THE UNIVERSITY
of EDINBURGH

Evidence evaluation and functional data analysis

Ya-Ting Chang

Doctor of Philosophy
University of Edinburgh
2019

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Ya-Ting Chang*)

Abstract

In forensic contexts, evidence gathered are valuable in suggesting possible crimes. The problems forensic scientists are often interested in are whether evidence found at the scene of the crime match those that are found related to some suspect. Prosecution and defense propositions are often put forward assuming the evidence came from the same source or they are not from the same source. A popular and objective measure of the value of evidence is the use of the likelihood ratios that is calculated as the ratio between the probabilities of observing the evidence given each proposition.

In this thesis, we will provide methodologies for the evaluation of the likelihood ratio when evidence are characterised by functional data such as mass spectrophotometry data. Three models will be developed based on fundamental functional data analysis and use of systems of basis functions for the decomposition of means. Each of the three models considers a different covariance structure for between- and within-group variations. They are independent and constant variances across groups, independent and constant within-group variances and auto-covariance. Two models that only make use of the data after dimension reduction are also developed. One is multivariate normal random-effects model with constant covariance matrix and the other one puts an inverse Wishart prior distribution on within-group covariance matrix. Both models consider two levels of variability, within- and between-group, for the mean.

All models will be used to calculate likelihood ratios for three sets of data and results will be compared using different measures of performances such of rates of misleading evidence, Tippett plots and empirical cross-entropy (ECE) plots. Sensitivity analysis is then done to test the effect of using different estimations of the hyperparameters on likelihood ratios. Furthermore, we also preprocessed the data in another way, that is taking first order differences and replace the original data to feed into the

models. Conclusions will be drawn based on the performances of each model on each dataset, including sensitivity analysis and more data preprocessing. Finally, guidances on how to choose the model for the calculation of likelihood ratios for other kinds of data will be provided.

Lay summary

It is forensic experts' job to comment on evidence in a court case. When a crime is committed and evidence gathered, forensic experts can use their expertise that comes from experiences to suggest how likely there is a connection between a suspect and the crime based on the evidence gathered. To do so, comparisons are to be made between two competing hypothesis; they are prosecution and defense or alternative hypotheses. Likelihood ratios can then be calculated as an objective measure of the strength of evidences in support of the prosecution hypothesis over the alternative hypothesis. It is the ratio between the probabilities of observing the evidence given the two hypotheses.

There are many ways of calculating likelihood ratios, all of which require modelling of the measurements obtained from the evidence and some databases with a collection of measurements of the same type of evidence from some relevant population. Methods for obtaining likelihood ratio were first developed for univariate continuous measurements but as there are more and more types of data becoming available, new ways of calculating likelihood ratios need to be developed. We will be focusing on analysing microspectrophotometry data that are functional data. They are multivariate and can be seen as observations of a smooth underlying function over a range of values. This is different from multivariate data as there are hidden structures between the points that makes them highly correlated. In the past, this types of data was compared visually or evaluated after many transformations. However, we are able to develop models that take into account all variabilities that are essential for distinguishing between evidence in one probabilistic model and produce likelihood ratios that are useful for the purpose of evaluating evidences. After all, our methodologies can be readily applied on different kinds of data with slight modifications.

Acknowledgements

First I would like to thank my supervisor, Professor Colin Aitken, for his constant support and guidance throughout my career as a PhD student. He has always been supportive of my ideas and positive regardless of the results we obtained. He always encouraged me to present our work at various conferences and workshops, this has been very beneficial for me. Moreover, he is always there when I need help. Overall, we had a wonderful time working together and I very much appreciate the opportunity to be his student for working with an amazing researcher on an interesting research project.

I would also like to thank Dr Ioannis Papastathopoulos for his guidance and valuable opinions on the writing of the thesis. My second supervisor Dr Amy Wilson also provided useful comments on many technical details used in this project. She has been of great support throughout my time as a student here.

Many thanks to Dr Grzegorz Zadora and Dr Patrick Buzzini for providing the data that are essential for this project and the University of Edinburgh for providing me scholarships to make studying here possible and easier.

More importantly, I would like to thank fellow PhD students, colleagues in the department and people I met during conferences and workshops for being there to have all the fun together and my best friends Jiaqi, Johan, Dave, Zhaoxun, Chunxiao and Jiawei for the amazing and wonderful time we share as PhD students. Lastly, I would also like to thank my parents and family members for all the support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to evidence evaluation

In forensic context, a problem of interest is to compare trace evidence, i.e., evidence that can possibly be used to suggest a crime. For example, blood stains, glass fragments, or gunshot residue found at the scene of a crime.

In cases where the source of evidence is of interest, or whether trace evidence found at different places suggest a connection, comparisons are made by assuming two competing propositions related to the origin of the evidence. For example, in a document examination problem where a document is suspected of being altered, the question of interest is whether it was produced by a suspect's pen, or ink. In such a problem, the prosecution proposition is called H_p and suggests possible connection and the defence proposition is called H_d and suggests otherwise. When comparing trace evidence, the one that is produced by the suspect is typically of known origin (the suspect) and the other one is of unknown origin. Here and throughout we call them the control and recovered evidence, respectively. The comparison between trace evidence will be based on these two propositions.

In order to compare H_p and H_d, a set of measurements has to be obtained for both control and recovered evidence. These measurements are used to represent features of evidence and are typically used to discriminate evidence for forensic purposes. The trace evidence are either discretely characterised, for example DNA profile of blood stain, or continuously characterised, for example measurements of refractive indices

and elemental concentrations of glass fragments.

To aid fact-finders in making decisions, likelihood ratios (Section 2.6) are widely accepted as an objective measure of the strength of the evidence in support of the proposition H_p over H_d. It is the ratio of the probabilities of observing the evidence given H_p and H_d, respectively.

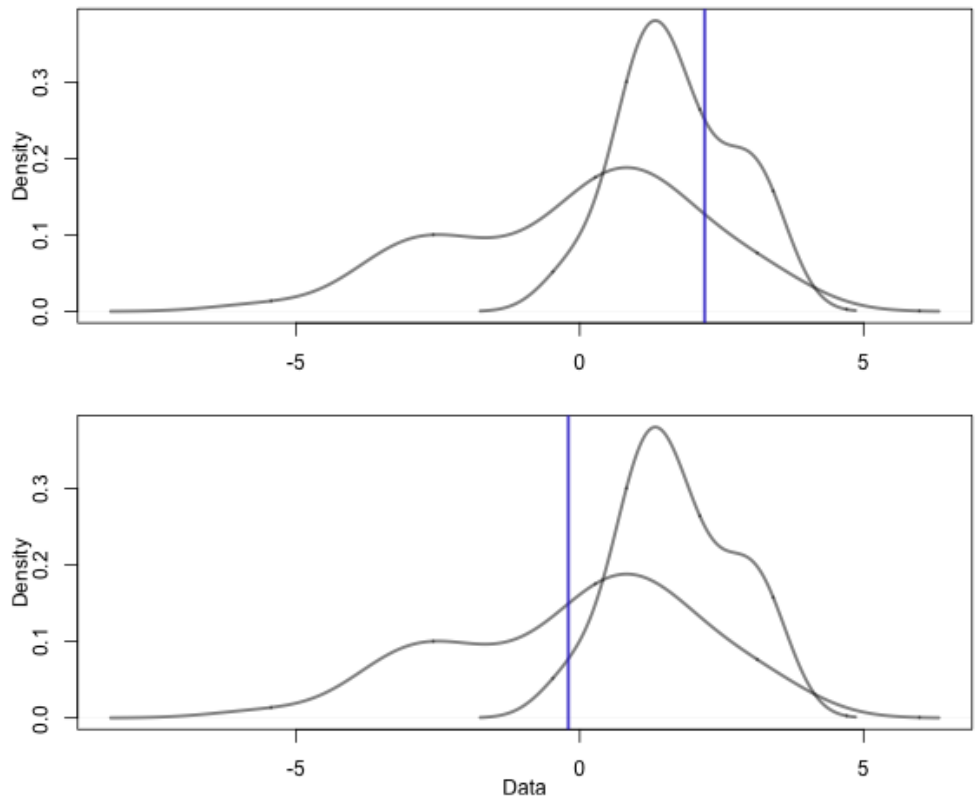


Figure 1.1: The evaluation of likelihood ratios for two scenarios under the two propositions H_p and H_d. The two propositions give two distinct distributions as illustrated by the densities drawn in both panels. The density under H_p typically has a smaller variation; hence is represented by the one towards the right. The purple lines represent some statistics of control and recovered evidence observed. Likelihood ratios are evaluated as the ratio between the two intersecting points with the numerator being the one intersecting with the density given H_p. The top panel gives a likelihood ratio that is greater than one, which supports H_p over H_d and the bottom panel gives a likelihood ratio that is less than one.

The evaluation of such probabilities requires knowledge of the distribution of the evidence given each proposition as shown in Figure 1.1, which requires the choice of a

relevant population. A database of relevant population is a collection of the same type of evidence that can be used for a given case. As an example, consider a murder case that happened in Central England where the suspect had left a bloodstain, the relevant population would not be the same as a murder case that happened in Spain regarding the ethnic composition of the general population in each region. Given common crimes and databases constructed, the use of likelihood ratio for the evaluation of evidence is well developed for evidence that are characterised by either discrete or continuous data (Aitken and Taroni, 2005). However, the use of likelihood ratio for evaluating evidence that are characterised by functional data where one or more variables is defined over a continuum, has not been well developed. Our primary aims are to develop statistical models that can be used to evaluate evidence that are characterised by functional data by building up existing methodologies used to analyse multivariate data and evidence evaluation with applications to ink and fibre data.

1.2 Ink and fibre data

Forensic ink examination have been performed for decades in aid of investigating forged documents. Although ink as an evidence does not have as big an impact in detecting and convicting crimes as other evidence such as fingerprint or shoe prints, it was used in many high profile cases since the U.S. Secret Service created the International Ink Library (Burfield et al., 2015). Burfield et al. (2015) also reviewed the use of functional data in characterising, comparing and classifying chemical data. It is found that functional data analysis is a powerful technique which enables to control the dimensionality and smoothness of a functional dataset but the implementation is complex when compared to multivariate analysis. While their work was on analysing ink chromatograms, we are interested in analysing microspectrophotometry data that can be obtained without destructing the evidence. Microspectrophotometry (MSP) data is the measure of colour which has many applications in forensic science as virtually everything has a colour. It is especially useful when we want to differentiate colours that look indifferent to the naked eyes.

Fibres are probably the most common form of evidence. They are used as evidence in a variety of cases (Frank and Sobol, 1990). Since the 1970's MSP has been used in

forensic science as an objective method for providing reproducible and discriminating analysis of the colour of a single fibre (Was-Gubala and Starczak, 2015).

Microspectrophotometry data of ink and fibre are the motivating examples of functional data our developed models are based on. To show their distinctive properties, a sample selected from each dataset is drawn. In Figure 1.2, some replicates of the same type are drawn as curves in the same panel for ink, wool and cotton separately.

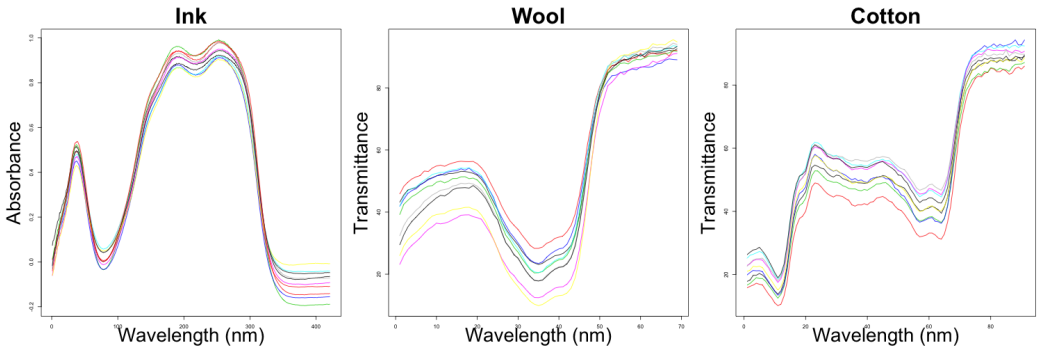


Figure 1.2: Three plots each showing multiple observations of microspectrophotometry (MSP) data by connecting m points $\{(w_j, y_{kij}), j = 1, \ldots, m\}$ of a type of ink, red woollen and red cotton.

Even though these are all microspectrophotometry data, they can look different for different materials. Microspectrophotometry data are useful in differentiating between colours that look indifferent to naked eyes so they can be used in forensic ink and fibre examinations.

By the visual representation of (functional) data of our interest, we will use the word $curve$ to indicate a set of pairs of observations that form a single unit of the data. Five models are introduced for the evaluation of evidence that are characterised by functional data. Three of which are proposed and two were pre-existing. The newly proposed models have an additive structure which assumes that each curve is mainly composed of an overall shape and errors at each point of observation. The shape is assumed to be representable by a linear combination of some basis functions and the errors follow certain distributions according to different assumptions for each model. They are constant and independent for all groups, constant and independent within groups, and constant and autocorrelated within groups. This method was unprece-

dented as it requires probabilistic models that can be complicated to evaluate for complex (functional) data. It is made possible through writing the overall shape as a linear combination of basis functions that serves as a medium for dimension reduction. However, it has the advantage of direct evaluation of likelihood ratios from the original data. For forensic ink or fibre evaluation this has never been tried before although methods for the evaluation of score-based likelihood ratios for ink data were proposed by Martyna et al. (2013). The pre-existing models are multivariate normal random-effects model (Aitken and Lucy, 2004) and the consideration of within-group variation on top of that (Bozza et al., 2008) but they had applications in the elemental composition of glass and handwriting evidence, respectively. To evaluate evidence using these models, we have to use common techniques of multivariate data analysis to first transform our data before a lower dimensional representation can be used in place of the original functional data of our interest.

Next, three datasets will be used to assess the performance of these models after basis selection and checking of assumptions. They are blue inks and red wool and cotton fibre data. Each one represents functional data with certain characteristics which will be introduced in Chapter 4. Tables and plots will be displayed to demonstrate the advantages and disadvantages of each model for different data. Discussion and conclusion will be drawn based on these results.

1.3 Chapter summary

Chapter 2 is mainly composed of two parts. Sections 2.2 to 2.5 and appendix D.1 introduce notation and methodologies that are fundamental for the understanding of the thesis. Section 2.6 introduces likelihood ratio for comparison problem in forensic context. Section 2.6.1 summarises the use of likelihood ratio in the evaluation of continuous data both univariate and multivariate, and Section 2.6.3 introduces score-based likelihood ratios. Section 2.6.4 gives references to evidence evaluation on ink and woollen and cotton fibre data which we use to evaluate the performances of our models and finally Section 2.6.5 introduces evidence evaluation for evidence characterised by functional data.

Chapter 3 introduces five models for the evaluation of likelihood ratio for evidence

characterised by functional data; three of them component-wise, that is, treating the curves as mainly composed of a function and some error for each measurement, and two of them dimension reduced, that is, only considering a representation of the original curves. These models are applicable to all functional data. They are introduced independently of data we use to assess the performance. Based on models specified in this chapter, the methods for evaluation with likelihood ratios are presented along with how estimates can be obtained from training data (relevant population).

Chapter 4 introduces three sets of data, each of them will be described in detail and various types of plots will be displayed for the ease of understanding. For the development and assessment of models to be used for likelihood ratio calculation, properties of these data will be examined and appropriate numbers of basis functions will be chosen using bases and criteria discussed in Section 2.3.

Chapter 5 examines model assumptions for each dataset before the models are used to evaluate likelihood ratios in Chapter 6. After fitting the models to each dataset, more data are simulated by assuming they are generated using the same procedure as our proposed models and the simulated data are compared with the original data as another way to assess the models' fit.

Chapter 6 contains likelihood ratios calculated for data introduced in Chapter 4 using models described in Chapter 3 and summarises in tables and plots to assess and compare the performance of each model on each dataset. For each dataset in Chapter 4 log likelihood ratios ($lLRs$) are calculated by splitting up the data into training and testing sets for estimation of model parameters and likelihood ratio calculation, respectively. For the purpose of performance evaluation, $lLRs$ are calculated for possible pairs of evidence so we are able to obtain a massive number of $lLRs$. Tables with values that summarise these log likelihood ratios including average log likelihood ratios and rates of misleading evidence are presented for different set-ups, i.e., different sizes of intervals (int) and number (n_s) of curves within a set ($\boldsymbol{Y}_c$ or $\boldsymbol{Y}_r$) to be used in one comparison. Tippett plots and empirical cross entropies are also shown and compared across models and datasets.

Chapter 7 includes sensitivity analysis of likelihood ratios obtained using different estimates of parameters when evaluating under a selection of models. A selection of two sets of curves will also be drawn to illustrate the cases that are failed to be

distinguished by lLR's under each model as a way to show the models' limitations.

Chapter 8 contains more results, for when data is pre-processed before feeding into the models. The process chosen is taking differences. It is done in order to to eliminate certain characteristics present within the original data that might contribute to some of the difficulties that some models introduced in Chapter 3 encounter while trying to distinguish evidence through the calculation of likelihood ratios. The likelihood ratios calculated using the processed data is presented in the same way as in Chapter 6 and compared with those in Chapter 6.

Chapter 9 summarises results presented so far and provides a list of future research directions including a guideline on selecting between proposed models for use on a new dataset.

Parts of this thesis are summarised in Aitken et al. (2019), they will be mentioned where appropriate.

Chapter 2

Background

2.1 Introduction

This chapter is mainly composed of two parts. Sections 2.2 to 2.5 and appendix D.1 introduce notations that are fundamental for the understanding of this thesis and methodologies that are used throughout. The rest of the chapter provides background and existing methodologies our work is built upon.

The following sections are organised so that Section 2.6 introduces likelihood ratios for comparison problems in a forensic context. Section 2.6.1 summarises the use of likelihood ratios in the evaluation of univariate continuous data and Section 2.6.2 extends the idea to multivariate continuous data, and Section 2.6.3 introduces score-based likelihood ratios. Section 2.6.4 gives references to evidence evaluation for ink and woollen and cotton fibre data which are the motivating examples of data of our interest and used for evaluating the performances of our models and finally, Section 2.6.5 introduces evidence evaluation for evidence characterised by functional data.

2.2 Some notation

Lower and upper case letters are used to denote a scalar or function with $p(\cdot)$, $f(\cdot)$ and $\pi(\cdot)$ commonly associated with probability density functions. The word sequence is used throughout to denote an indexed set, so that, for instance, a sequence $\{y_i\}$ of real numbers is a real-valued function on a certain index set $\{i\}$. If $\{y_i\}$ is a se-

quence of objects, each y_i is called an element of the sequence. Boldface lower case letters or numbers such as $\boldsymbol{v} \in \mathbb{R}^m$ are used to denote an m-dimensional column vector $(v_1, \ldots, v_m)^T$ and boldface upper case letters such as $\boldsymbol{M} \in \mathbb{R}^{m \times n}$ are used to denote an m by n matrix, which we sometimes write as $[\boldsymbol{v}_1\ \boldsymbol{v}_2 \ldots \boldsymbol{v}_{n-1}\ \boldsymbol{v}_n]$, i.e., a concatenation of n m-dimensional vectors. For a matrix $\boldsymbol{M}$ that can be written as a concatenation of vectors $\boldsymbol{v}_i \in \mathbb{R}^m$ for $i \in \{1 \ldots n\}$, $\boldsymbol{M}_{i,j}$ will be used to denote the $(i,j) - th$ entry of $\boldsymbol{M}$ or the $i - th$ element of $\boldsymbol{v}_j$.

By convention, arithmetic operations on vectors are applied component-wise. For example, if $f : \mathbb{R} \mapsto \mathbb{R}$, $\boldsymbol{v} \in \mathbb{R}^n$, then $f(\boldsymbol{v}) = (f(v_1), f(v_2), \ldots, f(v_n))^T$. Given functions f and g, the inner product $\langle f, g \rangle$ is defined as $\langle f, g \rangle = \int f(t)g(t)dt$.

2.3 Dimension reduction

Dimension reduction, or transformation of data of interest into a smaller space is a common technique when dealing with multivariate data primarily because it is easier to work with data in lower dimensions.

Dimension reduction can be achieved in two ways, either by feature selection or feature extraction. Feature selection relates to identifying subsets of important variables by some measure of predictive performance whereas in feature extraction raw variables are projected onto a lower dimensional Euclidean space or manifold in general. Feature selection constitutes an important aspect of the model building process that can be used to further improve performance or could emanate as a natural method for multivariate data analysis in many applications. However we will only consider feature extraction in this thesis.

2.3.1 Principal component analysis

Principal component analysis is a dimension reduction technique that can be used to approximate the data by introducing systems of eigenvectors that point towards directions with the largest variances.

Given a set of data $\boldsymbol{Y}$=$\{\boldsymbol{y}_j \in \mathbb{R}^m\colon\ j = 1, \ldots, n\}$, an n-by-B matrix $\boldsymbol{\Theta}$=$\boldsymbol{Y}\boldsymbol{V}$ consisting of the original data $\boldsymbol{Y}$ projected onto a new coordinate system that represents

directions with the maximum variances, can be obtained. This is based on the fact that any n-by-m matrix $\boldsymbol{Y}$ can be written as $\boldsymbol{UDV}^T$, where $\boldsymbol{U}$ is an n-by-n matrix consisting of left singular vectors of $\boldsymbol{Y}$, $\boldsymbol{D}$ is a n-by-m diagonal matrix consisting of singular values of $\boldsymbol{Y}$ and $\boldsymbol{V}$ is an m-by-m matrix consisting of right singular vectors of $\boldsymbol{Y}$. A n-by-m diagonal (rectangular) matrix $\boldsymbol{D}$ has nonzero elements only at its diagonal entries $\boldsymbol{D}_{i,i}$ for all $1 \leq i \leq \min\{m, n\}$. Multiplying $\mathbf{V}$ from the right in both left and right hand side of $\boldsymbol{Y} = \boldsymbol{UDV}^T$ gives the scores $\boldsymbol{\Theta} = \boldsymbol{YV} = \boldsymbol{UD}$.

The matrix $\boldsymbol{V} = [\boldsymbol{v}_1 \boldsymbol{v}_2 \dots \boldsymbol{v}_B] \in \mathbb{R}^{m \times B}$ is obtained as follows. First,

$$\begin{aligned}\boldsymbol{v}_1 &= \underset{||\boldsymbol{v}||=1}{\operatorname{argmax}} \sum_{i=1}^{n} |\boldsymbol{y}_i \cdot \boldsymbol{v}|^2 = \underset{||\boldsymbol{v}||=1}{\operatorname{argmax}} ||\boldsymbol{Yv}||^2 = \underset{\boldsymbol{v}}{\operatorname{argmax}} \frac{\boldsymbol{v}^T \boldsymbol{Y}^T \boldsymbol{Yv}}{\boldsymbol{v}^T \boldsymbol{v}} \\ &= \underset{\boldsymbol{v}}{\operatorname{argmax}} \frac{\boldsymbol{v}^T \boldsymbol{Sv}}{\boldsymbol{v}^T \boldsymbol{v}}.\end{aligned}$$

Since $\boldsymbol{S} = \boldsymbol{Y}^T \boldsymbol{Y}$ is positive semidefinite, it has a spectral decomposition $\boldsymbol{S} = \sum_{j=1}^{m} \lambda_j \boldsymbol{e}_j \boldsymbol{e}_j^T$ where $\lambda_j \geq 0 \, \forall j$, $||\boldsymbol{e}_j||$=1 and $\boldsymbol{e}_s^T \boldsymbol{e}_t = 0 \, \forall s \neq t$ so it follows that $\boldsymbol{v}_1 = \boldsymbol{e}_1$, the eigenvector that corresponds to the maximum eigenvalue λ_1. For $k > 1$, let $\tilde{\mathbf{Y}}$ be the projection of $\boldsymbol{Y}$ onto the remaining subspace, i.e., $\tilde{\boldsymbol{Y}}^{(k)} = \boldsymbol{Y} - \sum_{j=1}^{k-1} \boldsymbol{Y} \boldsymbol{v}_j \boldsymbol{v}_j^T = \boldsymbol{Y}(\mathbf{I} - \sum_{j=1}^{k-1} \boldsymbol{v}_j \boldsymbol{v}_j^T)$, then $\boldsymbol{v}_k$ is given by

$$\begin{aligned}\boldsymbol{v}_k &= \underset{||\boldsymbol{v}||=1}{\operatorname{argmax}} \sum_{i=1}^{n} |\tilde{\boldsymbol{Y}}_i^{(k)} \cdot \boldsymbol{v}|^2 = \underset{||\boldsymbol{v}||=1}{\operatorname{argmax}} ||\tilde{\boldsymbol{Y}}^{(k)} \boldsymbol{v}||^2 = \underset{\boldsymbol{v}}{\operatorname{argmax}} \frac{\boldsymbol{v}^T \tilde{\boldsymbol{Y}}^{(k)T} \tilde{\boldsymbol{Y}}^{(k)} \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}} \\ &= \underset{\boldsymbol{v}}{\operatorname{argmax}} \frac{\boldsymbol{v}^T (\mathbf{I} - \sum_{j=1}^{k-1} \boldsymbol{v}_j \boldsymbol{v}_j^T)^T \boldsymbol{Y}^T \boldsymbol{Y} (\boldsymbol{I} - \sum_{j=1}^{k-1} \boldsymbol{v}_j \boldsymbol{v}_j^T) \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}} \\ &= \underset{\boldsymbol{v}}{\operatorname{argmax}} \frac{\boldsymbol{v}^T (\boldsymbol{I} - \sum_{j=1}^{k-1} \boldsymbol{v}_j \boldsymbol{v}_j^T)^T \boldsymbol{Y}^T \boldsymbol{Y} (\boldsymbol{I} - \sum_{j=1}^{k-1} \boldsymbol{v}_j \boldsymbol{v}_j^T) \boldsymbol{v}}{\boldsymbol{v}^T (\boldsymbol{I} - \sum_{j=1}^{k-1} \boldsymbol{v}_j \boldsymbol{v}_j^T)^T (\boldsymbol{I} - \sum_{j=1}^{k-1} \boldsymbol{v}_j \mathbf{v}_j^T) \mathbf{v}} = \underset{\boldsymbol{v}}{\operatorname{argmax}} \frac{\tilde{\boldsymbol{v}}^T \boldsymbol{S} \tilde{\boldsymbol{v}}}{\tilde{\boldsymbol{v}}^T \tilde{\boldsymbol{v}}} = \boldsymbol{e}_k.\end{aligned}$$

Depending on whether $n > m$, a maximum of $\min\{m, n\}$ of these orthonormal vectors can be obtained, hence $B \leq \min\{m, n\}$. When $B = m$, $\boldsymbol{Y} = \boldsymbol{UDV}^T$ and $B < m$, an estimate for $\boldsymbol{Y}$ can be obtained by retaining selected (first B) columns of $\boldsymbol{D}$ so $\tilde{\boldsymbol{Y}}_B = \boldsymbol{UD}_B \boldsymbol{V}^T$. The scores are ordered by decreasing variance so by selecting only a subset of principal components, dimension reduction is achieved.

Using a two dimensional dataset $\boldsymbol{Y}$ as an example, the left panel in Figure 2.1 shows centered original data and two directions with the greatest variances indicated

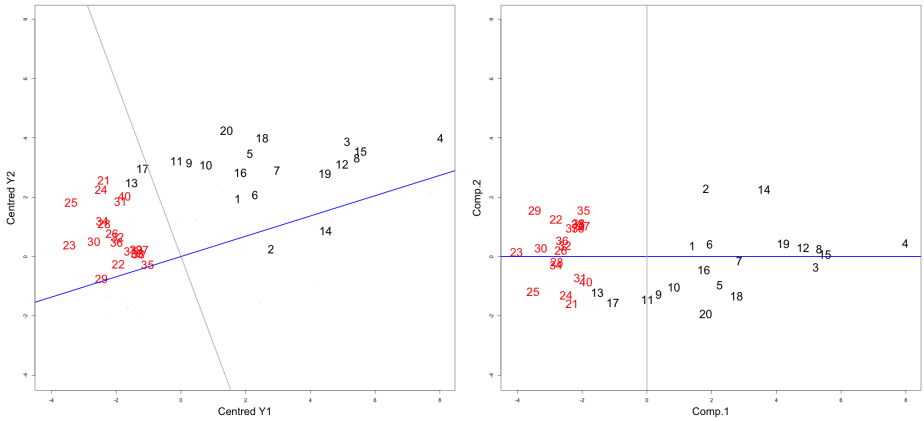


Figure 2.1: Example of principal component analysis on bivariate data $\boldsymbol{Y}$. For illustration purposes, this example shows transformation instead of dimension reduction. The bivariate data $\boldsymbol{Y}$ is constructed by putting two sets (clusters) of bivariate data together. They were generated using multivariate normal distribution with different means and covariance matrices and drawn as red and black numbers. The plot on the left shows the two dimensions of the data plotted against each other after subtracting their means, or centering. The plot on the right shows the first two principal components plotted against each other after the transformation. The blue and grey lines (in both plots) represent directions with the greatest and second greatest variations.

by blue and grey lines. The two lines are perpendicular. The right panel shows the transformed data, or principal components scores $\boldsymbol{\Theta}$.

2.3.2 Systems of basis functions

Any vector $\boldsymbol{y} \in \mathbb{R}^n$, for $n \in \mathbb{N}$, can be written as a linear combination of at most n linearly independent vectors $\mathbf{e}_i \in \mathbb{R}^n$, i.e., the vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ satisfy $c_1\boldsymbol{e}_1 + c_2\boldsymbol{e}_2 + \cdots + c_n\boldsymbol{e}_n = \mathbf{0}$ if and only if $\boldsymbol{c} = \mathbf{0}$. The set of vectors $\{\boldsymbol{e}_i\}_{i=1}^n$ forms a basis of $\mathbb{R}^n$, in the sense that the span of $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ is $\mathbb{R}^n$. A system of basis functions is a set of known functions $\{\phi_b : \mathbb{R} \mapsto \mathbb{R},\ b \in \{1, 2, \ldots, \infty\}\}$ whose span is equal to the space of functions $\mathbf{H}$ such that for any function $g : \mathbb{R} \mapsto \mathbb{R}$, there exists a sequence of scalars $\{\theta_b\}$ such that for all $x \in \mathbb{R}$, $g(x) = \sum_{b=1}^{\infty} \theta_b \phi_b(x)$. Therefore, in contrast to a discrete vector space of $\mathbb{R}^n$ where n linearly independent vectors $\{\boldsymbol{v}_i \in \mathbb{R}^n : 1 \leq i \leq n\}$ are needed to span the vector space, an infinitely many of these $\phi_b(x)$ are needed to form a basis that spans $\mathbf{H}$, we can think of $\mathbf{H}$ as infinite dimensional. In practice, an approximation based on truncating the infinite series and considering only the first B terms of the sum, $B \in \mathbb{N}$, is considered. An example of a system of basis function

is a collection of monomials $\{\phi_b(x) = x^{b-1} : b = 1, 2, \dots\}$ (Ramsay and Silverman, 2005).

To illustrate how systems of basis functions can be used to approximate another function, an example is shown in Figure 2.2. In Figure 2.2, the function $G(x)$ shown

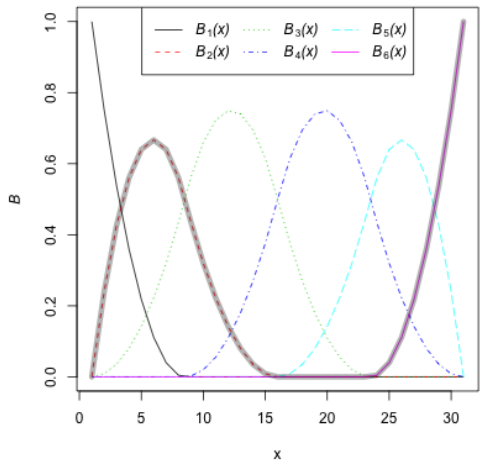


Figure 2.2: Example of basis functions $\{B_i : \mathbb{R} \mapsto [0, 1], i \in \{1, \dots, 6\}\}$ and a function G shown in shaded grey. The function $G(x)$ can be written as a linear combination of $B_1, \dots, B_6$, for example, $\theta_2 = \theta_6 = 1, \theta_1 = \theta_3 = \theta_4 = \theta_5 = 0$. The functions are plotted over $x \in [0, 30]$.

can be written as a linear combination of functions $B_2(x)$ and $B_6(x)$, i.e., $G(x) = B_2(x) + B_6(x)$ for $x \in [0, 30]$. The example is designed so that the function G equals the sum of only two functions but this is rarely the case since, in general, a finite number of basis functions might not be sufficient to fully reconstruct the function G.

2.3.3 Systems of B-spline basis functions

Spline functions are the most common choice of approximation system for functional data without cyclical or periodic patterns. A spline function defined on $[a, b]$ is a piecewise polynomial determined by the order $o \in \mathbb{N}$ that indicates polynomials of degree $o-1$ and a nondecreasing knot sequence $\{\tau_i\}_{i=0}^{N+1}$=$(\tau_0 = a, \tau_1, \dots, \tau_N, \tau_{N+1} = b)$ with N interior knots where adjacent polynomials pieces of order o meet. Since any linear combination of spline functions is still a spline function, it makes sense to make use of the system of basis functions that serve as the building block of these splines.

For a given number N of interior knots, a set of spline basis functions can be constructed as follows. Given $\boldsymbol{\tau} = (\tau_0, \tau_1, \dots, \tau_{N+1})$ where $\tau_0 \leq \tau_1 \leq \cdots \leq \tau_{N+1}$,

define the augmented knot sequence as $\boldsymbol{\tau}^* = (\tau_{-(o-1)}, \ldots, \tau_0, \ldots, \tau_{N+1}, \ldots, \tau_{N+o})$ with $\tau_{-(o-1)} = \cdots = \tau_0 \leq \tau_1 \leq \cdots \leq \tau_{N+1} = \cdots = \tau_{N+o}$ by appending boundary knots $o-1$ times. Re-index the augmented knot sequence as $\boldsymbol{\tau}^* = (\tau_0, \ldots, \tau_{N+2o-1})$ then a set of real-valued functions $B_{i,j}$ (for $i = 0, \ldots, N+2o-1$, $j = 1, \ldots, o$) can be obtained recursively by

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i^* \leq x < \tau_{i+1}^* \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

and,

$$B_{i,j+1}(x) = \alpha_{i,j+1}(x) B_{i,j}(x) + [1 - \alpha_{i+1,j+1}(x)] B_{i+1,j}(x) \tag{2.2}$$

where

$$\alpha_{i,j}(x) = \begin{cases} \frac{x - \tau_i^*}{\tau_{i+j-1}^* - \tau_i^*} & \text{if } \tau_{i+j}^* \neq \tau_i^* \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

Two examples for obtaining basis functions are to be described. They are the systems of B-spline basis functions used in our analysis. The first example is the set of basis functions used for fibre data where there are $N = 3$ equidistant interior knots with boundary knots at $(0, 4)$, and the order of the splines is $o = 3$. The augmented knot sequence used to construct the B-spline is $(0, 0, 0, 1, 2, 3, 4, 4, 4)$. The domains used here are for illustration only. The number of order $o = 3$ basis functions is $B = 6 = N + o$. The exact formula for the bases can be obtained using Equations (2.1) and (2.2) recursively as laid out in Figure 2.3 below. Every basis of order o is a linear combination of bases of degree $o - 1$. Only non-zero B's are shown.

The second example is the set of basis functions where there are $N = 6$ equidistant interior knots with boundary knots at $(0, 7)$, and the order of the splines is $o = 3$. The augmented knot sequence used to construct the B-spline is $(0, 0, 0, 1, 2, 3, 4, 5, 6, 7, 7, 7)$. Again, the boundary knots at $(0, 7)$ is for demonstration only. The number of basis functions (at order $o = 3$) is $B = 9 = N + o$. The detailed derivation of these basis functions can be found in Appendix B. In our models where B-splines are used, we

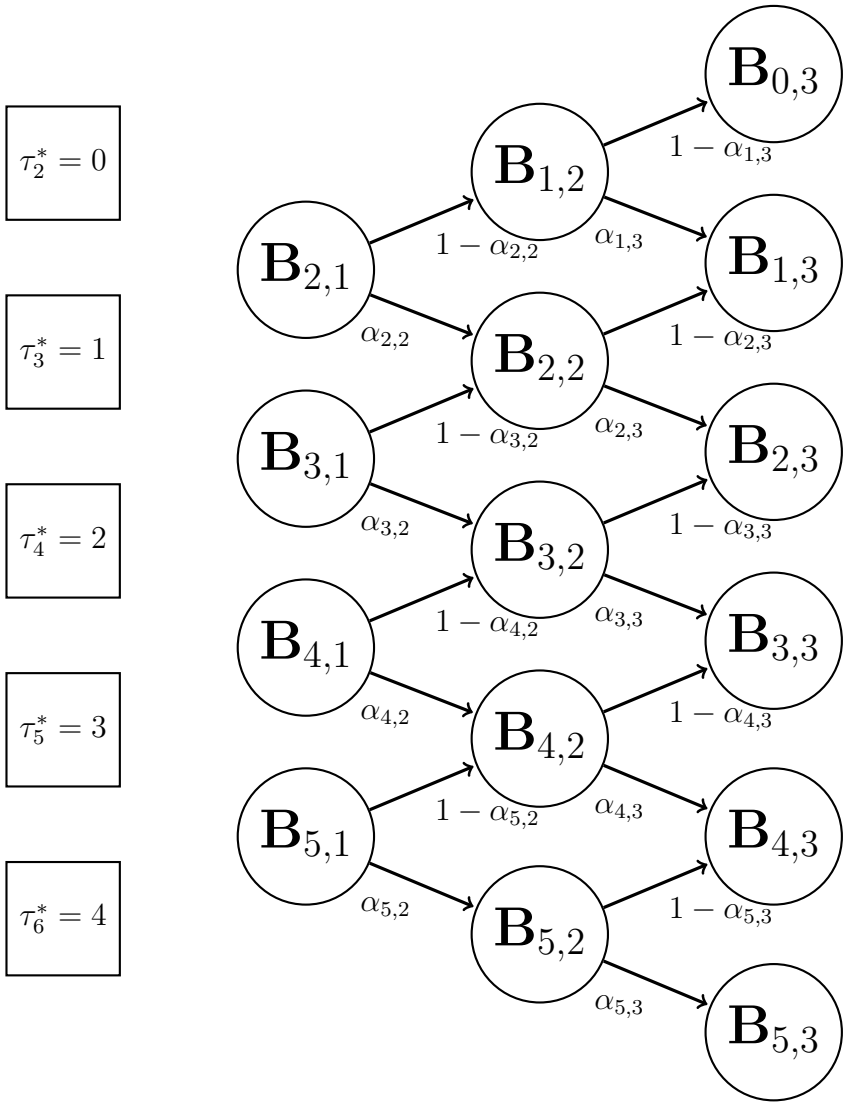
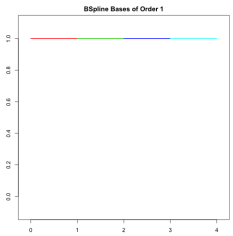


Figure 2.3: Example of derivation of a set of B-spline basis with $N = 3$ interior knots and $o = 3$ using Equations (2.1) and (2.2) recursively.

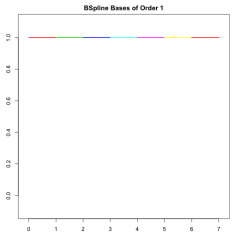


(a) Four bases of order 1.

(b) Five bases of order 2 that are linear combinations of bases of order 1.

(c) Six bases of order 3 that are linear combinations of bases of order 2.

Figure 2.4: Example of obtaining 6 bases of order 3 by assuming 5 equidistant knots (3 interior) from 4 bases of order 1.



(a) Seven bases of order 1.

(b) Eight bases of order 2 that are linear combinations of bases of order 1.

(c) Nine bases of order 3 that are linear combinations of bases of order 2.

Figure 2.5: Example of obtaining 9 bases of order 3 by assuming 8 equidistant knots (6 interior) from 7 bases of order 1.

define the set of basis functions by B and o instead of τ since equidistant knots are assumed. The problem now becomes selecting the right number of basis B of order o to fit the data. It will be discussed in Section 2.5. Part of this section is summarised in Aitken et al. (2019).

