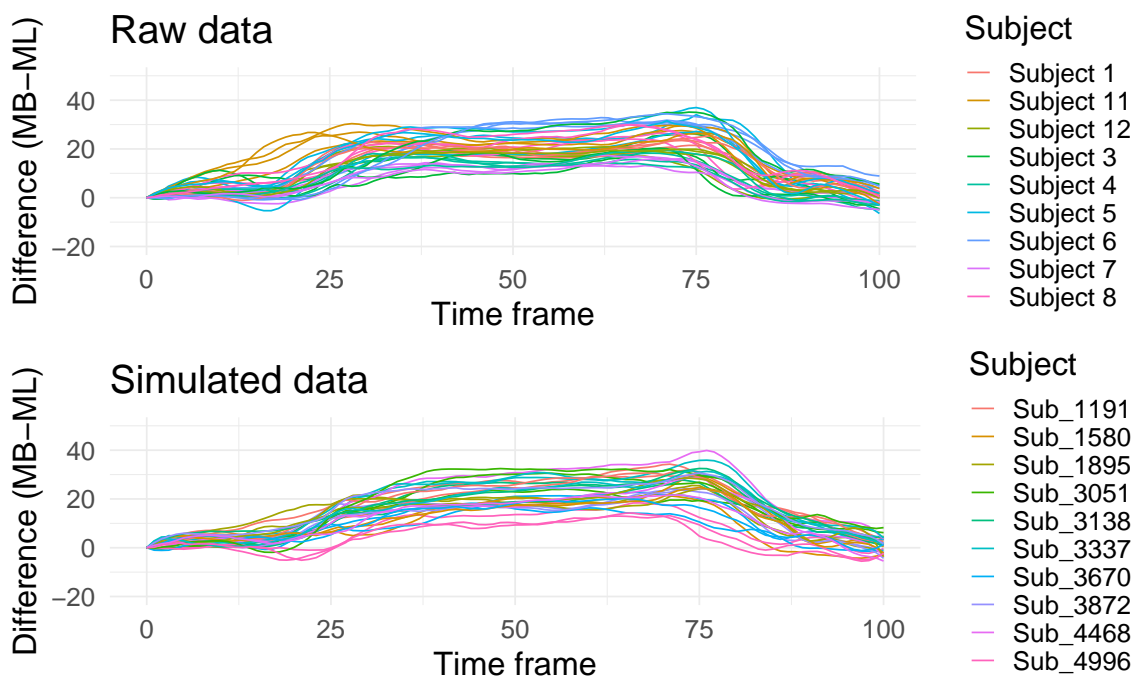


Figure 4.5: Observed and simulated data for studies with replicate measurements. Here MB is the marker-based and ML is the markerless method of measurement.

Figure 4.4 shows that the 95% fLoA calculated using the FDA and the LMM framework are almost the same for different sample sizes (Figure 4.4-A, B, C). It also shows that the coverage for both the frameworks are also almost the same (Figure 4.4-D, E, F). From these it can be concluded that both the FDA and LMM frameworks produce similar estimate of the 95% fLoA for studies with no replicates in the simulation strategy used.

4.5 Coverage and Computational Performance of a LMM in Studies with Replicates

In the previous section the performance of FDA and LMM frameworks to calculate 95% fLoA has been examined for studies with no replicates using a simulation study. In this section the performance, in terms of coverage, of the LMM framework to calculate 95% fLoA for studies with replicates will also be investigated using a simulation study. As in Chapter 3, both the B-spline basis and eigenbasis were used for the random-effects

4.5. Coverage and Computational Performance of a LMM in Studies with Replicates

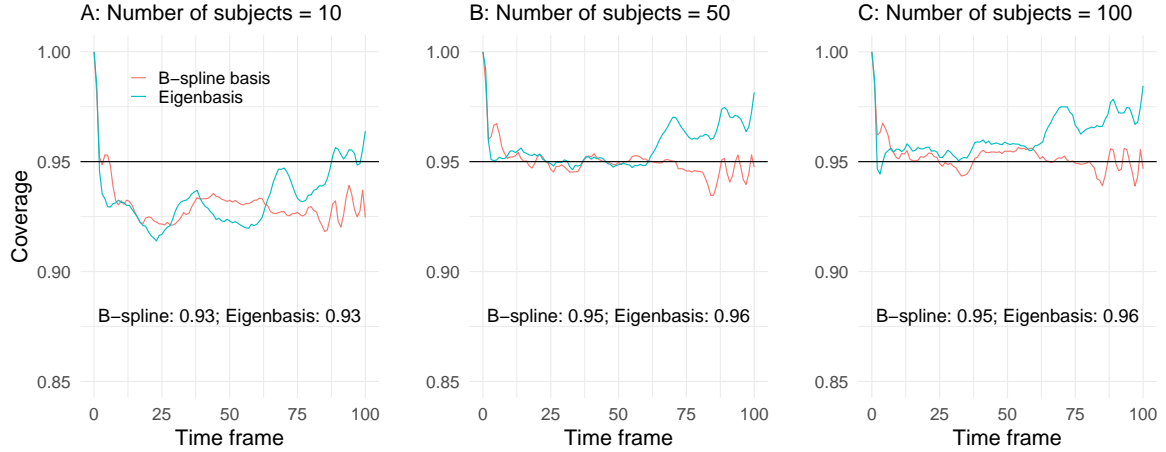


Figure 4.6: Simulation results to compare B-spline basis and eigenbasis for the random-effects regressor matrices using a linear mixed-effects modelling (LMM) framework to calculate 95% functional limits of agreement. Here, 100 different random samples were used to calculate the coverage curve for any given sample size. Overall coverage (average over the domain) corresponding to each framework are shown in panel-A, B and C.

regressor matrix. A comparison of the coverage of both frameworks will be evaluated. In addition a comparison of the time taken when using an eigenbasis compared to a B-spline basis for the random-effects regressor matrix will be made. The data generation procedure needed to simulate functional responses incorporating replicates at the individual level will be described first.

In this simulation, 5,000 subjects were considered as a population where, for each subject, three replicate functional responses were simulated. To simulate the functional responses, the observed difference curves of the right hip abduction angle from one measurement session were obtained. Following this, model (3.38) was fitted excluding the session specific level. For the fixed-effects, 12 B-spline basis functions were used. The same basis functions were used for the subject-specific and replicate specific random-effects regressor matrices. An ARMA(2, 1) was used for the error structure. After fitting the model mean curve, the variance-covariance matrix for the subject-specific deviation (from the mean curve) curve, the variance-covariance matrix for the replicate-specific deviation (from the subject-specific curve) curve, and the variance-covariance matrix of the residuals were obtained. Using the mean curve and the variance-covariance matrix of the subject specific deviation curve as the mean vector and variance-covariance ma-

4.5. Coverage and Computational Performance of a LMM in Studies with Replicates

trix of a multivariate Normal distribution, 5,000 subject-specific curves were obtained. For each of the subject specific curves, three replicate deviation curves were generated for each subject. To generate the replicate deviation curves, a mean of zero and the variance-covariance matrix of the replicate-specific deviation curves was used as the mean vector and variance-covariance matrix of a multivariate Normal distribution to generate three replicate curves for a given subject. For each replicate curve, an error curve was generated using a multivariate Normal distribution with zero as the mean vector and the variance-covariance matrix of the residual curves as the variance-covariance matrix. Adding the error curve to each subject specific replicate curve, a functional response for each replicate measurement was obtained for each subject. In total 15,000 functional responses were generated i.e. 3 from each of the 5,000 subjects. Figure 4.6 displays the observed data from the replicate measurements from nine different subjects and the simulated data of replicate measurements from ten different subjects.

4.5.1 Comparing Coverage When Using a B-spline and Eigenbasis for the Random-effects Regressor Matrices

After simulating data for studies with replicates, both the B-spline and eigenbasis for the random-effects regressor matrix were compared to evaluate the coverage of both frameworks. In addition, the time to fit both LMMs was also compared.

To compare the coverage for the B-spline and eigenbasis for the random-effects regressor matrix, the same simulation strategy was used (Figure 4.1). For a given sample size, 100 different samples were randomly chosen. Here a sample size of n means n different subjects were chosen with three replicate measurements for each subject. For a given sample, two separate LMMs were fitted: one with the B-spline basis and other with the eigenbasis for the random-effects regressor matrices. For both LMMs, 12 B-spline basis functions for the fixed-effects regressor matrix were chosen as outlined in Chapter 3. For the LMMs using B-spline basis functions for the random-effects regressor matrix, no error structures were included in the model to allow the `lme4` package to be used to estimate the unstructured variance-covariance matrix. For

4.5. Coverage and Computational Performance of a LMM in Studies with Replicates

Table 4.1: Time needed to fit LMMs for the two basis systems used for the random-effects regressor matrix.

Basis system	Error structure	Average ¹ time to fit the models (in minutes)		
		Sample size 10	Sample size 50	Sample size 100
B-spline basis	Not included	7.06	38.17	81.56
Eigenbasis	Included	3.70	22.56	50.36

¹ Averaged over 100 different model fits.

the LMMs with an eigenbasis for the random-effects regressor matrix, an ARMA(2,1) error structure was used as the models were estimated using the `nlme` package which accommodates this error structure.

It should be noted that a LMM model with an error structure needed more time to be fitted compared to the same model without a specified error structure. For the difference curves, incorporating an ARMA(2,1) error structure in a LMM removes the auto-correlation in the residuals. However, fitting a LMM with an unstructured variance-covariance matrices for the random-effects poses convergence issues while using the `nlme` package. For this reason when the B-spline basis functions were used for the random-effects, the `lme4` package was used. As the `lme4` package does not have the flexibility to incorporate an error structure in the model it was found, in Chapter 3, that the error structure did not influence the estimate of the standard deviation curve and hence the error structure was dropped for the LMMs with the B-spline basis functions for the random-effects regressor matrix in this simulation. For the LMMs with an eigenbasis for the random-effects regressor matrix, only diagonal variance-covariance matrices for the random-effects are needed. For this reason the `nlme` package was used as it allows an error structure to be specified in a LMM.

After fitting the LMMs, coverage of the 95% fLoA for the LMM with the B-spline and eigenbasis for the random-effects regressor matrix was compared (Figure 4.6). Both frameworks had similar coverage curves. However, the coverage curve of the 95% fLoA for a LMM with eigenbasis for the random-effects regressor matrix was slightly better. One possible explanation could be the inclusion of the ARMA(2,1) error structure in the LMMs which could be investigated further in a future study.

4.5. Coverage and Computational Performance of a LMM in Studies with Replicates

The time to fit LMMs using both the approaches was also compared (Table 4.1). The time to fit a LMM with an eigenbasis for the random-effects regressor matrices was always less compared to the time needed to fit a LMM with a B-spline basis for the random-effects regressor matrices. For sample size 10, the eigenbasis implementation is about 50% faster than the B-spline implementation for the random-effects regressor matrix. For sample size 50 and 10, the eigenbasis implementation is about 40% faster than the the B-spline implementation. From this it can be concluded that the eigenbasis implementation for the random-effects regressor matrix will be considerably faster than the B-spline implementation for the random-effects regressor matrix while calculating a 95% fLoA using the LMM framework. The reason is that a LMM with an eigenbasis for the random-effects regressor matrix allows a diagonal variance-covariance matrices for the random-effects to be used in the model while a B-spline basis for the random-effects regressor matrices needs an unstructured variance-covariance matrices for the random-effects. Since a LMM with a diagonal variance-covariance matrix requires fewer model parameters to be estimated compared to a LMM with an unstructured variance-covariance matrix, the eigenbasis implementation requires less time to fit a LMM.

In the beginning of this chapter a set of questions were posed regarding the use of the proposed LMM framework in method comparison studies with functional responses. Answers to those questions were obtained though a set of simulation studies. In particular, it was shown that in method comparison studies, one should interpret 95% fLoA as pointwise bands. This means that at each time point, 95% of the measurement values should be within the bands. For studies with no replicates, a simulation study found that the FDA and LMM frameworks give similar coverage for the different sample sizes considered. From this it can be concluded that for studies with no replicates, FDA and LMMs are comparable with respect to calculating a 95% fLoA. For studies with replicates, it was found that a LMM framework provides nominal coverage with respect to a 95% fLoA for different sample sizes. As a full B-spline implementation requires an unstructured variance-covariance matrix to be estimated for the random-effects, it is time consuming. An eigenbasis implementation of the random-effects regressor matrix was proposed. A simulation study found that both the eigenbasis and full B-spline

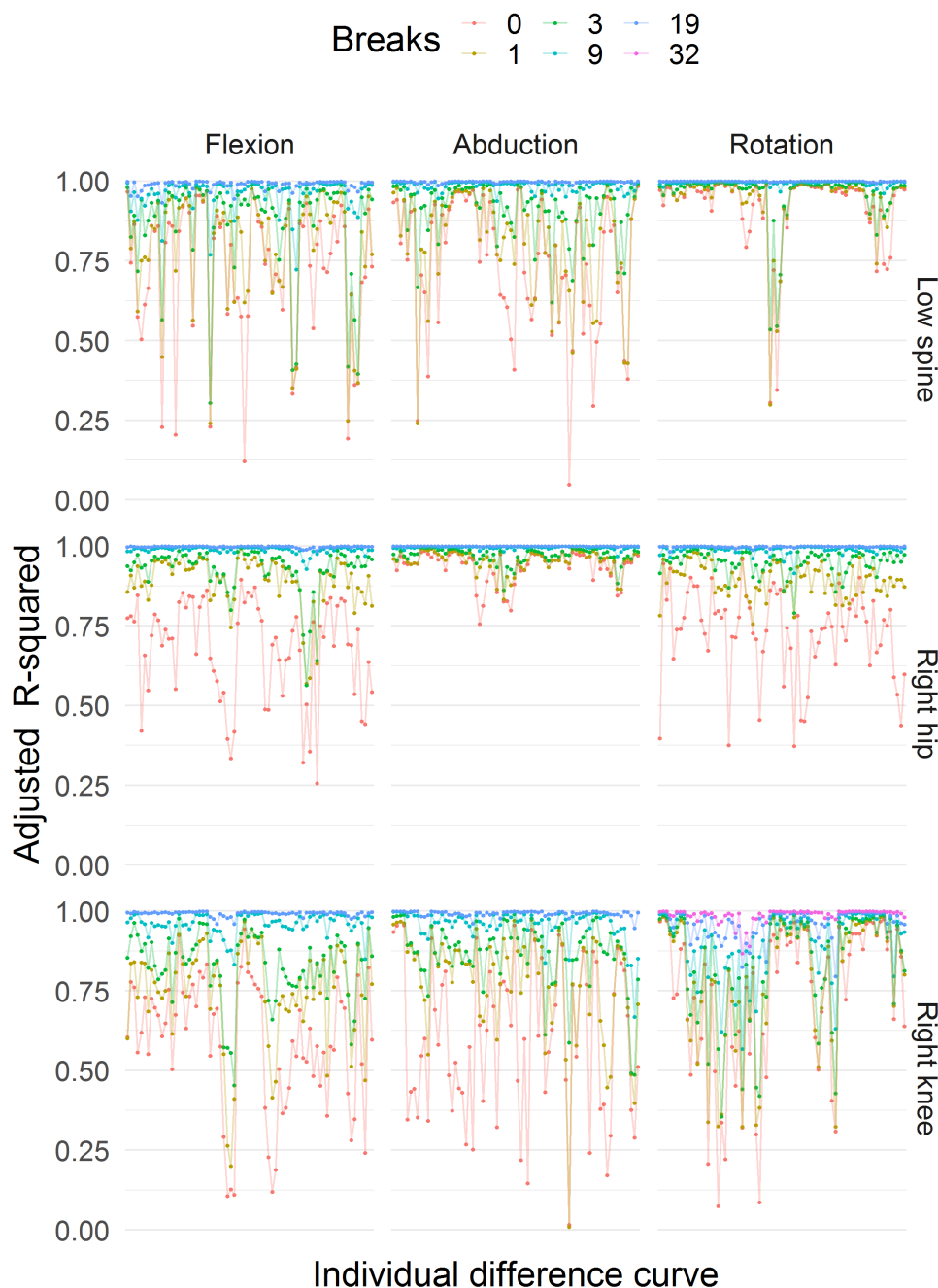


Figure 4.7: Adjusted R-squared values after fitting a linear regression model to each individual difference curve using B-spline basis functions with different numbers of equally spaced inner knots for different angles measured in the motion capture study. Here different points on x-axis are different difference curve. For each angle, 54 difference curves (9 athletes, 2 sessions, 3 replicates) were used measured by two different measurement methods: markerless and marker-based.

implementation give similar nominal coverage. However, the eigenbasis implementation provided slightly better coverage in the simulation setup used. It was found that an eigenbasis implementation is considerably faster when fitting a LMM compared to when using a full B-spline implementation.

The simulation studies suggest that a LMM framework to calculate a 95% fLoA provides valid results across the sample size and study design scenarios considered. This framework will now be applied to run the analyses needed for the other angles in the motion capture study which have not been considered so far.

4.6 Application: Motion Capture Study

The main aim of the motion capture study was to compare both the marker-based and markerless method of measurement where data were collected for a set of different angles using both the methods of measurement. So far in this thesis only the right hip abduction angle was used by way of example to demonstrate the various frameworks proposed to generate 95% fLoA. From the simulation studies, it was observed that the proposed framework of calculating a 95% fLoA using a LMM gave satisfactory results, in terms of coverage, for studies with and without replicate measurements at the individual level. The proposed methods will now be applied to calculate the 95% fLoA for all the other angles in the current study. Step by step procedures to calculate a 95% fLoA were proposed in Section 3.6 and 3.7. Both procedures can be used in practice. However, as the procedure proposed in Section 3.7 is faster with an added gain in terms of computational time and it allows an error structure to be specified in the LMM, it will be used to calculate the 95% fLoA for all the angles considered.