2.4 Functional data analysis

Ramsay and Silverman (2005) use the term functional data to describe a class of data with certain characteristics. A set of points $\{(w_j, \boldsymbol{y}_j) : j = 1, \ldots, m\}$ is said to be functional if the sequence $\{\boldsymbol{y}_j\}$ is considered to be samples of a smooth underlying function $\boldsymbol{x} : \mathbb{R} \mapsto \mathbb{R}^s$. Typically, the sequence $\{\boldsymbol{y}_j \in \mathbb{R}^s\}$ represents observations from a function x at points $\{w_j\}$ with error so that $y_j = \boldsymbol{x}(w_j) + \boldsymbol{\epsilon}_j$ where $\{w_j\}$ is a sequence of strictly increasing numbers. The set $\{w_j : 1 \leq j \leq m\}$ is commonly taken as time or wavelengths.

By convention, $\boldsymbol{y} = \{y_j\}$ will be used to represent a sequence of measurements that are taken at $\boldsymbol{w} = \{w_j\}$, or $\boldsymbol{y} = (y_1, \ldots, y_m)^T$ for $s = 1$ and analysed using ideas borrowed from multivariate data analysis. However, there are fundamental differences between multivariate and functional data; if $\boldsymbol{y}$ is treated as multivariate data, properties such as dimensionality and dependence of the elements y_j need to be taken care of. The following sections will explain how this is implemented by introducing techniques already common to multivariate data and how it will be used for analysing functional data.

2.4.1 System of basis functions and functional data

Let $\{y_j\}$ be observations of function f at $\{w_j\}$ with error. Now, suppose we have $\boldsymbol{w} = \{w_j : 1 \leq j \leq m\}$, a sequence of strictly increasing real numbers in the range $[0, 30]$ and $\boldsymbol{g} = \{g_j\} = G(\boldsymbol{w}) \in \mathbb{R}^m$, the function G evaluated at $\boldsymbol{w}$. Using the same example as in Figure 2.2, if we also have $\boldsymbol{b}_i = \{b_{ij}\} = B_i(\boldsymbol{w}) \in \mathbb{R}^m$ for all $i \in \{1, \ldots, 6\}$, $\mathbf{g}$ of dimension m can be represented by $\boldsymbol{\theta} = (0, 1, 0, 0, 0, 1)^T$, the coefficients of these basis functions, which is of a much smaller dimension. The vector of coefficients $\boldsymbol{\theta}$ is obtained as the unique solution of the system of linear equation so that $\boldsymbol{B\theta} = [\boldsymbol{b}_1\, \boldsymbol{b}_2\, \cdots\, \boldsymbol{b}_6]\boldsymbol{\theta} = \boldsymbol{g}$. Consider $\boldsymbol{B}^* = [\boldsymbol{b}_1\, \boldsymbol{b}_3\, \boldsymbol{b}_4\, \boldsymbol{b}_5]$, there are no $\boldsymbol{\theta}^*$ such that $\boldsymbol{B}^*\boldsymbol{\theta}^* = \boldsymbol{g}$ but an estimate for $\boldsymbol{\theta}^*$ can be obtained using least squares that minimises $||\boldsymbol{g} - \boldsymbol{B}^*\boldsymbol{\theta}^*||^2$, $\hat{\boldsymbol{\theta}}^* = (\boldsymbol{B}^{*T}\boldsymbol{B}^*)^{-1}\boldsymbol{B}^{*T}\boldsymbol{g}$. Note that as long as $\boldsymbol{g}, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_6$ are function evaluations at a given $\boldsymbol{w}$, then $\boldsymbol{w}$ can be suppressed without loss. Similarly, we only concern $\{y_j\}$ in the analysis of functional data.

2.4.2 Functional principal component analysis (fPCA)

In functional principal component analysis, we assume an underlying function x that our observations are based on, or $y_j = x(w_j) + e_j$ where $x(w_j)$ is centered at some true mean $\mu(w_j)$. We are interested in writing $x(w)$ as a linear combination of some functions $\{\phi_b(w) : b \in \{1, 2, \ldots, \infty\}\}$, or $x(w_j) = \mu(w_j) + \sum_{b=1}^{\infty} \theta_b \phi_b(w_j)$.

Since the measurements are taken at fixed and equally spaced intervals, the mean and variance of this function $x(w)$ can be estimated empirically by first estimating $\mu(w_j)$ by the sample mean $\frac{1}{nK}\sum_{i=1}^{nK} y_{ij}$ of all n replicates for all K groups and the covariance surface $\boldsymbol{\Sigma}$ by the sample covariance $\frac{1}{nK}\sum_{i=1}^{nK}(y_{is} - \hat{\mu}(w_s))(y_{it} - \hat{\mu}(w_t))$ for all $s \neq t$ and $Cov(s, s)$ is obtained by smoothing. The spectral decomposition gives $\boldsymbol{\Sigma}_{s,t} = \sum_{b=1}^{\infty} \lambda_b \phi_b(s)\phi_b(t) \approx \sum_{b=1}^{B} \lambda_b \phi_b(s)\phi_b(t)$ where $\{\phi_b(w) : \mathbb{R} \mapsto \mathbb{R} : b \in \{1, \ldots, B\}\}$ are the eigenfunctions of $x(w)$, specifically, $\int \phi_s(w)\phi_t(w)dw = 0$ for all $s \neq t$ and. Given the eigenfunctions, the scores are obtained by $\theta_b^{(i)} = \int (x_i(w) - \mu(w))\phi_b(w)dw$. So $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{Y} - \bar{\boldsymbol{Y}})^T(\boldsymbol{Y} - \bar{\boldsymbol{Y}}) = \boldsymbol{Y}_c^T\boldsymbol{Y}_c = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$ where $\boldsymbol{Y}_{i,j} = y_{ij}$ using the same decomposition as in Section 2.3.1 after centering and the scores can then be obtained by $\Theta = \boldsymbol{Y}_c\boldsymbol{V}$. This is equivalent as regressing $\boldsymbol{Y}_c$ onto $\boldsymbol{V}$ using ordinary linear regression $\Theta = \boldsymbol{Y}_c\boldsymbol{V}(\boldsymbol{V}^T\boldsymbol{V})^{-1} = \boldsymbol{Y}_c\boldsymbol{V}$ since $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ for $\boldsymbol{V}$ a collection of

eigenfunctions.

2.5 Selecting the number of basis functions

When dimension reduction is necessary for the analysis of data, choosing the right number of basis functions is crucial. The ultimate goal is to retain as least components as possible to avoid over-fitting but the reduced data have to be a proper representation of the original data so that no information is lost for the purpose of our work; i.e., reduced data will be sufficient for us to differentiate between propositions as we wish by the calculation of likelihood ratios. To this end we consider information criteria as means of assessing model fit while penalising model complexity in order to select the number of basis functions. Additionally, we also look at simple measures of goodness-of-fit tests such as residual sum of squares to aid model building.

For finite dimensional multivariate regression, the sums of squared residuals resulted from fitting a linear models always decreases as the number of independent variables increases so penalties need to be considered as otherwise complex models will always be preferred.

2.5.1 Information criteria

Statistical models are often constructed to find patterns based on limited amount of data in the hope of understanding the whole population of interest. When several models are proposed, criteria have to be considered for selecting the most appropriate model. In selecting the most appropriate models, methods such as maximum likelihood can be used for the estimation of parameters when the dimension and structure are specified. However, the problem is not as straightforward when the dimension is unknown as the method always favours the most complicated models for the fit always gets better as the model gets more complicated. In overcoming the shortcoming, Akaike proposed an information criterion as an extension to the maximum likelihood paradigm to select models without pre-specified dimension. For a given set of data y, it is assumed that it was generated by some mechanism or true model $g(y)$ that is unknown to us. The goal is to find a suitable model for y from a collection of candidate models with dimension

k such that it minimises the distance, that is, the Kullback-Leibler information defined as

$$I(\theta_k) = \mathbf{E}_{g(y)}\left[\log \frac{g(y)}{f(y|\theta_k)}\right].$$

Define $d(\theta_k) = \mathbf{E}\left[-2\log f(y|\theta_k)\right]$. Then we can write $2I(\theta_k) = d(\theta_k) - \mathbf{E}\left[-2\log g(y)\right]$ but $\mathbf{E}\left[-2\log g(y)\right]$ does not depend on θ_k so $d(\theta_k)$ alone is used instead of $I(\theta_k)$ and it is called the Kullback discrepancy. To measure the discrepancy, $d(\theta_k)$ would be evaluated at $\hat{\theta}_k$, the maximum likelihood estimates for the collection of models with dimension k. However, it is still not possible as $g(y)$ is unknown. Akaike suggested the use of $AIC = -2\log f(y|\theta_k) + 2k$ as it provides an asymptotically unbiased estimator of the expected Kullback discrepancy $\mathbf{E}\left[d(\hat{\theta}_k)\right]$. The term that includes the empirical log-likelihood $-2\log f(y|\theta_k)$ is called the goodness-of-fit term and $2k$ the penalty. While being asymptotically unbiased, Shibata (1980, 1981) claimed that AIC is not consistent. Several variants have been proposed such as CAIC (Bozdogan, 1987) and GIC (Konishi and Kitagawa, 1996) to correct for consistency and relax its assumptions, respectively. These variants differ mainly by their penalty terms. However, they all give a relative measure of the goodness of a model so if all of the candidate models fit the data terribly, they do not tell. For example, when comparing two models A and B which give AIC's of 5 and 10, respectively, we prefer model A over B for its smaller AIC but we do not know where a model that gives $AIC = 5$ stands in an absolute sense and how small these values can get. For smaller sample sizes, variants based on computationally intensive methods such as cross-validation, bootstrapping and Monte Carlo simulation tend to perform well.

Another commonly used criterion is Bayesian information criterion which selects the model that is a posteriori most probable. $BIC = -2\log f(y|\theta_k) + k\log n$ places a larger penalty on the number of parameters, favours lower-dimensional models compared to AIC.

For the purpose of choosing the number of components, information criteria AIC and BIC are often used. Ideally, we would pick the model with the lowest AIC or BIC but these values can decrease monotonically as the model gets larger so different techniques are used to choosing the optimal model based on these values including he use of scree plots.

2.5.2 Chi-squared like goodness of fit test

Since AIC and BIC are likelihood functions with some penalty, we consider another way of selecting models, that is, R^2E defined as

$$R^2E = \sum_{k,i,j} \frac{(y_{kij} - \hat{y}_{kij})^2}{|\hat{y}_{kij}|} = \sum_{k,i,j} \frac{\hat{r}_{kij}^2}{|\hat{y}_{kij}|}.$$

This is analogous to Chi-square goodness of fit test. It is better than AIC in the sense that there is division involved so the terms in the summation have limit. Since $y_{kij} = \hat{y}_{kij} + \hat{r}_{kij}$, given data y_{kij}, the closer $\hat{y}_{kij}$ is to y_{kij} the smaller $\hat{r}_{kij}$ so this is to be minimised as well. This provides another measure of fit compared to AIC and BIC.

2.6 Evidence evaluation and likelihood ratios

Likelihood ratio is a widely accepted measure for the evaluation of evidence (Lindley, 1977; Martyna et al., 2013; Aitken and Lucy, 2004). The prior odds in favour of H_p compared with H_d are updated to posterior odds so that evidential value is taken into account. To see this, let $p(\cdot)$ be the relevant probability density function and $\mathbf{E} = \{\mathbf{E}_c, \mathbf{E}_r\}$ be the set of controlled and recovered evidence. Then

$$\frac{p(H_p|\mathbf{E})}{p(H_d|\mathbf{E})} = \frac{p(\mathbf{E}|H_p)}{p(\mathbf{E}|H_d)} \times \frac{p(H_p)}{p(H_d)} = \textbf{LR} \times \textbf{prior odds}.$$

The likelihood ratio can be seen as a measure of the strength of evidence in support of H_p over H_d with a value greater than one supporting H_p over H_d, a value less than one supporting H_d over H_p and a value equal to one supporting both equally strongly. The evidential value is calculated by careful considerations of 1) the similarity of features observed for evidence being compared, 2) possible sources of variation including within- and between-group variations, 3) the dependency relations among features and 4) the rarity of the features. The value of evidence in support of the prosecution proposition should be stronger when measurements are similar and rare in the relevant population as opposed to similar but common. Likelihood ratio allows for an integrated evaluation given all of the aforementioned points (Martyna et al., 2013). Traditional significance test approach will also be mentioned briefly in the univariate continuous

case in Section 2.6.1 to show why it is not preferred.

2.6.1 Likelihood ratio and significance test for comparing evidence characterised by continuous data - the univariate case

Using glass fragments as example Lindley (1977) derived likelihood ratios under both normal and nonnormal assumptions for univariate measurements. Suppose glass fragments are found at a crime scene and on a suspect. We are interested in knowing whether they come from the same source, or not. Measurements are taken of the refractive indices of the fragments from the crime scene and the clothings of the suspect.

Let the fragments found at the crime scene be indexed by 1 to n_c and those found on the clothings of the suspect be indexed by 1 to n_r. Their measurements $x_1, x_2, \ldots, x_{n_c}$ and $y_1, y_2, \ldots, y_{n_r}$ are assumed to follow normal distributions with true values θ_c and θ_r as means and a known and constant variance σ^2. Denote the collection of measurements $x_1, x_2, \ldots, x_{n_c}$ and $y_1, y_2, \ldots, y_{n_r}$ as $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Their means $\bar{x}$ and $\bar{y}$ therefore follow normal distributions with means θ_c and θ_r and variances σ^2/n_c and σ^2/n_r. Under H_p where $\boldsymbol{x}$ and $\boldsymbol{y}$ are assumed to have the same origin, θ_c is assumed to equal to θ_r but not under H_d where the marginal density of $\boldsymbol{x}$ is assumed to be independent of that of $\boldsymbol{y}$. Like $\boldsymbol{x}$ and $\boldsymbol{y}$, the true values θ_c and θ_r are also assumed to follow normal distributions but with common mean μ and variance τ^2. Unlike a fully Bayesian approach, both of μ and τ^2 are assumed to be constant in the model and are to be estimated from some relevant population. Usually between-group variance τ^2 is assumed to be much greater than within-group variance σ^2 if not identical. This can be explained by the assumption that evidence of our interest are samples of the relevant population and measurements from within the same groups have smaller variation compared to those between different groups.

Likelihood ratio as defined previously in this section is evaluated as $p(\bar{x}, \bar{y}|H_p)/p(\bar{x}, \bar{y}|H_d)$. The numerator is evaluated by $p(\bar{x}, \bar{y}|H_p) = \int p(\bar{x}|\theta)p(\bar{y}|\theta)p(\theta)d\theta$ since under H_p, x and y are assumed to have the same origin; hence θ is common for measurements found at either places (crime scene and suspect) and $\bar{x}$ and $\bar{y}$ are independent given θ. The denominator, on the other hand, is evaluated by $p(\bar{x}, \bar{y}|H_d) = \int p(\bar{x}|\theta)p(\theta)d\theta \int p(\bar{y}|\theta)p(\theta)d\theta$ since under H_d, x and y are assumed to have independent origins; hence their joint

marginal density is the product of individual marginal densities. In evaluating the performance of this approach, comparisons are being carried out with a significance test under the frequentitst approach. Using a significance test, the comparison problem becomes testing of the null hypothesis that assumes the sets of evidence are similar or the difference between the means of their measurements follows a normal distribution with a known variance. The null hypothesis can then be rejected if the test statistic is in the critical region. Cases with different ratios between between- and within-group variances τ^2 and σ^2 are compared to show that the significance test approach fails to take into account the rarity of the measurements. It also only considers one hypothesis; the null hypothesis is assumed to be true until enough evidence is found to show otherwise at a pre-specified significance level that seems arbitrary. With these in mind, likelihood ratio is favoured over significance test in that it considers two propositions at once and it takes rarity into consideration when measuring the evidential value.

2.6.2 Likelihood ratio for comparing evidence characterised by continuous data - the multivariate case

It is not always possible to characterise evidence using univariate measurements and the advancement of technology especially in the computational power and storage capacities gave rise to a wide range of data becoming available for analysis which are typically multidimensional. Aitken and Lucy (2004) developed models for the calculation of likelihood ratios when the data is multivariate and normally distributed. Initially motivated by measurements of the concentrations in three elemental ratios, these multivariate measurements $\boldsymbol{y}_{c,1}, \cdots, \boldsymbol{y}_{c,n_c}$ and $\boldsymbol{y}_{r,1}, \cdots, \boldsymbol{y}_{r,n_r}$ are assumed to be normally distributed with the real values $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_r$ as means for control and recovered evidence with constant and known within-group variance covariance matrix $\boldsymbol{U}$. Similarly, $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_r$ are assumed to follow a normal distribution about $\boldsymbol{\mu}$ with variance covariance matrix $\boldsymbol{C}$, both constant and known. Similar to the univariate case we introduced earlier on in this section, $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ are also centred at $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_r$ but with variances $\boldsymbol{U}/n_c$ and $\boldsymbol{U}/n_r$, respectively. The likelihood ratio defined as $p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}|H_p)/p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}|H_d)$ has numerator that can be written as $p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}|H_p) = \int p(\bar{\boldsymbol{x}}|\boldsymbol{\theta}, \boldsymbol{U})p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{U})p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{C})d\boldsymbol{\theta}$ and denominator $p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}|H_d) = \int p(\bar{\boldsymbol{x}}|\boldsymbol{\theta}, \boldsymbol{U})p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{C})d\boldsymbol{\theta} \int p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{U})p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{C})d\boldsymbol{\theta}$.

It has a closed form solution due to the conjugacy nature of normal distributions.

It might be sensible to assume constant within-group variance covariance matrix in the above application where measurements are of concentrations in elemental ratios; however, in cases where there might be variations among within-group variance-covariance matrix, (Marquis et al., 2006) argued that the assumption of constant within-group variation might be unrealistic and can result in unknown uncertainty in likelihood ratios calculated since the variance still needs to be estimated. Therefore, Bozza et al. (2008) proposed a two-level model that take randomness of the covariance matrix into account and the original estimation problem of the covariance matrix becomes estimation of its hyperparameters.

By conjugacy, the within-group variance covariance matrix $\boldsymbol{U}$ is assumed to follow an inverse Wishart distribution with parameters $\boldsymbol{\Omega}$ and ν. The likelihood ratio defined as $p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}|H_p)/p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}|H_d)$ has numerator that can be written as $p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}|H_p) = \int\int p(\bar{\boldsymbol{x}}|\boldsymbol{\theta}, \boldsymbol{U})p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{U})p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{C})d\boldsymbol{\theta}p(\boldsymbol{U}|\boldsymbol{\Omega}, \nu)d\boldsymbol{U}$. However, this can not be evaluated analytically so approximations using Gibb's sampling is required. Using the relation $p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}|\boldsymbol{\Psi})p(\boldsymbol{\Psi}) = p(\boldsymbol{\Psi}|\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$ where $\boldsymbol{\Psi}$=$\{\boldsymbol{\theta}, \boldsymbol{U}\}$ represents the parameters of interest, the marginal densities we are interested in, $p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$, can be approximated using point estimates of $p(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}|\boldsymbol{\Psi})$, $p(\boldsymbol{\Psi})$ and $p(\boldsymbol{\Psi}|\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$ evaluated at a given value of $\boldsymbol{\Psi}^*$ that is usually taken to be the maximum likelihood estimate. The detailed algorithm can be found in Appendix C.5.

2.6.3 Score based likelihood ratios

The method for evidence evaluation laid out in Section 2.6.1 requires an exact specification of the probabilistic model for the data, which is sometimes hard to formulate, especially when the data generating mechanism is complex (Hepler et al., 2012). Even if the model is known, reliable evidential values rely on the estimation of parameters which are not always easy to obtain due to the uniqueness of each individual cases. Score based approaches can be used to overcome some of these difficulties by first calculating scores from features and then to evaluate the likelihood ratio using those scores as new features. The scores measure the similarity between features of the control and recovered evidence; they usually indicate the proximity under some choice of

distance.

Score based likelihood ratios are common in the evaluation of handwriting (Hepler et al., 2012) and forensic speaker recognition (Gonzalez-Rodriguez et al., 2006). However, it still requires the evaluation of likelihood ratios. A hybrid approach that combines chemometrics and likelihood ratio has been proposed by Martyna et al. (2016). In this work multiple multivariate scores are first calculated using linear discriminant analysis and subsequently, likelihood ratios of these scores are evaluated using the multivariate normal random-effects model proposed by Aitken and Lucy (2004) and finally score-based likelihood ratios are obtained by multiplying these individual likelihood ratios together. This is called naive likelihood models.

2.6.4 Evidence evaluation with reference to ink and fibre data

Forensic ink analysis is important for document examinations including identification of forgeries, counterfeit document, alternations to document, and determine the origin and dating of documents (White, 2004; Neumann et al., 2011). The examination of documents were mostly done visually (Thanasoulias et al., 2003; White, 2004). Visual examination of ink colour is very easily carried out and provides high discrimination power without being destructive to the samples. However, anyone intending to forge a document will try his best to match the colour of the ink; what looks similar with unaided eye might have substantially different chemical components or look very different under other lighting conditions (White, 2004). Therefore, objective colour comparison is needed and may be carried out by microspectrophotometric reflective measurements (Pfefferli, 1983). Martyna et al. (2013) used a hybrid approach for the evaluation of evidence as described in Section 2.6.3 on parameterised microspectrophotometry data using the so-called three colour systems.

Forensic fibre analysis is mostly done by colour measurements through thin layer chromatography (TLC), microspectrophotometry (MSP) or Raman spectroscopy (Buzzini and Massonnet, 2015; Massonnet et al., 2003). Among these, MSP is non-destructive and allows for measurements with small samples (De Wael et al., 2015; Was-Gubala and Starczak, 2015) so is often preferred. These spectra were used to be compared visually and discriminating power is computed (Smalldon and Moffat, 1973) as the

ratio between the number of distinguishable pairs to total number of pairs being compared. Natural fibres such as cottons and wools are harder to characterise using light microscopy due to limited number of morphological features (Buzzini and Massonnet, 2015) and variation within absorption and transmission curves within a fibre sample may occur due to uneven dye uptake that makes analysis more difficult. After all, no evidential values like likelihood ratios are currently used for forensic evaluation of fibres at the source level alone (Ray, 2016) as it is claimed that reporting this without case specific information is like communicating facts with no foundation of meaning. Nonetheless, the data available can still be used to test the performance of our models.

2.6.5 Evidence evaluation for evidence characterised by functional data

There are always chemical and physical features associated with forensic evidence that can be useful for comparisons. Chemical analyses that involve extraction of compositions usually provide more information as mixing components are identified. Examples of instrumental techniques include high performance liquid chromatography (HPLC) (Banas et al., 2010; Pfefferli, 1983; Kher et al., 2006), high performance thin layer chromatography (HPTLC) (Neumann et al., 2011), time-of-flight secondary ion mass spectrometry (ToF-SIMS) (Denman et al., 2010), direct analysis in real time (DART) (Cody et al., 2005) and desorption electrospray ionisation (DESI) (Takáts et al., 2004) for forensic ink analysis (Martyna et al., 2013). However, they are often destructive (Martyna et al., 2013; Zięba-Palus and Kunicki, 2006; Kher et al., 2006) so the samples would not be available for further analysis later in an investigation. Therefore, non-destructive methods such as Fourier transform infrared spectroscopy (FTIR) (Banas et al., 2010; Bojko et al., 2008; Martyna et al., 2015), Raman spectroscopy (Braz et al., 2013; de Souza Lins Borba et al., 2015; Massonnet et al., 2003) and microspectrophotometry (MSP) (De Wael et al., 2015) are often preferred (Martyna et al., 2013; Massonnet et al., 2003). These produce functional data (Section 2.4) that can be analysed using techniques borrowed from multivariate analysis so the evaluation of evidence based on these data usually focuses primarily on dimension reduction using a combination of chemometric techniques such as principal component analysis (Adam et al.,

2008), linear discriminant analysis (Kher et al., 2006) and statistical cluster analysis (Adam, 2008; Denman et al., 2010; Thanasoulias et al., 2002, 2003; Martyna et al., 2015, 2016; Roux et al., 1999). Since every case is unique and relevant background data that can be used to estimate the parameters are often small compared to the number of variables, independence is usually assumed where possible (Aitken et al., 2007).

Burfield et al. (2015) assessed the possible use of functional data analysis for comparing and classifying forensic ink chromatograms. Martyna et al. (2015) combined chemometric tools with the likelihood ratio approach on the evaluation of evidential value of FTIR spectra of polymers and Raman spectra of car paints, which was unprecedented despite the development of compressing multidimensional physicochemical data using wavelet transforms. Based on these works we took a step further to develop probabilistic models that are able to account for all variabilities despite data complexity to obtain likelihood ratios that can be used as a reliable measure of the strength of evidence in support of the propositions.

Chapter 3

Models for functional data

3.1 Introduction

In this chapter, five models are introduced for the evaluation of likelihood ratio for evidence characterised by functional data; three of them component-wise, that is, treating the curves as mainly composed of a function and some error for each measurement, and two of them dimension reduced, that is, only consider a representation of the original curves. These models are applicable for all functional data. They are introduced independently of the data we use to assess the performance. Based on models specified in this chapter, the methods for evaluation of likelihood ratios are presented along with how estimates can be obtained from training data.

3.2 Component-wise additive models for functional data

Given controlled and recovered evidence $\{\mathbf{E}_c\}$ and $\{\mathbf{E}_r\}$ that are in the form of functional data, we are interested in calculating likelihood ratios, to be written as the ratio between two $probabilities$, each concerning a proposition that has to do with the origin of the evidence. In particular, when evidence is in the form of continuous data, $probabilities$ means products of probability density functions.

Following the introduction to functional data in Section 2.4, each observation consists of measurements at m distinct values of w, denoted by $\{(w_j, y_j), j \in \{1, \ldots, m\}\}$, which will sometimes be called a curve. It is assumed that $\{y_j\}$ are observations

of a function x at $\{w_j\}$ with error, or $y_j = x(w_j) + e_j$. Since there are usually groups of observations and observations within groups, a set of data will be denoted by $\{(w_j, y_{kij}), j \in \{1, \ldots, m\}, i \in \{1, \ldots, n_k\}, k \in \{1, \ldots, K\}\}$ where k indicates the group the observation belongs to. Moreover, each component of an observation can be decomposed into an additive representation of a trend function x_k at w_{kij} that is dependent on group k, and measurement error e_{kij}, written as $y_{kij} = x_k(w_j) + e_{kij}$ or $\boldsymbol{y}_{ki} = x_k(\boldsymbol{w}) + \boldsymbol{e}_{ki}$ for the ith observation in group k.

Using systems of basis functions introduced in Section 2.3.2, $x_k(\boldsymbol{w})$ can be written as $x_k(\boldsymbol{w}) = \sum_{b=1}^{\infty} \theta_b^{(k)} \phi_b(\boldsymbol{w})$ and approximated (Section 2.3.2) by $\sum_{b=1}^{B} \theta_b^{(k)} \phi_b(\boldsymbol{w}) = \boldsymbol{\Phi}\boldsymbol{\theta}_k$ for $B < m < \infty$, where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1(\boldsymbol{w})\ \boldsymbol{\phi}_2(\boldsymbol{w}) \cdots \boldsymbol{\phi}_B(\boldsymbol{w})]$, a matrix of size $m \times B$ that consists of basis function evaluations and $\boldsymbol{\theta}_k = (\theta_1^{(k)}, \ldots, \theta_B^{(k)})^T$.

3.2.1 Problem definition

Our overall aim is to compare two sets of evidence, controlled and recovered, that are characterised by data $\boldsymbol{Y}_c = [\boldsymbol{y}_{c,1} \ldots \boldsymbol{y}_{c,n_c}]$ and $\boldsymbol{Y}_r = [\boldsymbol{y}_{r,1} \ldots \boldsymbol{y}_{r,n_r}]$ where n_c and n_r are the numbers of observations in the sets given data from some relevant population $\{\boldsymbol{Y}_k\} = \{[\boldsymbol{y}_{k,1} \ldots \boldsymbol{y}_{k,n_k}], k = 1, \ldots, K\}$. Note that $\boldsymbol{Y}_q$, $q \in \{c, r\}$ and $\boldsymbol{E}_q$, $q \in \{c, r\}$ are used interchangeably as we always refer to the data that characterise the evidence and we are only interested in the differentiability of data assuming evidence $\boldsymbol{E}_q$, $q \in \{c, r\}$ we want to differentiate have measurements $\boldsymbol{Y}_q$, $q \in \{c, r\}$ that are differentiable. That is to say, there might be cases where evidence from different sources share the same characteristics but that is not the scope of our work here. An example can be that two types of inks used by two different pens have the exact same colour; hence non-differentiable MSP, then we would consider them to be of the same source. For models to be introduced in Sections 3.2.2 to 3.2.4 we assume component-wise additive relation

$$\boldsymbol{y}_{ci} = x_c(\boldsymbol{w}) + \boldsymbol{e}_{ci} = \boldsymbol{\Phi}\boldsymbol{\theta}_c + \boldsymbol{r}_{ci}, \text{ and } \boldsymbol{y}_{ri'} = x_r(\boldsymbol{w}) + \boldsymbol{e}_{ri'} = \boldsymbol{\Phi}\boldsymbol{\theta}_r + \boldsymbol{r}_{ri'} \tag{3.1}$$

where $\boldsymbol{\Phi}$ is a matrix consisting of basis function evaluations and $\boldsymbol{r}$ represents the residual that consists of measurement error and error arising from using only B basis func-

tions.

For the purpose of evidence evaluation using calculation of likelihood ratios the comparison is to be made under the competing propositions that

$$H_p : \text{the sets of evidence have the same origin, and}$$
$$H_d : \text{the sets of evidence have different origins}$$

assuming $\boldsymbol{y}_{ci}, i = 1, \ldots, n_c$ and $\boldsymbol{y}_{ri'}, i' = 1, \ldots, n_r$ each follows a multivariate normal distribution with parameters $(\boldsymbol{\Phi\theta}_c, \boldsymbol{\Sigma}_c)$ and $(\boldsymbol{\Phi\theta}_r, \boldsymbol{\Sigma}_r)$ where $\boldsymbol{\Sigma}_q$ =var$(\boldsymbol{Y}_q|\boldsymbol{\theta}_q)$, $q \in \{c, r\}$ denotes the within-group covariance matrix. Under H_p, it is assumed that $(\boldsymbol{\Phi\theta}_c, \boldsymbol{\Sigma}_c)$ = $(\boldsymbol{\Phi\theta}_r, \boldsymbol{\Sigma}_r)$, denoted by $(\boldsymbol{\Phi\theta}, \boldsymbol{\Sigma})$ and under H_d, $(\boldsymbol{\Phi\theta}_c, \boldsymbol{\Sigma}_c)$ is statistically independent of $(\boldsymbol{\Phi\theta}_r, \boldsymbol{\Sigma}_r)$.

Random effects are considered for $\boldsymbol{\theta}_k$ using hierarchical models to take into account between-group variabilities for trend or shape. A natural candidate for the distribution of $\boldsymbol{\theta}_k$ is multivariate normal with mean η. In special cases we also consider random effect models for the within-group covariance structure. In what follows, we assume that the within-group covariance matrices are scalar multiples of a positive definite matrix $\boldsymbol{P}$, i.e., $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{P}$ with $(\sigma_1, \ldots, \sigma_K) \in (0, \infty)^K$. This modelling assumption can be classified further to the case of common between-group covariance given by $\{(\sigma_1, \ldots, \sigma_K) \in (0, \infty)^K : \sigma_1 = \cdots = \sigma_K\}$ and to the case of varying between group-covariance $(\sigma_1, \ldots, \sigma_K) \in (0, \infty)^K$. The matrix $\boldsymbol{P}$ is of size m-by-m so further modelling assumptions are required in order to reduce the number of free parameters and render estimation and inference possible.

The parameters in the models are estimated by using some relevant population $\{\boldsymbol{Y}_k\}$=$(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_K)$ that the evidence $(\boldsymbol{Y}_c, \boldsymbol{Y}_r)$ are thought to have come from. In cases where uncertainty is accounted for for either mean $\boldsymbol{\theta}$ or covariance matrices $\boldsymbol{\Sigma}$, estimation of the fixed parameters translates to estimation of their hyperparameters in their prior distributions under empirical Bayes approach. Due to the conjugacy of the chosen distributions, likelihood ratios evaluated under models specified in this section

have closed forms and these are obtained via

$$LR = \frac{f(\boldsymbol{Y}_c, \mathbf{Y}_r|H_p)}{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_d)} = \frac{\int_{\boldsymbol{\Sigma}}\int_{\boldsymbol{\theta}} f(\boldsymbol{Y}_c|\boldsymbol{\Phi\theta}, \boldsymbol{\Sigma}) f(\boldsymbol{Y}_r|\boldsymbol{\Phi\theta}, \boldsymbol{\Sigma}) f(\boldsymbol{\theta}, \boldsymbol{\Sigma}|H_p) d\boldsymbol{\theta} d\boldsymbol{\Sigma}}{\prod_{q\in\{c,r\}}\int_{\boldsymbol{\Sigma}}\int_{\boldsymbol{\theta}} f(\boldsymbol{Y}_q|\boldsymbol{\Phi\theta}_q, \boldsymbol{\Sigma}) f(\boldsymbol{\theta}, \boldsymbol{\Sigma}|H_d) d\boldsymbol{\theta} d\boldsymbol{\Sigma}} \quad (3.2)$$

where $f(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\cdot)$ indicates the joint prior density for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$; it is usually taken to be independent $f(\boldsymbol{\theta}|\cdot)f(\boldsymbol{\Sigma}|\cdot)$ or dependent $f(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \cdot)f(\boldsymbol{\Sigma}|\cdot)$ for models to be specified in this section.

3.2.2 CA-S Simplified multivariate normal random-effects model

The simplest case is considered here where $\boldsymbol{y}_{ki} = \boldsymbol{\Phi\theta}_k + \boldsymbol{r}_{ki}$ where $\boldsymbol{r}_{ki} = \sigma\boldsymbol{\epsilon}_{ki}$ for $\boldsymbol{\epsilon}_{ki}$ such that $Cov(\boldsymbol{\epsilon}_{ki}) = \boldsymbol{I}$. This is saying $Cov(\boldsymbol{r}_{ki}) = \boldsymbol{\Sigma}_k = \sigma^2\boldsymbol{I}_m$, that is, the variance at each component is independent and identically distributed for all components for all curves and for all groups.

The location parameter $\boldsymbol{\theta}_k$ follows a B-dimensional multivariate normal distribution with mean $\boldsymbol{\eta}$ and covariance matrix a diagonal matrix $\boldsymbol{D}$, denoted $\boldsymbol{\theta}_k \sim N_B(\boldsymbol{\eta}, \boldsymbol{D})$ for all k.

The within-source variation represented by $\boldsymbol{\Sigma}$ is assumed to be a multiple of identity matrix $\sigma^2\boldsymbol{I}_m$ where σ^2 is assumed to be constant over all groups, and will be estimated by the unbiased average mean squared error across all components for all curves for all groups obtained through the analysis of variance. That is calculated as

$$\hat{\sigma}^2 = \frac{1}{K(nm - B)}\sum_{k=1}^{K}\sum_{i=1}^{n_k} ||\boldsymbol{y}_{ki} - \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}_k||^2$$

where $\hat{\boldsymbol{\theta}}_k$ is the minimizer of $\sum_{i=1}^{n_k} ||\boldsymbol{y}_{ki} - \boldsymbol{\Phi\theta}_k||^2$, or $\hat{\boldsymbol{\theta}}_k = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\frac{1}{n_k}\mathbf{Y}_k\mathbf{1}_{n_k}$ where $\mathbf{1}_{n_k}$ is a length n_k vector of ones.

When random effect is considered for $\boldsymbol{\theta}_k$, it is assumed to be centered at $\boldsymbol{\eta}$ which is to be estimated by averaging over all estimated group means $\hat{\boldsymbol{\theta}}_k$, that is

$$\hat{\boldsymbol{\eta}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_k.$$

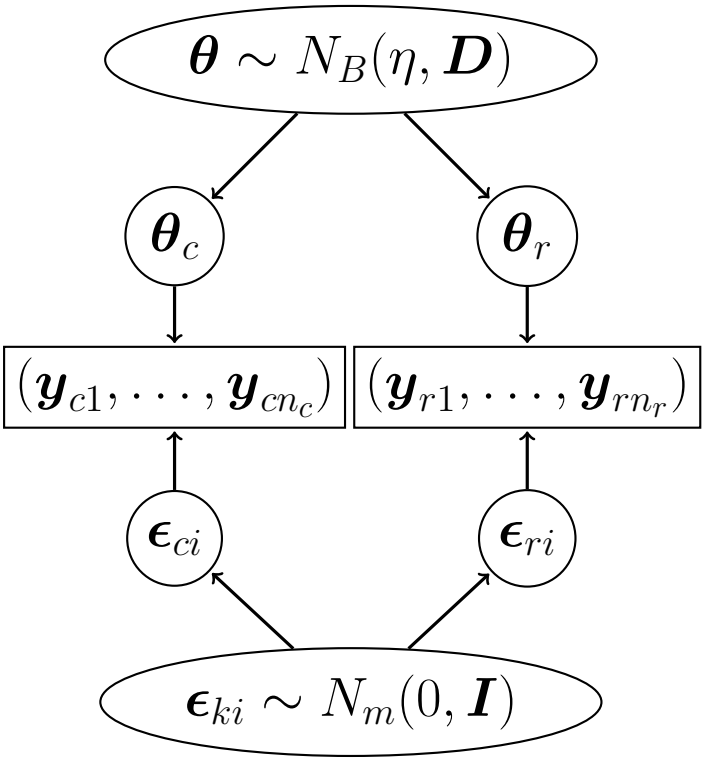


Figure 3.1: Schematic representation of simplified multivariate random-effects model for control and recovered evidence under the defense proposition where ($\boldsymbol{\theta}_c \neq \boldsymbol{\theta}_r$).

The diagonal elements of its covariance matrix D is then estimated by

$$\hat{D}_{bb} = \hat{\omega}_b = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}_b^{(k)} - \hat{\eta}_b)^2 - \frac{\hat{\sigma^2}}{n_k}\boldsymbol{\Phi}^T\boldsymbol{\Phi}.$$

using analysis of variance.

The integrand for the numerator of the likelihood ratio consists of three parts, the likelihoods $\prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m)$ and $\prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m)$, and a prior on $\boldsymbol{\theta}$, or between-group representation distribution $f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{D})$ as $\boldsymbol{\theta}$ is only the coefficient of the shape, or trend. Overall, there are two types of variation accounted, one within-group and one between-group. Putting all these together gives

$$LR = \frac{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_p)}{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_d)} = \frac{\int_{\boldsymbol{\theta}} \prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) \prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{D}) d\boldsymbol{\theta}}{\prod_{q\in\{c,r\}} \int_{\boldsymbol{\theta}} \prod_{i=1}^{n_q} f(\boldsymbol{y}_{qi}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{D}) d\boldsymbol{\theta}}$$

under this model. The resulting likelihood ratio can be simplified to

$$\frac{|\boldsymbol{\Sigma}_n^*|^{1/2}\text{exp}\left\{\frac{1}{2}\boldsymbol{\mu}_n^{*T}\boldsymbol{\Sigma}_n^{*-1}\boldsymbol{\mu}_n^*\right\}}{|\boldsymbol{\Sigma}_c^*|^{1/2}|\boldsymbol{\Sigma}_r^*|^{1/2}|\boldsymbol{D}|^{-1/2}\text{exp}\left\{-\frac{1}{2}\boldsymbol{\eta}^T\boldsymbol{D}^{-1}\boldsymbol{\eta}+\frac{1}{2}\boldsymbol{\mu}_c^{*T}\boldsymbol{\Sigma}_c^{*-1}\boldsymbol{\mu}_c^*+\frac{1}{2}\boldsymbol{\mu}_r^{*T}\boldsymbol{\Sigma}_r^{*-1}\boldsymbol{\mu}_r^*\right\}}$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_n^{*-1} &= \frac{n_c+n_r}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{D}^{-1}\\ \boldsymbol{\mu}_n^* &= \left(\frac{n_c+n_r}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{D}^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\left(\sum_{i=1}^{n_c}\boldsymbol{\Phi}^T\boldsymbol{y}_{ci}+\sum_{i=1}^{n_r}\boldsymbol{\Phi}^T\boldsymbol{y}_{ri}\right)+\boldsymbol{D}^{-1}\boldsymbol{\eta}\right)\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_q^{*-1} &= \frac{n_q}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{D}^{-1}\\ \boldsymbol{\mu}_q^* &= \left(\frac{n_q}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{D}^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n_q}\boldsymbol{\Phi}^T\boldsymbol{y}_{qi}+\boldsymbol{D}^{-1}\boldsymbol{\eta}\right) \text{ for } q\in\{c,r\}.\end{aligned}$$

The likelihood ratio can then be evaluated by plugging in estimates of hyperparameters from the relevant population.