For each angle a separate B-spline basis system for the corresponding fixed-effects regressor matrix is needed. The proposed R-squared criterion for selecting a B-spline basis system for the knot sequences was used. Using the R-squared criterion, it was found that an equally spaced knot sequence with 9 breaks was sufficient for the right hip flexion, right hip abduction, low spine rotation, and right hip rotation angles (Figure 4.7). For the low spine flexion, right knee flexion, low spine abduction, and right knee abduc-

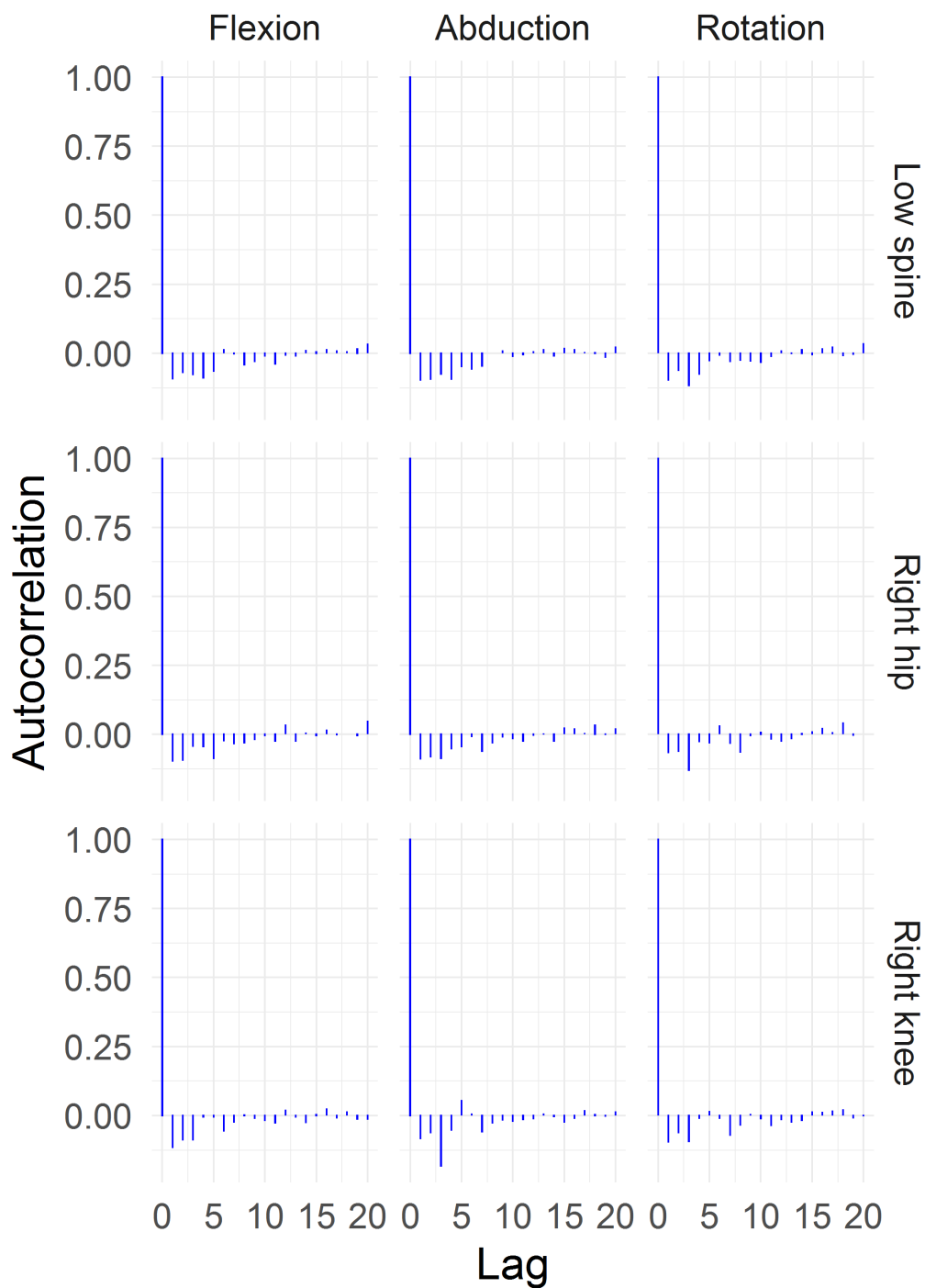


Figure 4.8: Autocorrelation plot of the residuals after fitting a LMM of the difference curves with autoregressive moving average ARMA(2,1) error structure for different angles during a lunge from the motion capture study.

4.6. Application: Motion Capture Study

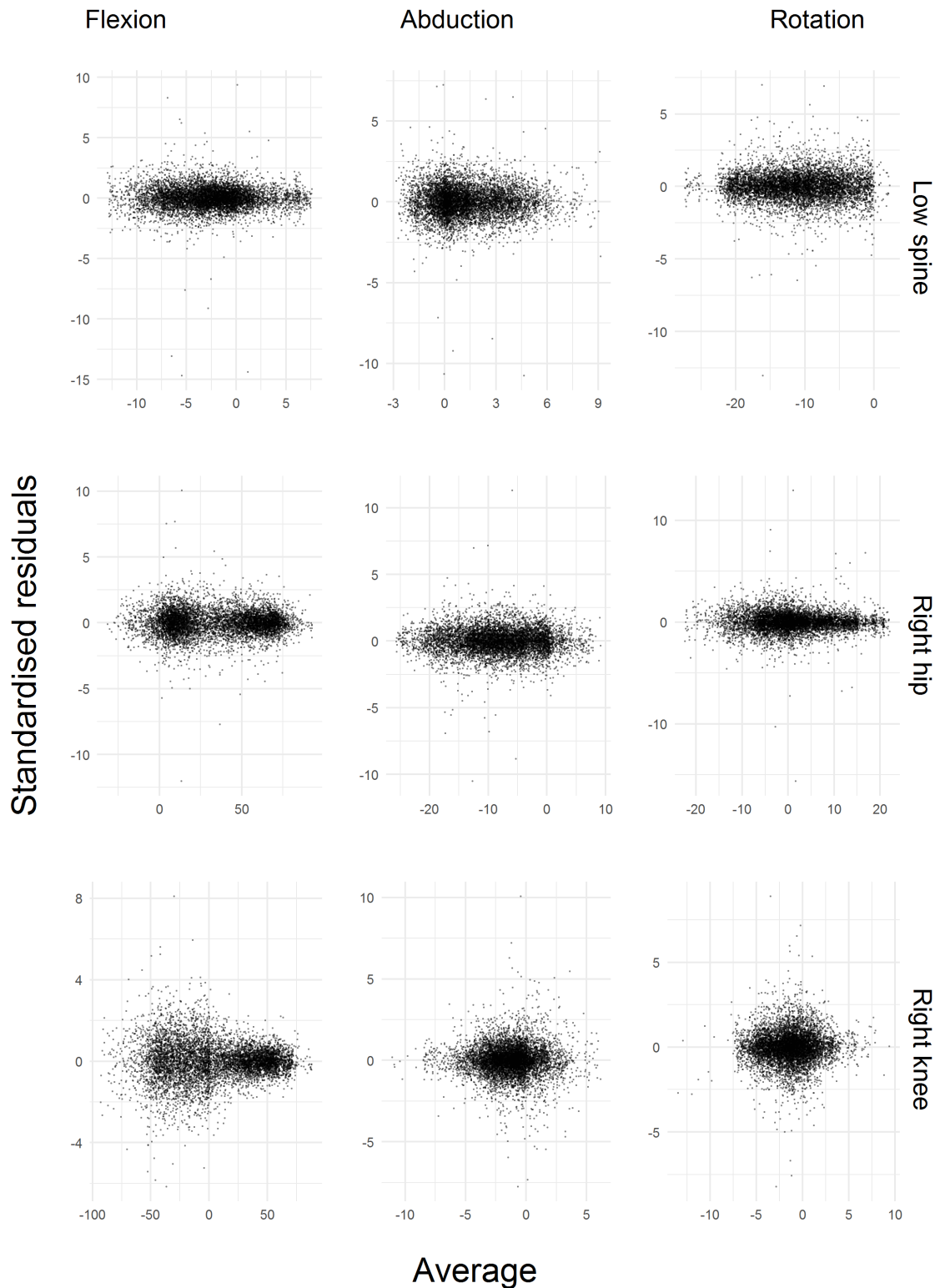


Figure 4.9: Residuals against average plot after fitting a linear mixed model of the difference curves for different angles during a lunge from the motion capture study.

Table 4.2: LMM specification for the difference curves by angles with the time taken to fit the model.

Angles	Fixed-effects (breaks)	Random-effects (eigenfunctions)	Error structure	Time to fit (in minutes)
Low spine flexion	19	12	ARMA(2,1)	22.92
Right hip flexion	9	12	ARMA(2,1)	20.14
Right knee flexion	19	15	ARMA(2,1)	31.45
Low spine abduction	19	12	ARMA(2,1)	27.35
Right hip abduction	9	12	ARMA(2,1)	18.86
Right knee abduction	19	14	ARMA(2,1)	34.10
Low spine rotation	9	12	ARMA(2,1)	20.10
Right hip rotation	9	12	ARMA(2,1)	21.93
Right knee rotation	32	14	ARMA(2,1)	31.68

tion 19 breaks for an equally spaced knot sequence was necessary (Figure 4.7). For the right knee rotation angle, 32 breaks for an equally spaced knot sequence was required (Figure 4.7).

After conducting an eigenvalue decomposition of the variance-covariance matrix of the difference curves, the required number of eigenvectors was chosen for each random-effects regressor matrix. Using the procedure described in Chapter 3, it was found that 12 eigenvectors were sufficient for the random-effects regressor matrices for the low spine flexion, right hip flexion, low spine abduction, right hip abduction, low spine rotation, and right hip rotation angle (Table 4.6). For the right knee abduction and right knee rotation 14 eigenvectors were sufficient (Table 4.6). For the right knee flexion 15 eigenvectors were needed (Table 4.6). For all the angles an ARMA(2,1) error structure was used which removed the autocorrelation from the residuals for all the LMMs (Figure 4.8). The model specification for the LMMs for each angle and the corresponding time to fit each model can be found in Table 4.6. The appropriateness of these model specifications can be justified by observation of the residual plots (Figure 4.9). Finally, the 95% fLoA were calculated using LMMs for each angle in order to assess the level of agreement, or lack of, and how this varied over the frame. (Figure 4.10).

From these results it is clear that the level of agreement between the marker-based and markerless method is poor in general (Figure 4.10). There is substantial bias

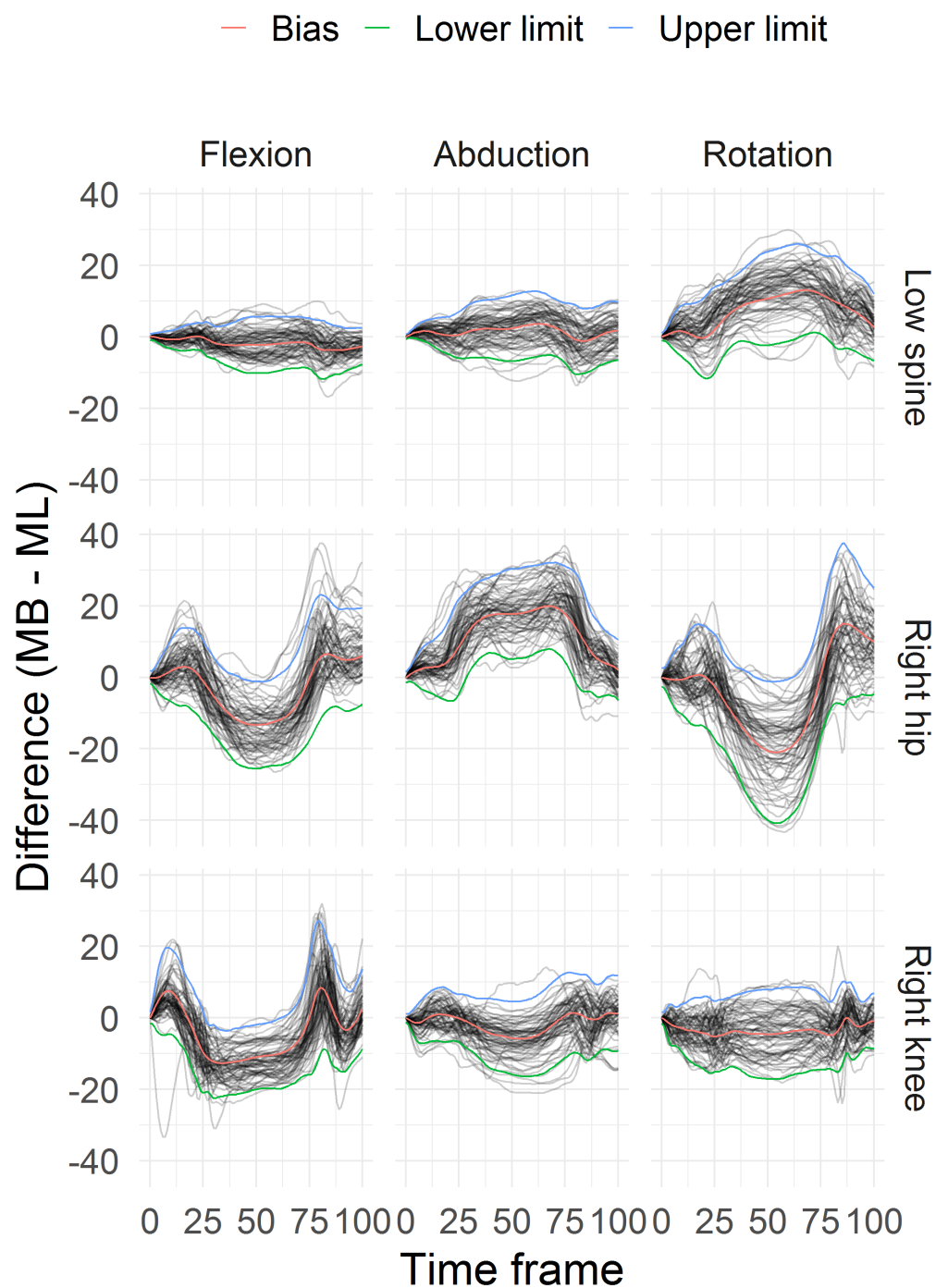


Figure 4.10: 95% functional limits of agreement using a linear mixed-effects model for all angles measured during a lunge in the motion capture study. For each angle, all the replicate measurements in the two measurement sessions were used to calculate the 95% functional limits of agreement. Here MB is the marker-based and ML is the markerless method of measurement.

between the two methods for all angles, except for the low spine flexion, low spine abduction, right knee abduction and right knee rotation angles. The width of the 95% fLoA is considerable for all angles and hence it can be concluded that the markerless system has a poor level of agreement if used to measure angles generated during a lunge as a replacement for the marker based system. It can also be noticed that the level of disagreement is more for rotation angles compared to the flexion and abduction angles. In the simulation study, it was found that in such studies with a sample size of 10 the coverage of calculated 95% fLoA was less than the 0.95. Given this observation the 95% fLoA in the motion capture study should be interpreted with caution. The reported 95% fLoA are likely to be narrower than they should be and a larger sample size is justified to provide more reliable inference. That said, it is clear from the estimated bias that the level of agreement is poor, in particular if the angle involves a rotation.

4.6.1 R Code to Fit a LMM to Calculate 95% fLoA

To illustrate how these models can be implemented in practice, further illustrating the attractiveness of using the mixed modelling firework, the R code needed to fit a LMM using the two R packages: `nlme` and `lme4` is given. Either function `lme` from the `nlme` package or the function `lmer` from the `lme4` package can be used to fit a LMM.

As an example, the code needed to fit a LMM for studies with no replicate measurements to calculate 95% fLoA using the `lme` function from the `nlme` package is as follows.

```
library(nlme)
lme_model_no_replicates <-
lme(dif ~ (X1 + X2 + X3 + X4 + X5 + X6 +
          X7 + X8 + X9 + X10 + X11 + X12)-1,
    random=list( sub = pdSymm( ~ (X1 + X2 + X3 + X4 + X5 + X6 +
                                X7 + X8 + X9 + X10 + X11 + X12)-1 )),
    correlation = corARMA(form=~frame| sub, p=2, q=1),
    data = dat,
```

4.6. Application: Motion Capture Study

```
control = lmeControl(msMaxIter=5000))
```

A simple extension is needed to the R code, in relation to how the correlation structure is specified, to fit a LMM for studies with replicate measurements to calculate 95% fLoA using the `lme` function from the `nlme` package as shown below.

```
library(nlme)
lme_model_replicates <-
lme( dif ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
      X10 + X11 + X12 - 1,
      random = list(sub = pdSymm(~ X1 + X2 + X3 + X4 + X5 + X6 +
                                  X7 + X8 + X9 + X10 + X11 + X12 - 1),
                    ses = pdSymm(~ X1 + X2 + X3 + X4 + X5 + X6 +
                                  X7 + X8 + X9 + X10 + X11 + X12 - 1),
                    rep = pdSymm(~ X1 + X2 + X3 + X4 + X5 + X6 +
                                  X7 + X8 + X9 + X10 + X11 + X12 - 1)),
      correlation = corARMA(form = ~frame | sub/ses/rep, p=2, q=1),
      data = dat,
      control = lmeControl(msMaxIter = 5000)
)
```

If there are any computational issues due to the `lme` function from the `nlme` package then the following code can be used for the `lmer` function from the `lme4` package. Note the limitation in using this package as an error structure cannot be used in the `lmer` function.

```
library(lme4)
lmer_model_replicates <- lmer( dif ~ (X1 + X2 + X3 + X4 + X5 + X6 +
                                     X7 + X8 + X9 + X10 + X11 + X12) - 1 +
                               (0 + (X1 + X2 + X3 + X4 + X5 + X6 + X7 +
                                     X8 + X9 + X10 + X11 + X12) | sub/ses/rep),
      data = dat)
```

4.7. Summary

When there is a need to use a computationally faster approach involving an eigenbasis system for the random-effects regressor matrix, the following code can be used for the `lme` function from the `nlme` package. Note that here a diagonal variance-covariance matrix was used for the random-effects.

```
library(nlme)
lme_model_replicates_eigenbasis <-
lme( dif ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
      X10 + X11 + X12 - 1,
      random = list(sub = pdDiag(~ X1.1 + X2.1 + X3.1 + X4.1 + X5.1 + X6.1 +
                                   X7.1 + X8.1 + X9.1 + X10.1 + X11.1 + X12.1 - 1),
                    ses = pdDiag(~ X1.1 + X2.1 + X3.1 + X4.1 + X5.1 + X6.1 +
                                   X7.1 + X8.1 + X9.1 + X10.1 + X11.1 + X12.1 - 1),
                    rep = pdDiag(~ X1.1 + X2.1 + X3.1 + X4.1 + X5.1 + X6.1 +
                                   X7.1 + X8.1 + X9.1 + X10.1 + X11.1 + X12.1 - 1)),
      correlation = corARMA(form = ~frame| sub/ses/rep, p=2, q=1),
      data = dat_eig,
      control = lmeControl(msMaxIter = 5000)
)
```

4.7 Summary

In this chapter, a simulation strategy was designed to mirror situations where functional responses were generated from a study with no replicate measurements and with replicate measurements. The data collected in the motion capture study was used to guide the simulation model to simulate a pseudo population.

The correct interpretation of functional limits of agreement was provided using a simulation study. From the results, one can see that the 95% fLoA is best interpreted as a pointwise band that contains 95% of the measurements at each time point. However, a 95% global band can also be achieved using a grid search of different pointwise bands using the method presented in section 4.4.1.

For each scenario considered, the achieved coverage for each approach was calculated and compared to the nominal coverage (0.95) required.

For studies with no replicate measurements, it was found that the 95% fLoA using the proposed LMM framework gave reasonable coverage for different sample sizes. It was also found that to calculate the 95% fLoA, both the FDA and LMM framework are comparable for different sample sizes.