3.2.3 CA-const. Constant within-group variance model

In this model, we relax the constant variance across groups assumption and assume curves from different groups have different variances but constant within group. This is taking $\boldsymbol{\Sigma}_k$ to be $\sigma_k^2\boldsymbol{I}_m$, again a diagonal matrix but dependent on k. Also, in contrast to the simplified multivariate normal random-effects model, we assume that the covariance for the coefficient vector $\boldsymbol{\theta}_k$ for group k is a multiple of a positive definite matrix. The multiple is taken to be the same as the within group variance, or σ_k^2. Over all they follow an inverse gamma distribution, denoted $\sigma^2 \sim \mathbf{IG}(\gamma,\delta)$.

In contrary to the simplified multivariate normal random-effects model, the constant within-group variance model assumes variation in within-group variance-covariance matrix where σ_k^2 will be estimated by the mean squared error within groups, that is,

$$\hat{\sigma}_k^2 = RSS_k/(mn_k-B) = \frac{1}{mn-B}\sum_{i=1}^{n_q}||\boldsymbol{y}_{ki}-\boldsymbol{\Phi}\boldsymbol{\theta}_k||^2 = 1/\hat{\lambda}_k$$

where $\hat{\boldsymbol{\theta}}_k$ is the minimizer of $\sum_{i=1}^{n_k} ||\boldsymbol{y}_{ki} - \boldsymbol{\Phi}\boldsymbol{\theta}_k||^2$, or

$$\hat{\boldsymbol{\theta}}_k = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\frac{1}{n_k}\boldsymbol{Y}_k\mathbf{1}_{n_k}$$

with $\mathbf{1}_{n_k}$ being a length n_k vector of ones. An inverse gamma prior on σ_k^2 is equivalent as a gamma prior on λ_k. Using the expectation and variance of gamma distribution; $\mathbf{E}[\lambda] = \gamma/\delta$ and $Var(\lambda) = \gamma/\delta^2$, the parameters are estimated by $\hat{\delta} = \hat{\lambda}/s_\lambda^2$ and $\hat{\gamma} = \hat{\lambda}^2/s_\lambda^2$ where $\hat{\lambda} = \frac{1}{K}\sum_{k=1}^{K}\lambda_k$ and $s_\lambda^2 = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\lambda}_k - \hat{\lambda})^2$.

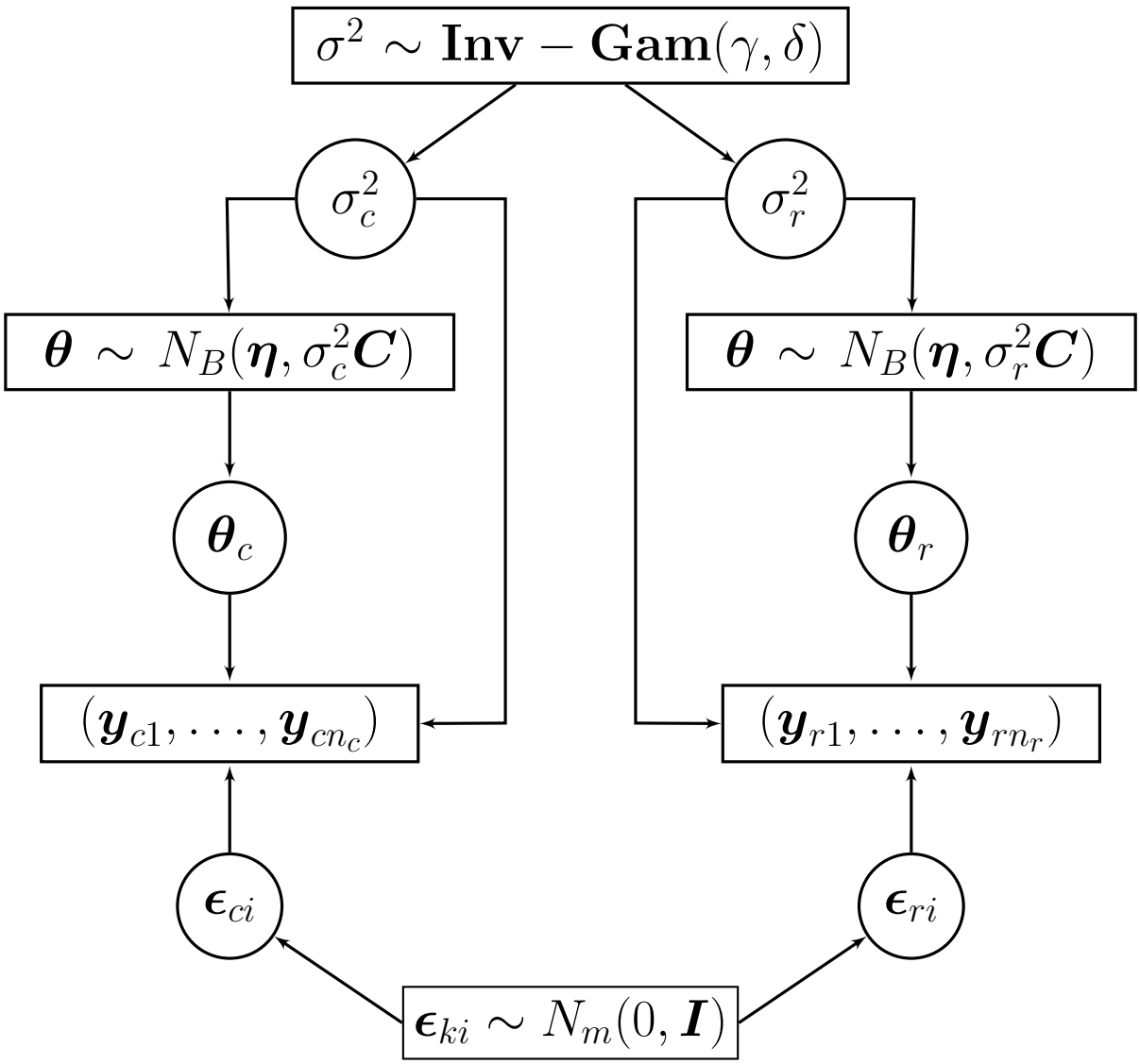


Figure 3.2: Schematic representation of constant within-group variance model for control and recovered evidence under the defense proposition where ($\boldsymbol{\theta}_c \neq \boldsymbol{\theta}_r$).

The between-source variance covariance matrix is to be estimated as

$$\hat{\boldsymbol{C}} = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\eta}})^2 / \frac{\sum_{k=1}^{K}\hat{\sigma}_k^2}{K} - \frac{1}{n}\boldsymbol{\Phi}^T\boldsymbol{\Phi}$$

using the analysis of variance.

Under this model, there is one extra term in each of the integrals comparing to simplified multivariate normal random-effects model.

$$\begin{aligned} LR &= \frac{f(\boldsymbol{Y}_c, \mathbf{Y}_r|H_p)}{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_d)} \\ &= \frac{\int_{\sigma^2}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m)\prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \sigma^2\boldsymbol{C}) f(\sigma^2|\gamma, \delta) d\boldsymbol{\theta} d\sigma^2}{\prod_{q\in\{c,r\}}\int_{\sigma^2}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_q} f(\boldsymbol{y}_{qi}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \sigma^2\boldsymbol{C}) f(\sigma^2|\gamma, \delta) d\boldsymbol{\theta} d\sigma^2} \end{aligned}$$

under this model, which can be simplified to

$$LR = \frac{\Gamma(\gamma)|\boldsymbol{C}|^{1/2}}{\delta^{\gamma}} \frac{\Gamma(\gamma^*)}{\Gamma(\gamma_c^*)\Gamma(\gamma_r^*)} \frac{\delta_c^{*\gamma_c^*}\delta_r^{*\gamma_r^*}}{\delta^{*\gamma^*}} \frac{|(n_c+n_r)\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}|^{-1/2}}{|n_c\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}|^{-1/2}|n_r\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}|^{-1/2}}$$

where

$$\begin{aligned} \gamma_q^* &= \gamma + \frac{n_q m}{2}, \\ \delta_q^* &= \delta + \frac{1}{2}\left[\sum_{i=1}^{n_q}\boldsymbol{y}_{qi}^T\boldsymbol{y}_{qi} + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}\right. \\ &\quad \left. - \left(\eta^T\boldsymbol{C}^{-1} + \sum_{i=1}^{n_q}\boldsymbol{y}_{qi}^T\boldsymbol{\Phi}\right)\left(n_q\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}\right)^{-1}\left(\boldsymbol{C}^{-1}\boldsymbol{\eta} + \boldsymbol{\Phi}^T\sum_{i=1}^{n_q}\boldsymbol{y}_{qi}\right)\right], \\ \gamma^* &= \gamma + \frac{n_c m}{2} + \frac{n_r m}{2}, \text{ and} \\ \delta^* &= \delta + \frac{1}{2}\left(\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{y}_{ci} + \sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{y}_{ri} + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta} - \boldsymbol{\mu}_n^{*T}\boldsymbol{\Sigma}_n^{*-1}\boldsymbol{\mu}_n^*\right) \end{aligned}$$

with

$$\begin{aligned} \boldsymbol{\mu}_n^* &= \left(\boldsymbol{C}^{-1}\boldsymbol{\eta}^T + \boldsymbol{\Phi}^T\left(n_c\bar{\boldsymbol{y}}_c + n_r\bar{\boldsymbol{y}}_r\right)\right) \\ \boldsymbol{\Sigma}_n^* &= \left((n_c+n_r)\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}\right)^{-1}. \end{aligned}$$

3.2.4 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model

In this model we relax the assumption of independent within group variance and assume an autoregressive structure on $\boldsymbol{P}$. For data $\boldsymbol{y}_{ki} = \boldsymbol{\Phi\theta}_k + \boldsymbol{r}_{ki} = \boldsymbol{\Phi\theta}_k + \sigma_k\boldsymbol{\epsilon}_{ki}$ we assume a lag 1 autoregressive structure for residuals

$$r_{kij} = \psi r_{ki,j-1} + \omega_{kij} \text{ and } \omega_{kij} \sim N(0, \tau_k^2) \text{ for all } k, i, j \tag{3.3}$$

where $\boldsymbol{r}$ is independent of $\boldsymbol{\omega}$. The positive definite matrix $\boldsymbol{P}$ is again var$(\boldsymbol{\epsilon}_{ki})$ but with non-zero off-diagonal elements. The $(j, j-s)$th element of $\boldsymbol{P}$, when multiplied by plugging in σ_k^2, gives $\sigma_k^2\boldsymbol{P}_{j,j-s} = \text{cov}(r_{kij}, r_{ki,j-s}) = \text{cov}(\psi r_{ki,j-1} + \omega_{kij}, r_{ki,j-s}) = \psi\text{cov}(r_{ki,j-1}, r_{ki,j-s}) = \psi^s\sigma_k^2$ for $0 \leq s \leq j$; therefore, $\boldsymbol{P}$ is simply a function of ψ and $\hat{\boldsymbol{P}}$ will be estimated by plugging in $\hat{\psi}$.

This can be obtained using regression by assuming a linear relation $\boldsymbol{y}_{ki}^* = \psi\boldsymbol{x}_{ki}^* + \boldsymbol{w}_{ki}$ where $\boldsymbol{y}_{ki}^* = [r_{ki2}r_{ki3}\ldots r_{kim}]^T$ and $\boldsymbol{x}_{ki}^* = [r_{ki1}r_{ki2}\ldots r_{ki,m-1}]^T$ for all i for all k. Since $\boldsymbol{P}$ is assumed to be constant and fixed for all k, $\hat{\psi}$ is simply taken to be the average of all $\hat{\psi}_k$'s obtained using distinct k's.

$$\hat{\boldsymbol{P}} = \begin{pmatrix} 1 & \hat{\psi} & \hat{\psi}^2 & \cdots & \hat{\psi}^{m-1} \\ \hat{\psi} & 1 & \hat{\psi} & & \\ \hat{\psi}^2 & \hat{\psi} & 1 & & \\ \vdots & & & \ddots & \vdots \\ \hat{\psi}^{m-1} & & & \cdots & 1 \end{pmatrix}$$

Using the relation $r_{kij} = \sigma_k\epsilon_{kij}$ together with Equation (3.3) gives $\sigma_k\epsilon_{kij} = \psi\sigma_k\epsilon_{ki,j-1} + \omega_{kij}$. Taking variance on both sides then gives $\sigma_k^2 = \psi_k^2\sigma_k^2 + \tau_k^2$ and $\hat{\sigma}_k^2 = \frac{\hat{\tau}_k^2}{1-\hat{\psi}_k^2}$ for all k. To obtain $\hat{\sigma}_k^2$ we also need $\hat{\tau}_k^2$. Given $\hat{\psi}$, $\hat{\tau}_k^2$ can be estimated by the mean residual sum of squares $\sum_i ||\boldsymbol{y}_{ki}^* - \hat{\psi}\boldsymbol{x}_{ki}^*||^2/(n(m-1)-1)$ and $\hat{\sigma}_k^2$ can then be estimated using $\hat{\sigma}_k^2 = \frac{\hat{\tau}_k^2}{1-\hat{\psi}^2}$ for a given k.

The rest of the hyperparameters can be estimated using the same way as the constant within-group variance model. Details can be found in Appendix C.

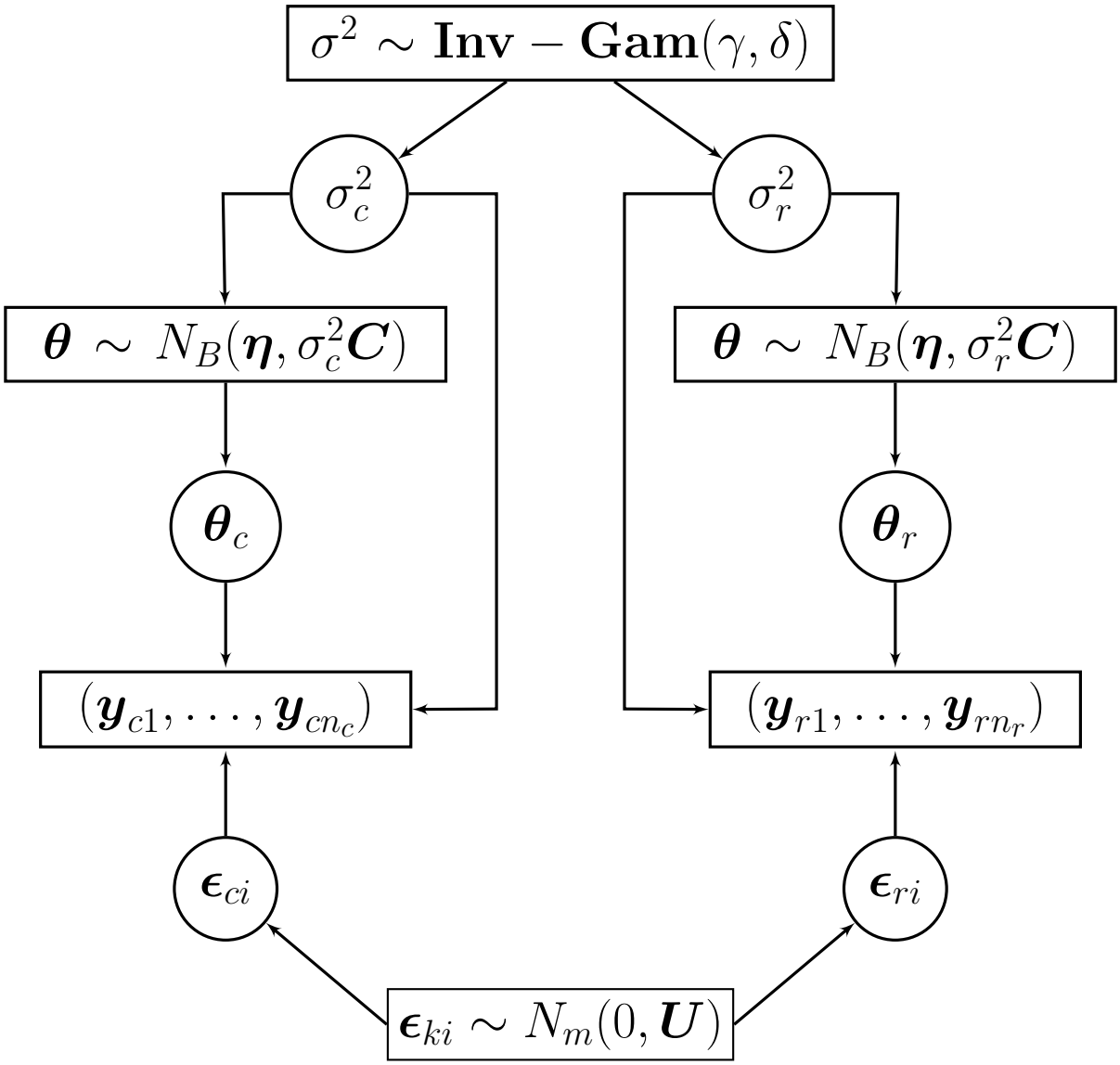


Figure 3.3: Schematic representation of multivariate normal random-effects with autoregressive within-group covariance model for control and recovered evidence under the defense proposition where ($\boldsymbol{\theta}_c \neq \boldsymbol{\theta}_r$).

Likelihood ratio defined as

$$\begin{aligned}&\frac{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_p)}{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_d)}\\ =&\frac{\int_{\sigma^2}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\theta}, \sigma^2\boldsymbol{P})\prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\theta}, \sigma^2\boldsymbol{P})f(\boldsymbol{\theta}|\boldsymbol{\eta}, \sigma^2\boldsymbol{C})f(\sigma^2|\gamma, \delta)d\boldsymbol{\theta}d\sigma^2}{\prod_{q\in\{c,r\}}\int_{\sigma^2}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_q} f(\boldsymbol{y}_{ci}|\boldsymbol{\theta}, \sigma^2\boldsymbol{P})f(\boldsymbol{\theta}|\boldsymbol{\eta}, \sigma^2\boldsymbol{C})f(\sigma^2|\gamma, \delta)d\boldsymbol{\theta}d\sigma^2}\end{aligned}$$

under this model can be simplified to

$$LR = \frac{\Gamma(\gamma)|\boldsymbol{C}|^{1/2}}{\delta^{\gamma}}\frac{\Gamma(\gamma^*)}{\Gamma(\gamma_c^*)\Gamma(\gamma_r^*)}\frac{\delta_c^{*\gamma_c^*}\delta_r^{*\gamma_r^*}}{\delta^{*\gamma^*}}\frac{|(n_c+n_r)\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\boldsymbol{C}^{-1}|^{-1/2}}{|n_c\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\boldsymbol{C}^{-1}|^{-1/2}|n_r\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\boldsymbol{C}^{-1}|^{-1/2}}$$

where

$$\begin{aligned}\gamma_q^* &= \gamma + \frac{n_q m}{2},\\ \delta_q^* &= \delta + \frac{1}{2}\left[\sum_{i=1}^{n_q}\boldsymbol{y}_{qi}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{qi} + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}\right.\\ &\quad\left. - \left(\boldsymbol{\eta}^T\boldsymbol{C}^{-1} + \sum_{i=1}^{n_q}\boldsymbol{y}_{qi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}\right)\left(n_q\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\boldsymbol{C}^{-1}\right)^{-1}\left(\boldsymbol{C}^{-1}\boldsymbol{\eta} + \boldsymbol{\Phi}^T\mathbf{P}^{-1}\sum_{i=1}^{n_q}\boldsymbol{y}_{qi}\right)\right],\\ \gamma^* &= \gamma + \frac{n_c m}{2} + \frac{n_r m}{2}, \text{ and}\\ \delta^* &= \delta + \frac{1}{2}\left[\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ci} + \sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ri} + \boldsymbol{\eta}^T\mathbf{C}^{-1}\boldsymbol{\eta} - \boldsymbol{\mu}_n^{*T}\boldsymbol{\Sigma}_n^{*-1}\boldsymbol{\mu}_n^*\right]\end{aligned}$$

with

$$\begin{aligned}\boldsymbol{\mu}_n^* &= \left(\boldsymbol{C}^{-1}\boldsymbol{\eta}^T + \boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\left(n_c\bar{\boldsymbol{y}}_c + n_r\bar{\boldsymbol{y}}_r\right)\right)\\ \boldsymbol{\Sigma}_n^* &= ((n_c+n_r)\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\boldsymbol{C}^{-1})^{-1}.\end{aligned}$$

3.3 Models with dimension reduction

Again, our aim is to compare two sets of evidence, controlled and recovered, that are characterised by data $\boldsymbol{Y}_c = [\boldsymbol{y}_{c,1}\dots\boldsymbol{y}_{c,n_c}]$ and $\boldsymbol{Y}_r = [\boldsymbol{y}_{r,1}\dots\boldsymbol{y}_{r,n_r}]$ where n_c and n_r are the numbers of observations in the sets given data from some relevant population $\{\boldsymbol{Y}_k\} = \{[\boldsymbol{y}_{k,1}\dots\boldsymbol{y}_{k,n_k}], k = 1,\dots,K\}$. For models to be introduced in Sections 3.3.1 and 3.3.2 we work with the dimension reduced data $\boldsymbol{Z}_k = \boldsymbol{\Phi}^T\boldsymbol{Y}_k$ where

$\boldsymbol{\Phi}$ is a m by B matrix consisting of function evaluations of B basis functions (Section 2.3.2) at m points. Comparing to the component-wise additive models for functional data, this $\boldsymbol{Z}_k$ has mean equal to the estimate of $\boldsymbol{\theta}_k$ assuming $\mathbf{E}[\boldsymbol{Y}_{ki}] = \boldsymbol{\Phi}\boldsymbol{\theta}_k$. So our likelihood ratio will be evaluated by assuming

$$LR = \frac{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_p)}{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_d)} = \frac{f(\boldsymbol{Z}_c, \boldsymbol{Z}_r|H_p)}{f(\boldsymbol{Z}_c, \boldsymbol{Z}_r|H_d)}.$$

In dealing with dimension reduced data only, we are only modeling the shape or trend parameter and ignoring the variances and residuals, that is, $\boldsymbol{r}$ in Equation 3.1, by assuming they are negligible. This simplifies the problem a lot by working with a lower dimensional representation of our data $\boldsymbol{Y}$ but there will also be loss of information. A dimension reduced multivariate normal random-effects model has been published in Aitken et al. (2019) together with some results that will be presented in Chapter 6.

3.3.1 DR-S Dimension reduced multivariate normal random-effects model

Let $\boldsymbol{\theta}_k$ be the group mean, $\boldsymbol{U}$ be the within-group covariance matrix, $\boldsymbol{\eta}$ be the overall mean and $\boldsymbol{C}$ be the between-group covariance matrix. We assume the dimension reduced data $\boldsymbol{z}_{ki}$, that is, $\boldsymbol{\Phi}^T\boldsymbol{y}_{ki}$ follows a multivariate normal distribution with mean $\boldsymbol{\theta}_k$ and covariance $\boldsymbol{U}$. The group $\boldsymbol{\theta}_k$ follows a multivariate normal distribution with mean $\boldsymbol{\eta}$ and covariance $\boldsymbol{C}$. A special case is considered where the within-group covariance $\boldsymbol{U}$ and between-group covariance $\boldsymbol{C}$ are assumed to be diagonal and $\boldsymbol{D}$ will be used to denote the diagonal between-group covariance.

The overall mean is estimated by the average of K group means, or $\hat{\boldsymbol{\eta}} = \sum_{k=1}^{K} \hat{\boldsymbol{\theta}}_k/K$ where $\hat{\boldsymbol{\theta}}_k$ is the minimiser of $\sum_{i=1}^{n_k} ||\boldsymbol{z}_{ki} - \boldsymbol{\theta}_k||^2$. The within-group covariance is estimated by $\hat{\boldsymbol{U}} = \hat{\text{var}}(\boldsymbol{z}_{ki}) = \sum_{i=1}^{n_k}(\boldsymbol{z}_{ki} - \hat{\boldsymbol{\theta}})(\boldsymbol{z}_{ki} - \hat{\boldsymbol{\theta}})^T/(Kn - K)$ and the between-group covariance is estimated by $\hat{\boldsymbol{C}} = \hat{\text{var}}(\boldsymbol{\theta}_k) = \sum_{i=1}^{n_k}\left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\eta}}\right)\left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\eta}}\right)^T/(K-1) - \hat{\boldsymbol{U}}/n$ using the analysis of variance.

For the special case where we assume diagonal variance-covariance matrices, the

between-group covariance will be estimated by

$$\hat{\sigma}^2 = \frac{1}{KB(n-1)} \sum_{k=1}^{K} \sum_{i=1}^{n} ||\boldsymbol{z}_{ki} - \hat{\boldsymbol{\theta}}_k||^2$$

and between-group covariance matrix with diagonal values estimated by

$$\hat{\omega}_i^2 = \frac{1}{K-1} \sum_{k=1}^{K} \left(\hat{\theta}_i^{(k)} - \hat{\eta}_i \right)^2 - \frac{\hat{\sigma}^2}{n} \boldsymbol{\Phi}^T \boldsymbol{\Phi}.$$

The likelihood ratio under this model can be written as

$$LR = \frac{\int_{\boldsymbol{\theta}} \prod_{i=1}^{n_c} f(\boldsymbol{z}_{ci}|\boldsymbol{\theta}, \boldsymbol{U}) \prod_{i=1}^{n_r} f(\boldsymbol{z}_{ri}|\boldsymbol{\theta}, \boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta}}{\prod_{q \in \{c,r\}} \int_{\boldsymbol{\theta}} \prod_{i=1}^{n_q} f(\boldsymbol{z}_{ci}|\boldsymbol{\theta}, \boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta}}$$

where all probability density functions $f(\cdot)$ are multivariate normal. The numerator of the likelihood ratio can be shown to simplify to

$$|2\pi \boldsymbol{U}|^{-(n_c+n_r)/2} |2\pi \boldsymbol{C}|^{-1/2} |2\pi \left((n_c + n_r)\boldsymbol{U}^{-1} + \boldsymbol{C}^{-1} \right)^{-1}|^{1/2} \exp \left\{ -\frac{1}{2} \left(\boldsymbol{H}_1 + \boldsymbol{H}_2 + \boldsymbol{H}_3 \right) \right\}$$

where

$$\begin{aligned}
H_1 &= \sum_{i=1}^{n_c} (\boldsymbol{z}_{ci} - \bar{\boldsymbol{z}}_c)^T \boldsymbol{U}^{-1} (\boldsymbol{z}_{ci} - \bar{\boldsymbol{z}}_c) + \sum_{i=1}^{n_r} (\boldsymbol{z}_{ri} - \bar{\boldsymbol{z}}_r)^T \boldsymbol{U}^{-1} (\boldsymbol{z}_{ri} - \bar{\boldsymbol{z}}_r) \\
&= \text{tr} \left(\sum_{i=1}^{n_c} (\boldsymbol{z}_{ci} - \bar{\boldsymbol{z}}_c)^T \boldsymbol{U}^{-1} (\boldsymbol{z}_{ci} - \bar{\boldsymbol{z}}_c) \right) + \text{tr} \left(\sum_{i=1}^{n_r} (\boldsymbol{z}_{ri} - \bar{\boldsymbol{z}}_r)^T \boldsymbol{U}^{-1} (\boldsymbol{z}_{ri} - \bar{\boldsymbol{z}}_r) \right)
\end{aligned}$$

due to the dimensions (1 by 1) of the expressions

$$\begin{aligned}
&= \text{tr} \left(\sum_{i=1}^{n_c} (\boldsymbol{z}_{ci} - \bar{\boldsymbol{z}}_c)(\boldsymbol{z}_{ci} - \bar{\boldsymbol{z}}_c)^T \boldsymbol{U}^{-1} \right) + \text{tr} \left(\sum_{i=1}^{n_r} (\boldsymbol{z}_{ri} - \bar{\boldsymbol{z}}_r)(\boldsymbol{z}_{ri} - \bar{\boldsymbol{z}}_r)^T \boldsymbol{U}^{-1} \right) \\
&= \text{tr} \left(\boldsymbol{S}_c \boldsymbol{U}^{-1} \right) + \text{tr} \left(\boldsymbol{S}_r \boldsymbol{U}^{-1} \right), \\
H_2 &= (\boldsymbol{z}^* - \boldsymbol{\eta})^T \left(\frac{\boldsymbol{U}}{n_c + n_r} + \boldsymbol{C} \right)^{-1} (\boldsymbol{z}^* - \boldsymbol{\eta}), \text{ and} \\
H_3 &= (\bar{\boldsymbol{z}}_c - \bar{\boldsymbol{z}}_r)^T \left(\frac{\boldsymbol{U}}{n_c} + \frac{\boldsymbol{U}}{n_r} \right)^{-1} (\bar{\boldsymbol{z}}_c - \bar{\boldsymbol{z}}_r)
\end{aligned}$$

with

$$\boldsymbol{z}^* = \frac{n_c \bar{\boldsymbol{z}}_c + n_r \bar{\boldsymbol{z}}_r}{n_c + n_r}.$$

The denominator, on the other hand, has two similar and independent terms, each can be simplified as

$$|2\pi \boldsymbol{U}|^{-n_q/2}|2\pi \boldsymbol{C}|^{-1/2}|2\pi \left(n_q \boldsymbol{U}^{-1} + \boldsymbol{C}^{-1}\right)^{-1}|^{1/2} \text{exp}\left\{-\frac{1}{2}\left(\boldsymbol{H}_{1q} + \boldsymbol{H}_{4q}\right)\right\}$$

where

$$H_{1q} = \text{tr}\left(\boldsymbol{S}_q \boldsymbol{U}^{-1}\right), \text{ and}$$
$$H_{4q} = (\boldsymbol{z}_q - \boldsymbol{\eta})^T \left(\frac{\boldsymbol{U}}{n_q} + \boldsymbol{C}\right)^{-1} (\boldsymbol{z}_q - \boldsymbol{\eta}).$$

LR can then be evaluated as

$$\begin{aligned}
&\frac{|2\pi \boldsymbol{U}|^{-(n_c+n_r)/2}\,|2\pi \boldsymbol{C}|^{-1/2}\left|2\pi\left((n_c+n_r)\boldsymbol{U}^{-1} + \boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2} \text{exp}\left\{-\frac{1}{2}\left(\boldsymbol{H}_1 + \boldsymbol{H}_2 + \boldsymbol{H}_3\right)\right\}}{\prod_{q\in\{c,r\}}|2\pi \boldsymbol{U}|^{-n_q/2}\,|2\pi \boldsymbol{C}|^{-1/2}\left|2\pi\left(n_q \boldsymbol{U}^{-1} + \boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2} \text{exp}\left\{-\frac{1}{2}\left(\boldsymbol{H}_{1q} + \boldsymbol{H}_{4q}\right)\right\}} \\
=&\frac{\left|\left((n_c+n_r)\boldsymbol{U}^{-1} + \boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2} \text{exp}\left\{-\frac{1}{2}\left(\boldsymbol{H}_2 + \boldsymbol{H}_3\right)\right\}}{|\boldsymbol{C}|^{-1/2}\left|\left(n_c \boldsymbol{U}^{-1} + \boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2}\left|\left(n_r \boldsymbol{U}^{-1} + \boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2} \text{exp}\left\{-\frac{1}{2}\left(\boldsymbol{H}_{4c} + \boldsymbol{H}_{4r}\right)\right\}}.
\end{aligned}$$

3.3.2 DR-C Multivariate normal random-effects with non constant within-group covariance

For this model we relax the constant within-group covariance assumption from Section 3.3.1 and assume that the within-group covariance follows an inverse Wishart $(\boldsymbol{\Omega}, \nu)$ distribution. This was first proposed by Bozza et al. (2008) to take into account non-constant within-group covariance for hand-writing data. The likelihood ratio we would like to evaluate has one extra term in the integrals compared to the model defined in Section 3.3.1 and can be written as

$$LR = \frac{f(\boldsymbol{Z}_c, \boldsymbol{Z}_r|H_p)}{f(\boldsymbol{Z}_c, \boldsymbol{Z}_r|H_d)} = \frac{\int_{\boldsymbol{U}}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_c} f(\boldsymbol{z}_{ci}|\boldsymbol{\theta}, \boldsymbol{U})\prod_{i=1}^{n_r} f(\boldsymbol{z}_{ri}|\boldsymbol{\theta}, \boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta} f(\boldsymbol{U}|\boldsymbol{\Omega}, \nu) d\boldsymbol{U}}{\prod_{q\in\{c,r\}}\int_{\boldsymbol{U}}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_q} f(\boldsymbol{z}_{ci}|\boldsymbol{\theta}, \boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta} f(\boldsymbol{U}|\boldsymbol{\Omega}, \nu) d\boldsymbol{U}}$$

While the likelihood ratio derived under this model cannot be evaluated directly, the calculation relies heavily on the conjugacy of the distributions.

The numerator of the likelihood ratio is evaluated under the proposition that the

data for the recovered curve and the control curve come from the same origin, or $\boldsymbol{\theta}_r = \boldsymbol{\theta}_c$ and $\boldsymbol{U}_r = \boldsymbol{U}_c$.

We are interested in

$$f(\boldsymbol{Z}_c, \boldsymbol{Z}_r|H_p) = \int\int \prod_{i=1}^{n_c} f(\boldsymbol{z}_{ci}|\boldsymbol{\theta}, \boldsymbol{U}) \prod_{i=1}^{n_r} f(\boldsymbol{z}_{ri}|\boldsymbol{\theta}, \boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta} f(\boldsymbol{U}|\boldsymbol{\Omega}, \nu) d\boldsymbol{U}$$

which is difficult to evaluate analytically. However, using Bayes' Theorem as in Chib (1995), the marginal likelihood can be written as

$$f(\boldsymbol{Z}|H_p) = \frac{f(\boldsymbol{Z}|\boldsymbol{\Psi}, H_p)\pi(\boldsymbol{\Psi}|H_p)}{\pi(\boldsymbol{\Psi}|\boldsymbol{Z}, H_p)}$$

where $\boldsymbol{\Psi} = (\boldsymbol{\theta}, \boldsymbol{U})$ and $\boldsymbol{Z} = \{\boldsymbol{Z}_c, \boldsymbol{Z}_r\}$. Denoting the maximum likelihood estimate as $\boldsymbol{\Psi}^*$, the estimate of the marginal density on logarithmic scale is

$$ln\{\hat{f}(\boldsymbol{Z}|H_p)\} = ln\{f(\boldsymbol{Z}|\boldsymbol{\Psi}^*, H_p)\} + ln\{\pi(\boldsymbol{\Psi}^*|H_p)\} - ln\{\hat{\pi}(\boldsymbol{\Psi}^*|\boldsymbol{Z}, H_p)\} \quad (3.4)$$

where $\hat{\pi}(\boldsymbol{\Psi}^*|\mathbf{Z}, H_p)$, the posterior joint density given data can be estimated using samples drawn from Gibbs sampling algorithm described in Bozza et al (2008).

The density function of the observation, or profile likelihood is given by

$$f(\boldsymbol{Z}|\boldsymbol{\Psi}, H_p) = \prod_{i=1}^{n_c} f(\mathbf{z}_{ci}|\boldsymbol{\theta}, \boldsymbol{U}) \prod_{i=1}^{n_r} f(\mathbf{z}_{ri}|\boldsymbol{\theta}, \boldsymbol{U}) \quad (3.5)$$

$$= \prod_{q\in\{c,r\}} \prod_{i=1}^{n_q} (2\pi)^{-p/2} |\boldsymbol{U}|^{-1/2} \text{exp}\left\{-\frac{1}{2}(\boldsymbol{z}_{qi} - \boldsymbol{\theta})^T \boldsymbol{U}^{-1} (\boldsymbol{z}_{qi} - \boldsymbol{\theta})\right\}. \quad (3.6)$$

The prior density for Ψ is given by

$$f(\boldsymbol{\Psi}|H_p)$$
$$= (2\pi)^{-p/2} |\boldsymbol{C}|^{-1/2} \text{exp}\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\eta})^T \boldsymbol{C}^{-1} (\boldsymbol{\theta} - \boldsymbol{\eta})\right\} \frac{|\boldsymbol{\Omega}|^{\frac{\nu}{2}} |\boldsymbol{U}|^{-\frac{\nu+p+1}{2}}}{2^{\nu p/2} \Gamma_p(\nu/2)} \text{exp}\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}\boldsymbol{U}^{-1}\right)\right\}.$$

The complete conditional density of $\boldsymbol{\theta}$ is then

$$f(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{U}) = \frac{f(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{U})f(\boldsymbol{\theta}|\boldsymbol{\eta},\boldsymbol{C})}{\int f(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{U})f(\boldsymbol{\theta}|\boldsymbol{\eta},\boldsymbol{C})d\boldsymbol{\theta}}$$
$$\propto \exp\left\{-\frac{1}{2}\left[\sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{qi}-\boldsymbol{\theta}) + (\boldsymbol{\theta}-\boldsymbol{\eta})^T\boldsymbol{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\eta})\right]\right\}$$

which can be shown to be still of type normal with parameters $(\boldsymbol{\eta}^*, \boldsymbol{C}^*)$, where

$$\boldsymbol{C}^* = \left(\sum_{l=1}^{2}\sum_{i=1}^{n_l}\boldsymbol{U}^{-1} + \boldsymbol{C}^{-1}\right)^{-1}$$
$$\boldsymbol{\eta}^* = \boldsymbol{C}^*\left(\boldsymbol{C}^{-1}\boldsymbol{\eta} + \sum_{l=1}^{2}\sum_{i=1}^{n_l}\boldsymbol{U}^{-1}\boldsymbol{z}_{li}\right).$$

The complete conditional density of $\boldsymbol{U}$ would be

$$f(\boldsymbol{U}|\mathbf{Z},\boldsymbol{\theta})$$
$$\propto |\boldsymbol{U}|^{-(n_c+n_r)/2}|\boldsymbol{U}|^{-(\nu+p+1)/2}\exp\left\{-\frac{1}{2}\left[\sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}(\mathbf{z}_{qi}-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\mathbf{z}_{qi}-\boldsymbol{\theta}) + \text{tr}\left(\boldsymbol{\Omega}\boldsymbol{U}^{-1}\right)\right]\right\}$$
$$\propto |\boldsymbol{U}|^{-(n_c+n_r+\nu+p+1)/2}\exp\left\{-\frac{1}{2}\left[\text{tr}\left(\sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta})(\boldsymbol{z}_{qi}-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}\right) + \text{tr}\left(\boldsymbol{\Omega}\boldsymbol{U}^{-1}\right)\right]\right\}$$
$$\propto |\boldsymbol{U}|^{-(n_c+n_r+\nu+p+1)/2}\exp\left\{-\frac{1}{2}\left[\text{tr}\left(\left(\boldsymbol{\Omega} + \sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta})(\boldsymbol{z}_{qi}-\boldsymbol{\theta})^T\right)\boldsymbol{U}^{-1}\right)\right]\right\}$$

which can be shown to be still of type inverse-Wishart with parameters (Ω^*, ν^*), where

$$\boldsymbol{\Omega}^* = \boldsymbol{\Omega} + \sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta})(\mathbf{z}_{qi}-\boldsymbol{\theta})^T$$
$$\nu^* = \nu + n_c + n_r.$$

The algorithm is then

1. Estimate $\hat{\boldsymbol{\eta}}$, $\hat{\boldsymbol{C}}$, and $\hat{\boldsymbol{\Omega}}$ from background (relevant population).

2. Sample $\boldsymbol{\theta}^g \sim N_B(\boldsymbol{\eta}^*, \boldsymbol{C}^*)$ and $\boldsymbol{U}^g \sim \mathcal{IW}(\boldsymbol{\Omega}^*, \nu^*)$, $g = 1, \ldots, G$ alternatively.