For studies with replicate measurements, the LMM framework also provided reasonable coverage for the different sample sizes considered. Both the B-spline basis system and the eigenbasis system for the random-effects regressor matrices provided comparable coverage for a 95% fLoA. However, fitting an LMM with an eigenbasis for the random-effects regressor matrices is considerably faster compared to when using a B-spline basis system.

The proposed LMM framework was applied to calculate a 95% fLoA to assess the agreement between the two methods of measurement for all the angles collected during a lunge in the motion capture study. The sample size used in the study is small and the study may best be considered as a pilot study. It was found however that in general the agreement between the two methods of measurement is poor and replacing the marker-based system with the newly proposed markerless system is not to be recommended, in particular for angles that involve a rotation.

The thesis concludes in the next chapter where an overall summary of the thesis and of the case studies used are given. The strengths and limitations of the work presented in this thesis and interesting areas of further work are discussed.

Chapter 5

Conclusions and Further Work

5.1 Introduction

Often new methods of measurement emerge to provide a convenient and/or cheaper solution to the existing gold standard measurement method in use (Altman and Bland, 1983; Bland and Altman, 1986). Before using a new measurement method, a scientific investigation, a method comparison study, is required to establish the comparability of the new method with the existing one (Altman and Bland, 1983) by estimating the degree, or level of agreement, between the two methods. If the level of agreement is considered adequate for practical purposes the newer system can be used. In this thesis problems of this nature were considered, in particular for functional response variables. New approaches, graphical and analytical, for method comparison studies for both univariate and functional responses are proposed. A summary of the results of these approaches when applied to the two case studies is given followed by a discussion of potential areas of further work and an overall conclusion.

A brief summary of the results of each study is now given.

5.2 CRP Analyser Study Summary and Conclusion

As discussed in Chapter 1, the measurement of C-Reactive Protein (CRP) is used to monitor the inflammatory states of elite athletes. The standard laboratory method

requires a venous blood samples drawn via the antecubital vein in a serum separator tube for the measurement of CRP. As frequent collection and transportation of blood samples to a laboratory is challenging, longitudinal studies examining CRP biomarkers in elite sports are rare. Measuring CRP using a point of care (POC) test of a capillary blood sample removes such challenges and provide a practical solution for rapid result reporting. A simple and small volume POC test provides a number of advantages over a laboratory test in the elite high performance environment. However, before using a POC test to identify meaningful changes in CRP measurement, it has to be established that the laboratory method and a POC test are comparable to justify its use.

The level of agreement between the standard laboratory test and a POC test can be assessed by estimating the 95% LoA between the two methods. These limits provide guidance on the presence (and form) of bias between the two methods of measurement on average, and how different the measurements by these two methods for each individual measurement are likely to be. However, the 95% LoA are only interpretable when it can be assumed that the mean and standard deviation of the difference between the measurements made by the two methods are constant over the magnitude of the value being measured. These assumptions can be verified visually using the so called Bland-Altman plot. The classical Bland-Altman plot displays the differences on the y-axis and the value of the average on the x-axis and a visual inspection is often used to decide whether or not the assumption seems justified. As this is a subjective approach, error might occur in some situations where one might think that the assumptions are justified but in reality they are not. To remove this subjectivity, this thesis proposed a new graphical modification to the Bland-Altman plot. The approach uses a LOESS line on the Bland-Altman plot to aid in making the judgement more objective. Using this new graphical approach, it is found that the assumptions do not seem justified for the CRP analyser study. It appears there is a non-linear bias exist between the two methods of measurement. A literature search found no existing analytical approach to calculate 95% LoA in the presence of non-linear bias. For this reason this thesis proposed a novel analytical approach to calculate the 95% LoA in the presence of non-linear bias. The analytical approach uses a nonparametric LMM to model the relationship between the

5.2. CRP Analyser Study Summary and Conclusion

differences and the averages. After fitting the model it was found that the non-linear bias was adequately modelled using the nonparametric LMM approach. However, this approach was not sufficient enough to remove the heteroscedasticity. After taking the logarithm transform of the CRP measurements the nonparametric LMM approach was able to remove the heteroscedasticity. After calculating the 95% LoA using the new approach, proposed in this thesis, it is found that there is a substantial bias between the measurements made by the two methods of measurement when the magnitude of the value is below about 1 mg/L. However, since log CRP was used for the analysis, the bias between these two methods are multiplicative.

For example, for an average log CRP value of -0.96 the relative bias is 1.69 with the 95% LoA (1.08, 2.31). After taking the anti-log, the average log CRP of -0.96 should be interpreted as a geometric mean of 0.38 mg/L. This suggests that on average a POC method could produce measurements which are 5.42 times higher than the measurement produced by the laboratory method while the 95% LoA (1.08, 2.31) should be interpreted as the POC method could produce measurements 2.94 to 10.07 times higher than the measurement produced by the laboratory method when the true value of the measurement is 0.38 mg/L (geometric mean).

Although the existence of bias below a measurement of CRP value 1 mg/L may seem not relevant for clinical purpose, it may have important consequences in terms of performance in elite sports. For this reason it can be concluded from the analysis that the two method do not agree for low values of CRP and the POC test cannot be considered as interchangeable with the standard laboratory method for such values. It could also be the case that there is a threshold below which the POC test can not detect CRP accurately.

A summary of the main findings of the motion capture study now follows. To start, a short reminder of the aim of the trial and the design used is given.

5.3 Motion Capture Study Summary and Conclusion

Human movement analysis can identify weaknesses in functional movement in elite athletes and hence can be used as a guidance for developing athletic development program tailored for individual needs to ensure long term development of the athletes (Lloyd et al., 2016). It can also be used to avoid injuries in competitive sporting environment (Lloyd et al., 2015; Chorba et al., 2010).

To quantify the functional movement, Cardan (Euler) sequence of angles were measured for different joint segments along the X, Y, and Z-axis in the anatomical body planes during a right leg lunge (Glasoe et al., 2014). This produces flexion, abduction and rotational angle curves over the duration of the exercise along the three different axis. These angle curves over time are the main functional response for the motion capture study. A hierarchical study design was used where measurements were collected in two different sessions and for each session, three functional replicates were collected from each athlete.

The existing practice in biomechanical research is to measure the flexion, abduction and rotational functional response using a marker-based (MB) motion capture system. This system puts reflective markers on bony landmarks and by tracking the trajectories of these marker using cameras the position and orientation of body segments in the the 3D coordinate system can be calculated. However, the application of markers to anatomical bony points is a time consuming process, constrained by inter-, intra-tester and session reliability issues and also suffers from soft tissue artefacts.

The emergence of novel markerless motion capture systems without the use of reflective markers is a rapidly growing field which has an attractive future advancement in motion analysis. A markerless motion capture system offers a fully automatic, non-invasive, markerless approach, which would ultimately provide a major breakthrough for research and if it is deemed reliable then that could remove the difficulty of quantitatively assessing movement quality in elite academy footballers and ultimately have wider scope in the general population. However, to use the markerless motion capture

system researchers need evidence that the system is comparable with the marker-based system.

In this study functional responses were collected for nine different angles during a lunge. In this thesis functional responses collected for the right hip abduction angle during a lunge was used for the purpose of illustration of the proposed LMM framework. Then, using the proposed framework, a comparison of the methods of measurement was performed for each of the angles. The level of agreement between the two methods of measurement was based on whether the magnitude of the 95% fLoA was precise enough i.e. the range of the limit of agreement was smaller than what would be deemed as a meaningful change. It was found that there is a substantial disagreement between the two methods of measurements for different angles during a right lunge as the width of the agreement limits was considered large for a major portion of the duration of the lunge exercise. This call relied on domain knowledge from the biomechanists involved in the study.

It is found that the markerless motion capture system over-estimated kinematic measurement for most angles. There is a substantial bias between the two methods of measurement for most of the angles during a lunge. To use the markerless system for motion capture one must calibrate the system to remove this bias. The results of this analysis can be used to provide the frame by frame bias adjustment as needed. However, the difference between these two methods of measurement for an individual measurement is substantially higher for all the angles measured. From this, it can be concluded that the markerless cannot be used instead of marker-based motion capture system, in particular when the angle in question represents an internal/external rotation.

Now overall summary of the thesis is given in the next section.

5.4 Summary of the Thesis

The statistical approaches to deal with method comparison study with functional responses are different to those used for univariate responses. This area of statistical methodology is relatively new and there is room for further methodological develop-

ment. In this thesis a new approach is presented to calculate 95% functional limits of agreement using a LMM framework. This approach can easily handle functional replicates as well as data generated from a hierarchical study design.

When fitting a LMM in any context, there are many aspects of the model that must be considered in order to model the response appropriately. All of these aspects will differ from study to study. The same applies when using a LMM to model a functional response in method agreement studies. For example, to calculate the 95% fLoA using a LMM, one needs to fit a nonparametric linear mixed-effects model with an appropriate fixed-effects regressor matrix, random-effects regressor matrix, structure of variance-covariance of the random-effects and the correlation structure for the error. The choices for each will differ from study to study. In this thesis, to calculate the 95% fLoA using the proposed LMM framework, a set of guidelines for the model specifications are given:

- For the fixed-effects regressor matrix, choose a cubic B-spline basis system with a knot sequence suggested by the adjusted R-squared criterion.
- Choose the same B-spline basis system for the random-effects regressor matrix.
- The variance-covariance matrix for the random-effects must be unstructured.
- Based on autocorrelation plot choose a appropriate correlation structure for the error.