3. Obtain maximum likelihood approximation of $\Psi^* = (\boldsymbol{\theta}^*, \boldsymbol{U}^*)$ by

$$\Psi^* = \max_{\boldsymbol{\Psi}^g} f(\boldsymbol{Z}|\boldsymbol{\Psi}^g, H_p).$$

4. Compute

$$\hat{\pi}(\boldsymbol{U}^*|\boldsymbol{Z}) = \sum_{g=1}^{G} \frac{\pi(\boldsymbol{U}^*|\boldsymbol{Z}, \boldsymbol{\theta}^g)}{G}$$

5. Posterior is then given by $\hat{\pi}(\boldsymbol{\Psi}^*|\boldsymbol{Z}) = \pi(\boldsymbol{\theta}^*|\boldsymbol{U}^*, \boldsymbol{Z})\hat{\pi}(\boldsymbol{U}^*|\boldsymbol{Z})$.

The marginal likelihood (on logarithmic scale) can then be estimated using equation (3.4). A similar procedure can be carried out for the denominator of LR. The two independent integrals can be estimated by replacing $\boldsymbol{Z}$ with $\boldsymbol{Z}_c$ and $\boldsymbol{Z}_r$.

Chapter 4

Data description and selection of basis functions

4.1 Introduction

We are mainly interested in evaluating evidence that are characterised by functional data (Ramsay and Silverman, 2005), that is, it is assumed to be samples of an underlying function of another variable (explained in detail in Section 2.4). Previously, these kind of data was compared visually, our work aims to develop a systematic way of comparing them more objectively by the calculation of likelihood ratios.

Likelihood ratios calculated based on different models or assumptions can vary. Data to be used for the evaluation of performance of our models are described and presented here. They are the motivating examples that represent data of interest for the development of our models.

Three sets of data will be introduced, each of them will be described in detail and various types of plots will be displayed for the ease of understanding. The same sets of data are included in Aitken et al. (2019). Since dimension reduction is essential in all cases and all of our proposed models have a dimension reduction component; the component-wise additive models all assume data $\boldsymbol{Y}_k$ is centred at $\boldsymbol{\Phi\theta}_k$ given $\boldsymbol{\theta}_k$ and the dimension reduced models are applicable to data after a transformation $\boldsymbol{Z}_k = \boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{Y}_k$, the data will first be fitted to choose the most appropriate model for this purpose.

When modelling $\boldsymbol{Y}$, we are interested in its mean and variances. In this chapter, we

focus on modeling the mean. Since our data are functional, it makes sense to assume a functional mean. For dimension reduction purpose, we will use approximation, with an intention to represent the data using a smaller dimensional representation. To do so, basis functions are natural choices. There are many systems of basis functions that can be used to approximate functions. For example, it is natural to use basis functions with periodic boundary conditions in order to reconstruct and estimate seasonal cycles in data. Some common choices of basis functions that are independent of the data are Fourier series, splines, wavelets, exponential and power bases. An example of basis functions that are constructed from the data are empirical orthogonal functions, or eigenfunctions. Based on the properties of the data of our interest in this thesis (to be specified later on in Chapter 4), we will only be using B-spline basis and eigenfunctions for our data.

Given the sample sizes of our datasets (to be specified in the relevant section for each dataset), we will only consider the number of basis (B) to be between 5 and 10 and order of basis (o) between 2 and 4 inclusive. For each of these combinations of B and o, AIC and R^2E (Section 2.5) will be calculated and plots will be drawn for original data alongside fitted mean curves and residuals for the same choices of B and o. The optimal configuration will be chosen based on both numerical (AIC and R^2E values) and visual fit. Once B and o are selected for a given dataset, the fit by using the same numbers of eigenfunctions will also be plotted against those by using B-spline basis with selected order o for comparison. Both of these choices of basis functions, that is, B-spline basis functions and eigenfunctions obtained using fPCA, will be used for all models.

4.1.1 Selecting the number of B-spline basis functions

We would like to find the most appropriate set of basis functions for dimension reduction. Numerically, it is done by assuming control and recovered curves $\boldsymbol{Y}_c = \{\boldsymbol{y}_{ci}, \dots \boldsymbol{y}_{cn}\}$ and $\boldsymbol{Y}_r$ follow a multivariate normal distribution centred at $\boldsymbol{\Phi\theta}_c^{(M)}$ and $\boldsymbol{\Phi\theta}_r^{(M)}$, respectively where M specifies the model, or choices of number of B-spline basis functions or eigenfunctions (principal components) used for the purpose of dimension reduction here. Since we do not know the structure of var($\boldsymbol{y}_{ci}$) or var($\boldsymbol{y}_{ri}$),

AIC and R^2E are calculated based on assuming $\boldsymbol{y}_{ki} \sim N_m(\boldsymbol{\Phi}\boldsymbol{\theta}_k^{(M)}, \sigma_k^{2(M)}\boldsymbol{I}_m)$ for all $i \in \{1, \ldots, n\}$ for all $k \in \{1, 2, \ldots, K\}$, the variance is independent and identical for all points on the curves for a given group where K is the number of groups in the relevant population. AIC introduced in Section 2.5.1 translates to

$$AIC = -2\sum_k \sum_i \log f(\boldsymbol{y}_{ki}|\boldsymbol{\Phi}\boldsymbol{\theta}_k, \sigma_k^2) + 2KB$$

where $f(\boldsymbol{y}_{ki}|\boldsymbol{\theta}_k, \sigma_k^2)$ is multivariate normal density function and

$$R^2E = \sum_k \sum_i \sum_j \frac{(y_{kij} - \hat{y}_{kij})^2}{|\hat{y}_{kij}|}$$

where $\hat{y}_{kij} = \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}_k(j)$, the j-th component of the fitted curve $\boldsymbol{\Phi}\hat{\boldsymbol{\theta}}_k$.

Maximum decrement will be used as our main criteria to select the number and order of basis functions when AIC and R^2E are calculated if there is no obvious minimum. This is similar to selecting the number of components to retain using scree test in principal component analysis Cattell (1966).

For visual fit test, three sets of figures will be plotted for each dataset. Each with a different order starting from $o = 2$. The optimal combination will then be chosen based on numerical and visual criteria. Some of these figures are used in Aitken et al. (2019). After the number (and order) of basis functions is chosen, plots will be drawn to show the fits using different choices of basis, that is, B-spline basis functions and eigenfunctions obtained from fPCA.

4.1.2 Functional principal component analysis

In contrast to B-spline basis functions which are independent of data, eigenfunctions as introduced in Section 2.4.2 are empirical basis functions constructed using data so they are different fundamentally. We will check the fits of these basis functions and make comparison with fits using B-spline basis functions. To show the effect of eigenfunctions being empirical orthogonal functions, it is compared with fits of B-spline basis functions when the number of basis functions is small.

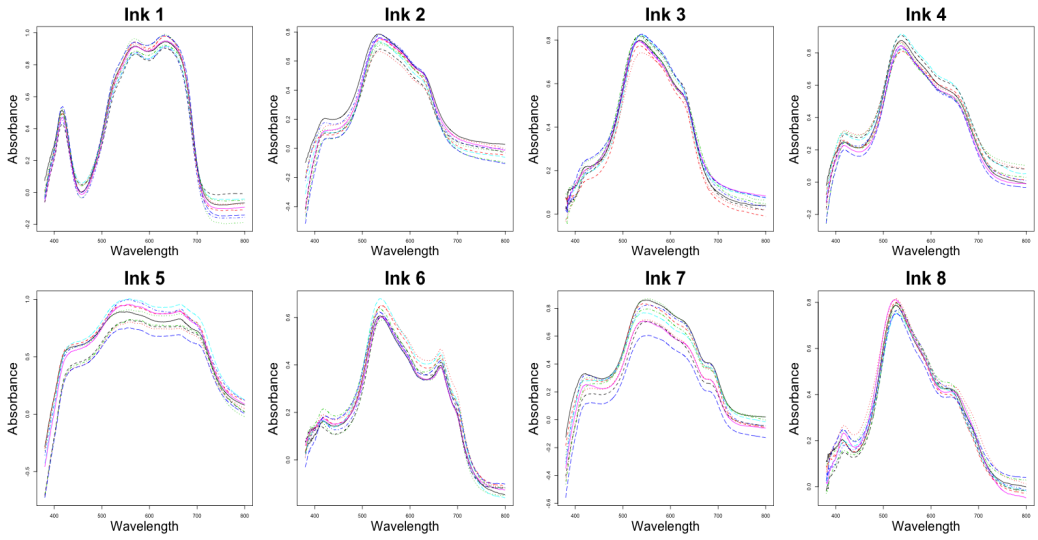
4.2 Pen ink

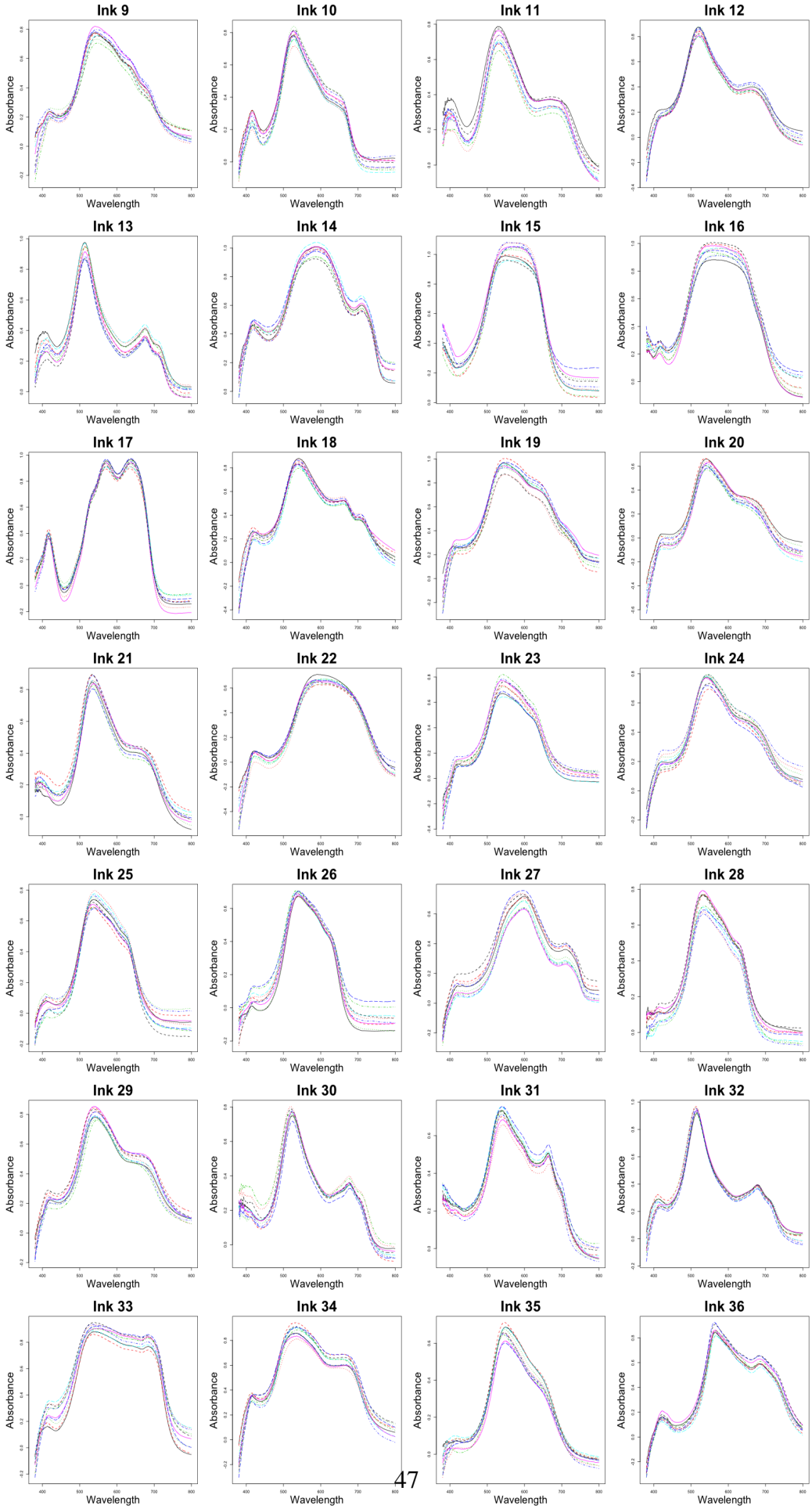
The data was provided by Institute of Forensic Research in Krakow, Poland. Forty blue inks that were collected primarily from the Polish market were analysed. One line was drawn by each and 10 observations were taken using microspectrophotometer (MSP) Zeiss Axioplan 2 with a J&M Tidas Diode Array Detector (DAD; MCS/16 1024/100-1, Germany), which was configured for the VIS range (380-800 nm) analyses.

Each observation consists of m measurements of absorbance y_j at wavelength w_j ranging from 380 to 800 nanometers. Absorbance is calculated as $y = log(I_o/I)$ where I_o and I are the intensities of the electromagnetic beam before and after contact with the sample. Data collection is described in detail in Martyna et al. (2013) where re-parameterised data was analysed.

Forty diagrams each showing n_k = 10 observations of MSP of the same type k of ink are shown. Every colour dashed line is drawn by connecting $m = 421$ points $\{(w_{kij}, y_{kij}), j = 1, \ldots, m\}$ in $\mathbb{R}^2$ for visualisation of an observation of a sample (one sample for each type). For a given dataset, $\{w_j\}$ is fixed for all types k and observations i, i.e., $\{w_j\}$=$\{380, \ldots, 800\}$. For ink data, the intervals $w_{ki,j+1} - w_{kij}$, or the difference in wavelength at which the measurements are taken is fixed at 1 nm. Different intervals (int) will also be considered in analysing the data where only every int-th point are used.

Figure 4.1: Forty plots each showing $n_k = 10$ observations of MSP by connecting $m = 421$ points $\{(w_j, y_{kij}), j = 1, \ldots, m\}$ of a type of ink.




Ink 9
Ink 10
Ink 11
Ink 12
Ink 13
Ink 14
Ink 15
Ink 16
Ink 17
Ink 18
Ink 19
Ink 20
Ink 21
Ink 22
Ink 23
Ink 24
Ink 25
Ink 26
Ink 27
Ink 28
Ink 29
Ink 30
Ink 31
Ink 32
Ink 33
Ink 34
Ink 35
Ink 36
Absorbance
Wavelength

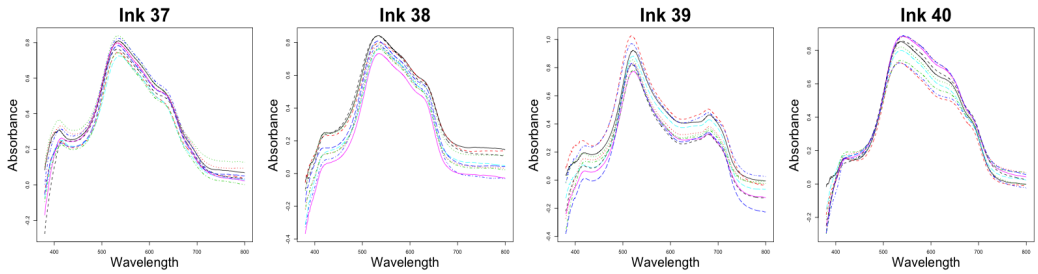


Figure 4.2: Each block shows $n_k = 10$ measurements of a type k of ink.

It can be seen that the curves (spectra) are quite smooth as there are generally no sharp edges, that is, non-differentiable points, when being drawn by connecting the points $\{(w_j, y_{kij}),\ j = 1, \ldots, m\}$ that are assumed to be samples of an underlying function $x(w)$. Also, the general shapes are different for each group (type) of ink and overall the shapes consist of 1 to 3 major peaks with 1 of them being highest and 2 shoulder-like.

Different types of ink can be categorised into 3 or 4 kinds of shapes, for example, 1 and 17 are very similar and 15 and 16 are of a similar type of shape; however, 5 and 33 are quite unique, and 8, 10, 21 and 35 are also similar in shape. Other than visually distinguishable shapes there are usually vertical separation of curves that is possibly caused by the difference in the concentration of ink (dye) measured (Was-Gubala and Starczak, 2015), which contribute to within-group variation.

4.2.1 Choosing the number of B-spline basis functions

The resulting AIC and R^2E values for ink data are summarised in Tables 4.1 and 4.2 with most optimal values shaded pink. They indicate the most favouring choices based on the decrements and magnitudes in comparison with values nearby.

Based on Table 4.1 there is an overall decrease of AIC as B increases and there are large drops as B increases from 8 to 9 regardless of o and these drops are larger than when B is increased from 5 to 6. However, as B is increased to 10, AIC goes up for $o = 4$ so $B = 9$ is more optimal than $B = 10$. Similar pattern can be seen in Table 4.2 for $B = 9$ where the next drop happens at $B = 13$ but not as much.

The extreme cases where there are spikes of R^2E at basis functions of order 4 is due to the fact that the ratio r_{kij}^2 to $|\hat{y}_{kij}|$ gives more weight to r_{kij}^2 with smaller values

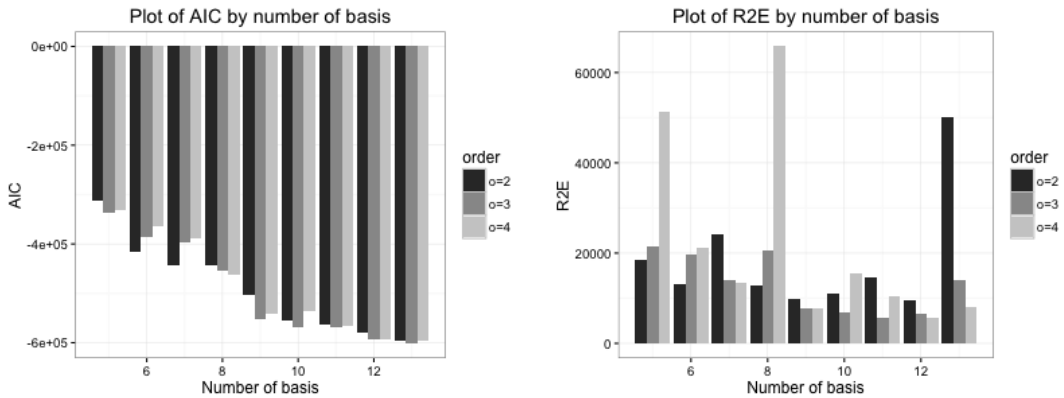
$B \setminus o$	2	3	4
5	-313539	-337762	-331881
6	-415318	-386114	-364320
7	-442129	-396490	-388058
8	-442268	-453792	-461255
9	-502492	-551178	-540106
10	-554339	-568836	-534948
11	-563310	-569831	-566362
12	-579995	-592041	-592956
13	-596833	-601354	-596841

Table 4.1: AIC values for ink data

$B \setminus o$	2	3	4
5	18413	21510	51454
6	13223	19821	21069
7	24067	14099	13566
8	12824	20710	65786
9	9899	7818	7784
10	10918	6879	15570
11	14481	5581	10471
12	9448	6596	5519
13	50110	14094	8090

Table 4.2: R^2E values for ink data

of $|\hat{y}_{kij}|$; therefore, slightly worse fit for smaller fitted values have a great effect in the over all R^2E's. Based on these numerical results, we found that B might not need to exceed 10 so the fits are plotted for B between 5 and 10. Before we draw some plots of the data, it is easier to compare AIC and R^2E by plotting them against number of B-spline basis functions used.



(a) AIC values for ink data in barplot.

(b) R^2E values for ink data in barplot.

Figure 4.3: AIC and R^2E values for ink data in barplots.

We can see from Tables 4.1 and 4.2 that given the same number of basis functions, higher orders do not always perform better especially for $B = 4, 5, 8, 10$ in Figure 4.3b. In cases where higher orders perform better, there is only a small improvement. This effect can be explained using Figure 2.3; same number but higher order basis functions are obtained by fewer order 1 (independent) basis functions.

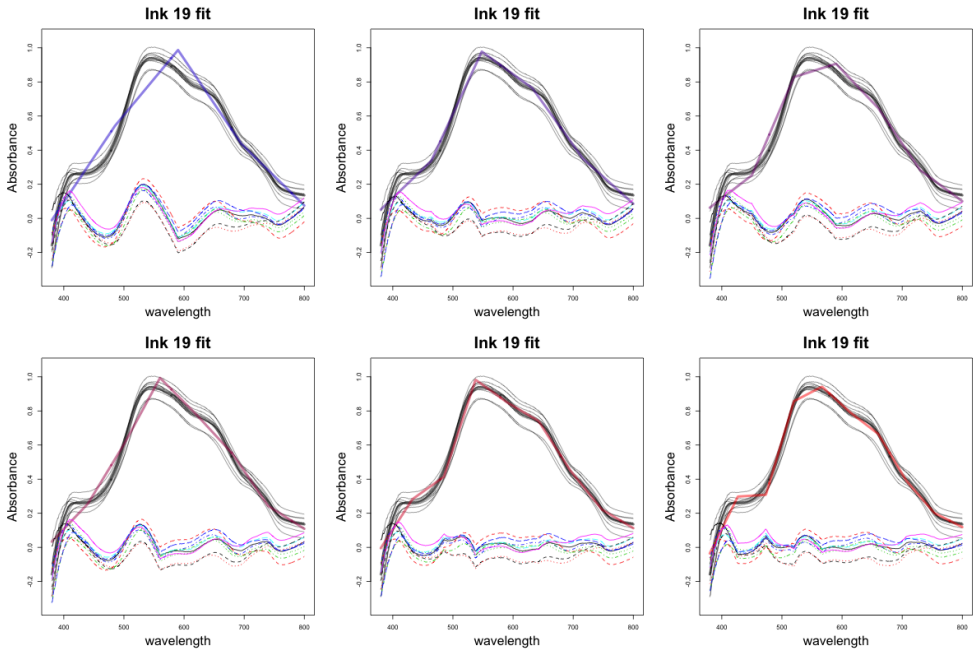


Figure 4.4: Fitting of a type of ink using different number of B-spline basis functions of order 2. From top left to bottom right, the number of basis functions used are between 5 and 10 (inclusive). The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

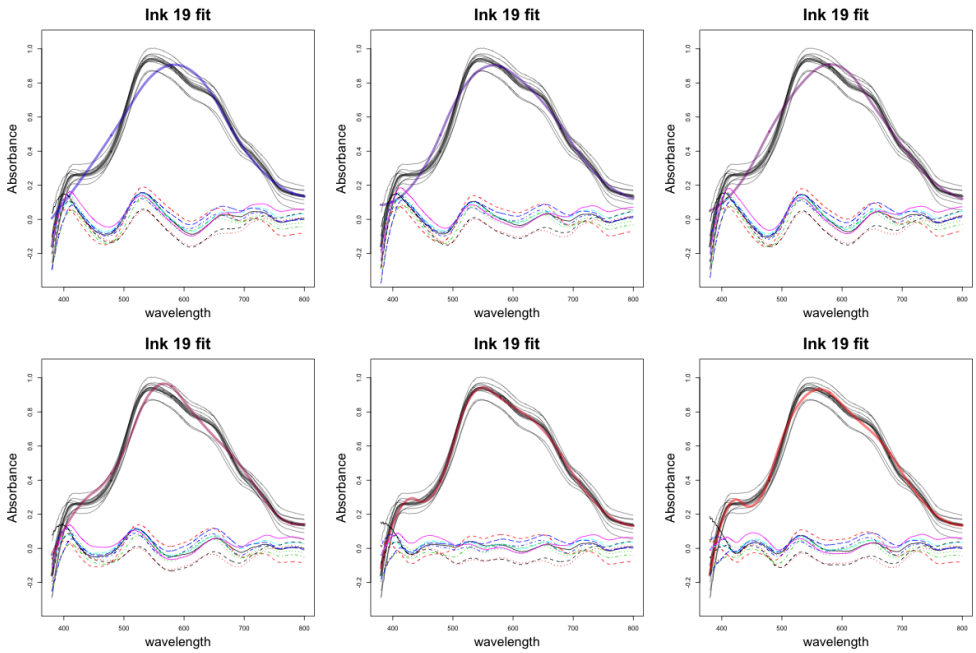


Figure 4.5: Fitting of a type of ink using different number of B-spline basis functions of order 3. From top left to bottom right, the number of basis functions used are between 5 and 10 (inclusive). The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

Based on Figure 4.4 while the fits seem to follow the shape of our curves and errors are small, there are some edges due to basis functions are of degree 1. These fits do not seem to resemble our data, which fails to represent these curves on average so they are not ideal. Moreover, Table 4.1 also suggested the same; $o = 3$ generally outperforms $o = 2$ given B. For order 3, 9 and 10 B-spline basis functions seem to fit the data quite well and there is no significant improvement as number of basis functions increases from 9 to 10.

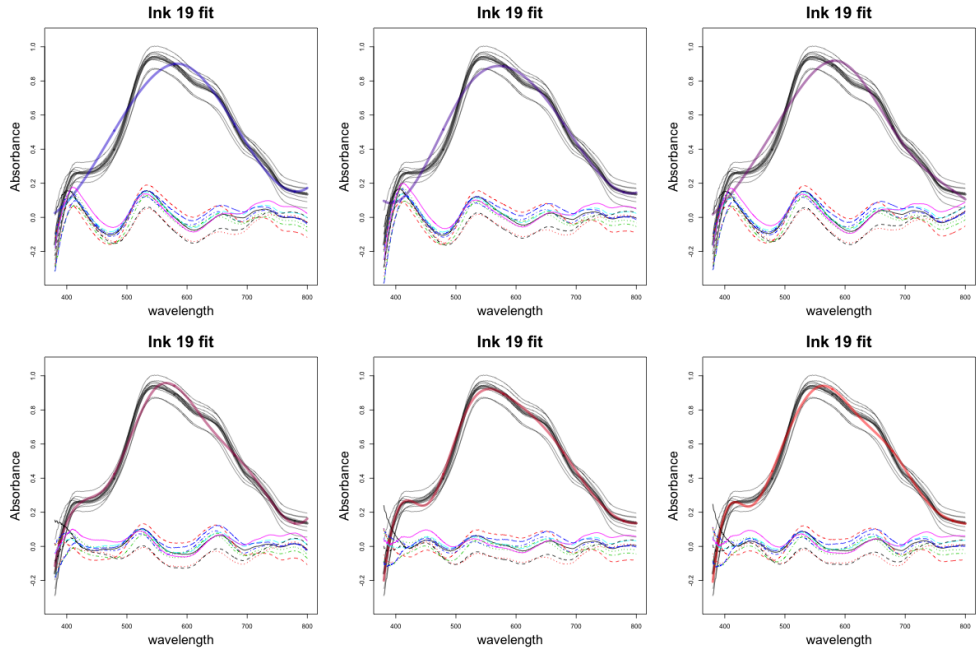


Figure 4.6: Fitting of a type of ink using different number of B-spline basis functions of order 4. From top left to bottom right, the number of basis functions used are between 5 and 10 (inclusive). The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

For order 4, 9 and 10 B-spline basis functions also seem to fit the data well. In this case $B = 9$ and $o = 3$ seems reasonable for ink data.

4.2.2 Functional principal component analysis

Once B is selected along with o for B-spline basis functions, the fit of using B-spline basis functions will be used to be compared with the use of eigenfunctions.

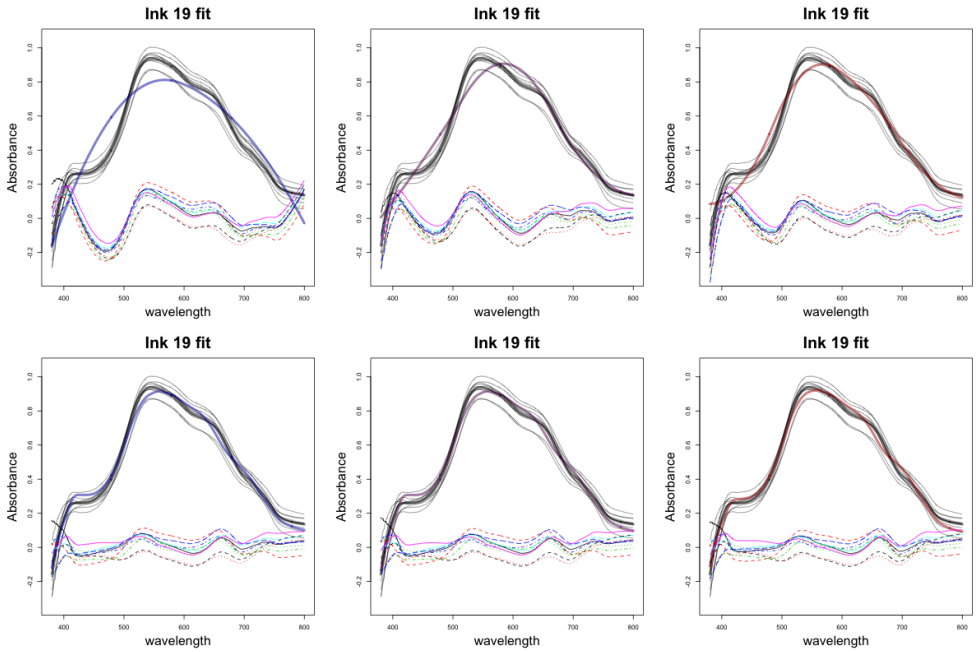


Figure 4.7: Compare fittings of a type of ink using same numbers of B-spline basis functions of order 3 and eigenfunctions obtained from functional principal component analysis. From left to right, the number of basis functions used are between 4 and 6 inclusive. The first row shows the use of B-spline basis functions and second row show the use of eigenfunctions as basis functions. The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

Since eigenfunctions are constructed empirically, or tailored to the data, it has greater fit for smaller numbers of basis functions, as expected. A number of basis as low as 5 provides reasonably well fit.

4.2.3 Conclusion

We managed to select the optimal number and order of basis for using B-spline basis functions. The order chosen is 3 and number of basis functions chosen is 9 for both choices of basis functions. This will give us more options when fitting models for the evaluation of likelihood ratios. However, modeling only the mean is never enough but it is essential for this applications. We will look at variance covariance structure later on.

4.3 Red wool fibre data

Other than ink data, we also have red wool and cotton data. Both datasets consist of 20 samples. Nine replicates of MSP spectra were collected for each sample. The red wool dataset includes data of spectra ranging from 350 to 690 nm (visible spectral range) with intervals of 5 nm.

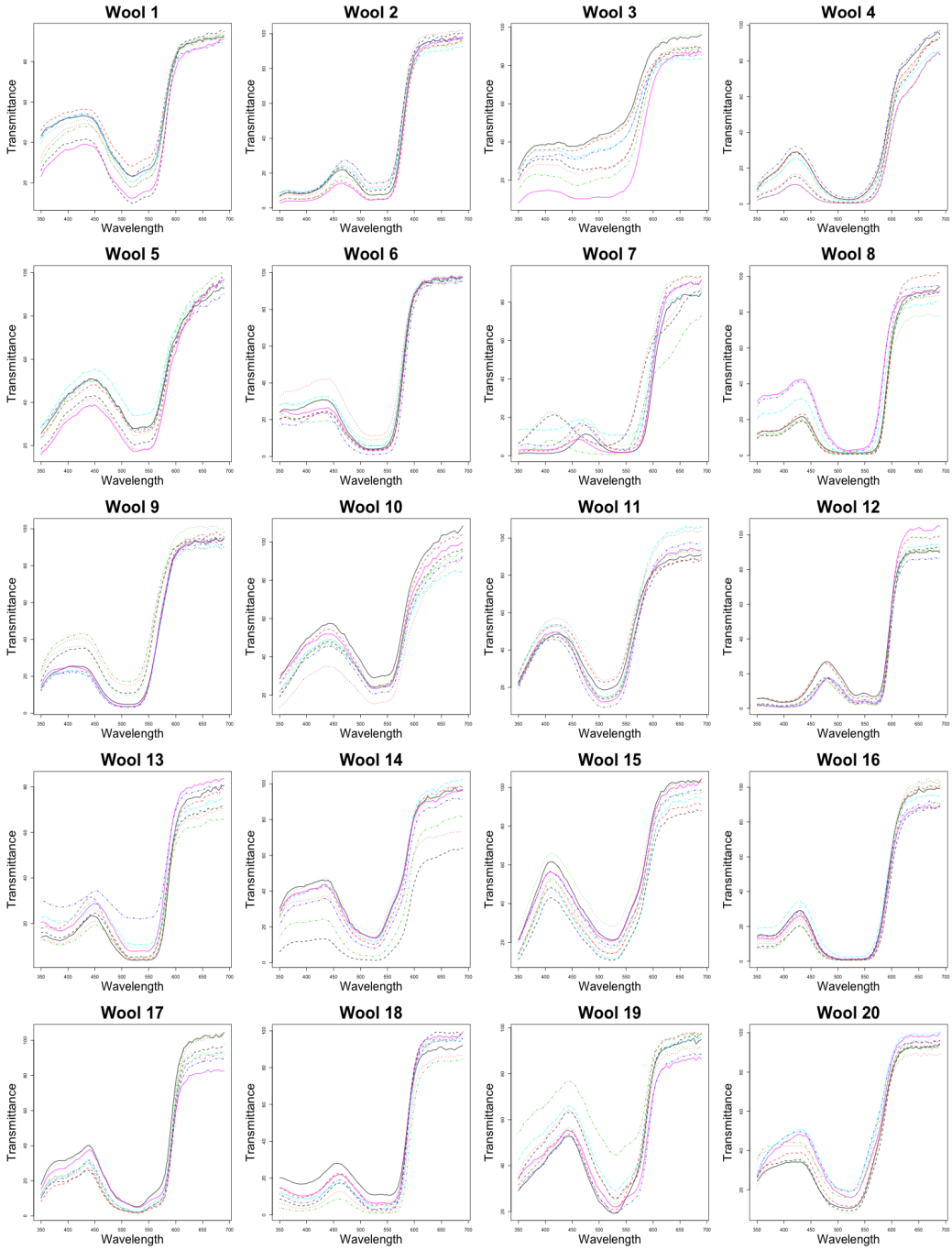


Figure 4.8: Each block shows $n_k = 9$ measurements of a type k of wool.

Compared to ink data, woollen fibre data have greater within-sample variation. Other than vertical separations, the slopes and gradients seem to be dependent on the locations. Taking wool 13 as an example, the vertical separation that occurs between 450 and 550 nm results in flatter curve for the curve at the top (drawn as blue dotted) and sharper for red and purple solid curve (drawn by solid line); this can easily be seen for wool 3 as well. In other words, the separation causes more variation in the shapes which makes it hard to distinguish between groups due to the similarities of the shapes of all types of wool; smaller between-group variation compared to within-group variations. They are all spoon-like with a big drop in transmittance roughly between 400 and 550 nm, depending on the group. Overall, the similarities among different groups together with noticeable within-group variations makes it harder to distinguish than ink data.

4.3.1 Choosing the number of B-spline basis functions

The resulting AIC and R^2E values for wool data are summarised in Section 4.3.1 and table 4.4 with most optimal values shaded pink. They indicate the most favouring choices based on the decrements and magnitudes in comparison with values nearby.

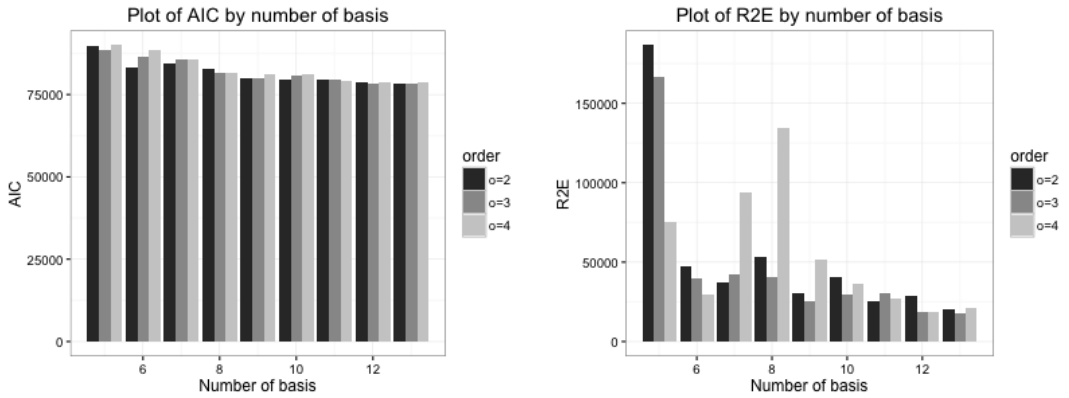
$B \setminus o$	2	3	4
5	89579	88357	89948
6	83183	86655	88485
7	84469	85680	85457
8	83002	81524	81631
9	79775	79929	81362
10	79419	80591	81145
11	79612	79455	79286
12	78753	78419	78816
13	78291	78254	78775

Table 4.3: AIC values for wool data

$B \setminus o$	2	3	4
5	186554	167051	75120
6	47276	39569	29793
7	37476	42671	94039
8	52994	40847	134110
9	30450	25675	51837
10	40390	29623	36142
11	25283	30714	27015
12	29059	18762	18505
13	20396	17889	21373

Table 4.4: R^2E values for wool data

For wool data, we can see $o = 4$ doesn't outperform $o < 4$ if not worse. Given $o = 3$, $B = 6, 8$ seem to be good choices using the same criteria: maximum decrement but when considering R^2E together with AIC, $B = 6$ looks more optimal.



(a) AIC values for wool data in barplot (b) R^2E values for wool data in barplot

Figure 4.9: AIC and R^2E values for wool data in barplots

Based on AIC alone in Figure 4.9, it is hard to pick B but $B = 6$ is clearly the best given R^2E as there is no more big drops for the values as B increases. $B = 6$ gives the best R^2E for all orders and the next best number of basis functions would be 9 but the decrement from $B = 8$ to $B = 9$ is much smaller compared to from $B = 5$ to $B = 6$. We will make a decision after checking the fits.

The combination of $B = 9$, $o = 2$ also looks good but the decrease from $B = 8$ is not as large as those from $B = 5$ to $B = 6$. Moreover, we only have 9 replicates for each group of wool data so $B = 6$ is favourable for parameter estimation purpose.