The 95% LoA proposed by Bland and Altman relies on two important assumptions; the mean and standard deviation of the differences are constant over the range of the value being measured. However, this assumption is not necessary for situations when a regression model is used to estimate the 95% LoA as the modelling approach can accommodate the scenario where the mean and standard deviation of the differences vary over the measurement range. In this case one needs to check the usual assumptions required for the specific modelling approach being used. A simple graphical approach is also proposed in this thesis to check if a LMM is sufficient to handle the situation when the mean and standard deviations of the differences vary over the range of measurement. This proposed graphical approach involves plotting the residuals of a LMM against the

average value of the measurements for both the methods ignoring the time frame. If it is noticed that there is no clear pattern in the residuals then it can be concluded that the LMM is sufficient to model the bias and standard deviation over the range of values of the measurement.

A simulation study was conducted to compare the results obtained using both the FDA and LMM frameworks for studies with no replicate measurements. With the simulation situation considered, it was found that both the frameworks produces the similar estimates of the 95% fLoA. For the studies with replicates, it was also found that a LMM framework provides satisfactory results for the calculation of the 95% fLoA.

Although only a specific scenario was considered in the simulation study (i.e. the right hip abduction angle) to compare the two frameworks, it is likely that the same results would be obtained in other situations with similar functional responses over a fixed grid. The results provide empirical evidence that both the FDA and LMM approaches are comparable in the designs considered here with no replicates. This result provides a big advantage for method comparison study since using a LMM framework as one can estimate the 95% fLoA in the presence of replicates, if the study design is hierarchical, and if there is a need to adjust for covariates quite easily. This is not to say that FDA is not an appropriate method to use in method comparison studies in more complex designs. From a practical point of view, the use of LMM is quite widespread in applied statistics and is popular in method comparison studies with univariate responses (Carstensen et al., 2008). It could be advantageous to use the same modelling framework for method agreement studies involving functional responses. Being able to extend this framework to accommodate functional responses is an attractive proposition.

Fitting a nonparametric LMM with an unstructured variance-covariance matrix is however a time consuming process. This may appear as an obstacle to calculate the 95% fLoA using a LMM framework. In this thesis a new computational approach was proposed to remove this issue. The proposed computational approach involves choosing a different basis system for the random-effects regressor matrix. This new basis system consists of a selected number of eigenvectors of the sample variance-covariance matrix of the functional responses. As the eigenvectors are orthogonal to each other,

one can now use a diagonal variance-covariance matrix for the random-effects in the LMM. This diagonal variance-covariance matrix for the random-effects in LMM provides huge computational advantages since less parameters need to be estimated for the variance-covariance matrix for the random-effects compared to the number of parameters needed to be estimated when the variance-covariance matrix for the random-effects is unstructured. To use this computationally faster approach to calculate the 95% fLoA the following guidelines for a LMM is given:

- For the fixed-effects regressor matrix, choose a cubic B-spline basis system with a knot sequence suggested by the adjusted R-squared criterion.
- Calculate variance-covariance matrix for functional responses and obtain eigenvalue-eigenvector decomposition of the matrix.
- Choose the first few eigenvectors that explain at least 99% of the variation of the responses.
- Use a diagonal variance-covariance matrix for the random-effects is sufficient.
- Choose a appropriate correlation structure for the based on the autocorrelation plot .

A simulation study was conducted to assess the performance and computational time of this new approach. It was found that the approach of using an eigenbasis for the random-effects regressor matrix provides the same performance compared to a full B-spline implementation for a LMM. However, computational time required to fit a LMM with an eigenbasis for the random-effects regressor matrix is substantially less compared to the time required to fit a LMM with a full B-spline implementation.

5.5 Open Questions and Further Work

The statistical approaches and analysis conducted in this thesis can further be extended.

For method comparison study, replicates measurement are essential to estimate within subject variability which contains the error of measurement for a given method

5.5. Open Questions and Further Work

of measurement. Without collecting replicate measurements it is not possible to estimate within- and between- subject variability separately. In the CRP analyser study, replicates were taken, however they were taken at different time points usually one month apart from each replicates. Although this allowed an estimate of the within- and between- subject variability separately, the issue is that the source of within-subject variability was not only from the measurement error. The true value of the CRP for a given subject varies from month to month which was also part of the within-subject variability. Since correct estimation of the variability of measurement error is crucial to estimate reliability indices of a method of measurement, for the CRP analyser study it was not possible to estimate reliability indices for both the standard laboratory test and the POC test. This issues could have been avoided if multiple replicates were taken at each time points. For future method comparison study to correctly estimate reliability indices one needs to collect multiple replicates at each time point.

The 95% LoA is a population quantity and which is estimated using a sample from the population. Since it is a sample statistic, interval estimates are needed to assess the uncertainty of the estimate. However it is rare in the literature that the estimate of the 95% LoA are accompanied with corresponding interval estimates. As the 95% LoA consist of two limits (lower limits of agreement and upper limits of agreement), confidence interval for these limits would provide an assessment of the uncertainty in each estimate. Although it is possible to estimate the confidence intervals for both limits, it is only essential to report the lower number of the confidence interval of the lower limit and the upper number of the confidence interval for the upper limit. A future study is needed to calculate confidence interval for the 95% fLoA.

A sample size calculation is a key component of all study designs to ensure that proposed sample size has sufficient power. A sample size calculation was not used for either of the case studies presented and both studies run the risk of making a Type II error. The summary statistics from each study could be used to design a more robust trial if needed.

In the motion capture study three functional responses, flexion, abduction and rotation angle curve, were collected to measure the motion of each body segment. These

data are inherently multivariate functional responses. To assess the agreement of the ML method with the gold-standard MB method, a multivariate analysis of the functional responses is desirable. This is an area of further work worth considering.

In this thesis, the performance of the proposed linear mixed-effects modelling framework was only evaluated for studies with no replicates and studies with replicates. Only one specific situation was considered for the simulation study where data were simulated based on the right hip abduction angle curves. Further simulation studies could be carried out involving more complex study designs, considering different sampling rates for the frame, in the presence of covariates, and what aspects of the model are robust to misspecification.

Appendix

R code for Chapter 2

Estimating reliability indices using a LMM

```
library(tidyverse)
library(lubridate)
library(nlme)
library(splines)

crp_lme <- lme(CRP ~ method + factor(month.season) - 1,
  random = list(athlete.id = pdDiag(~ method - 1)),
  weight = varIdent(form = ~ 1 | method), crp_long
)

poc_between <- 0.1471273^2
lab_between <- 0.2315653^2

poc_within <- 0.5908963^2
lab_within <- poc_within * 1.270238

poc_total <- poc_between + poc_within
lab_total <- lab_between + lab_within

poc_icc <- poc_between / poc_total
lab_icc <- lab_between / lab_total

poc_sem <- sqrt(poc_within)
lab_sem <- sqrt(lab_within)
```

Estimating CCC using a LMM

```
crp_lme_ccc <- lme(CRP ~ method + factor(month.season) - 1,  
  random = list(athlete.id = ~1),  
  weight = varIdent(form = ~ 1 | method), crp_long  
)
```

```
ccc_var_alpha <- 0.2567833^2  
ccc_var_poc <- 0.5647438^2  
ccc_var_lab <- 1.287454 * ccc_var_poc
```

```
ccc_beta_poc <- 0.95729081  
ccc_beta_lab <- 0.70014796
```

```
ccc <- (2 * ccc_var_alpha) /  
  ((ccc_var_alpha + ccc_var_poc) +  
   (ccc_var_alpha + ccc_var_lab) +  
   (ccc_beta_poc - ccc_beta_lab)^2)
```

95% LoA using log transformed response

```
log_ave <- unique(crp$log.ave)[order(unique(crp$log.ave))]  
X_log <- model.matrix(~ bs(log_ave, 4))
```

```
log.crp.model <- lme(log.diff ~ bs(log.ave, 4),  
  random = ~ 1 | athlete.id,  
  data = crp  
)
```

```
log_beta <- coef(summary(log.crp.model))[, 1]  
log_fit <- X_log %*% log_beta
```

```
log_var_between <- log.crp.model %>%
  getVarCov() %>%
  c()
log_var_within <- sigma(log.crp.model)^2
log_sd <- sqrt(log_var_between + log_var_within)

log_loa_dat <-
  data.frame(
    log_ave = log_ave,
    fit = log_fit,
    sd = log_sd,
    ul = log_fit + 2 * log_sd,
    ll = log_fit - 2 * log_sd
  ) %>%
  pivot_longer(c(fit, ul, ll))
```

R code for Chapter 3

95% fLoA using FDA for a Study with no Replicates

```
library(tidyverse)
library(fda)

dataMat <-
  dif_curve_dat %>%
  dplyr::select(sub, frame, dif) %>%
  pivot_wider(names_from = sub, values_from = dif) %>%
  dplyr::select(-frame) %>%
  as.matrix()

tobs <- seq(0, 100) / 100
nobs <- length(tobs)
knots <- tobs
```

```
nknots <- length(knots)
norder <- 4
nbasis <- length(knots) + norder - 2
basis <- create.bspline.basis(c(min(tobs),
                               max(tobs)), nbasis, norder, knots)

fdPar <- fdPar(basis, 2, 1e-5)
smoothlist <- smooth.basis(tobs, dataMat, fdPar)

fitfd <- smoothlist$fd
biasfd <- mean.fd(fitfd)
sdfd <- sd.fd(fitfd)

biasvec <- c(eval.fd(tobs, biasfd))
sdvec <- eval.fd(tobs, sdfd)

loadat <-
  data.frame(time = tobs, bias = biasvec, sd = sdvec) %>%
  mutate(ul = bias + 2 * sd) %>%
  mutate(ll = bias - 2 * sd) %>%
  pivot_longer(cols = c("bias", "ul", "ll"), names_to = "var")
```

95% fLoA using a LMM for a Study with no Replicates

```
library(tidyverse)
library(nlme)
library(splines)

fm1_dif <-
lme(dif ~ (X1 + X2 + X3 + X4 + X5 + X6 +
           X7 + X8 + X9 + X10 + X11 + X12) - 1,
    random = list(sub = pdSymm(~ (X1 + X2 + X3 + X4 + X5 + X6 +
                                   X7 + X8 + X9 + X10 + X11 + X12) - 1)),
    correlation = corARMA(form = ~ frame | sub, p = 2, q = 1),
```

```
data = dat,
control = lmeControl(msMaxIter = 5000)
)

bias_curve <- predict(fm1_dif, level = 0)[1:101] %>%
  unname()
sd_curve <- getVarCov(fm1_dif, type = "marginal")[[1]] %>%
  as.matrix() %>%
  diag() %>%
  sqrt()

mem_loa <-
  data.frame(time = tobs, bias = bias_curve, sd = sd_curve) %>%
  mutate(ul = bias + 2 * sd) %>%
  mutate(ll = bias - 2 * sd) %>%
  pivot_longer(cols = c("bias", "ul", "ll"), names_to = "var") %>%
  mutate(approach = "LMM")
```

95% fLoA using a LMM for Studies with Replicates (B-spline Basis Implementation)

```
library(tidyverse)
library(splines)
library(lme4)

inner_knots_fix <- seq(10,90, 10)
dat <- data.frame(dif_dat,
                 bs(dif_dat$frame, knots=inner_knots_fix))

lmer_model <-
lmer(dif ~ (X1 + X2 + X3 + X4 + X5 + X6 +
           X7 + X8 + X9 + X10 + X11 + X12) - 1 +
      (0 + ( X1 + X2 + X3 + X4 + X5 + X6 +
             X7 + X8 + X9 + X10 +
```

```
      X11 + X12) | sub/ses/rep),
  data = dat)

# mean curve
coef_fixed <- c("X1", "X2", "X3", "X4", "X5",
               "X6", "X7", "X8", "X9", "X10",
               "X11", "X12")

beta <- coef(summary(lmer_model))[coef_fixed, 1]
frame <- (0:100)
bias_curve <- bs(frame, knots = inner_knots_fix) %*% beta %>% c()

# standard deviation curve

Zmat <- bs(frame, knots = inner_knots_fix)
var_cov_mat <- VarCorr(lmer_model)
PSI_sub <- var_cov_mat$sub
PSI_ses <- var_cov_mat$`ses:sub`
PSI_rep <- var_cov_mat$`rep:(ses:sub)`
Imat <- diag(rep(1, 101))
Sigma <- Imat * sigma(lmer_model)^2

Vmat <- (Zmat %*% PSI_sub %*% t(Zmat)) +
        (Zmat %*% PSI_ses %*% t(Zmat)) +
        (Zmat %*% PSI_rep %*% t(Zmat)) +
        Sigma

sd_curve <- sqrt(diag(Vmat))

loa_dat <-
  data.frame(
    frame = frame,
    bias = bias_curve,
    ll = (bias_curve - 2 * sd_curve),
    ul = (bias_curve + 2 * sd_curve)
  ) %>%
```

```
pivot_longer(cols = c("bias", "ll", "ul"))
```

95% fLoA using a LMM for Studies with Replicates (Eigenbasis Implementation)

```
library(tidyverse)
library(nlme)
library(splines)
```

```
datMat <-
  dif_dat %>%
  unite("gr", sub, ses, rep) %>%
  dplyr::select(gr, frame, dif) %>%
  pivot_wider(names_from = frame, values_from = dif) %>%
  mutate(gr = NULL) %>%
  as.matrix()
```

```
SIGMA <- cov(datMat)
```

```
lambda <- eigen(SIGMA)$values
Z <- eigen(SIGMA)$vectors[, 1:12]
```

```
inner_knots_fix <- seq(10, 90, 10)
```

```
dat_eig <- data.frame(dif_dat, bs(dif_dat$frame,
knots = inner_knots_fix), Z)
```

```
fm1_dif <-
lme(dif ~ (X1 + X2 + X3 + X4 + X5 + X6 +
          X7 + X8 + X9 + X10 + X11 + X12) - 1,
    random =
    list(sub = pdDiag(~(X1.1 + X2.1 + X3.1 + X4.1 + X5.1 + X6.1 +
                        X7.1 + X8.1 + X9.1 + X10.1 + X11.1 +
                        X12.1) - 1 ),
         ses = pdDiag(~(X1.1 + X2.1 + X3.1 + X4.1 + X5.1 + X6.1 +
                        X7.1 + X8.1 + X9.1 + X10.1 + X11.1 +
```

```

                                X12.1) - 1 ),
rep = pdDiag(~(X1.1 + X2.1 + X3.1 + X4.1 + X5.1 + X6.1 +
              X7.1 + X8.1 + X9.1 + X10.1 + X11.1 +
              X12.1) - 1 )),
correlation = corARMA(form=~frame| sub/ses/rep, p=2, q=1),
data = dat_eig,
control = lmeControl(msMaxIter=500000))

# mean curve
coef_fixed <- c("X1", "X2", "X3", "X4",
               "X5", "X6", "X7", "X8", "X9",
               "X10", "X11", "X12")
beta <- coef(summary(fm1_dif))[coef_fixed, 1]
frame <- (0:100)
bias_curve <- bs(frame, knots = inner_knots_fix) %*% beta %>% c()

# standard deviation curve
coef_random <- c("X1.1", "X2.1", "X3.1", "X4.1",
                 "X5.1", "X6.1", "X7.1", "X8.1", "X9.1",
                 "X10.1", "X11.1", "X12.1")
Zmat <- Z

PSI_sub <- fm1_dif$modelStruct$reStruct$sub %>% as.matrix()
PSI_ses <- fm1_dif$modelStruct$reStruct$ses %>% as.matrix()
PSI_rep <- fm1_dif$modelStruct$reStruct$rep %>% as.matrix()
Imat <- diag(rep(1, 101))

Vmat <- (Zmat %*% PSI_sub %*% t(Zmat)) +
        (Zmat %*% PSI_ses %*% t(Zmat)) +
        (Zmat %*% PSI_rep %*% t(Zmat)) + Imat

sd_curve <- Vmat %>%
  diag() %>%
  sqrt() * sigma(fm1_dif) %>% c()
```

```
loa_dat <-  
  data.frame(  
    frame = frame,  
    bias = bias_curve,  
    ll = (bias_curve - 2 * sd_curve),  
    ul = (bias_curve + 2 * sd_curve)  
  ) %>%  
  pivot_longer(cols = c("bias", "ll", "ul"))
```

Bibliography

- J. H. Ahlberg, E. N. Nilson, and J. L. Walsh. *The theory of splines and their applications*. Academic Press, 1967.
- D. Altman and M. Bland. Measurement in medicine: the analysis of method comparison studies. *Journal of the royal statistical society: series D (The statistician)*, 32(3):307–317, 1983.
- J. J. Bartko. The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11, 1966.
- D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.
- M. Bland and D. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.
- M. Bland and D. Altman. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in biology and medicine*, 20(5):337–340, 1990.
- M. Bland and D. Altman. Comparing methods of measurement: why

- plotting difference against standard method is misleading. *The lancet*, 346(8982):1085–1087, 1995.
- M. Bland and D. Altman. Measuring agreement in method comparison studies. *Statistical methods in medical research*, 8(2):135–160, 1999.
- M. Borhani, A. H. McGregor, and A. M. Bull. An alternative technical marker set for the pelvis is more repeatable than the standard pelvic marker set. *Gait & posture*, 38(4):1032–1037, 2013.
- A. Cappozzo, F. Catani, U. Della Croce, and A. Leardini. Position and orientation in space of bones during movement: anatomical frame definition and determination. *Clinical biomechanics*, 10(4):171–178, 1995.
- A. Cappozzo, U. Della Croce, A. Leardini, and L. Chiari. Human movement analysis using stereophotogrammetry: Part 1: theoretical background. *Gait & posture*, 21(2):186–196, 2005.
- J. L. Carrasco and L. Jover. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, 59(4):849–858, 2003.
- B. Carstensen, J. Simpson, and L. C. Gurrin. Statistical models for assessing agreement in method comparison studies with replicate measurements. *The international journal of biostatistics*, 4(1), 2008.
- A. Cereatti, T. Bonci, M. Akbarshahi, K. Aminian, A. Barré, M. Begon, D. L. Benoit, C. Charbonnier, F. Dal Maso, S. Fantozzi, et al. Stan-

- standardization proposal of soft tissue artefact description for data sharing in human motion measurements. *Journal of biomechanics*, 62:5–13, 2017.
- E. Ceseracciu, Z. Sawacha, and C. Cobelli. Comparison of markerless and marker-based motion capture technologies through simultaneous data collection during gait: proof of concept. *PloS one*, 9(3):e87640, 2014.
- R. S. Chorba, D. J. Chorba, L. E. Bouillon, C. A. Overmyer, and J. A. Landis. Use of a functional movement screening tool to determine injury risk in female collegiate athletes. *North American journal of sports physical therapy: NAJSPT*, 5(2):47, 2010.
- W. G. Cochran. Errors of measurement in statistics. *Technometrics*, 10(4):637–666, 1968.
- S. L. Colyer, M. Evans, D. P. Cosker, and A. I. Salo. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports medicine-open*, 4(1):1–15, 2018.
- L. Coyne. *Evaluation of a markerless motion capture system in a Premier League Football Academy*. PhD thesis, National University of Ireland, Galway, 2021.
- A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