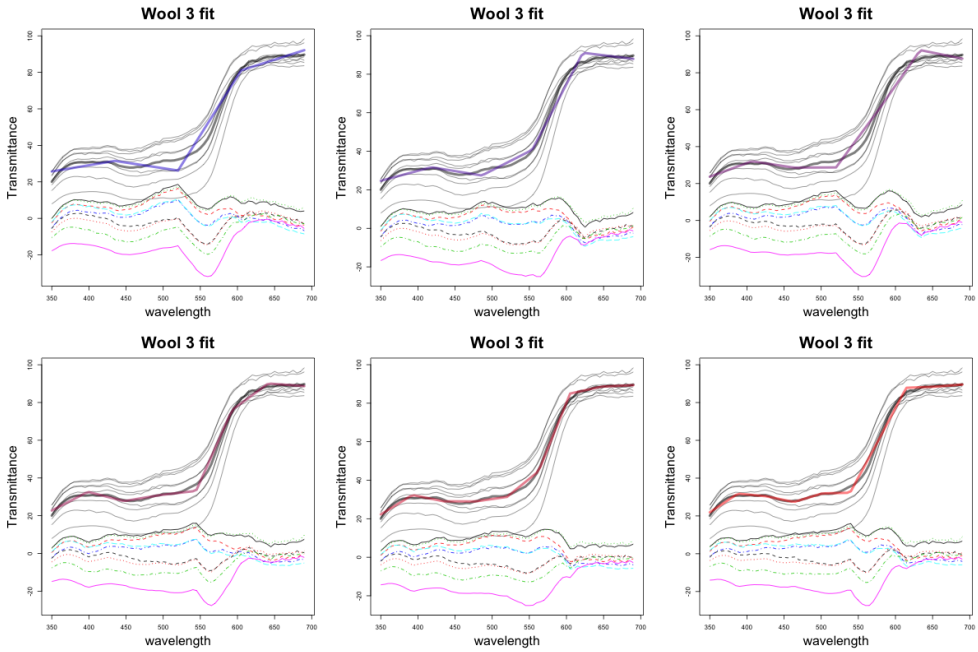


Figure 4.10: Fitting of a type of wool using different number of B-spline basis functions of order 2. From top left to bottom right, the number of basis functions used are between 5 and 10 (inclusive). The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

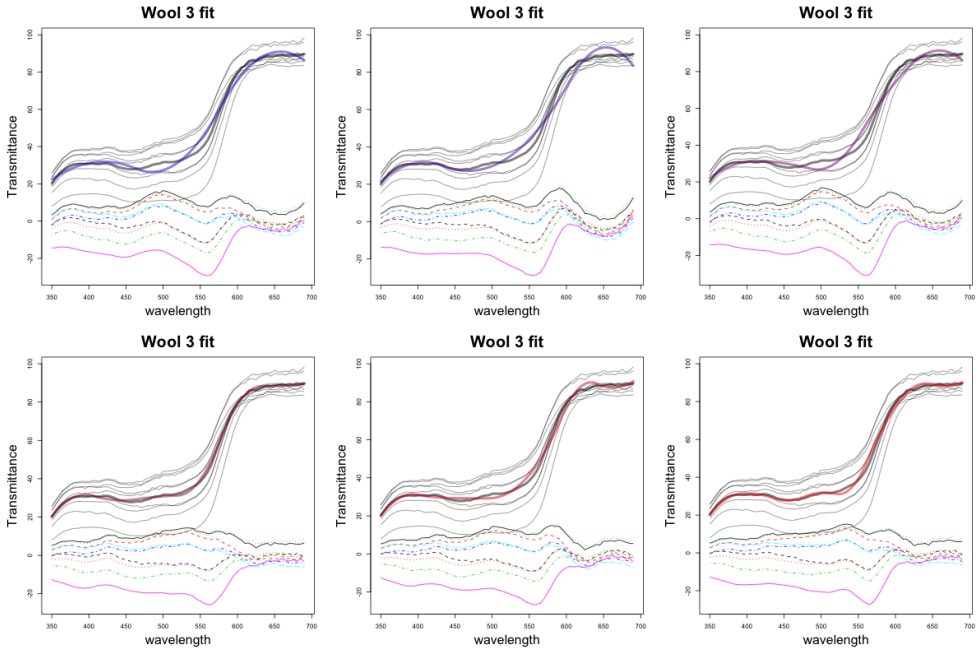


Figure 4.11: Fitting of a type of wool using different number of B-spline basis functions of order 3. From top left to bottom right, the number of basis functions used are between 5 and 10 (inclusive). The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

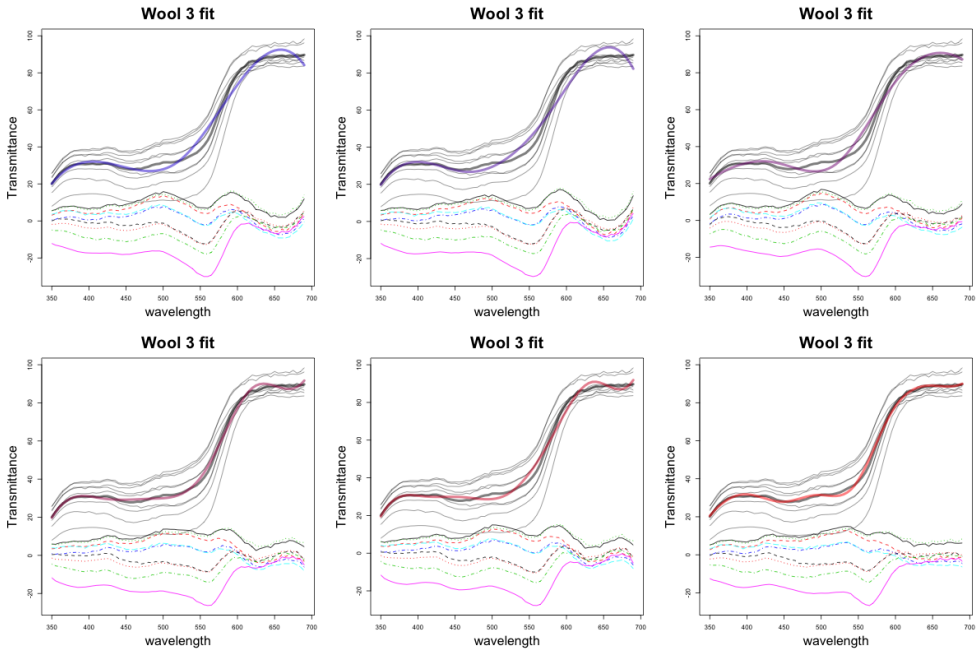


Figure 4.12: Fitting of a type of wool using different number of B-spline basis functions of order 4. From top left to bottom right, the number of basis functions used are between 5 and 10 (inclusive). The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

Our choice of $B = 6$, $o = 3$ using AIC and R^2E does not seem to have the best fit by looking at the purple or red shaded lines (fitted mean curve); however, when looking at residuals, that are the curves underneath them, there is no significant differences among different choices of B and o so we will stick with $B = 6$, $o = 3$.

4.3.2 Functional principal component analysis

Once B is selected along with o for B-spline basis functions, the fit of using B-spline basis functions will be used to be compared with the use of eigenfunctions.

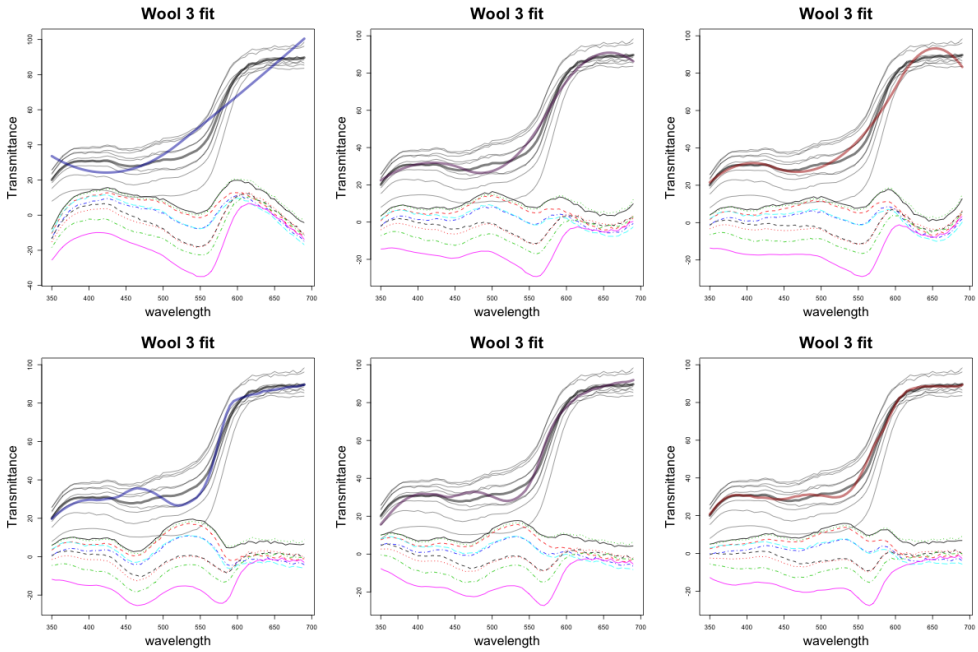


Figure 4.13: Compare fittings of a type of wool using same numbers of B-spline basis functions of order 3 with eigenfunctions obtained from functional principal component analysis. From left to right, the number of basis functions used are between 4 and 6 (inclusive). The first row shows the use of B-spline basis functions and second row show the use of eigenfunctions as basis functions. The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

Based on the fits on one type of wool, using eigenfunctions obtained from fPCA does not seem to outperform that of using B-splines basis functions for $B < 6$.

4.3.3 Conclusion

We choose $B = 6$, $o = 3$ primarily based on AIC and R^2E and the size of our dataset. However, regarding the fits, there are still vertical separation to account for.

4.4 Red cotton fibre data

Both wool and cotton fibre data consist of 20 samples. Nine replicates of MSP spectra were collected for each sample. The red cottons dataset includes data of spectra from 240 to 690 nm (UV-visible spectral range) with intervals of 5 nm. This is different

from wool data due to the fact that wool absorbs UV radiation and therefore there is no informative signal in the UV range comprised here between 240 and 350 nm.

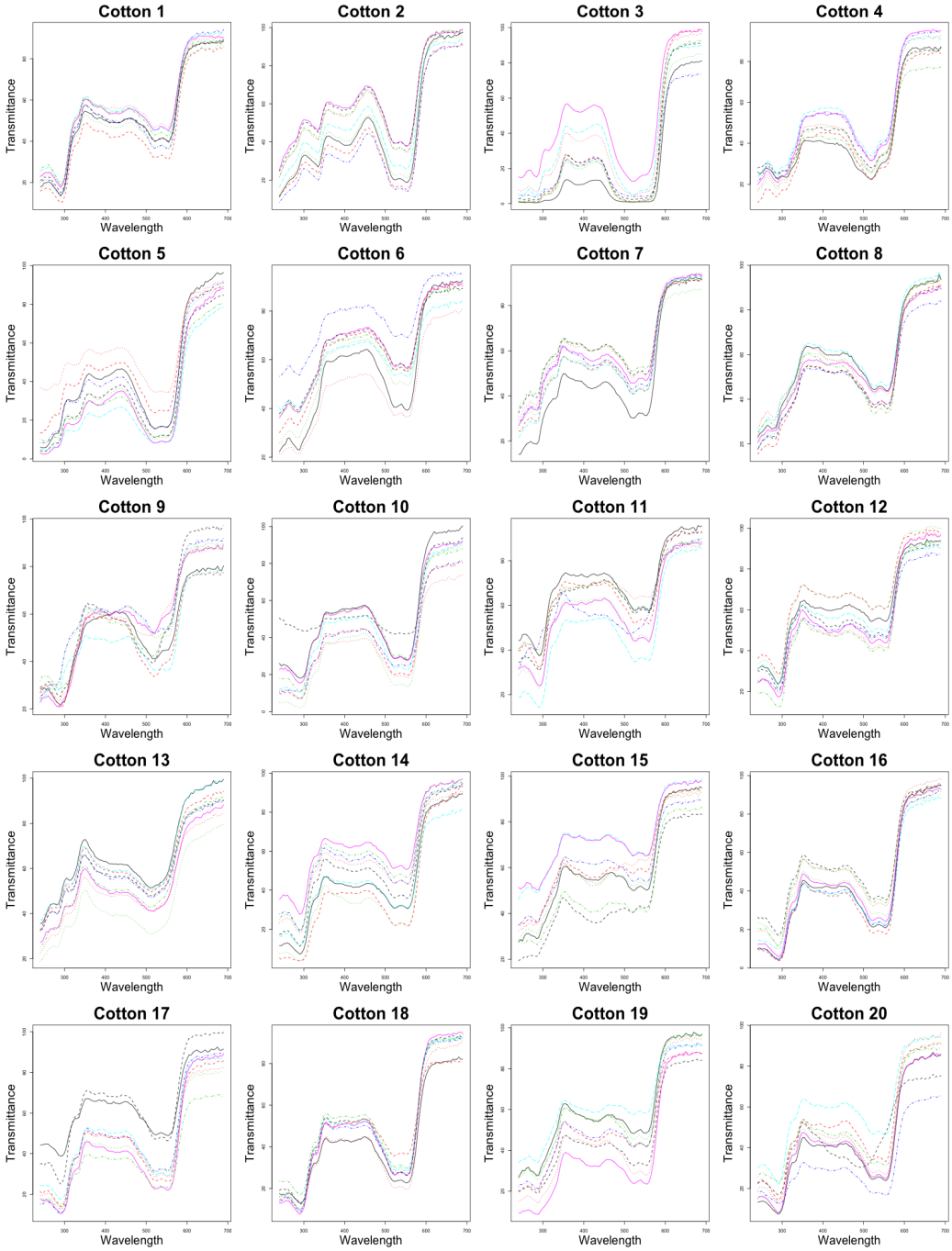


Figure 4.14: Each block shows $n_k = 9$ measurements of a type k of cotton.

Similar to ink data, there is some between-group variation but not as large, and also some within-group variations other than vertical separation. For cotton data, it is harder to distinguish between curves as the main shapes all look similar just like wool data; there is smaller between-group variation in terms of overall shape. All

groups have curves that look like a hat roughly between 240-550 nm with a neck and head from 550 nm onward. These are the main traits we should make note of when modeling the data.

4.4.1 Choosing the number of B-spline basis functions

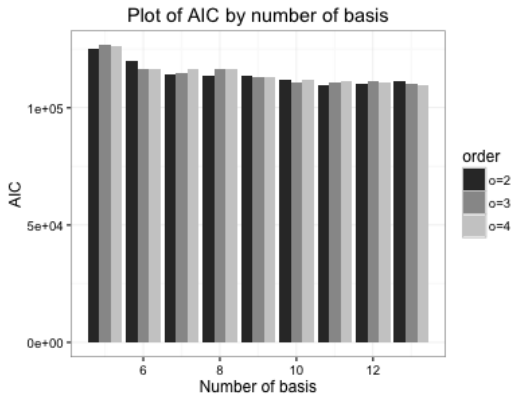
The resulting AIC and R^2E values for cotton data are summarised in Section 4.4.1 and table 4.6 with most optimal values shaded pink. They indicate the most favouring choices based on the decrements and magnitudes in comparison with values nearby.

$B \setminus o$	2	3	4
5	124929	126587	126558
6	120131	116580	116386
7	113982	114936	116700
8	113782	116434	116693
9	113655	113140	112838
10	112126	110965	112094
11	109865	110873	111541
12	110098	111257	110699
13	111078	110444	109671

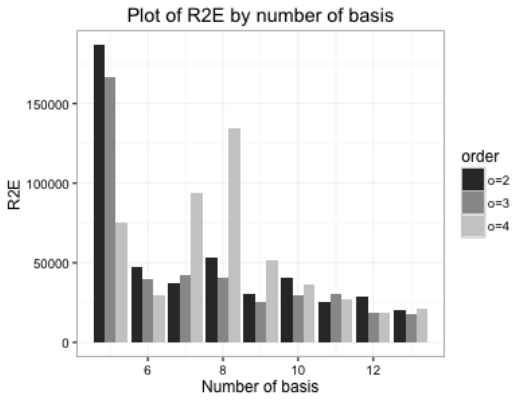
Table 4.5: AIC values for cotton data

$B \setminus o$	2	3	4
5	56373	79508	78746
6	42768	37452	33627
7	31037	28929	30794
8	27233	30651	44482
9	28194	32089	30325
10	29598	27196	29103
11	28240	79758	26426
12	24093	25007	28013
13	25147	26906	27413

Table 4.6: R^2E values for cotton data



(a) AIC values for cotton data in barplot



(b) R^2E values for cotton data in barplot

Figure 4.15: AIC and R^2E values for cotton data in barplots

Again for cotton data, it is quite clear to see $B = 6$, $o = 3$ is optimal given the decrement and consensus given by AIC and R^2E. We will make a decision after checking the fits.

There are some drops at $B = 9$ for R^2E as can be seen in Figure 4.15. However, this is not agreed by AIC and not favourable considering our sample size so we will stick with $B = 6$, $o = 3$.

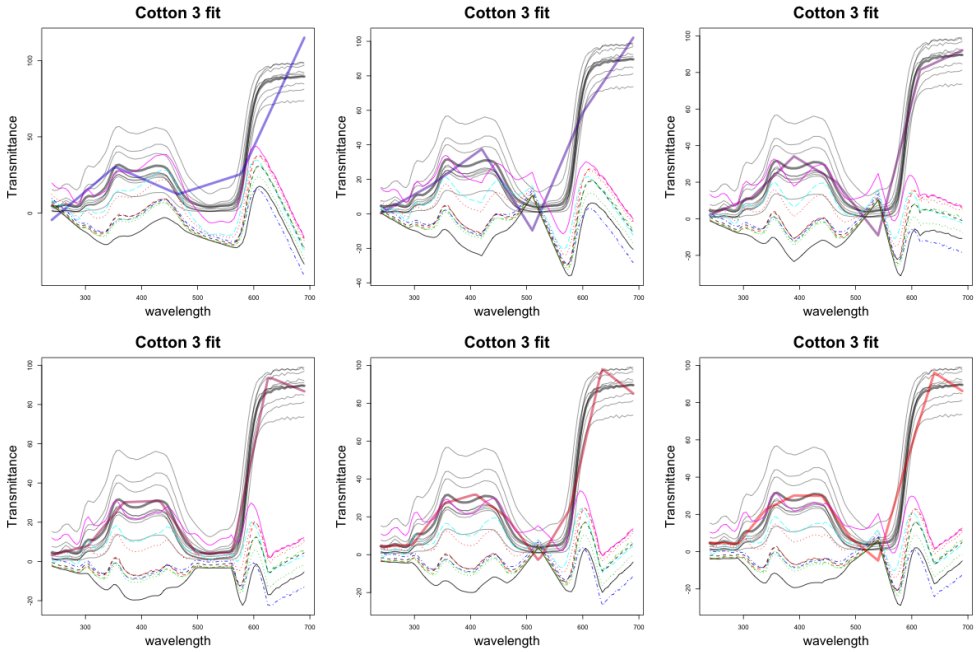


Figure 4.16: Fitting of a type of cotton using different number of B-spline basis functions of order 2. From top left to bottom right, the number of basis functions used are between 5 and 10 (inclusive). The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

Overall fits of order 4 B-spline basis functions do not outperform order 3 B-spline basis functions given the same number of basis. For order 3, different numbers $B = 6, 7, 8, 9, 10$ of B-spline basis all perform similarly so $B = 6$ is still optimal.

4.4.2 Functional principal component analysis

Once B is selected along with o for B-spline basis functions, the fit of using B-spline basis functions will be used to be compared with the use of eigenfunctions.

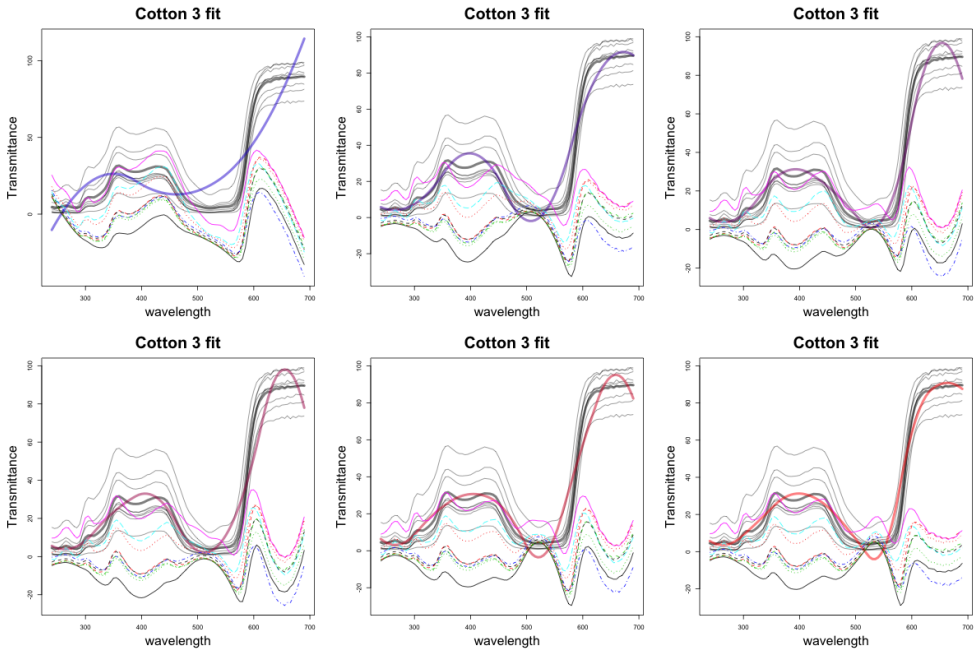


Figure 4.17: Fitting of a type of cotton using different number of B-spline basis functions of order 3. From top left to bottom right, the number of basis functions used are between 5 and 10 (inclusive). The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

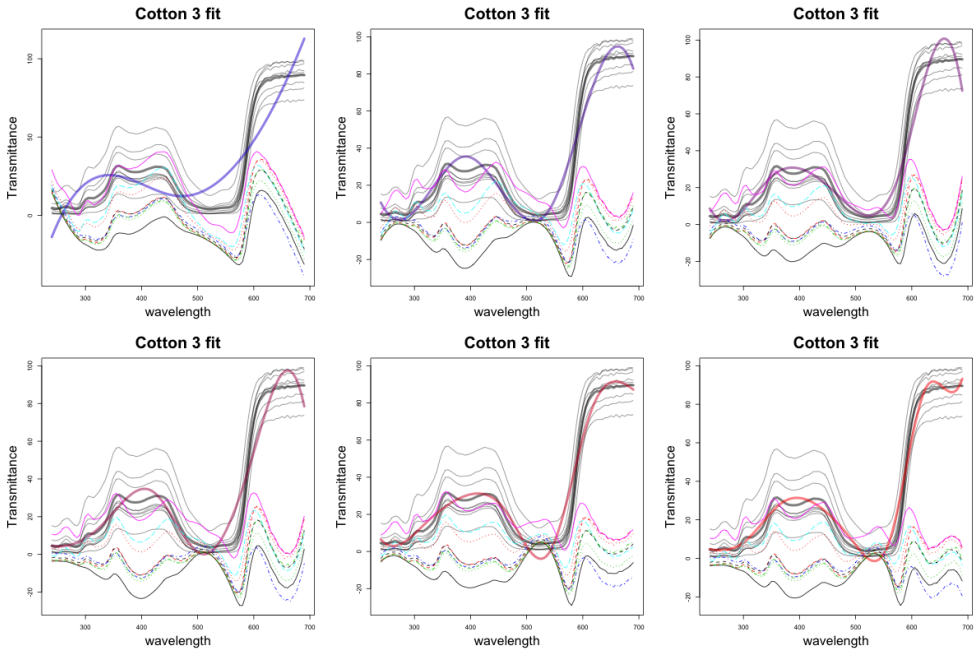


Figure 4.18: Fitting of a type of cotton using different number of B-spline basis functions of order 4. From top left to bottom right, the number of basis functions used are between 5 and 10 (inclusive). The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

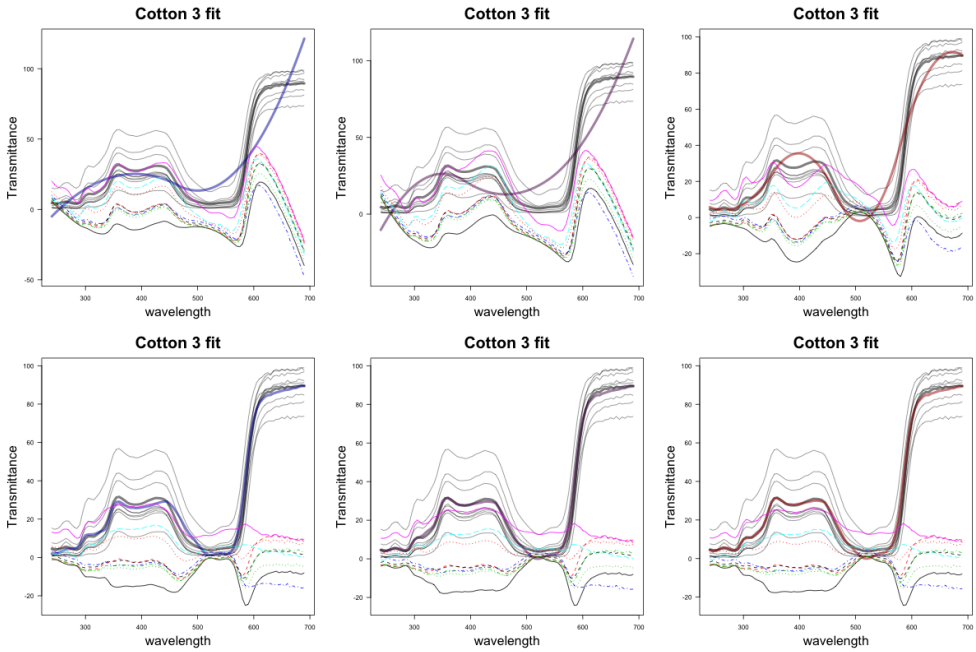


Figure 4.19: Compare fittings of a type of cotton using same numbers of B-spline basis functions of order 3 with eigenfunctions obtained from functional principal component analysis. From left to right, the number of basis functions used are between 4 and 6 (inclusive). The first row shows the use of B-spline basis functions and second row show the use of eigenfunctions as basis functions. The curves underneath are residuals after the fitted (purple) curves are subtracted from the original curves above.

The fits of eigenfunctions obtained from fPCA outperform that of B-spline basis functions in terms of smaller residuals overall as can be seen from 600 nm onward.

4.4.3 Conclusion

The choices we make for cotton data is also $B = 6$, $o = 3$, same as for wool data.

4.5 Conclusion

There are limitations in the number of basis functions we can pick due to the size of our dataset so we do not always pick the ones with possibly the best fits.

Chapter 5

Model fitting and simulations

5.1 Introduction

All models introduced in Chapter 3 will be used to calculate likelihood ratios for all datasets introduced in Chapter 4. Before likelihood ratios are evaluated, various methods are used to check the fits of the models to each dataset including simulation. Model fitting includes checking the distributions of residuals since basis function fitting for the group mean has been done in Chapter 4 for the purpose of dimension reduction. A few methods will be used to check the assumptions of the proposed model. These include boxplots and Chi-squared Q-Q plots presented for both choices of basis functions. After the assumptions have been checked, simulations will be used to replicate datasets based on the parameters estimated under each model and they will be compared visually with the original data for similarities and differences.

5.2 Data exploration

Since our proposed hierarchical models differ primarily by variance-covariance structures and between-group (shape) distributions are all assumed to be multivariate normal, we are interested in checking whether these assumptions are indeed true for the datasets. We will use the notations

$$\boldsymbol{y}_{ki} = x(\boldsymbol{w}) + \boldsymbol{e}_{ki} = \boldsymbol{\Phi}\boldsymbol{\theta}_{ki} + \boldsymbol{r}_{ki}$$

to represent a curve for $1 \leq i \leq n_k = n$ observations for each group k where $1 \leq k \leq K$ and $\boldsymbol{y}_{ki} \in \mathbb{R}^m$. The matrix $\boldsymbol{\Phi}$ is a collection of basis functions $\phi_1 \ldots \phi_b$ evaluated at $\boldsymbol{w}$. The parameter θ_{ki} depends sometimes on k alone and between-group distributions for θ_k are assumed to be multivariate normal for all proposed models.

5.2.1 Within-group covariance and residuals

When assuming $\boldsymbol{y}_{ki} = \boldsymbol{\Phi}\boldsymbol{\theta}_k + \boldsymbol{r}_{ki}$ it is not possible to estimate $\boldsymbol{\Sigma}$ =var($\boldsymbol{y}_{ki}$) for small datasets since $\boldsymbol{\Sigma}$ is a positive semi-definite matrix of size m by m and m is usually much greater than sample size n.

Boxplots of $\hat{\boldsymbol{r}}_{kij}$ will be drawn both by element (j) and group (k) separately to see the distributions within the same groups and across different groups.

5.2.2 Between-group distribution

First $\boldsymbol{\theta}_k$ will be estimated using ordinary least square, i.e., $\hat{\boldsymbol{\theta}}_k = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T \sum_{i=1}^n \boldsymbol{y}_{ki}/n = \boldsymbol{\Phi}^T \sum_{i=1}^n \frac{\boldsymbol{y}_{ki}}{n}$ for orthonormal basis functions used. Boxplots and pairwise scatter plots will be used to show the magnitudes and distribution of each element of $\hat{\boldsymbol{\theta}}_{ki}^{(b)}$ in comparison with one another. Chi-squared Q-Q plots will also be used. They are multivariate version of qq-plots that plots the squared Mahalanobis distance, that is the multivariate generalization of z-score, against its Chi-squared quantile. A straight line is expected for data that follows a multivariate normal distribution.

5.3 Simulation

Simulations are used to check the fit of the models by replicating the generating mechanism that our datasets are assumed to have come from. Thus, given accurate estimates of model parameters, if the models are valid, should produce data that are similar to our original data. In order to simulate these datasets, estimation of parameters are obtained using formulas given in relevant sections in Chapter 3 and data are generated using processes as indicated by the schematic representations, also in the relevant sections in Chapter 3 for the same numbers of replicates n and types K as the original data. For the purpose of examining model fit in this chapter, only a selection of

4 groups of the simulated data will be drawn for illustration purposes, they will be named as [data] g in the figures where g $\in \{1, 2, 3, 4\}$ indicate curves' group membership. These have nothing to do with original data with the same group number as drawn in Chapter 4. Using models introduced in Chapter 3, datasets will be generated in the following orders specified for each model.

5.3.1 Simplified multivariate normal random-effects model

Under this model, datasets will be generated by first simulate group means $\boldsymbol{\theta}_k \sim N(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{D}})$ then $\boldsymbol{y}_{ki} \sim N(\boldsymbol{\Phi}\boldsymbol{\theta}_k, \hat{\sigma}^2\boldsymbol{I}_m)$ for $1 \leq k \leq K$ and $1 \leq i \leq n_k$. Details of this model can be found in Section 3.2.2 and CR-S will be used to refer to this model.

5.3.2 Constant within-group variance model

Under this model, datasets will be generated by first simulate group variances $\sigma_k^2 \sim Inv{-}Gam(\hat{\gamma}, \hat{\delta})$ then group means $\boldsymbol{\theta}_k \sim N(\hat{\boldsymbol{\eta}}, \sigma_k^2\hat{\boldsymbol{C}})$ and finally, $\boldsymbol{y}_{ki} \sim N(\boldsymbol{\Phi}\boldsymbol{\theta}_k, \sigma_k^2\boldsymbol{I}_m)$ for $1 \leq k \leq K$ and $1 \leq i \leq n_k$. Details of this model can be found in Section 3.2.3 and CR-const. will be used to refer to this model.

5.3.3 Multivariate normal random-effects with autoregressive within-group covariance model

Under this model, datasets will be generated by first simulate group variances $\sigma_k^2 \sim Inv{-}Gam(\hat{\gamma}, \hat{\delta})$ then group means $\boldsymbol{\theta}_k \sim N(\hat{\boldsymbol{\eta}}, \sigma_k^2\hat{\boldsymbol{C}})$ and finally, $\boldsymbol{y}_{ki} \sim N(\boldsymbol{\Phi}\boldsymbol{\theta}_k, \sigma_k^2\boldsymbol{P})$ for $1 \leq k \leq K$ and $1 \leq i \leq n_k$. Details of this model can be found in Section 3.2.4 and CR-ar will be used to refer to this model.

5.3.4 Dimension reduced multivariate normal random-effects model

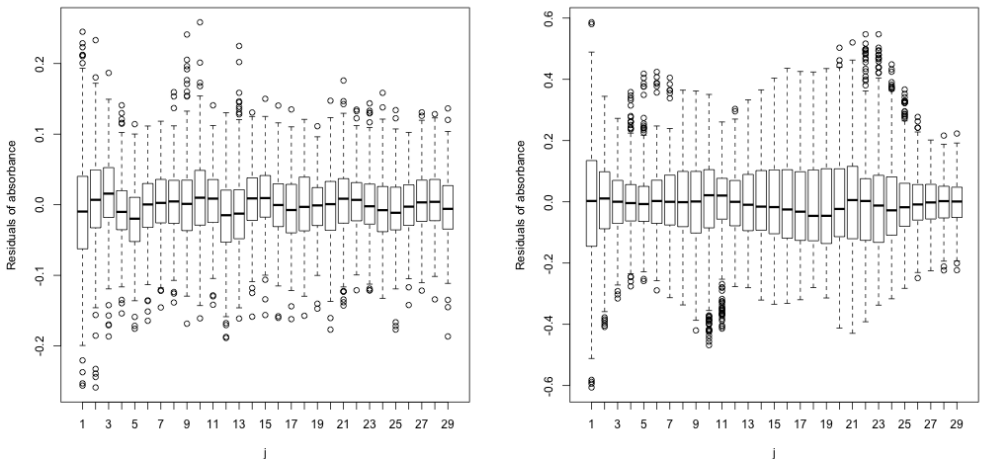
Under this model, datasets will be generated by first simulate group means $\boldsymbol{\theta}_k \sim N(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{C}})$ then $\boldsymbol{z}_{ki} \sim N(\boldsymbol{\theta}_k, \hat{\boldsymbol{U}})$ for $1 \leq k \leq K$ and $1 \leq i \leq n_k$. To compare with original data we will reconstruct $\hat{\boldsymbol{y}}_{ki}$ as $\boldsymbol{\Phi}\boldsymbol{z}_{ki}$. Details of this model can be found in Section 3.3.1 and DR-S will be used to refer to this model.

5.4 Ink data

Sample of ink data consists of $K = 40$ groups of $n = n_k = 10$ MSP measurements of absorbance $\boldsymbol{y}_{ki}$ versus wavelength for $1 \leq i \leq n$ for all k. Absorbance are measured at wavelengths ranging from 380-800 nm with intervals of 1nm so using all the points, that is, taking interval or $int = 1$, the total number of points, the dimension of our data, is $m = 421$.

5.4.1 Residuals

Elementwise residuals are displayed when different basis functions are used to show difference in the distributions. The use of $int = 15$, or taking every 15^{th} point, results in 29 points left from $m = 421$ for the ease of illustration.



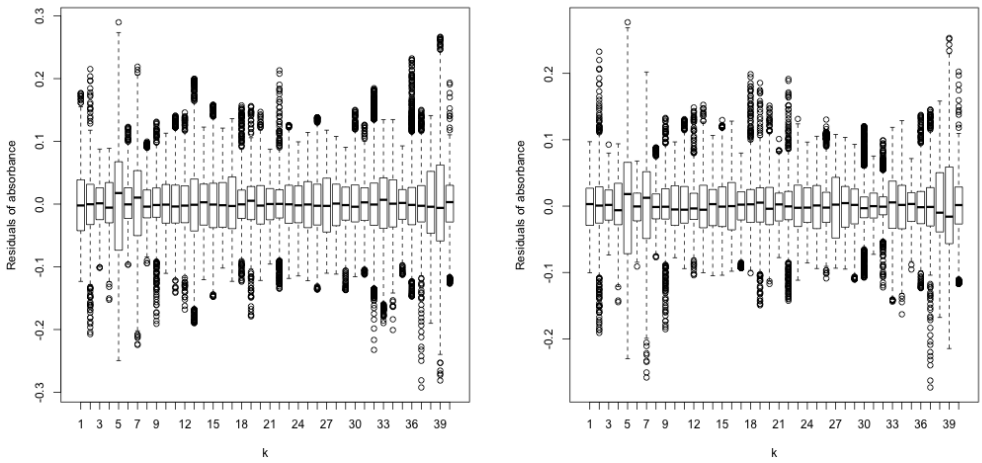
(a) Boxplots for elements of $\hat{\boldsymbol{r}}_{ki}$ for all Kn curves for 9 B-spline basis functions and interval $int = 15$.

(b) Boxplots for elements of $\hat{\boldsymbol{r}}_{ki}$ for all Kn curves for 9 eigenfunctions and interval $int = 15$.

Figure 5.1: Boxplots of elements of $\hat{\boldsymbol{r}}_{ki}$ for 9 basis functions of different choice where $int = 15$.

Each boxplot in Figure 5.1 is for an element of $\boldsymbol{r}_{ki}$ for all $1 \leq i \leq n = 10$ for all $1 \leq k \leq K = 40$ (400 each). Each boxplot in Figure 5.2 is for all m elements of $\boldsymbol{r}_{ki}$ for all $1 \leq i \leq n = 10$ curves for a given group k where $1 \leq k \leq K = 40$ ($10 \times m$ each). Based on Figure 5.1 it seems that variances of elements of $\hat{\boldsymbol{r}}_{ki}$ vary

more across different groups when eigenfunctions obtained from fPCA are used in comparison with B-spline basis functions used. However, Figure 5.2 shows quite the opposite but the differences in variation are smaller. This can be explained by the better fits in Figure 4.7.



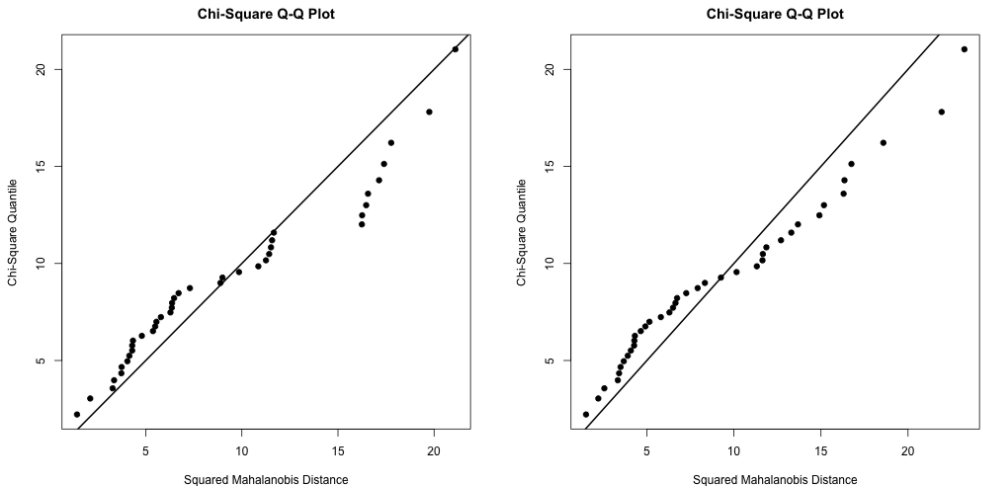
(a) Boxplots of $\hat{\boldsymbol{r}}_{ki}$ for all m points on n curves within group k for 9 B-spline basis functions and interval $int = 1$.

(b) Boxplots of $\hat{\boldsymbol{r}}_{ki}$ for all m points on n curves within group k for 9 eigenfunctions and interval $int = 1$.

Figure 5.2: Boxplots of $\hat{\boldsymbol{r}}_{ki}$ by group for 9 basis functions of different choices where $int = 1$.

5.4.2 Between-group distribution for ink data

Chi-squared Q-Q plots are multivariate version of qq-plots that plots the squared Mahalanobis distance, that is the multivariate generalization of z-score, against its Chi-squared quantile. A straight line is expected for data that follows a multivariate normal distribution.

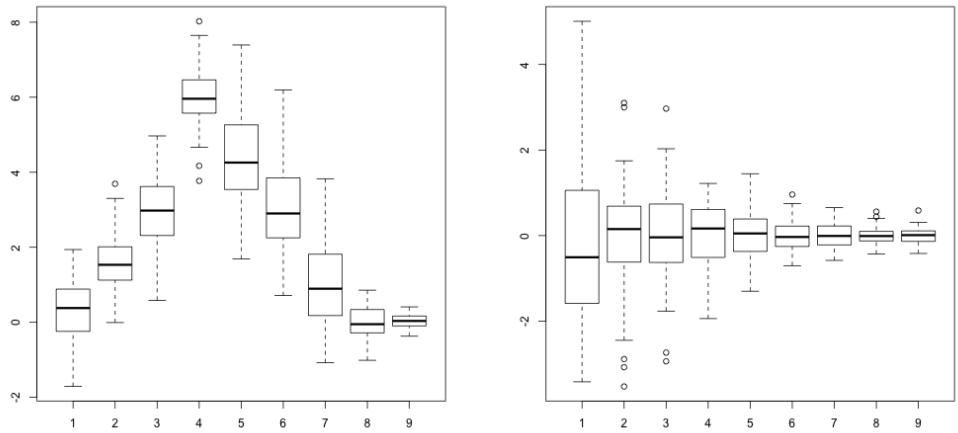


(a) Chi-squared Q-Q plot of fitted $\boldsymbol{\theta}_k$ for 9 B-spline basis functions of order 3 where $int = 1$.

(b) Chi-squared Q-Q plot of fitted $\boldsymbol{\theta}_k$ for 9 eigenfunctions obtained from fPCA used where $int = 1$.

Figure 5.3: Chi-squared Q-Q plots of fitted $\boldsymbol{\theta}_k$ for 9 basis functions of different choices used whertr interval $int = 1$.

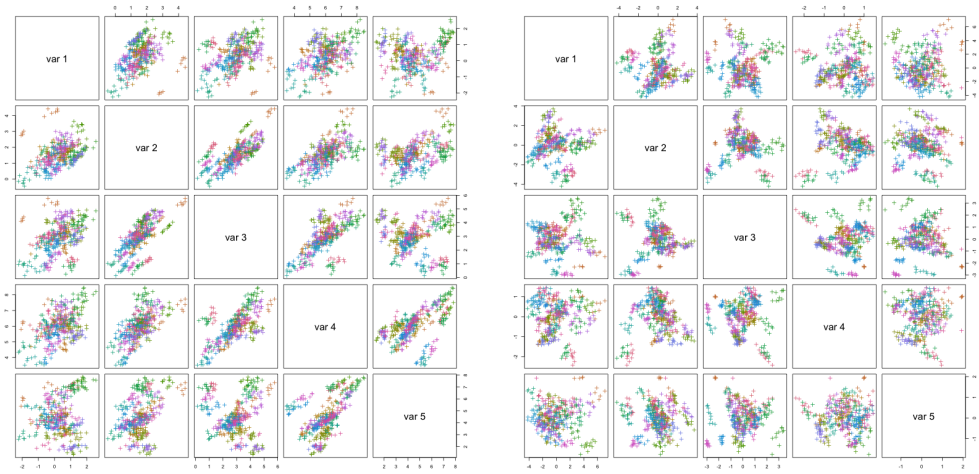
Based on Figure 5.3, the fitted $\boldsymbol{\theta}_k$ might not follow multivariate normal distribution.



(a) Boxplots of fitted $\boldsymbol{\theta}_k$ for 9 B-spline basis functions of order 3 where $int = 1$.

(b) Boxplots of fitted $\boldsymbol{\theta}_k$ for 9 eigenfunctions obtained from fPCA where $int = 1$.

Figure 5.4: Boxplots of fitted $\boldsymbol{\theta}_k$ for 9 basis functions of different choices where $int = 1$.



(a) Pairwise scatter plot of the first 5 elements of the fitted $\boldsymbol{\theta}_k$ when 9 B-spline basis functions of order 3 are used.

(b) Pairwise scatter plot of the first 5 elements of the fitted $\boldsymbol{\theta}_k$ when 9 eigenfunctions obtained from fPCA are used.

Figure 5.5: Pairwise scatter plot of the first 5 elements in $\boldsymbol{Z}_k$ for 6 basis functions of different choices where $int = 1$.

From the Chi-squared Q-Q plot and pairwise scatterplots it is pretty clear that fitted $\boldsymbol{\theta}_k$ for different choices of basis functions do not follow the same distribution.

5.4.3 Simulation - CA-S for ink data

Simulated ink data using CA-S with parameters estimated from the original data and data generated for the same number of replicates.

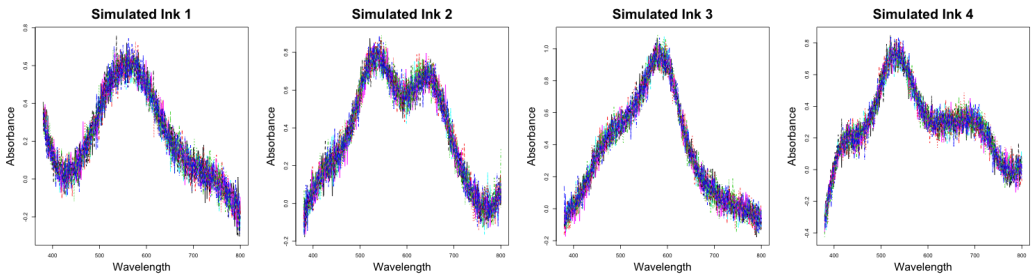


Figure 5.6: Each block shows $n_k = 10$ measurements of a type k of ink data simulated under simplified multivariate normal random-effects model. Refer to Section 5.3.1 for details.

Curves within the same group are centred at the same mean (pointwise) which does not resemble original data that have vertical separations. However, in terms of shapes where there are peaks and shoulders, these simulated data do resemble those of the original data.

5.4.4 Simulation - CA-const. for ink data

Simulated ink data using CA-const. with parameters estimated from the original data and data generated for the same number of replicates.

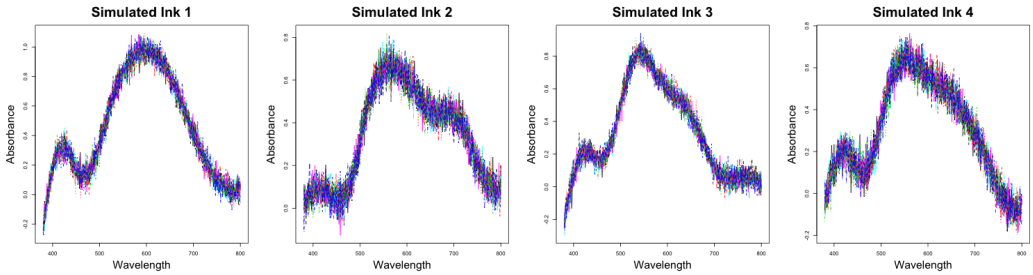


Figure 5.7: Each block shows $n = 10$ measurements for each of 4 types of ink data simulated under constant within-group variance model. Refer to Section 5.3.2 for details.

This model differs to CA-S primarily by the relaxation of constant between-group variances but this difference is not very obvious from these plots. These plots also show that the model fails to model the separation of curves within groups.

5.4.5 Simulation - CA-ar for ink data

Simulated ink data using CA-ar with parameters estimated from the original data and data generated for the same number of replicates.

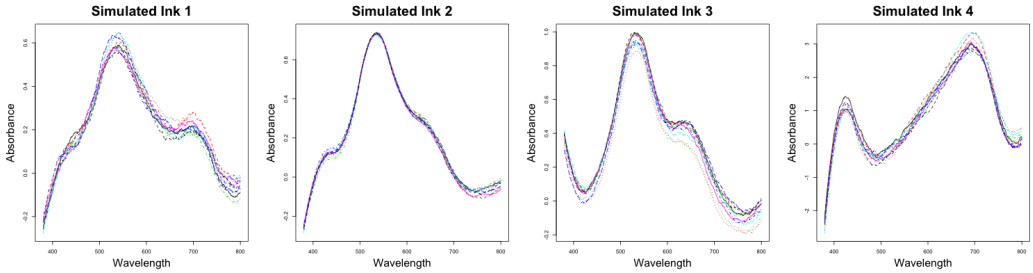


Figure 5.8: Each block shows $n = 10$ measurements for each of 4 types of ink data simulated under multivariate normal random-effects with autoregressive within-group covariance model. Refer to Section 5.3.3 for details.

By assuming autoregressive lag-1 errors, the error part looks more continuous due to high correlations thus resulted in separation of curves that make them resemble the original data (ink).

5.4.6 Simulation - DR-S for ink data

Simulated ink data using DR-S with parameters estimated from the original data and data generated for the same number of replicates.

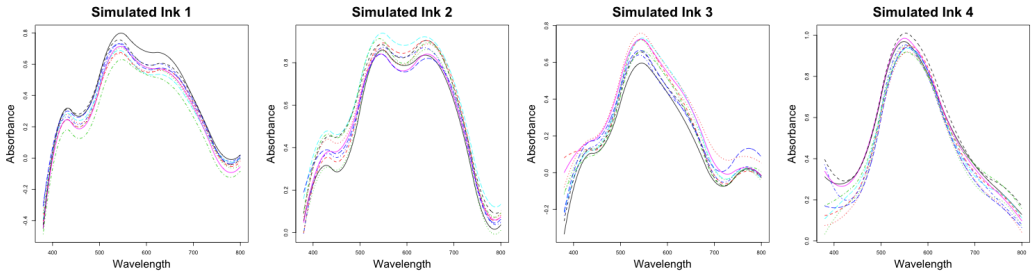


Figure 5.9: Each block shows $n = 10$ measurements for each of 4 types of ink data simulated under dimension reduced multivariate normal random-effects model. Refer to Section 5.3.4 for details.

This model successfully captures the overall shape and separation of curves that is somehow dependent on the position or magnitude of absorbance. Since this model generates curves similar to our original data, we expect it to give likelihood ratios that can be helpful in distinguishing between groups.

5.4.7 Conclusion

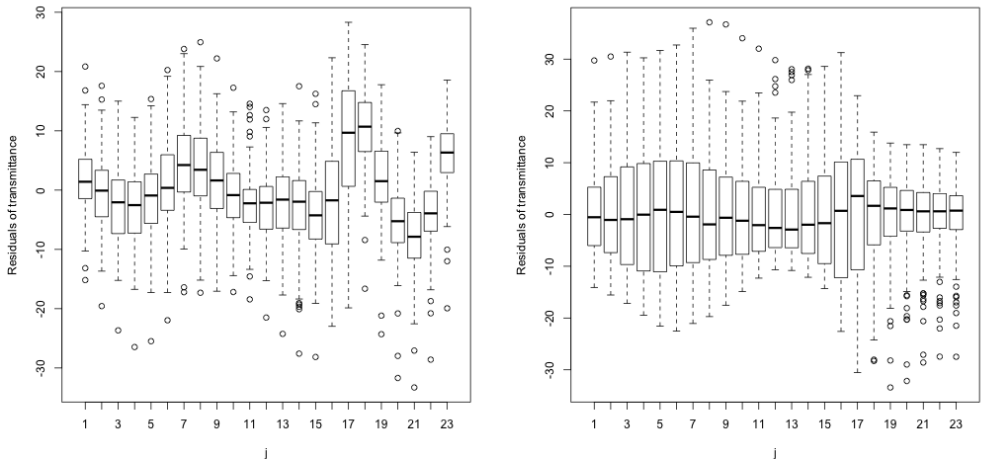
Simulated ink data successfully capture the shape and the variations in overall shape that can be used to distinguish between groups. However, simplified multivariate normal random-effects model and constant within-group variance model fail to model within-group variations as expected as independent variances at each point is assumed. However, multivariate normal random-effects with autoregressive within-group covariance model successfully models the slight separation of curves within the same groups and behaves like dimension reduced multivariate normal random-effects model which only models the shape. Overall, modeling of the variations within groups do not look necessary for ink data.

5.5 Wool data

Sample of wool data consists of $K = 20$ groups of $n = n_k = 9$ MSP measurements of transmittance $\boldsymbol{y}_{ki}$ versus wavelength for $1 \leq i \leq n$ for all k. Transmittance are measured at wavelengths ranging from 350-690 nm with intervals of 5 nm so using all the points, that is, taking interval or $int = 1$, the total number of points, the dimension of our data, is $m = 69$. An interval int of 2 means every observation is 10 nm apart.

5.5.1 Residuals

Elementwise residuals are displayed when different basis functions are used to show difference in the distributions. The use of $int = 3$, or taking every $3^r d$ point, results in 31 points left from $m = 69$ for the ease of illustration. Each boxplot in Figure 5.10 is for an element of $\boldsymbol{r}_{ki}$ for all curves $1 \leq i \leq n = 9$ for all groups $1 \leq k \leq K = 20$. Each boxplot in Figure 5.11 is for all elements of $\boldsymbol{r}_{ki}$ for all curves $1 \leq i \leq n = 10$ for a given group $1 \leq k \leq K = 40$.

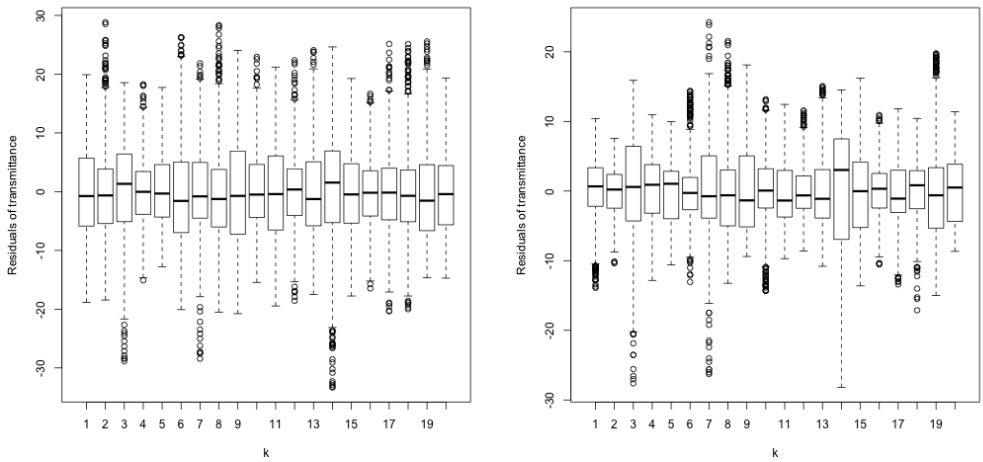


(a) Boxplots for elements of $\hat{\boldsymbol{r}}_{ki}$ for all Kn curves for 6 B-spline basis functions and interval $int = 3$.

(b) Boxplots for elements of $\hat{\boldsymbol{r}}_{ki}$ for all Kn curves for 6 eigenfunctions from fPCA and interval $int = 3$.

Figure 5.10: Boxplots of elements of $\hat{\boldsymbol{r}}_{ki}$ for 9 basis functions of different choice where $int = 1$.

Like ink data, elementwise boxplots show greater variations when eigenfunctions obtained from fPCA are used compared to B-spline basis functions used.



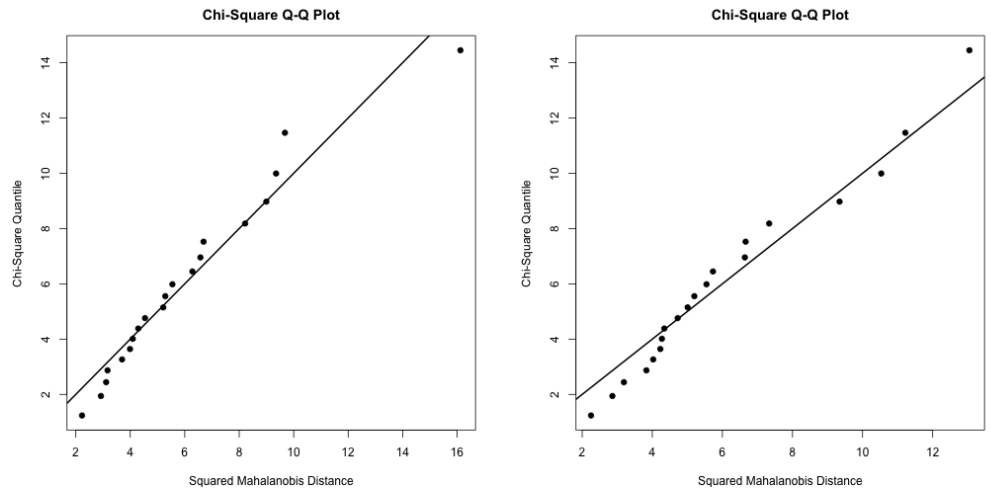
(a) Boxplots of $\hat{\boldsymbol{r}}_{ki}$ for all m points on n curves within group k for 6 B-spline basis functions where $int = 1$.

(b) Boxplots of $\hat{\boldsymbol{r}}_{ki}$ for all m points on n curves within group k for 6 eigenfunctions from fPCA where $int = 1$.

Figure 5.11: Boxplots of $\hat{\boldsymbol{r}}_{ki}$ by group for 6 basis functions of different choice where $int = 1$.

5.5.2 Between-group distribution for wool data

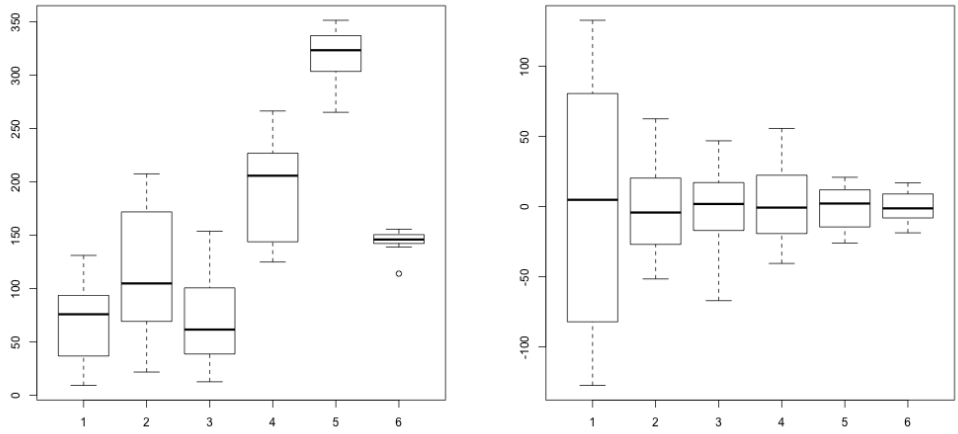
Chi-squared Q-Q plots are multivariate version of qq-plots that plots the squared Mahalanobis distance, that is the multivariate generalization of z-score, against its Chi-squared quantile. A straight line is expected for data that follows a multivariate normal distribution.



(a) Chi-squared Q-Q plot of fitted $\boldsymbol{\theta}_k$ for 6 B-spline basis functions of order 3 where $int = 1$.

(b) Chi-squared Q-Q plot of fitted $\boldsymbol{\theta}_k$ for 6 eigenfunctions obtained from fPCA used where $int = 1$.

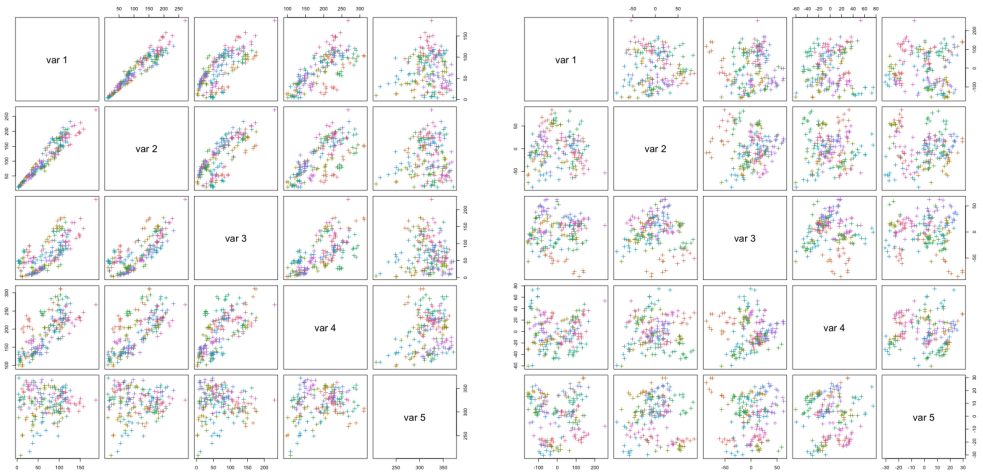
Figure 5.12: Chi-squared Q-Q plots of fitted $\boldsymbol{\theta}_k$ for 6 basis functions of different choices used whertr interval $int = 1$.



(a) Boxplots of fitted coefficients $\boldsymbol{\theta}_k$ for 6 B-spline basis functions of order 3 where $int = 1$.

(b) Boxplots of fitted coefficients $\boldsymbol{\theta}_k$ for 6 eigenfunctions from fPCA and interval $int = 1$.

Figure 5.13: Boxplots of fitted coefficients $\boldsymbol{\theta}_k$ when $B = 6$ for different choices of basis functions and $int = 1$.



(a) Pairwise scatter plot of the first 5 elements of the fitted $\boldsymbol{\theta}_k$ when 6 B-spline basis functions of order 3 are used.

(b) Pairwise scatter plot of the first 5 elements of the fitted $\boldsymbol{\theta}_k$ when 6 eigenfunctions obtained from fPCA are used.

Figure 5.14: Pairwise scatter plot of the first 5 elements of fitted $\boldsymbol{\theta}_k$ for different choices of basis functions.

From the Chi-squared Q-Q plot and pairwise scatterplots it can be seen that the fitted

$\boldsymbol{\theta}_k$ probably follows multivariate normal distribution. However, the fitted coefficients obtained using different choices of basis functions follow different distributions.

5.5.3 Simulation - CA-S for wool data

Simulated wool data using CA-S with parameters estimated from the original data and data generated for the same number of replicates.

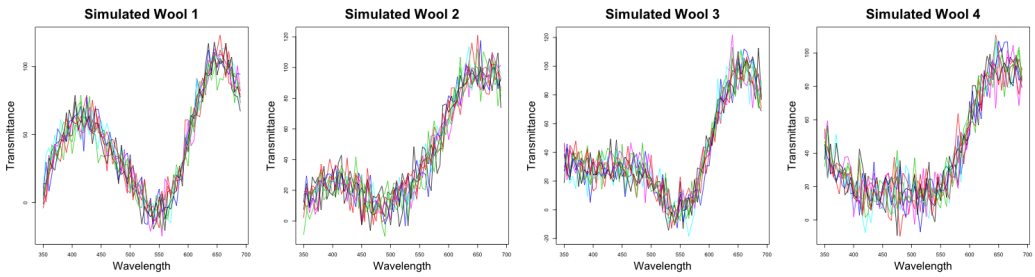


Figure 5.15: Each block shows $n = 9$ measurements for each of 4 types of wool data simulated under simplified multivariate normal random-effects model. Refer to Section 5.3.1 for details.

This model seems to capture the shape of wool data, that is, a check mark; however, the curves all seem to be centred at one place with variation around it at each point (wavelength). This is consistent with our model assumption but not our data, as expected. There are some variation between groups in terms of shape.

5.5.4 Simulation - CA-const. for wool data

Simulated wool data using CA-const. with parameters estimated from the original data and data generated for the same number of replicates.

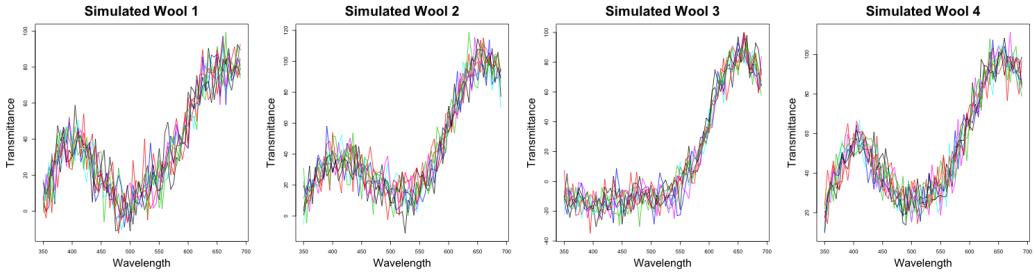


Figure 5.16: Each block shows $n = 9$ measurements for each of 4 types of wool data simulated under constant within-group variance model. Refer to Section 5.3.2 for details.

This model is very similar to CA-S with not quite noticeable change of variance at each point between different groups. The variation among shapes are similar to that of CA-S so there is no noticeable effect of using a diagonal between-group variance-covariance structure.

5.5.5 Simulation - CA-ar for wool data

Simulated wool data using CA-ar with parameters estimated from the original data and data generated for the same number of replicates.

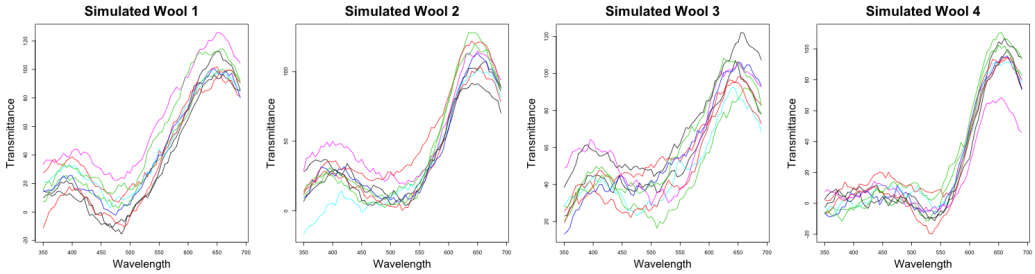


Figure 5.17: Each block shows $n = 9$ measurements for each of 4 types of wool data simulated under multivariate normal random-effects with autoregressive within-group covariance model. Refer to Section 5.3.3 for details.

There are some separation of curves as we want.

5.5.6 Simulation - DR-S for wool data

Simulated wool data using DR-S with parameters estimated from the original data and data generated for the same number of replicates.

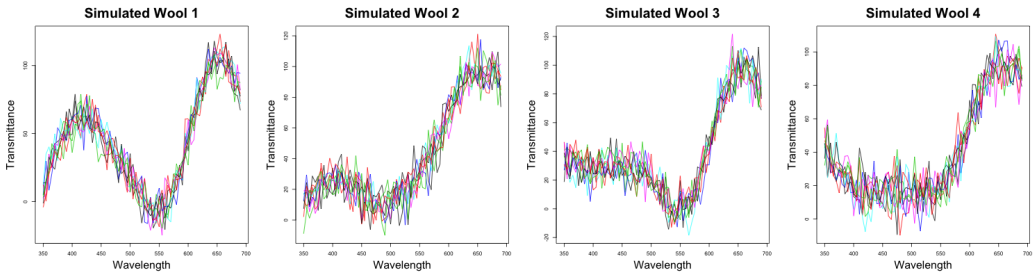


Figure 5.18: Each block shows $n = 9$ measurements for each of 4 types of wool data simulated under dimension reduced multivariate normal random-effects model. Refer to Section 5.3.4 for details.

All curves are separated with larger between-group variations comparing to the previous models, which is also as expected given our model assumptions.

5.5.7 Conclusion

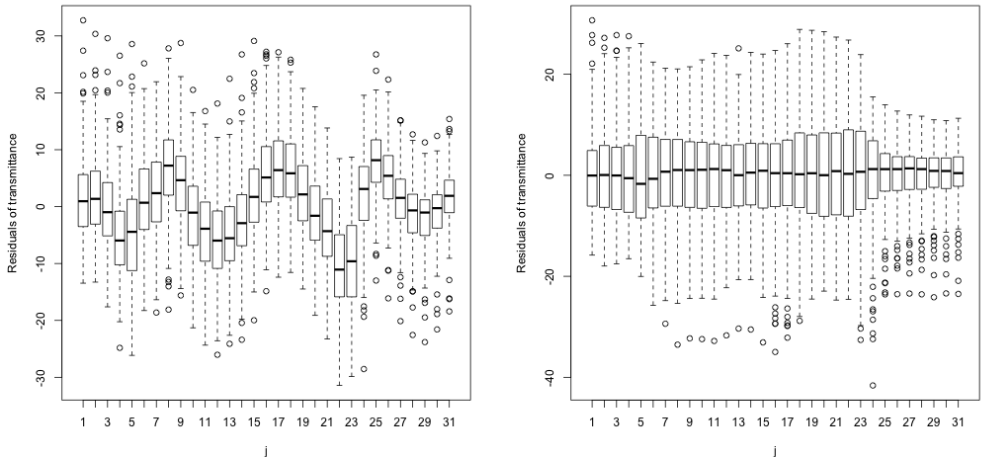
The models perform as expected; however, the larger variation within-groups compared to between-groups for wool data makes it harder to distinguish in comparison with ink data. Moreover, all models seem to fail to model the shape of the curves at wavelength above 650 nm. This could be due to the B-spline basis functions used to model the shape.

5.6 Cotton data

Sample of cotton data consists of $K = 20$ groups of $n = n_k = 9$ MSP measurements of transmittance $\boldsymbol{y}_{ki}$ versus wavelength for $1 \leq i \leq n$ for all k. Transmittance are measured at wavelengths ranging from 240-690 nm with intervals of 5 nm so using all the points, that is, taking interval or $int = 1$, the total number of points, the dimension of our data, is $m = 91$. An interval int of 2 means every observation is 10 nm apart.

5.6.1 Residuals

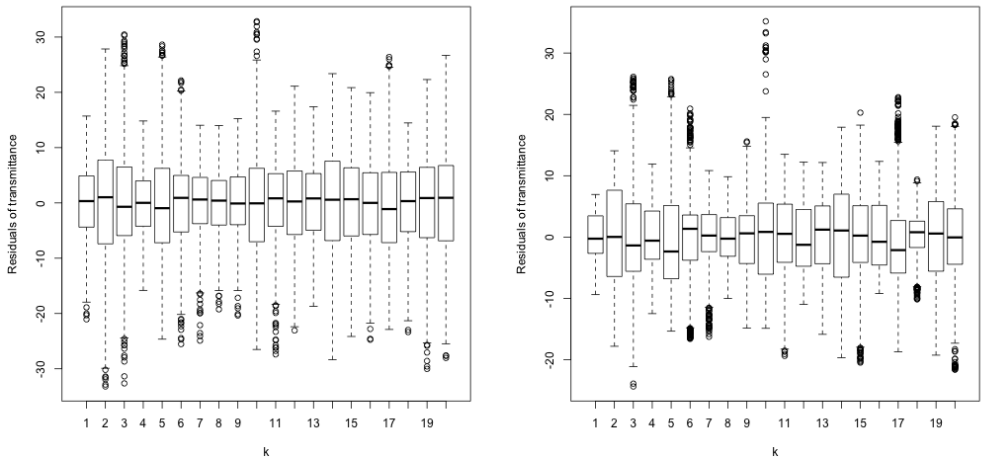
Elementwise residuals are displayed when different basis functions are used to show difference in the distributions. The use of $int = 3$, or taking every $3^r d$ point, results in 31 points left from $m = 91$ for the ease of illustration.



(a) Boxplots for elements of $\hat{\boldsymbol{r}}_{ki}$ for all Kn curves for 6 B-spline basis functions where $int = 3$.

(b) Boxplots for elements of $\hat{\boldsymbol{r}}_{ki}$ for all Kn curves for 6 eigenfunctions obtained from fPCA where $int = 3$.

Figure 5.19: Boxplots of $\hat{\boldsymbol{r}}_{ki}$ when 6 B-spline basis functions of order 3 are used.



(a) Boxplots of $\hat{\boldsymbol{r}}_{ki}$ for all m points on n curves within group k for 6 B-spline basis functions where $int = 1$.

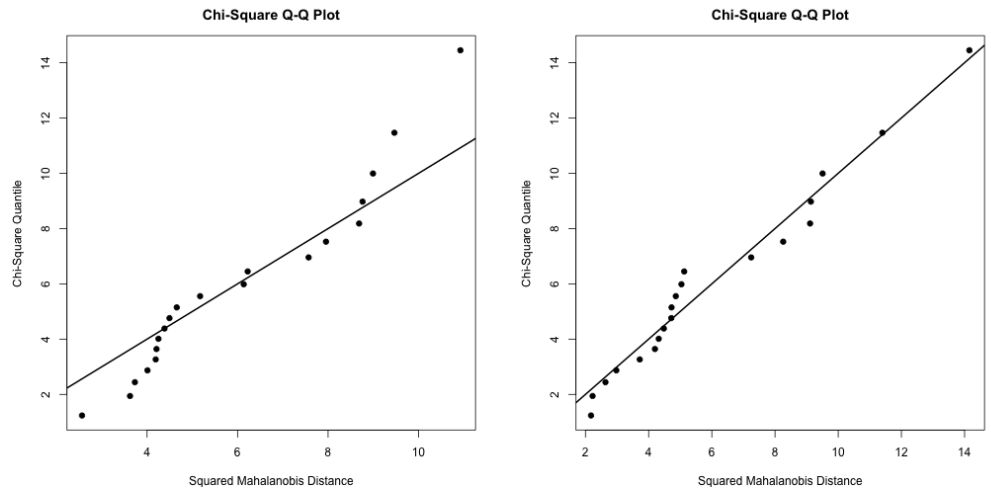
(b) Boxplots of $\hat{\boldsymbol{r}}_{ki}$ for all m points on n curves within group k for 6 eigenfunctions obtained from fPCA where $int = 1$.

Figure 5.20: Boxplots of $\hat{\boldsymbol{r}}_{ki}$ by group for 6 basis functions of different choice where $int = 1$.

Each boxplot within Figure 5.19 is for an element of $\boldsymbol{r}_{ki}$ for all $1 \leq i \leq n = 9$ for all $1 \leq k \leq K = 20$. From this plot we can tell variances varies within-group. Each boxplot within Figure 5.20 is for an element of $\boldsymbol{r}_{ki}$ for all $1 \leq i \leq n = 10$ for all $1 \leq k \leq K = 40$.

5.6.2 Between-group distribution for cotton data

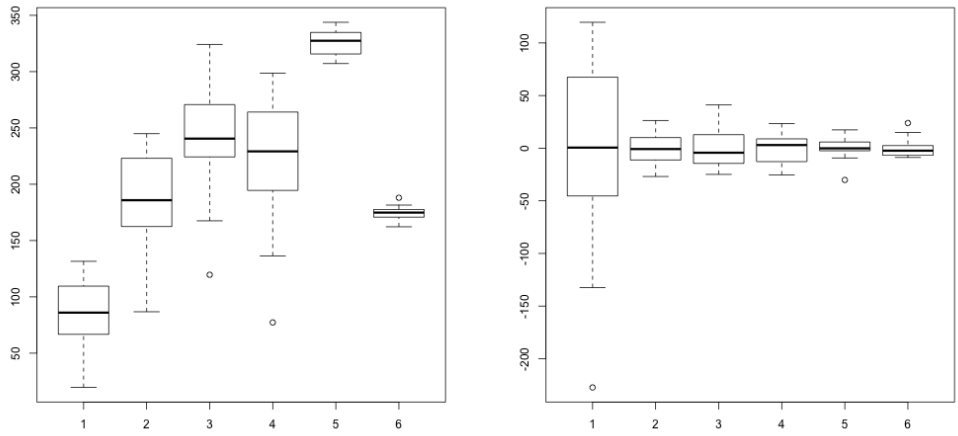
Chi-squared Q-Q plots are multivariate version of qq-plots that plots the squared Mahalanobis distance, that is the multivariate generalization of z-score, against its Chi-squared quantile. A straight line is expected for data that follows a multivariate normal distribution.



(a) Chi-squared Q-Q plot of fitted $\boldsymbol{\theta}_k$ for 6 B-spline basis functions of order 3 used where $int = 1$.

(b) Chi-squared Q-Q plot of fitted $\boldsymbol{\theta}_k$ for 6 eigenfunctions obtained from fPCA used where $int = 1$.

Figure 5.21: Chi-squared Q-Q plots of fitted $\boldsymbol{\theta}_k$ for 6 basis functions of different choices used where interval $int = 1$.

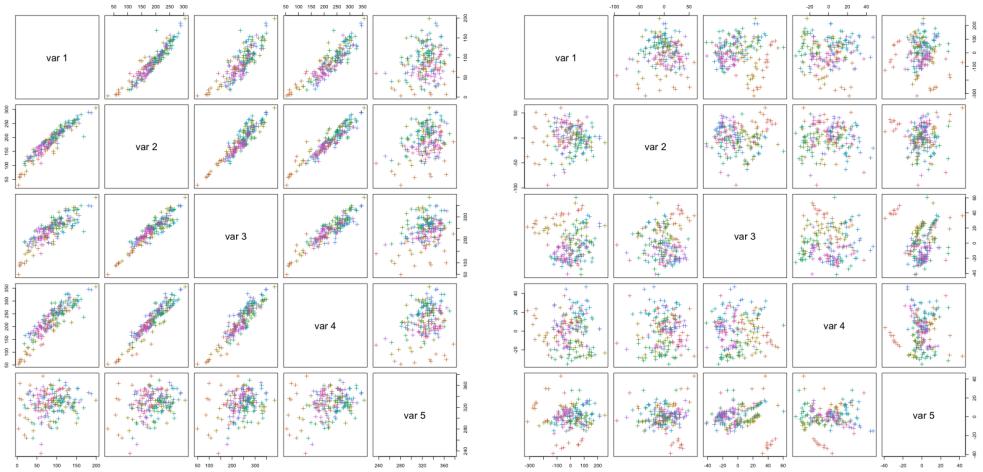


(a) Boxplots of fitted coefficients $\boldsymbol{\theta}_k$ for 6 B-spline basis functions of order 3 where $int = 1$.

(b) Boxplots of fitted coefficients $\boldsymbol{\theta}_k$ for 6 eigenfunctions from fPCA where $int = 1$.

Figure 5.22: Boxplots of fitted coefficients $\boldsymbol{\theta}_k$ when $B = 6$ for different choices of basis functions and $int = 1$.

Chi-squared Q-Q plots and boxplots of fitted coefficients $\boldsymbol{\theta}_k$ suggest that they might follow a multivariate normal distribution for when eigenfunctions obtained from fPCA are used.



(a) Pairwise scatter plot of the first 5 elements of the fitted $\boldsymbol{\theta}_k$ when 6 B-spline basis functions of order 3 are used where $int = 1$.

(b) Pairwise scatter plot of the first 5 elements of the fitted $\boldsymbol{\theta}_k$ when 6 eigenfunctions obtained from fPCA are used where $int = 1$.

Figure 5.23: Pairwise scatter plots of fitted $\boldsymbol{\theta}_k$ under different basis functions where $int = 1$.

From the boxplots and pairwise scatterplots it is pretty clear that fitted $\boldsymbol{\theta}_k$ for different choice of basis functions used do not follow the same distribution. There are higher correlations for those fitted when B-spline basis functions are used and the other ones (fitted coefficients when eigenfunctions from fPCA are used) are centred around zero.

5.6.3 Simulation - CA-S for cotton data

Simulated cotton data using CA-S with parameters estimated from the original data and data generated for the same number of replicates.

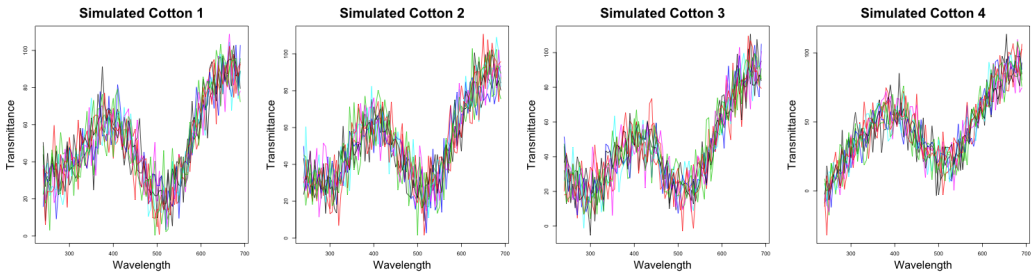


Figure 5.24: Each block shows $n = 9$ measurements for each of 4 types of cotton data simulated under simplified multivariate normal random-effects model.

The simulations show that this model does not capture the shapes of the curves exactly and it fails to model one of the most important features of the data, within-group variations primarily expressed as separation of the curves. These curves only resemble original data at local minimums and maximums.

5.6.4 Simulation - CA-const. for cotton data

Simulated cotton data using CA-const. with parameters estimated from the original data and data generated for the same number of replicates.

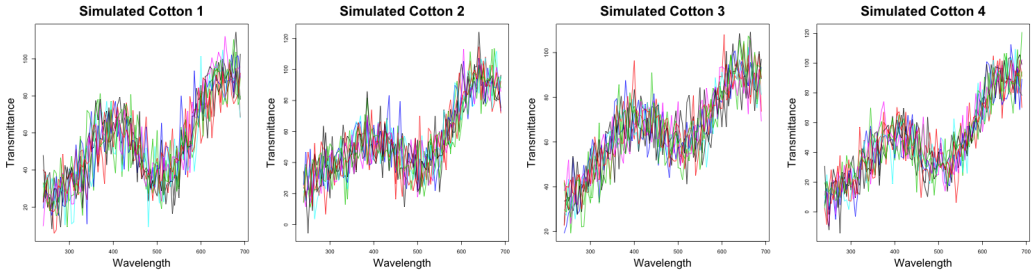


Figure 5.25: Each block shows $n = 9$ measurements for each of 4 types of cotton data simulated under constant within-group variance model.

Data simulated from this model look similar to those simulated from the previous model (CA-S). Some variations for within-group variance can be seen but not too obvious. Neither is it easy to see the effect of relaxing the diagonal covariance assumption for the coefficients (for the shape). It is still not modeling the within-group variation as expected.

5.6.5 Simulation - CA-ar for cotton data

Simulated cotton data using CA-ar with parameters estimated from the original data and data generated for the same number of replicates.

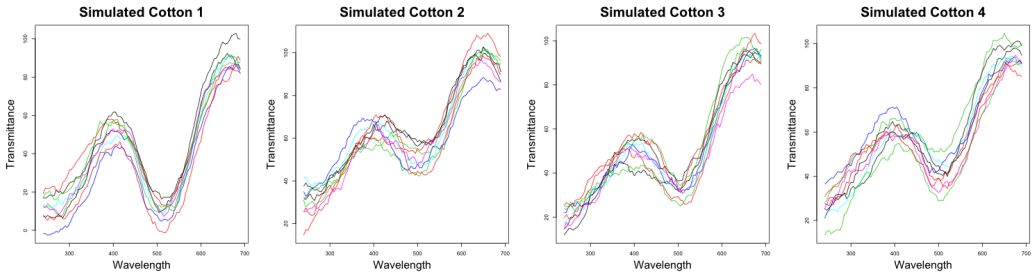


Figure 5.26: Each block shows $n = 9$ measurements for each of 4 types of cotton data simulated under multivariate normal random-effects with autoregressive within-group covariance model.

Data simulated under this model show an improvement in modeling the within-group variation as represented by the separation of curves. Since the measurements are taken at smaller intervals for our ink data, there are greater dependencies between consecutive points on the curve.