- C. de Boor. *A practical guide to splines*. springer-verlag New York, 2001.
- P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger. *Analysis of longitudinal data*. Oxford university press, 2002.
- O. A. Donoghue, A. J. Harrison, N. Coffey, and K. Hayes. Functional data analysis of running kinematics in chronic achilles tendon injury. *Medicine and science in sports and exercise*, 40(7):1323–1335, 2008.
- T. W. Du Clos. Function of c-reactive protein. *Annals of medicine*, 32(4):274–278, 2000.
- E. Fehrenbach and M. E. Schneider. Trauma-induced systemic inflammatory response versus exercise-induced immunomodulatory effects. *Sports medicine*, 36(5):373–384, 2006.
- K. R. Ford, G. D. Myer, and T. E. Hewett. Reliability of landing 3d motion analysis: implications for longitudinal analyses. *Medicine and science in sports and exercise*, 39(11):2021–2028, 2007.
- D. A. Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.
- W. M. Glasoe, F. A. Pena, and V. Phadke. Cardan angle rotation sequence effects on first-metatarsophalangeal joint kinematics: implications for measuring hallux valgus deformity. *Journal of foot and ankle research*, 7(1):1–5, 2014.
- S. Harsted, A. Holsgaard-Larsen, L. Hestbæk, E. Boyle, and H. H. Lauridsen. Concurrent validity of lower extremity kinematics and jump

- characteristics captured in pre-school children by a markerless 3d motion capture system. *Chiropractic & manual therapies*, 27(1):1–16, 2019.
- W. G. Hopkins. Measures of reliability in sports medicine and science. *Sports medicine*, 30(1):1–15, 2000.
- T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.
- J. A. Hunter Bennett, K. Norton, and K. Davison. Are we really “screening” movement? the role of assessing movement quality in exercise settings. *Journal of sport and health science*, 9(6):489, 2020.
- I. Ispirlidis, I. G. Fatouros, A. Z. Jamurtas, M. G. Nikolaidis, I. Michailidis, I. Douroudos, K. Margonis, A. Chatzinikolaou, E. Kalistratos, I. Katrabasas, et al. Time-course of changes in inflammatory and performance responses following a soccer game. *Clinical journal of sport medicine*, 18(5):423–431, 2008.
- C. L. Kimberlin and A. G. Winterstein. Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy*, 65(23):2276–2284, 2008.
- P. Komdeur, F. E. Pollo, and R. W. Jackson. Dynamic knee motion in anterior cruciate impairment: a report and case study. In *Baylor Uni-*

- versity Medical Center Proceedings*, volume 15, pages 257–259. Taylor & Francis, 2002.
- N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.
- I. Lawrence and K. Lin. Assay validation using the concordance correlation coefficient. *Biometrics*, pages 599–604, 1992.
- J. Lee, D. Koh, and C. Ong. Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Computers in biology and medicine*, 19(1):61–70, 1989.
- T. Lencioni, I. Carpinella, M. Rabuffetti, A. Marzegan, and M. Ferrarin. Human kinematic, kinetic and emg data during different walking and stair ascending and descending tasks. *Scientific data*, 6(1):1–10, 2019.
- L. Li, L. Zeng, Z.-J. Lin, M. Cazzell, and H. Liu. Tutorial on use of intraclass correlation coefficients for assessing intertest reliability and its application in functional near-infrared spectroscopy-based brain imaging. *Journal of biomedical optics*, 20(5):050801, 2015.
- L. I.-K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.
- R. S. Lloyd, J. L. Oliver, J. M. Radnor, B. C. Rhodes, A. D. Faigenbaum, and G. D. Myer. Relationships between functional movement screen scores, maturation and physical performance in young soccer players. *Journal of sports sciences*, 33(1):11–19, 2015.

- R. S. Lloyd, J. B. Cronin, A. D. Faigenbaum, G. G. Haff, R. Howard, W. J. Kraemer, L. J. Micheli, G. D. Myer, and J. L. Oliver. National strength and conditioning association position statement on long-term athletic development. *Journal of strength and conditioning research*, 30(6):1491–1509, 2016.
- M. L. Mackenzie, C. Donovan, and B. McArdle. Regression spline mixed models: A forestry example. *Journal of agricultural, biological, and environmental statistics*, 10(4):394–410, 2005.
- J. L. Markström, L. Schelin, and C. K. Häger. A novel standardised side hop test reliably evaluates landing mechanics for anterior cruciate ligament reconstructed persons and controls. *Sports biomechanics*, 2018.
- J. L. McGinley, R. Baker, R. Wolfe, and M. E. Morris. The reliability of three-dimensional kinematic gait measurements: a systematic review. *Gait & posture*, 29(3):360–369, 2009.
- S. G. McLean, K. Walker, K. Ford, G. Myer, T. Hewett, and A. J. van den Bogert. Evaluation of a two dimensional analysis method as a screening and evaluation tool for anterior cruciate ligament injury. *British journal of sports medicine*, 39(6):355–362, 2005.
- G. Micula and S. Micula. *Handbook of splines*. Springer Science & Business Media, 1999.
- E. Miller, K. Kaufman, T. Kingsbury, E. Wolf, J. Wilken, and M. Wy-

- att. Mechanical testing for three-dimensional motion analysis reliability. *Gait & posture*, 50:116–119, 2016.
- C. E. Milner, C. G. Westlake, and J. J. Tate. Test–retest reliability of knee biomechanics during stop jump landings. *Journal of biomechanics*, 44(9):1814–1816, 2011.
- J. Newell, T. Aitchison, and S. Grant. *Statistics for sports and exercise science: a practical approach*. Routledge, 2014.
- T. d. P. Oliveira, J. Hinde, and S. S. Zocchi. Longitudinal concordance correlation function based on variance components: an application in fruit color analysis. *Journal of agricultural, biological and environmental statistics*, 23(2):233–254, 2018.
- E. Olsen, T. Pfau, and C. Ritz. Functional limits of agreement applied as a novel method comparison tool for accuracy and precision of inertial measurement unit derived displacement of the distal limb in horses. *Journal of biomechanics*, 46(13):2320–2325, 2013.
- T. C. Pataky, M. A. Robinson, and J. Vanrenterghem. Vector field statistical analysis of kinematic and force trajectories. *Journal of biomechanics*, 46(14):2394–2401, 2013.
- M. A. Perrott, T. Pizzari, J. Cook, and J. A. McClelland. Comparison of lower limb and trunk kinematics between markerless and marker-based motion capture systems. *Gait & posture*, 52:57–61, 2017.

- J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS*. Springer science & business media, 2000.
- A. Pini, J. L. Markström, and L. Schelin. Test–retest reliability measures for curve data: An overview with recommendations and supplementary code. *Sports biomechanics*, pages 1–22, 2019.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL <https://www.R-project.org/>.
- K. Rácz, Z. Palya, M. Takács, G. Nagymáté, and R. M. Kiss. Evaluation of anatomical landmark calibration accuracy of a motion capture based analysis protocol. *Materials today: proceedings*, 5(13):26538–26543, 2018.
- J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2002.
- J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2005.
- J. A. Rice and C. O. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57(1):253–259, 2001.
- C. Richter, B. Marshall, K. Moran, et al. Comparison of discrete-point vs. dimensionality-reduction techniques for describing performance-related aspects of maximal vertical jumping. *Journal of biomechanics*, 47(12):3012–3017, 2014.

- D. G. E. Robertson, G. E. Caldwell, J. Hamill, G. Kamen, and S. Whittlesey. *Research methods in biomechanics*. Human kinetics, 2013.
- J. Røislien, L. Rennie, and I. Skaaret. Functional limits of agreement: a method for assessing agreement between measurements of gait curves. *Gait & posture*, 36(3):495–499, 2012.
- A. Roy. An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of biopharmaceutical statistics*, 19(1):150–173, 2009.
- D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*. Cambridge university press, 2003.
- D. Ryan, C. Lewin, S. Forsythe, and A. McCall. Developing world-class soccer players: An example of the academy physical development program from an english premier league team. *Strength & conditioning journal*, 40(3):2–11, 2018.
- S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*. John Wiley & Sons, 2006.
- G. A. Seber and A. J. Lee. *Linear regression analysis*. John Wiley & Sons, 2012.
- P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.
- A. Souglis, G. Bogdanis, I. Giannopoulou, C. Papadopoulos, and N. Apostolidis. Comparison of inflammatory responses and muscle

- damage indices following a soccer, basketball, volleyball and handball game at an elite competitive level. *Research in sports medicine*, 23(1): 59–72, 2015.
- G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer, New York, 2000.
- P. Watson and A. Petrie. Method agreement analysis: a review of correct methodology. *Theriogenology*, 73(9):1167–1179, 2010.
- J. P. Weir. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *The Journal of strength & conditioning research*, 19(1):231–240, 2005.
- G. Wu, S. Siegler, P. Allard, C. Kirtley, A. Leardini, D. Rosenbaum, M. Whittle, D. D D’Lima, L. Cristofolini, H. Witte, et al. Isb recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part i: ankle, hip, and spine. *Journal of biomechanics*, 35(4):543–548, 2002.
- S. X. Yang, M. S. Christiansen, P. K. Larsen, T. Alkjær, T. B. Moeslund, E. B. Simonsen, and N. Lynnerup. Markerless motion capture systems for tracking of persons in forensic biomechanics: an overview. *Computer methods in biomechanics and biomedical engineering: imaging & Visualization*, 2(1):46–65, 2014.
- M. Żuk and C. Pezowicz. Kinematic analysis of a six-degrees-of-freedom model based on isb recommendation: a repeatability analysis and com-

parison with conventional gait model. *Applied bionics and biomechanics*, 2015, 2015.