5.6.6 Simulation - DR-S for cotton data

Simulated cotton data using DR-S with parameters estimated from the original data and data generated for the same number of replicates.

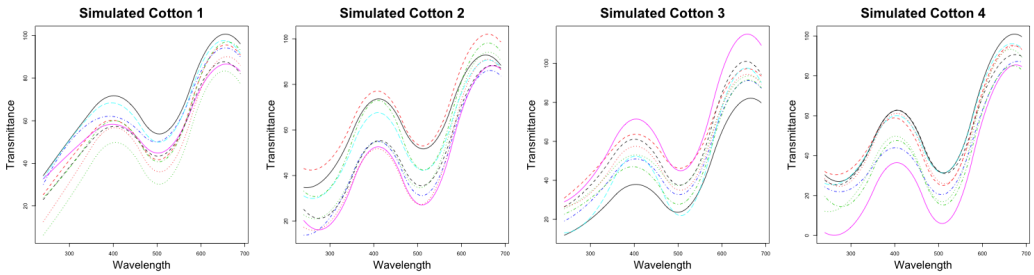


Figure 5.27: Each block shows $n = 9$ measurements for each of 4 types of cotton data simulated under dimension reduced multivariate normal random-effects model.

The variations that are dependent on the location of the curves in the original data can be seen modeled here but the overall shape does not resemble that of original data. This can be explained by the poor fit given by our choice of the number of basis functions.

5.6.7 Conclusion

The models perform as expected. There are limitations for each model and these are expected given the model assumptions.

5.7 Conclusion

We will see via the results in the next chapter whether each of the limitations as set out by the assumptions of the models have an impact in whether the likelihood ratios estimated can distinguish between curves from different groups.

Chapter 6

Results and interpretations

6.1 Introduction

In this chapter, all models introduced in Chapter 3 are used to evaluate likelihood ratios for datasets introduced in Chapter 4.

Likelihood ratios are calculated as follow. Suppose there are K groups for each dataset, each having n measurements of m points per curve for a total of Kn curves. Each comparison that gives one likelihood ratio is obtained by first picking 2 sets of n_s curves to represent control and recovered evidence. Denote these sets of n_s curves by $\boldsymbol{Y}_c$ and $\boldsymbol{Y}_r$, respectively where $\boldsymbol{Y}_c$ can be equivalent to $\boldsymbol{Y}_r$. The hyperparameters are estimated using the rest of the K groups by assuming they represent the relevant population. Given $\boldsymbol{Y}_c, \boldsymbol{Y}_r$ and the estimation of the hyperparameters, likelihood ratios are calculated by the formulas given in the relevant section for each model in Chapter 3. Likelihood ratios will be reported after a log base 10 transformation for the ease of comparison given the scale of their magnitude. They will be denoted by

$$lLR_{M,\Xi_M}(c,r) = log_{10}(LR) = log_{10}\frac{p_{M,\Xi_M}(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_p)}{p_{M,\Xi_M}(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_d)}.$$

where M is the model used and Ξ_M is the set of parameters used for evaluation under chosen M including int and n_s. The $lLR's$ are expected to be greater than 0 for $c = r$ and less than 0 otherwise. For each setup (int, n_s, B, M) a table of lLR will be calculated. The choices and combinations of these are dependent on models and data. The table can be decomposed as in Figure 6.1.

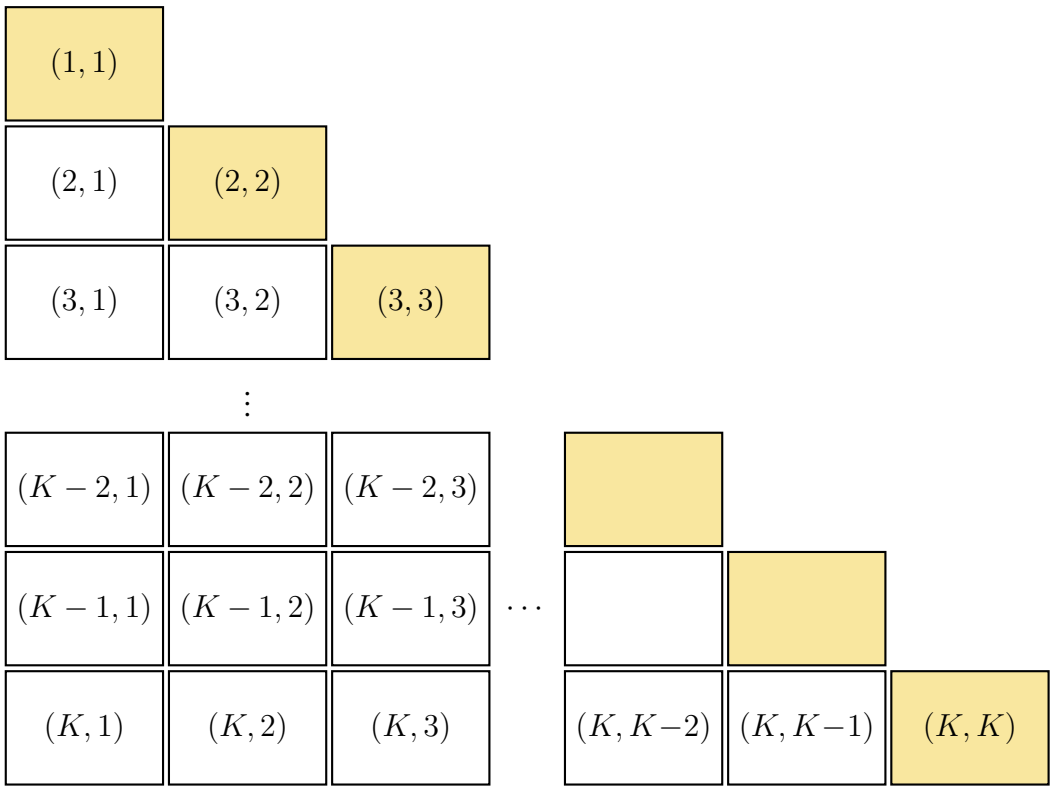


Figure 6.1: Table of lLR calculated given each setup. Each block represent a subtable for comparisons between 2 (identical or distinct) groups. The dimension of the tables are dependent on n and n_s, that are, the number of repeated measurements within a group $k = 1, \ldots, K$ and the number of curves to be used in a comparison. The shaded blocks on the diagonal represent tables of lLR from within-group comparisons and they are lower triangular. The rest of the blocks represent tables of lLR from between groups and they are full.

Results of these lLR's obtained are assessed in various ways including tables, and Tippett and empirical cross entropy plots. Summary tables are used to compare lLR's obtained for different setups and chosen basis functions. The choices for basis functions are B-spline basis functions (subsection 2.3.3) and eigenfunctions obtained using functional principal component analysis (subsection 2.4.2). These tables consist of four numbers S, D, FP and FN to summarise results. The notation S denotes average $lLR_{M,\Xi_M}(c, r)$ for sets of curves $\{\boldsymbol{Y}_c, \boldsymbol{Y}_r\}$ where $c = r$, and D denotes average $lLR_{M,\Xi_M}(c, r)$ for $c \neq r$, they are averages taken from diagonal and off diagonal tables in Figure 6.1, respectively. The notation $FP(FN)$ denotes percentage of misleading evidence when a lLR greater (less) than zero is obtained for a between (within) group comparison. Figures are rounded to the nearest second decimal place. Tippett plots of

$p(log_{10}(LR) > x) = \frac{\sum_{lLRs} I(lLR>x)}{|lLR|}$, the proportion of lLR's that are greater than the value at the $x - axis$ will be drawn to show the empirical (inverse) cumulative proportion for some of the best models along with empirical cross entropy (ECE) plots. The setups selected for plots might not be the same as those selected based on the summary tables due to different criteria used. Summary tables and selected Tippett plots obtained from dimension reduced multivariate normal random-effects model for three sets of data are published in Aitken et al. (2019) using the same procedure described above.

Although we are only interested in $p(\mathbf{E}|H_p)/p(\mathbf{E}|H_d)$, to assess the performance of these models, it would be helpful to consider some cost functions in comparing between the models since we not only want the signs of these lLR's to be correct, the magnitudes of these likelihood ratios should be large in either direction for the support of the propositions. Empirical cross entropy plots consider all cases of prior probabilities and the notion of penalty. It is calculated as

$$ECE = \frac{P(H_p)}{N_s}\sum_{i=1}^{N_s}\log_2\left[1+\frac{P(H_d)}{LR_iP(H_p)}\right] + \frac{P(H_d)}{N_d}\sum_{i=1}^{N_d}\log_2\left[1+\frac{LR_iP(H_p)}{P(H_d)}\right]$$

where N_s and N_d are the number of within- and between-group comparisons, respectively. Each ECE plot consists of three (ECE) lines. A black dotted line represents the null likelihood ratio where there is no information given by the observations and that the likelihood is always equal to one. The blue dashed line represents the ECE for calibrated likelihood ratios. This shows the best set of ECE for when there is no loss of information due to calibrations. It represents the best likelihood ratio values set for all other sets that give the same discriminating power Martyna et al. (2013); Zadora et al. (2013) and it is calculated by using the Pool Adjacent Violators algorithm Cover and Thomas (2005). Finally, the red line is the ECE for our observed likelihood ratios.

6.2 Summary of models

Model		Setup		C.F.	Hyperparameters	Setup	Note
Component-wise additive models (CA)							
CA-S	Simplified multivariate normal random-effects model	$\boldsymbol{y}_{ki}$ $\boldsymbol{\theta}_k$	$\sim N_m(\boldsymbol{\Phi\theta}_k, \sigma^2\boldsymbol{I}_m)$ $\sim N_B(\boldsymbol{\eta}, \boldsymbol{D})$	✓	σ^2, $\boldsymbol{\eta}$, $\boldsymbol{D}$	int, n_s	Section 3.2.2
CA-const.	Constant within-group variance model	$\boldsymbol{y}_{ki}$ $\boldsymbol{\theta}_k$ λ_k	$\sim N_m(\boldsymbol{\Phi\theta}_k, \sigma_k^2\boldsymbol{I}_m = \frac{1}{\lambda_k}\boldsymbol{I}_m)$ $\sim N_B(\boldsymbol{\eta}, \boldsymbol{C}/\lambda_k)$ $\sim Gamma(\gamma, \delta)$	✓	$\boldsymbol{\eta}$, $\boldsymbol{C}$, γ, δ	int, n_s	Section 3.2.3
CA-ar	Multivariate normal random-effects with autoregressive within-group covariance model	$\boldsymbol{y}_{ki}$ $\boldsymbol{\theta_k}$ λ_k	$\sim N_m(\boldsymbol{\Phi\theta}_k, \sigma_k^2\boldsymbol{P} = \boldsymbol{P}/\lambda_k)$ $\sim N_B(\boldsymbol{\eta}, \boldsymbol{C}/\lambda_k)$ $\sim Gamma(\gamma, \delta)$	✓	$\boldsymbol{\eta}$, $\boldsymbol{C}$, $\boldsymbol{U}$, γ, δ	int, n_s	Section 3.2.4
Dimension reduced model (DR)							
DR-S	Dimension reduced multivariate normal random-effects model	$\boldsymbol{z}_{ki}$ $\boldsymbol{\theta}_k$	$\sim N_B(\boldsymbol{\theta}_k, \boldsymbol{U})$ $\sim N_B(\boldsymbol{\eta}, \boldsymbol{C})$	✓	$\boldsymbol{U}$, $\boldsymbol{\eta}$, $\boldsymbol{C}$	B, n_s	Section 3.3.1
DR-C	Multivariate normal random-effects with non constant within-group covariance model	$\boldsymbol{z}_{ki}$ $\boldsymbol{\theta}_k$ $\boldsymbol{U}$	$\sim N_B(\boldsymbol{\theta}_k, \boldsymbol{U}_k)$ $\sim N_B(\boldsymbol{\eta}, \boldsymbol{C})$ $\sim \mathcal{W}^{-1}(\boldsymbol{\Omega}, \nu)$		ν, $\boldsymbol{\Omega}$, $\boldsymbol{\eta}$, $\boldsymbol{C}$	B, n_s	Section 3.3.2

Table 6.1: Summary of Models

6.3 Ink

Sample of ink data consists of $K = 40$ groups of $n = n_k = 10$ MSP measurements of absorbance $\boldsymbol{y}_{ki}$ versus wavelength for $1 \leq i \leq n$ for all k. Absorbance are measured at wavelengths ranging from 380-800 nm with intervals of 1nm so using all the points, that is, taking interval or $int = 1$, the total number of points, the dimension of our data, is $m = 421$.

6.3.1 Summary tables for ink data

For each model, three tables of results will be reported for ink data for 3 distinct values of n_s. The three values are 1, 3 and 5. Since we have 10 measurements of one sample for each of the 40 different types of ink, there are $10 \times 11 \div 2 = 55$ within-group and $10 \times 10 = 100$ between-group lLR's for comparisons between 40 and $40 \times 39 \div 2 = 780$ pairs of groups for $n_s = 1$. For $n_s = 3$, lLR's are obtained for comparing sets of $n_s = 3$ measurements with another (mutually exclusive) set of $n_s = 3$ measurements so there are $\lfloor \frac{10}{3} \rfloor \times (\lfloor \frac{10}{3} \rfloor + 1) \div 2 = 6$ within-group and $\lfloor \frac{10}{3} \rfloor \times \lfloor \frac{10}{3} \rfloor = 9$ between-group lLR's for comparisons between 40 and $40 \times 39 \div 2 = 780$ pairs of groups. For $n_s = 5$ there are $\lfloor \frac{10}{5} \rfloor \times (\lfloor \frac{10}{5} \rfloor + 1) \div 2 = 3$ within-group (including with itself) and $\lfloor \frac{10}{5} \rfloor \times \lfloor \frac{10}{5} \rfloor = 4$ between-group lLR's for comparisons between 40 and $40 \times 39 \div 2 = 780$ pairs of groups. Within-group comparisons include comparisons with the same group of curve(s) itself.

For component-wise additive models, the number B of basis functions is set to 9 as chosen in Chapter 4 for both B-spline basis functions and eigenfunctions obtained from fPCA. The order o of B-spline basis functions is always set to 3. The intervals int considered are 1, 5 and 15 to take into account situations where data available is limited. It is also used to represent cases where there might be different structures among data due to data collected at different intervals. For dimension reduced models, the chosen B are between 4 and 9 inclusive for B-spline basis functions used and 2 to 9 inclusive for eigenfunctions from fPCA used.

6.3.2 CA-S Simplified multivariate normal random-effects model - ink data

Log likelihood ratios calculated using the simplified multivariate normal random-effects model for ink data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	-45.41	-853.43	0.04	72.09	-67.44	-1202.19	0.02	75.86
1	5	-2.63	-161.96	0.66	40.18	7.05	-10.09	17.94	6.55
1	15	2.88	-48.64	2.93	15.45	0.38	-0.01	49.32	4.09
3	1	-42.50	-2458.89	0.00	44.17	-64.31	-3458.58	0.00	45.83
3	5	-0.65	-478.51	0.19	22.08	15.35	-49.13	7.46	2.92
3	15	4.71	-150.92	0.83	12.08	2.69	-0.30	45.91	1.67
5	1	-50.18	-4098.20	0.00	32.50	-75.40	-5759.94	0.00	33.33
5	5	-1.50	-803.10	0.00	22.50	19.94	-95.05	4.94	3.33
5	15	5.01	-256.94	0.45	12.50	6.10	-1.22	43.56	1.67

Table 6.2: Summary table of lLR's obtained using simplified multivariate normal random-effects model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 9 and order of basis used is 3 for B-spline basis functions.

Based on Table 6.2, some D are enormous in magnitude. We can see from the table that given n_s, as int increases S and FP generally goes up and D and FN goes down drastically. This is true for both B-spline basis functions and eigenfunctions from fPCA used. There are usually trade-offs between FP and FN. Given int, as n_s increases, FN rate generally declines as well. Patterns of results obtained using B-spline basis functions differ a lot from results obtained using eigenfunctions from fPCA; when B-spline basis functions are used, the performance improves as int or n_s increases. However, when eigenfunctions from fPCA are used, performance is always optimal at $int = 5$ given n_s and worsen int either increases or decreases. The performance is highly dependent on setup, i.e., $n_s = 5,\ int = 5$ with eigenfunctions obtained from functional principal component analysis, gives the best results in terms of FP and FN for ink data under this model (CA-S).

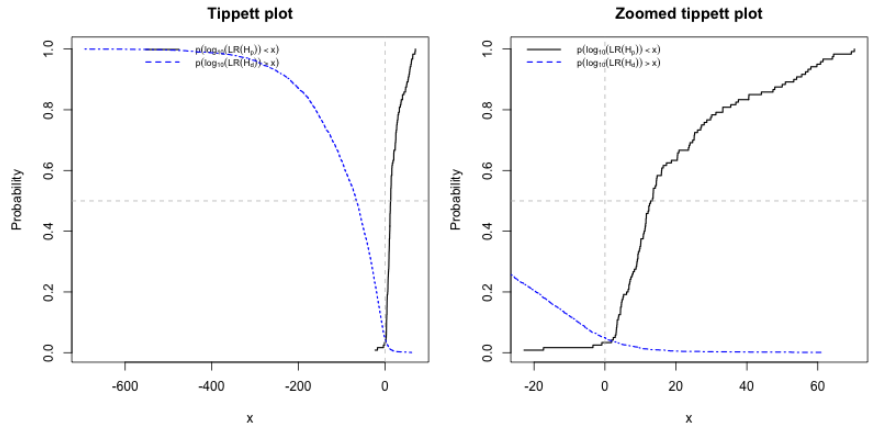


Figure 6.2: Tippett plot for ink data with setup $n_s = 5$, $int = 5$ under model CA-S when eigenfunctions from fPCA are used.

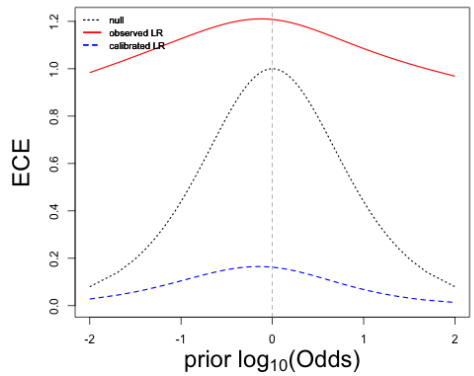


Figure 6.3: ECE plot for ink data with setup $n_s = 5$, $int = 5$ under model CA-S when eigenfunctions from fPCA are used.

Although the Tippett plot as drawn in Figure 6.2 shows only a small percentages of overlap between the two sets (between- and within-group comparisons) of likelihood ratios, the ECE as drawn in Figure 6.3 is saying this model is not giving any useful information as the loss of information is greater than likelihood ratios all equal to one.

6.3.3 CA-const. Constant within-group variance model - ink data

Log likelihood ratios calculated using the constant within-group variance model for ink data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	-101.19	-472.61	0.01	78.45	-202.88	-689.62	0.00	80.59
1	5	-8.14	-80.36	0.27	57.23	9.77	-21.60	8.93	10.23
1	15	1.65	-21.98	1.06	30.32	1.04	-1.74	31.97	5.36
3	1	-61.52	-1024.49	0.00	47.08	-86.31	-1241.68	0.00	49.58
3	5	-3.42	-196.49	0.14	34.17	30.08	-98.10	3.52	4.58
3	15	3.79	-61.35	0.47	17.08	4.84	-6.24	25.23	1.25
5	1	-50.84	-1579.49	0.00	33.33	-69.64	-1879.77	0.00	33.33
5	5	-1.43	-309.11	0.06	27.50	44.91	-192.05	1.92	4.17
5	15	4.73	-99.65	0.26	15.00	10.01	-12.10	23.59	1.67

Table 6.3: Summary table of lLR's obtained using constant within-group variance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 9 and order of basis used is 3 for B-spline basis functions.

Based on Table 6.3, the magnitudes of D decreases from those obtained under CA-S. Using B-spline basis functions, FN rates are always too high but decreases as either n_s and int increases. Using eigenfunctions from fPCA, FP and FN are similar to those obtained under CA-S. Since FN increases compared to CA-S, FP rates are generally smaller. Finally, the best setup is still $n_s = 5,\ int = 5$ when eigenfunctions from fPCA are used with lowered FP compared to CA-S.

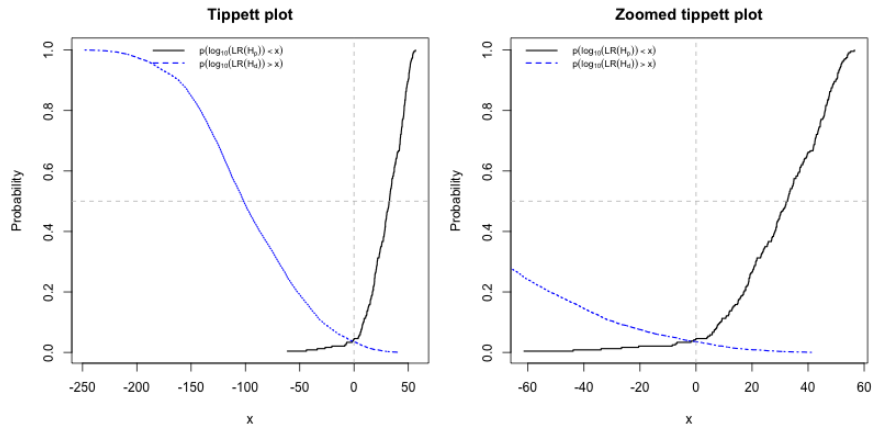


Figure 6.4: Tippett plot for ink data with setup $n_s = 3$, $int = 5$ under model CA-const. when eigenfunctions obtained from fPCA are used.

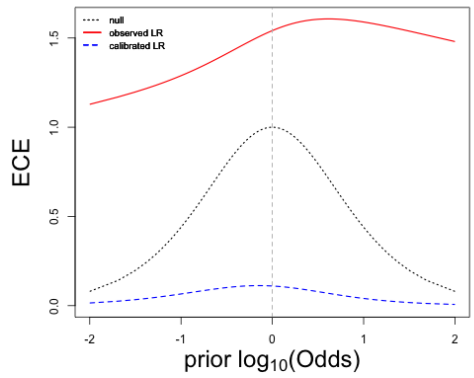


Figure 6.5: ECE plot for ink data with setup $n_s = 3$, $int = 5$ under model CA-const. when eigenfunctions obtained from fPCA are used.

Based on Figure 6.5, there are even more loss of information than CA-S.

6.3.4 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model - ink data

Log likelihood ratios calculated using the multivariate normal random-effects with autoregressive within-group covariance model for ink data are summarised in tables and plots for assessing the performance are drawn for one selection of setups. Overall the magnitudes of S and D drop according to Table 6.4. The patterns for FP and FN stays the same; however, just by changing the within-group covariance matrix FN rates decreased drastically, especially for $int = 1$ meaning the assumption of autocorrelation is somewhat important at least for $int = 1$, which is consistent with our assumptions. The best setup is still $n_s = 5,\ int = 5$ with eigenfunctions from fPCA just like CA-const.. Even though overall this model outperforms CA-const., its best performing setup is not better than that obtained under CA-const. in terms of sum of FP and FN rates. Under this setup, S, D, FP, and FN are all of similar magnitude to those obtained using CA-const.. More setups that give reasonably excellent results include $n_s = 3,\ int = 5$ with eigenfunctions from fPCA and $n_s = 3,\ int = 15$ with B-spline basis functions.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	0.63	-86.98	0.50	32.64	-3.99	-141.09	0.16	49.09
1	5	3.39	-24.51	1.39	13.82	9.27	-14.37	9.07	3.91
1	15	3.26	-9.48	3.86	8.18	1.09	-1.25	35.92	2.59
3	1	1.45	-260.01	0.14	26.67	-2.71	-404.51	0.04	32.50
3	5	5.16	-79.21	0.56	14.17	26.10	-73.31	3.80	2.92
3	15	5.12	-33.32	0.98	7.92	4.65	-5.01	21.89	0.83
5	1	0.76	-437.12	0.06	22.50	-3.57	-671.62	0.00	25.83
5	5	5.57	-135.80	0.35	15.83	38.62	-147.77	2.05	4.17
5	15	5.88	-58.18	0.58	10.83	9.80	-10.09	18.62	0.83

Table 6.4: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 9 and order of basis used is 3 for B-spline basis functions.

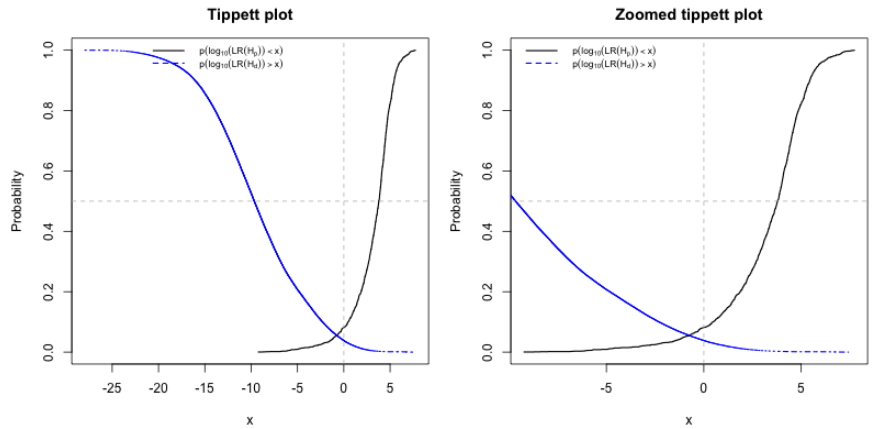


Figure 6.6: Tippett plot for ink data with setup $n_s = 1$, $int = 15$ under model CA-ar.

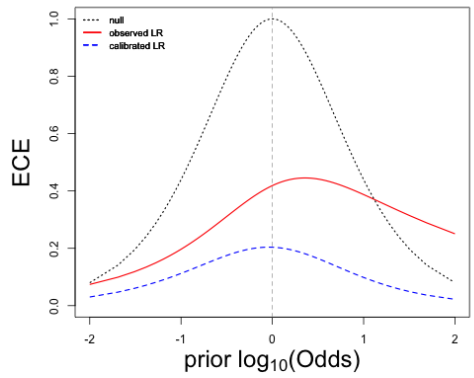


Figure 6.7: ECE plot for ink data with setup $n_s = 1$, $int = 15$ under model CA-ar.

Including an autocorrelation structure to the covariance matrix makes the loss in information much smaller compare to CA-const. according to their ECE plots.

6.3.5 DR-S Dimension reduced multivariate random-effects model - ink data

Log likelihood ratios calculated using the dimension reduced multivariate random-effects model for ink data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	1.05	-8.47	17.11	4.23
3	-	-	-	-	1.74	-15.89	10.17	3.50
4	1.98	-20.95	10.94	3.77	2.42	-28.65	6.42	3.09
5	2.63	-28.34	6.25	2.73	2.91	-33.90	5.24	1.91
6	3.21	-42.39	5.17	1.68	3.42	-45.12	4.28	1.82
7	3.86	-49.76	4.07	0.95	4.04	-55.08	4.01	1.14
8	4.47	-56.25	3.46	0.86	4.58	-62.08	3.56	0.82
9	4.95	-63.87	3.21	0.73	5.24	-67.86	3.00	0.59

Table 6.5: Summary table of lLR's obtained using dimension reduced multivariate random-effects model for comparing sets of size $n_s = 1$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

Both S and D fell to more reasonable magnitudes as indicated in Table 6.5 compared to component-wise additive models. The performance looks good for B as small

as 4 as both FP and FN rates are much smaller comparing to component-wise additive models.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	1.57	-27.28	8.11	2.08
3	-	-	-	-	2.46	-50.89	4.29	2.92
4	2.91	-66.82	4.59	3.75	3.39	-90.67	2.44	3.33
5	3.83	-89.89	2.34	3.33	4.16	-107.12	1.97	2.92
6	4.68	-133.31	2.02	2.08	4.88	-141.81	1.48	2.50
7	5.63	-156.68	1.44	1.67	5.82	-172.78	1.50	2.50
8	6.56	-177.52	1.10	1.67	6.68	-194.87	1.30	2.08
9	7.38	-200.96	1.11	1.25	7.64	-213.84	1.04	1.25

Table 6.6: Summary table of lLR's obtained using dimension reduced multivariate random-effects model for comparing sets of size $n_s = 3$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

For $n_s = 3$ there is still no trade-off between FP and FN. As n_s increases from 1 to 3, smaller B gives smaller FP and FN and larger B gives smaller FP and larger FN. This is the first model so far that given n_s, both FP and FN decline as B

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	1.75	-46.61	5.83	3.33
3	-	-	-	-	2.72	-86.76	2.40	5.00
4	3.31	-113.28	2.50	4.17	3.81	-153.35	1.63	3.33
5	4.39	-152.95	1.47	5.00	4.67	-181.83	1.44	5.00
6	5.31	-225.79	1.12	4.17	5.53	-240.22	1.09	3.33
7	6.33	-265.58	0.87	3.33	6.58	-292.62	0.93	3.33
8	7.37	-300.67	0.87	3.33	7.52	-329.94	0.74	3.33
9	8.31	-341.00	0.64	2.50	8.62	-362.05	0.67	2.50

Table 6.7: Summary table of lLR's obtained using dimension reduced multivariate random-effects model for comparing sets of size $n_s = 5$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

increases. Just by considering within- and between-group variations for the representation of the overall shape of the curves, huge improvements are being made as FP and FN declined drastically compared to component-wise additive models. From $B = 7$ onward there is no noticeable improvement of performance for either choice of basis functions.

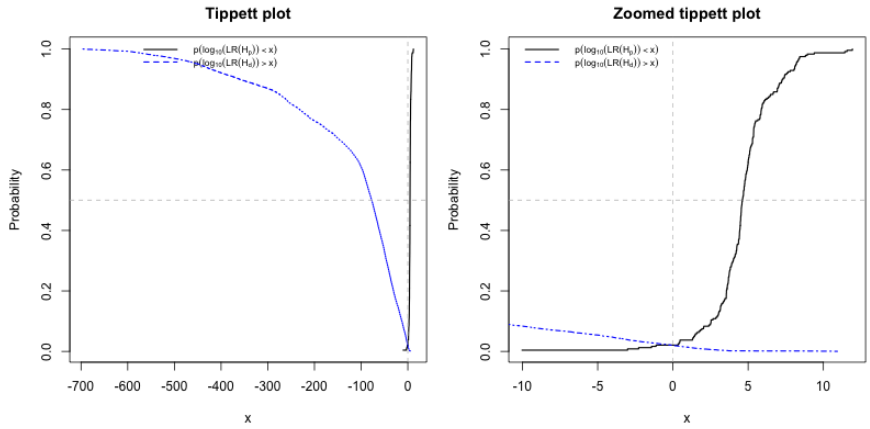


Figure 6.8: Tippett plot for ink data with setup $n_s = 3$, $B = 6$ under model DR-S.

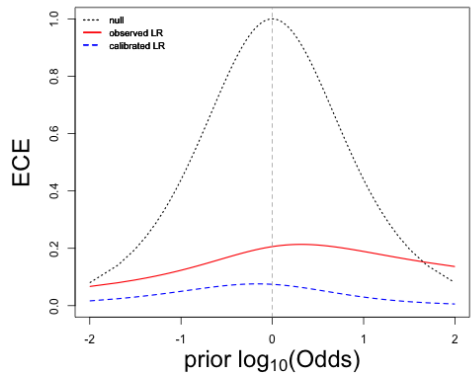


Figure 6.9: ECE plot for ink data with setup $n_s = 3$, $B = 6$ under model DR-S.

According to Figure 6.9 this model is not as good calibrated at $\log_{10}(Odds) > 1.6$ but the loss of information is small when $\log_{10}(Odds) < 1.6$. This model is so far the best performing model compared to all component-wise additive models with much lower FP and FN even at smaller B and indicated by the ECE plot in Figure 6.9.

6.3.6 DR-C Multivariate normal random-effects model with non constant within-group covariance model - ink data

Log likelihood ratios calculated using the multivariate normal random-effects model with non constant within-group covariance model for ink data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	1.07	-3.08	17.07	4.14
3	-	-	-	-	1.76	-4.17	10.50	2.91
4	2.03	-4.35	11.51	3.18	2.47	-5.38	7.17	1.95
5	2.71	-5.10	6.99	1.41	2.97	-5.54	6.17	1.09
6	3.31	-5.67	6.39	0.73	3.50	-5.83	5.43	0.68
7	3.97	-5.65	5.40	0.32	4.14	-5.89	5.54	0.36
8	4.59	-5.59	5.33	0.32	4.70	-5.77	5.35	0.32
9	5.13	-5.23	5.41	0.27	5.37	-0.47	5.22	0.18

Table 6.8: Summary table of lLR's obtained using multivariate normal random-effects model with non constant within-group covariance model for comparing sets of size $n_s = 1$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

Looking at $n_s = 1$ alone, this is one of the best performing models for ink data. Like DR-S, both FP and FN decrease as B increases. Comparing to DR-S, there is a decline in FN and increase of FP, which results in an overall worse performance in terms of sums of FP and FN.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	1.67	-6.58	8.25	2.08
3	-	-	-	-	2.66	-8.51	4.44	2.50
4	3.25	-8.61	5.03	2.92	3.75	-10.21	2.99	2.08
5	4.36	-9.75	2.72	2.08	4.66	-10.39	2.38	1.67
6	5.42	-10.48	2.44	0.42	5.62	-10.72	2.05	1.25
7	6.65	-10.40	2.15	0.42	6.75	-10.69	2.11	0.00
8	7.88	-10.38	2.02	0.00	8.03	-10.55	2.09	0.00
9	9.02	-10.01	1.89	0.42	9.31	-10.29	1.89	0.00

Table 6.9: Summary table of lLR's obtained using multivariate normal random-effects model with non constant within-group covariance model for comparing sets of size $n_s = 3$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

Both FP and FN generally decrease as B increases. As n_s increases from 1 to 3, there is a drop in FP, which makes this model better than DR-S.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	1.98	-9.41	5.71	3.33
3	-	-	-	-	3.18	-11.96	2.72	5.00
4	4.09	-12.09	2.72	4.17	4.61	-14.03	1.99	3.33
5	5.53	-13.60	1.83	2.50	5.81	-14.37	1.67	2.50
6	6.91	-14.57	1.47	1.67	7.14	-14.83	1.35	1.67
7	8.53	-14.54	1.22	0.83	8.72	-14.88	1.41	0.00
8	10.18	-14.71	1.28	0.83	10.31	-14.92	1.19	0.00
9	11.89	-14.43	1.25	0.00	12.21	-15.09	1.15	0.00

Table 6.10: Summary table of lLR's obtained using multivariate normal random-effects model with non constant within-group covariance model for comparing sets of size $n_s = 5$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

As n_s increases from 3 to 5, FP all decrease. This model differs from DR-S only in the consideration of variation in within-group variance-covariance matrix and the results clearly indicate this by the declines in FN from DR-S that might be due to within-group variation. However, this is offset by a slight increase in FP as tradeoff.

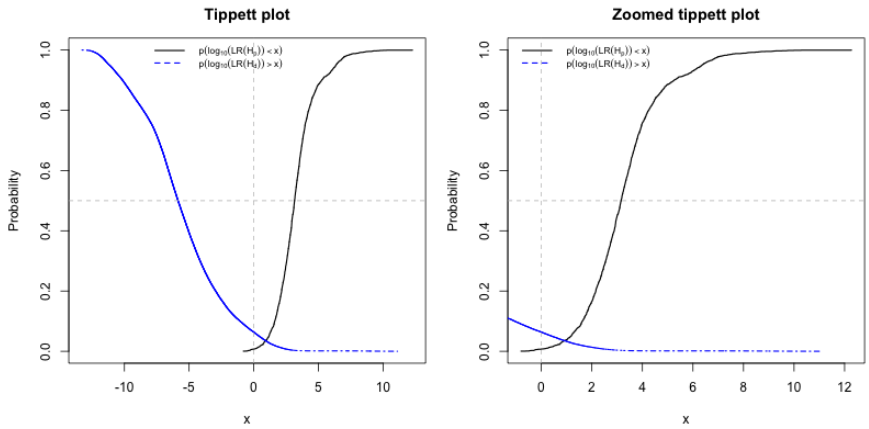


Figure 6.10: Tippett plot for ink data with setup $n_s = 1$, $B = 6$ under model DR-C.

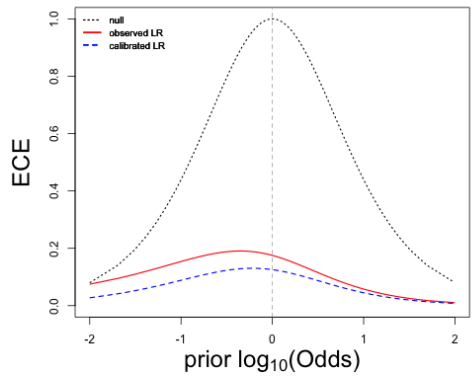


Figure 6.11: ECE plot for ink data with setup $n_s = 1$, $B = 6$ under model DR-C.

We can tell from the ECE plots that DR-C is the best performing model although this is not clear from the summary tables or Tippett plots when comparing with DR-S.

6.3.7 Conclusion

Based on these results, we can tell there are variation among the overall shape of the curves, both within- and between-groups and when a model takes into account both of these variations, it performs well in terms of lowered FP and FN. The ECE further suggests DR-C outperforms DR-S.

6.4 Wool data

Sample of wool data consists of $K = 20$ groups of $n = n_k = 9$ MSP measurements of transmittance $\boldsymbol{y}_{ki}$ versus wavelength for $1 \leq i \leq n$ for all k. Transmittance are measured at wavelengths ranging from 350-690 nm with intervals of 5 nm so using all the points, that is, taking interval or $int = 5$, the total number of points, the dimension of our data, is $m = 69$.

6.4.1 Summary table for wool data

For each model, three tables of results will be reported for wool data for 3 distinct values of n_s. The three values are 1, 2 and 3. Since we have 9 measurements of one

sample for each of the 20 different types of wool fibres, there are $9 \times 10 \div 2 = 45$ within-group and $9 \times 9 = 81$ between-group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups for $n_s = 1$. For $n_s = 2$, lLR's are obtained for comparing sets of $n_s = 2$ measurements with another (mutually exclusive) set of $n_s = 2$ measurements so there are $\lfloor \frac{9}{2} \rfloor \times (\lfloor \frac{9}{2} \rfloor + 1) \div 2 = 10$ within group and $\lfloor \frac{9}{2} \rfloor \times \lfloor \frac{9}{2} \rfloor = 16$ between group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups. For $n_s = 3$ there are $\lfloor \frac{9}{3} \rfloor \times (\lfloor \frac{9}{3} \rfloor + 1) \div 2 = 6$ within-group and $\lfloor \frac{9}{3} \rfloor \times \lfloor \frac{9}{3} \rfloor = 9$ between-group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups.

6.4.2 CA-S Simplified multivariate normal random-effects model - wool data

Log likelihood ratios calculated using the simplified multivariate normal random-effects model for wool data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	-1.90	-33.09	5.42	40.33	-7.96	-70.88	1.45	52.67
1	2	0.47	-14.58	11.56	27.78	0.60	-13.57	13.55	27.22
1	3	0.97	-9.42	16.09	21.22	1.31	-3.95	32.63	15.89
2	1	-0.77	-60.87	2.60	31.00	-6.28	-127.24	0.86	43.50
2	2	1.32	-27.96	6.51	20.50	1.91	-26.54	7.11	18.50
2	3	1.72	-18.62	9.24	15.00	2.39	-8.89	24.54	8.00
3	1	-2.70	-98.31	1.75	33.33	-10.86	-202.61	0.29	40.83
3	2	0.65	-46.07	4.15	21.67	1.67	-44.49	4.85	22.50
3	3	1.38	-31.14	6.26	19.17	3.03	-15.96	17.78	10.00

Table 6.11: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

Based on Table 6.11 the use of eigenfunctions from fPCA behaves differently compared to using B-spline basis functions. Using B-spline basis functions gives high FN

compared to FP for smaller n_s and declines as either int or n_s increases. On the other hand, the use of eigenfunctions from fPCA always performs best when $int = 2$ given n_s and either FP or FN increases as int increases or decreases. Overall, for any setup (choice of n_s and int), this model always gives high (greater than 15%) FP or FN rate, which can possibly be explained by its overall roughness and high within-group variations of its original data. Comparing to ink data, this models generally performs better for smaller n_s and worse as either n_s or int increases. The best performing setup is $n_s = 2,\ int = 2$ using eigenfunctions from fPCA.

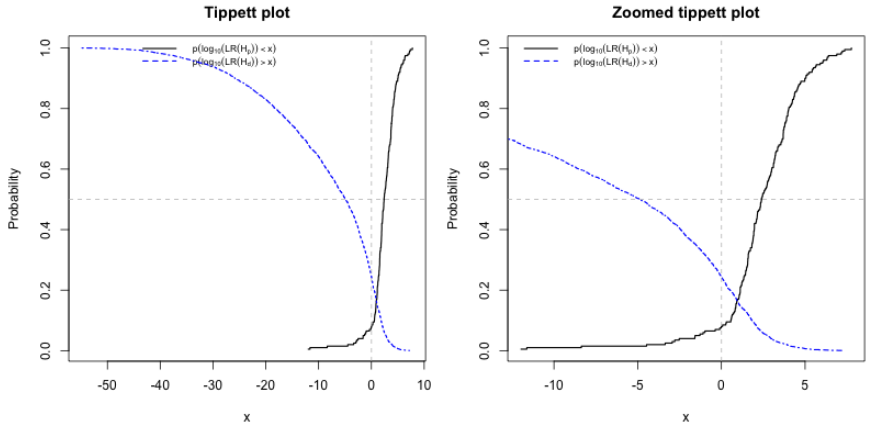


Figure 6.12: Tippett plot for wool data with setup $n_s = 2$, $int = 3$ under model CA-S when eigenfunctions obtained from fPCA are used.

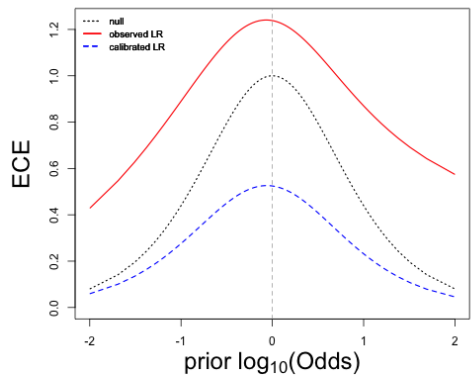


Figure 6.13: ECE plot for wool data with setup $n_s = 2$, $int = 3$ under model CA-S when eigenfunctions obtained from fPCA are used.

Like for ink data, there is too much loss of information suggested by ECE plot as shown in Figure 6.13 that the model is not giving any useful information.

6.4.3 CA-const. Constant within-group variance model - wool data

Log likelihood ratios calculated using the constant within-group variance model for wool data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	-3.02	-26.74	2.83	47.44	-13.73	-61.69	0.38	61.44
1	2	-0.13	-11.55	6.89	35.56	1.48	-11.83	11.65	28.56
1	3	0.46	-7.31	10.52	29.56	1.60	-3.47	29.82	17.78
2	1	-1.76	-47.21	1.32	40.00	-11.19	-96.32	0.26	52.50
2	2	0.83	-21.36	3.95	26.50	3.89	-25.60	5.62	22.50
2	3	1.33	-13.83	5.89	21.00	3.66	-8.74	20.69	10.00
3	1	-3.87	-72.28	0.64	40.00	-15.84	-138.64	0.06	45.83
3	2	0.02	-33.56	2.22	30.83	4.00	-43.17	3.74	26.67
3	3	0.91	-22.05	3.74	25.00	5.06	-16.06	14.80	11.67

Table 6.12: Summary table of lLR's obtained using constant within-group variance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

The use of eigenfunctions from fPCA also behaves differently compared to using B-spline basis functions. Using B-spline basis functions give higher FN compared to FP and declines as n_s increases. The results obtained by using eigenfunctions from fPCA exhibit similar pattern as CA-S, that is, better when $int = 2$ given n_s and worsen as int increases or decreases. Like ink data when modeled using CA-const., there are always trade-offs between FP and FN as int increases given n_s but not necessarily as n_s increases for a given int. Overall, FP rates decreased and FN rates increased from those obtained under CA-S by the consideration of variation in the constant within-group variances and relaxation of diagonal between-group covariance matrix. Comparing to ink data, using either eigenfunctions from fPCA or B-spline basis functions gives lower FN for smaller n_s and higher FN for larger n_s. Like CA-

S, this model favours larger n_s and int when using B-spline basis functions. The best performing setup is $n_s = 3,\ int = 3$ usng eigenfunctions from fPCA.

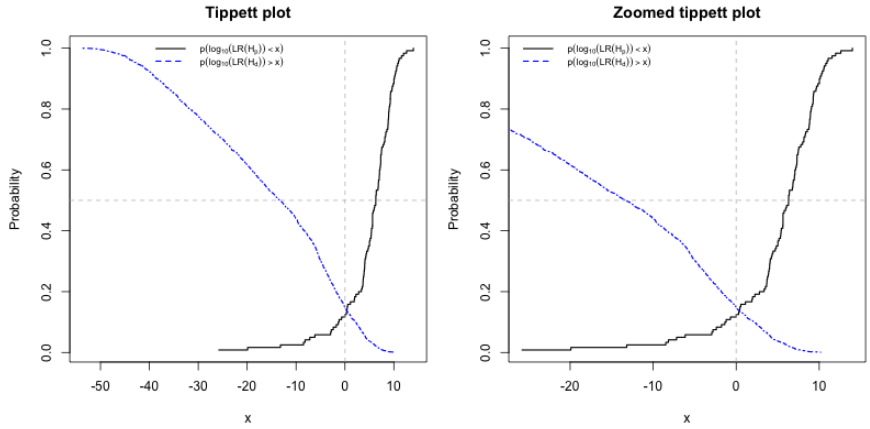


Figure 6.14: Tippett plot for wool data with setup $n_s = 3$, $int = 3$ under model CA-const. when eigenfunctions obtained from fPCA are used.

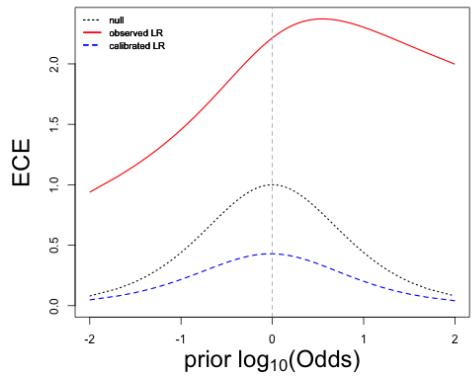


Figure 6.15: ECE plot for wool data with setup $n_s = 3$, $int = 3$ under model CA-const. when eigenfunctions obtained from fPCA are used.

Based on Figure 6.15 more loss of information is seen when taking into account variation in within-group variances comparing to CA-S.

6.4.4 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model - wool data

Log likelihood ratios calculated using the multivariate normal random-effects with autoregressive within-group covariance model for wool data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	1.96	-1.50	31.51	2.78	1.31	-22.24	1.45	21.56
1	2	1.30	-0.87	34.35	3.44	2.93	-6.74	7.55	8.78
1	3	1.06	-1.09	33.43	6.22	2.06	-1.82	25.24	5.44
2	1	3.01	-4.65	16.64	3.00	1.71	-45.46	0.86	19.00
2	2	2.18	-2.68	19.77	3.00	4.89	-17.51	3.29	6.50
2	3	1.84	-2.85	19.97	4.00	4.22	-5.91	12.04	4.00
3	1	3.68	-8.45	9.47	3.33	0.80	-70.67	0.47	22.50
3	2	2.72	-5.15	12.81	3.33	5.95	-30.02	2.63	9.17
3	3	2.23	-5.32	12.81	7.50	5.99	-11.38	7.08	5.00

Table 6.13: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

Like other component-wise additive models, use of eigenfunctions from fPCA behaves differently compared to using B-spline basis functions. The magnitudes and signs of S and D look promising as they are of signs we expect to see and not too large in magnitude. Using B-spline basis functions overall gives high FP and low FN rates but using eigenfunctions from fPCA gives better and satisfactory results when $n_s = 2,\ int = 2$ and performance worsen as either n_s or int increases or decreases. Comparing to CA-const., just by considering autoregressive structure on within-group covariance matrix we are able to bring down FN; however, there is also an increase of FP but for larger n_s, there are overall decrease in the sum of FP and FN when B-spline basis functions are used. On the other hand, using eigenfunctions from fPCA always give lower FP and FN comparing to those obtained under CA-const.. The best performing setup is $n_s = 2,\ int = 2$ using eigenfunctions from fPCA.

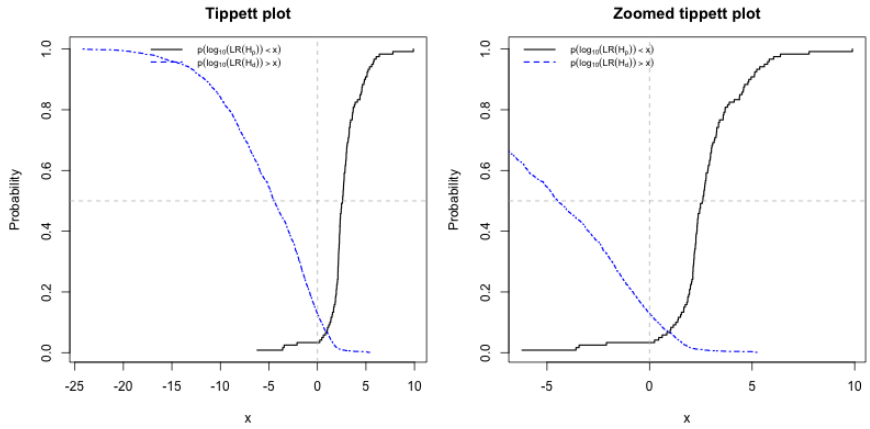


Figure 6.16: Tippett plot for wool data with setup $n_s = 3$, $int = 2$ under model CA-ar.

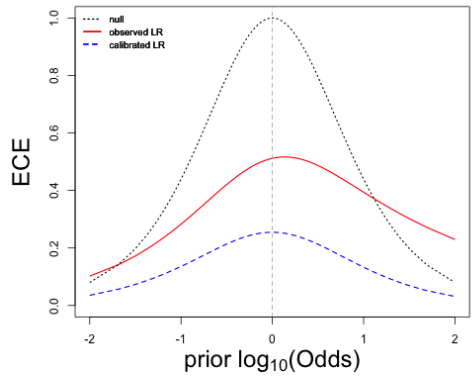


Figure 6.17: ECE plot for wool data with setup $n_s = 3$, $int = 2$ under model CA-ar.

According to Figure 6.17 there is much smaller loss of information compared to other CA models.

6.4.5 DR-S Dimension reduced multivariate random-effects model - wool data

Log likelihood ratios calculated using the dimension reduced multivariate random-effects model for wool data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	0.44	-0.80	34.13	10.56
3	-	-	-	-	0.87	-2.04	22.36	5.78
4	3.21	-16.05	10.70	7.67	2.65	-13.81	11.59	7.89
5	1.63	-7.96	8.40	9.22	1.48	-7.71	9.01	7.00
6	1.48	-10.21	6.17	9.33	1.94	-10.47	5.50	7.78
7	1.88	-13.47	5.05	8.78	1.76	-15.09	4.09	8.67
8	2.10	-16.24	3.72	9.33	2.04	-16.43	3.78	9.11

Table 6.14: Summary table of lLR's obtained using dimension reduced multivariate random-effects model for comparing sets of size $n_s = 1$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

Just looking at $n_s = 1$, the magnitudes and signs of S and D look promising with both FP and FN generally decrease as B increases. However, FN reaches a minimum and increases afterwards.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	0.67	-1.86	25.95	7.00
3	-	-	-	-	1.23	-4.76	14.84	4.00
4	4.19	-35.45	6.64	7.50	3.27	-31.16	7.17	6.50
5	2.18	-17.54	4.67	9.00	1.82	-17.48	4.90	8.00
6	1.74	-22.44	3.03	9.50	2.41	-23.35	2.70	7.00
7	2.36	-29.58	2.73	7.50	2.10	-32.44	2.04	8.50
8	2.65	-35.64	1.55	8.50	2.52	-35.38	1.78	7.00

Table 6.15: Summary table of lLR's obtained using dimension reduced multivariate random-effects model for comparing sets of size $n_s = 2$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

When n_s increases from 1 to 2, FP rates generally decrease and FN decreases for small B. The magnitude of D increases comparing to $n_s = 1$ and the effect is reflected in decreased FP.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	0.75	-3.32	21.81	10.00
3	-	-	-	-	1.36	-7.88	10.47	8.33
4	4.34	-55.87	4.56	10.83	3.02	-48.96	5.03	9.17
5	2.27	-27.86	3.27	14.17	1.70	-27.40	3.27	9.17
6	1.28	-35.78	1.93	13.33	2.30	-36.85	1.81	10.83
7	1.78	-47.04	1.99	14.17	1.41	-51.65	0.99	10.00
8	1.92	-56.41	1.05	14.17	1.90	-56.38	0.94	10.83

Table 6.16: Summary table of lLR's obtained using dimension reduced multivariate random-effects model for comparing sets of size $n_s = 3$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

When n_s increases from 2 to 3, FN rates generally increase. The magnitude of D increases again comparing to $n_s = 2$ and the effect is reflected in slight decrease of FP; however, this is offset by a larger increase in FN. The best performing setup is $n_s = 2,\ B = 8$. Overall, the improvement of DR-S from component-wise additive models explain there are also both within- and between-group variations among the representation of the overall shape of the curves for wool data.

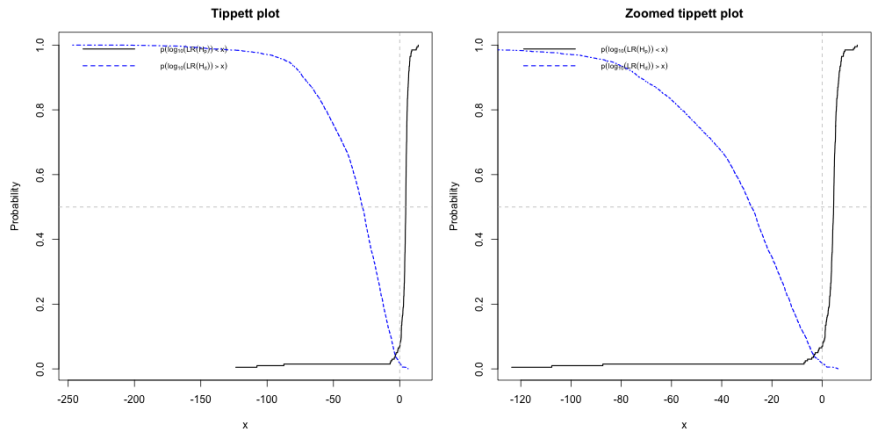


Figure 6.18: Tippett plot for wool data with setup $n_s = 2$, $B = 8$ under model DR-S when eigenfunctions obtained from fPCA are used.

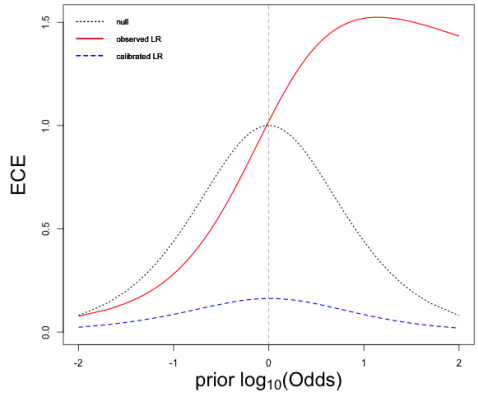


Figure 6.19: ECE for wool data with setup $n_s = 2$, $B = 8$ under model DR-S when eigenfunctions obtained from fPCA are used.

Comparing to CA-ar, this model have larger loss of information at $\log_{10}(Odds) > -0.1$ which possibly suggests there are within-group variations that need to be modeled.

6.4.6 DR-C Multivariate normal random-effects with non constant within-group covariance model - wool data

Log likelihood ratios calculated using the multivariate normal random-effects with non constant within-group covariance model for wool data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	0.45	-0.56	33.22	11.11
3	-	-	-	-	0.90	-1.16	22.16	5.78
4	1.39	-6.97	10.70	7.67	1.16	-5.19	12.72	7.11
5	1.75	-2.68	9.56	8.11	1.82	-2.58	10.10	6.56
6	2.14	-2.92	7.85	7.89	2.32	-3.00	6.81	6.67
7	2.80	-3.05	7.93	7.00	2.70	-3.50	5.84	7.56
8	3.20	-3.10	6.71	7.67	3.15	-3.42	6.08	6.56

Table 6.17: Summary table of lLR's obtained using multivariate normal random-effects model with non constant within-group covariance model for comparing sets of size $n_s = 1$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

The magnitudes of D are smaller comparing to those obtained using DR-S. Comparing to $n_s = 1$ results obtained from DR-S, there is an overall decrease of FN.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	0.73	-1.24	25.49	7.50
3	-	-	-	-	1.34	-2.38	14.90	4.00
4	1.82	-15.39	6.64	7.50	1.45	-11.87	7.76	6.00
5	2.55	-4.58	6.05	6.00	2.65	-4.57	5.53	5.50
6	3.19	-4.91	4.24	5.50	3.38	-5.07	3.62	6.00
7	4.20	-5.03	4.31	5.00	4.06	-5.55	3.19	4.50
8	4.88	-4.98	3.59	5.00	4.69	-5.51	3.06	5.50

Table 6.18: Summary table of lLR's obtained using multivariate normal random-effects model with non constant within-group covariance model for comparing sets of size $n_s = 2$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

The setup $n_s = 2$ outperforms $n_s = 1$ completely in terms of decreased FP and FN. Comparing to DR-S, FN rates are generally smaller for larger B.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	0.87	-2.01	21.40	10.00
3	-	-	-	-	1.55	-3.58	10.41	8.33
4	1.89	-24.26	4.56	10.83	1.39	-18.74	5.44	9.17
5	2.95	-6.30	4.09	11.67	3.04	-6.22	3.80	9.17
6	3.75	-6.69	3.10	8.33	3.94	-6.86	2.34	8.33
7	4.98	-6.90	3.33	6.67	4.70	-7.41	1.87	7.50
8	5.95	-6.60	2.16	7.50	5.74	-7.43	1.75	7.50

Table 6.19: Summary table of lLR's obtained using multivariate normal random-effects model with non constant within-group covariance model for comparing sets of size $n_s = 3$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

As expected, FN declined compared to DR-S as a result of considering variation for within-group variance-covariance matrix. However, the decrease is offset by a slight increase in FP. The best performing setup is $n_s = 2,\ B = 7$.

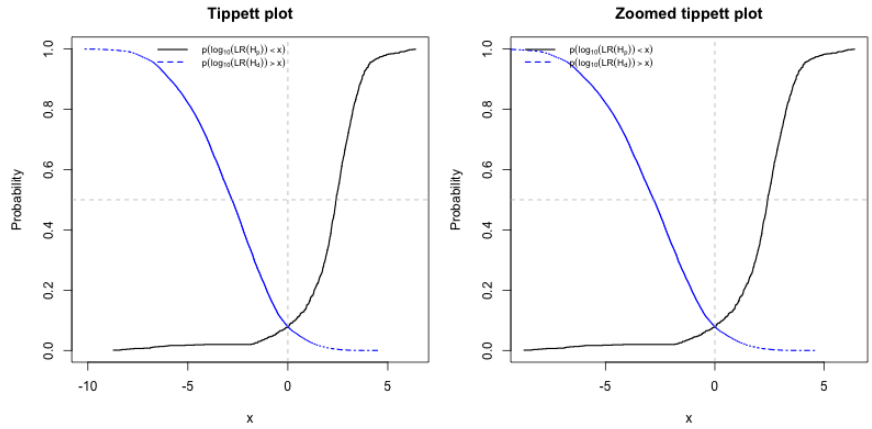


Figure 6.20: Tippett plot for wool data with setup $n_s = 1$, $B = 6$ under model DR-C.

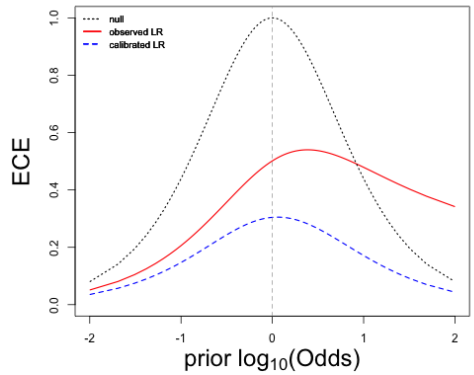


Figure 6.21: ECE for wool data with setup $n_s = 1$, $B = 6$ under model DR-C.

According to Figure 6.21 this model performs similar to that of CA-ar, which could possibly suggest the models pretty much capture the same amount of information through modeling of within-group variability.

6.4.7 Conclusion

The results are quite different for wool data compared to ink data for component-wise additive models. This can possibly be explained by the difference in between- and within-group variations present in the original datasets. However, dimension reduced models consistently perform well for either datasets suggesting there might be no need to model component-wise variances for either ink or wool data.

6.5 Cotton data

Sample of ink data consists of $K = 20$ groups of $n = n_k = 9$ MSP measurements of transmittance $\boldsymbol{y}_{ki}$ versus wavelength for $1 \leq i \leq n$ for all k. Transmittance are measured at wavelengths ranging from 240-690 nm with intervals of 5 nm so using all the points, that is, taking interval or $int = 5$, the total number of points, the dimension of our data, is $m = 91$.

6.5.1 Summary tables for cotton data

For each model, results will be reported for cotton data for 3 distinct values of n_s. The three values are 1, 2 and 3. Since we have 9 measurements of one sample for each of the 20 different types of cotton fibres, there are $9 \times 10 \div 2 = 45$ within-group and $9 \times 9 = 81$ between-group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups for $n_s = 1$. For $n_s = 2$, lLR's are obtained for comparing sets of $n_s = 2$ measurements with another (mutually exclusive) set of $n_s = 2$ measurements so there are $\lfloor \frac{9}{2} \rfloor \times (\lfloor \frac{9}{2} \rfloor + 1) \div 2 = 10$ within group and $\lfloor \frac{9}{2} \rfloor \times \lfloor \frac{9}{2} \rfloor = 16$ between group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups. For $n_s = 3$ there are $\lfloor \frac{9}{3} \rfloor \times (\lfloor \frac{9}{3} \rfloor + 1) \div 2 = 6$ within-group and $\lfloor \frac{9}{3} \rfloor \times \lfloor \frac{9}{3} \rfloor = 9$ between-group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups.

6.5.2 CA-S Simplified multivariate normal random-effects model - cotton data

Log likelihood ratios calculated using the simplified multivariate normal random-effects model for cotton data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

We can see from S and D in Table 6.20 that the results are not optimal as their signs are not as expected and effects of these are reflected in high FP or FN rates. Use of B-spline basis functions gives very similar results from using eigenfunctions from fPCA; they generally give high FN rates and slightly lower FP yet still quite high. There are always trade-offs between FN and FP given n_s. Using B-spline basis functions, results worsen as int increases given n_s in terms of sum of FP and FN.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	-6.08	-28.38	13.13	51.44	-13.26	-49.47	3.91	63.44
1	2	-1.68	-12.51	22.16	41.11	-2.71	-14.77	17.15	48.89
1	3	-0.44	-7.49	28.60	35.00	0.06	-5.64	28.32	38.56
2	1	-5.23	-50.58	7.93	43.50	-11.71	-84.67	2.11	48.50
2	2	-1.06	-23.37	16.64	35.50	-2.05	-27.11	11.35	40.50
2	3	0.12	-14.58	23.32	33.00	0.45	-12.04	23.52	34.50
3	1	-4.92	-72.99	4.85	31.67	-11.27	-120.56	1.40	43.33
3	2	-0.75	-34.40	10.99	28.33	-1.60	-39.72	6.90	29.17
3	3	0.42	-21.83	18.36	23.33	0.75	-18.61	18.60	23.33

Table 6.20: Summary table of lLR's obtained using simplified multivariate normal random-effects model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

Using eigenfunctions from fPCA, results are best when $int = 2$ given n_s and worsen as int either decreases or increases; however, given int, results are generally better as n_s increases for either basis functions used. There must be features of the data that are essential for distinguishing between groups not being captured here.

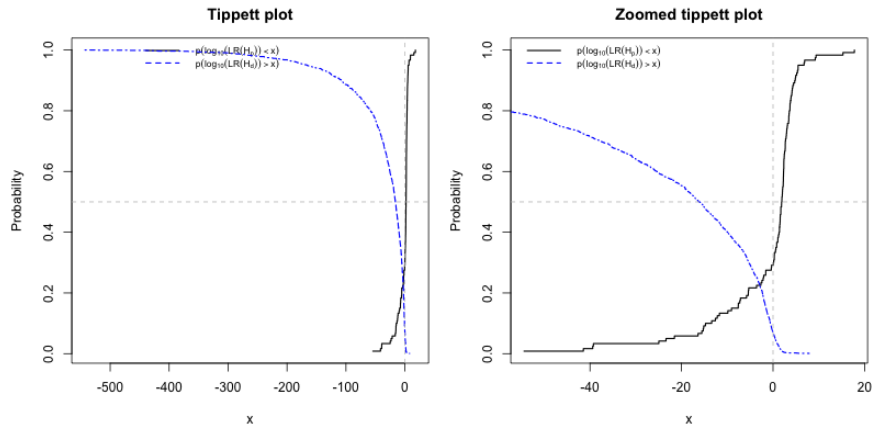


Figure 6.22: Tippett plot for cotton data with setup $n_s = 3$, $int = 2$ under model CA-S when eigenfunctions obtained from fPCA are used.

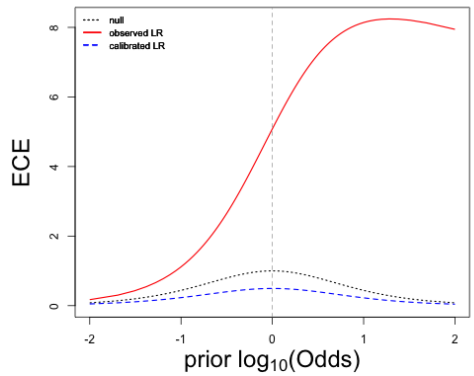


Figure 6.23: ECE plot for cotton data with setup $n_s = 3$, $int = 2$ under model CA-S when eigenfunctions obtained from fPCA are used.

According to ECE plot in Figure 6.23 this model does not provide information useful for distinguishing between curves.

6.5.3 CA-const. Constant within-group variance model - cotton data

Log likelihood ratios calculated using the constant within-group variance model for cotton data are summarised in tables and plots for assessing the performance are drawn for one selection of setups. In Table 6.21 the signs of S are also alarming and the effect is reflected in high FN for all set-ups for either choices of basis functions. The FN rates are higher comparing to those obtained using CA-S for all set-ups; however, there is trade-offs between FP and FN. The performance improves as large increase in FN is completely offset by even larger decline in FP using B-spline basis functions but not as much when eigenfunctions from fPCA are used.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	-9.21	-28.54	5.00	61.89	-25.48	-62.85	1.05	68.67
1	2	-3.28	-12.41	11.43	53.00	-4.55	-17.25	10.58	53.78
1	3	-1.69	-7.58	16.86	48.78	-0.90	-7.12	21.23	43.89
2	1	-8.93	-46.48	1.81	49.00	-21.88	-86.15	0.16	56.00
2	2	-3.05	-21.28	5.16	47.50	-4.31	-30.52	3.55	46.00
2	3	-1.42	-13.36	8.62	44.00	-0.54	-14.65	12.37	40.50
3	1	-9.31	-64.38	1.70	46.67	-23.44	-114.92	0.41	49.17
3	2	-3.20	-30.17	3.45	40.83	-4.87	-44.33	2.40	41.67
3	3	-1.47	-19.20	5.73	37.50	-0.62	-22.58	8.54	34.17

Table 6.21: Summary table of lLR's obtained using constant within-group variance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

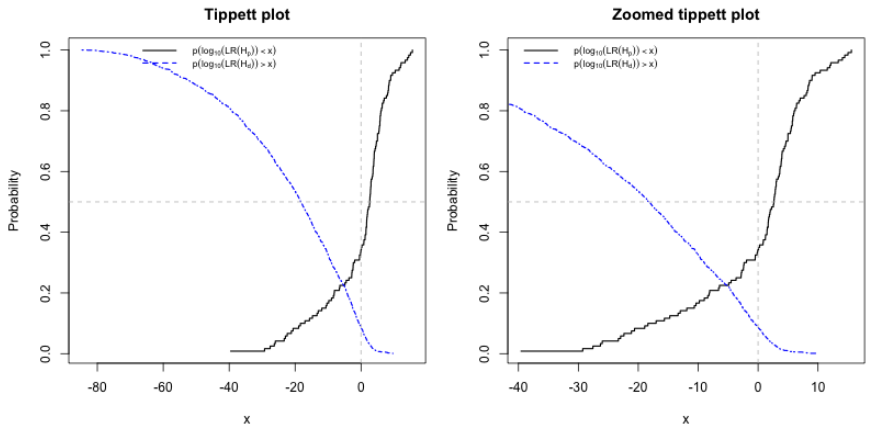


Figure 6.24: Tippett plot for cotton data with setup $n_s = 3$, $int = 3$ under model CA-const. when eigenfunctions obtained from fPCA are used.

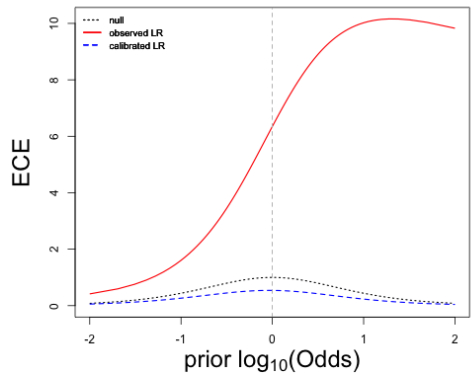


Figure 6.25: ECE plot for cotton data with setup $n_s = 3$, $int = 3$ under model CA-const. when eigenfunctions obtained from fPCA are used.

The high FP and FN rates together with the ECE as plotted in Figure 6.25 suggests that this model does not capture any information useful for distinguishing between curves for cotton data.

6.5.4 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model - cotton data

Log likelihood ratios calculated using the multivariate normal random-effects with autoregressive within-group covariance model for cotton data are summarised in tables and plots for assessing the performance are drawn for one selection of setups. The pattern of in Table 6.22 resembles more of that obtained using CA-S than CA-const. when using B-spline basis functions. However, comparing with CA-const., this models performs better in terms of sums of FP and FN when B-spline basis functions are used but larger FP is not favoured. The use of eigenfunctions from fPCA generally results in declined of FN and not as large increase in FP, which is better than CA-const.. When B-spline basis functions are used, increasing int results in increase of FP given n_s. From the results above (from all 3 models) it can be seen that component-wise models have their limits in the ability to distinguish between groups for cotton data. Moreover, considering variation for within-group variances might worsen the performance but a correct structure of variance-covariance matrix helps a lot in terms of lowered FP and FN. The best performing setup is $n_s = 3,\ int = 2$.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	0.69	-1.66	29.69	15.78	0.01	-18.25	4.73	32.44
1	2	0.56	-0.94	35.04	13.89	1.68	-8.30	11.77	21.22
1	3	0.46	-0.80	36.86	15.89	2.02	-3.12	21.33	12.67
2	1	1.23	-3.78	20.95	8.00	1.16	-35.09	2.27	27.00
2	2	0.96	-2.27	27.24	8.50	2.85	-18.48	5.82	18.00
2	3	0.79	-1.91	28.88	9.00	3.61	-8.52	15.62	12.00
3	1	1.47	-6.26	14.74	9.17	1.44	-52.46	1.52	26.67
3	2	1.11	-3.87	20.88	12.50	3.46	-29.51	3.33	19.17
3	3	0.92	-3.23	23.39	10.83	4.74	-14.88	10.07	10.83

Table 6.22: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

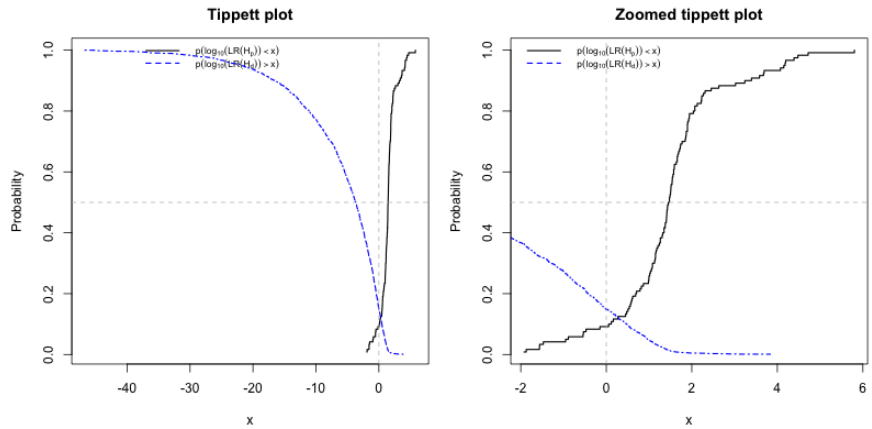


Figure 6.26: Tippett plot for cotton data with setup $n_s = 3$, $int = 1$ under model CA-ar.

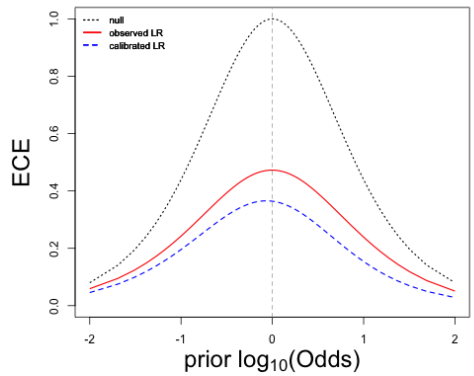


Figure 6.27: ECE plot for cotton data with setup $n_s = 3$, $int = 1$ under model CA-ar.

Surprisingly with high FP and FN rates, the ECE as plotted in Figure 6.27 is showing that the model is reasonably well (not perfectly) calibrated with much smaller loss of information.

6.5.5 DR-S Dimension reduced multivariate random-effects model - cotton data

Log likelihood ratios calculated using the dimension reduced multivariate random-effects model for cotton data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.15	-0.14	51.98	17.78
2	-	-	-	-	0.21	-0.25	48.57	17.00
3	-	-	-	-	0.42	-0.65	38.36	14.00
4	0.48	-0.90	34.89	12.33	0.58	-1.04	29.34	12.00
5	0.62	-1.20	29.53	12.67	0.70	-1.21	30.18	10.89
6	0.71	-2.37	20.80	10.78	0.76	-2.16	23.53	11.11
7	1.16	-3.82	18.35	9.00	0.80	-4.47	16.76	11.56
8	1.41	-5.87	13.22	10.33	1.54	-8.47	12.70	10.00

Table 6.23: Summary table of lLR's obtained using dimension reduced multivariate random-effects model for comparing sets of size $n_s = 1$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.27	-0.41	46.48	12.50
2	-	-	-	-	0.37	-0.66	41.88	13.00
3	-	-	-	-	0.74	-1.59	29.97	7.50
4	0.83	-2.22	23.59	10.00	0.98	-2.54	19.87	9.50
5	1.09	-2.94	19.28	12.50	1.16	-3.01	18.85	8.00
6	1.17	-5.41	12.24	10.00	1.23	-5.02	15.16	10.50
7	1.87	-8.76	10.69	10.00	1.27	-9.73	9.61	13.50
8	2.14	-13.23	6.84	12.50	2.29	-18.31	7.50	11.00

Table 6.24: Summary table of lLR's obtained using dimension reduced multivariate random-effects model for comparing sets of size $n_s = 2$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

The signs and magnitudes of S and D look promising. Like ink and wool data, FP and FN generally decrease at the same time as B increases for DR-S. When n_s increases from 1 to 2, magnitudes of S and D increase slightly and both FP and FN generally decrease for $B < 7$.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.35	-0.72	43.16	10.00
2	-	-	-	-	0.43	-1.17	37.31	10.83
3	-	-	-	-	0.88	-2.69	25.15	10.00
4	1.02	-3.71	17.08	11.67	1.20	-4.25	14.74	12.50
5	1.39	-4.89	13.92	13.33	1.47	-4.93	11.70	10.83
6	1.56	-8.77	8.54	9.17	1.75	-7.84	9.82	10.83
7	2.52	-14.06	7.08	10.83	1.86	-14.99	6.78	11.67
8	2.83	-20.84	4.50	12.50	2.98	-28.96	5.50	11.67

Table 6.25: Summary table of lLR's obtained using dimension reduced multivariate random-effects model for comparing sets of size $n_s = 3$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

It looks like larger n_s and B are required for capturing the features used to distinguish between groups as sums of FP and FN generally decline as n_s increases. Surprisingly, FP and FN do not keep decreasing as B increases and this probably suggest that $B = 6$ is optimal, which is consistent with our selection in chapter 4.

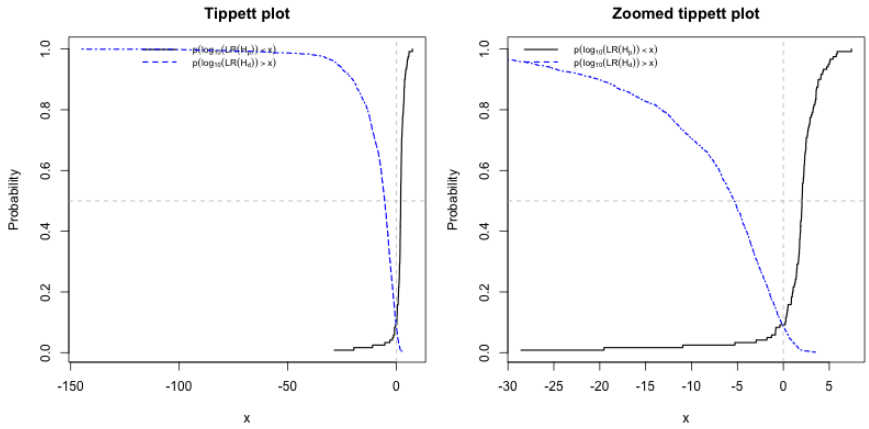


Figure 6.28: Tippett plot for cotton data with setup $n_s = 3$, $B = 6$ under model DR-S.

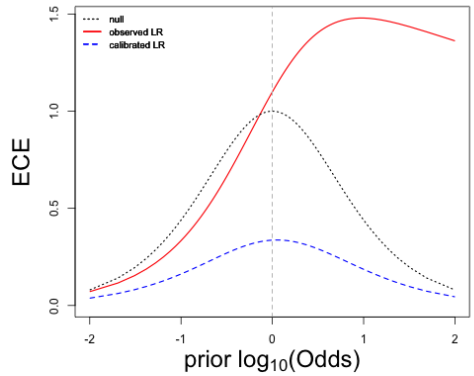


Figure 6.29: ECE for cotton data with setup $n_s = 3$, $B = 6$ under model DR-S.

Comparing to CA-ar., there are some loss of information which suggests there might be within-group variations that are essential for distinguishing between curves.

6.5.6 DR-C Multivariate normal random-effects model with non constant within-group covariance model - cotton data

Log likelihood ratios calculated using the multivariate normal random-effects model with non constant within-group covariance model for cotton data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	0.22	-0.19	46.47	19.33
3	-	-	-	-	0.45	-0.42	36.30	15.89
4	0.57	-0.50	32.42	13.11	0.67	-0.64	27.89	12.89
5	0.80	-0.61	27.92	12.22	0.91	-0.64	29.17	10.67
6	1.12	-0.99	20.55	9.67	1.16	-1.04	23.21	10.89
7	1.68	-1.21	20.79	8.11	1.52	-1.67	17.20	10.22
8	2.14	-1.57	15.46	8.44	2.41	-1.96	14.83	8.56

Table 6.26: Summary table of lLR's obtained using multivariate normal random-effects model with non constant within-group covariance model for comparing sets of size $n_s = 1$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

The signs of S and D look promising. There is a general decrease of FP and FN as B increases without tradeoffs. There is really no improvement from DR-S.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	0.41	-0.48	39.84	13.00
3	-	-	-	-	0.86	-0.96	28.88	8.00
4	1.12	-1.14	21.81	8.00	1.23	-1.42	19.34	9.00
5	1.57	-1.36	18.36	7.50	1.67	-1.45	18.36	6.50
6	2.11	-1.98	12.27	6.00	2.15	-2.07	15.36	7.50
7	3.03	-2.28	12.43	5.50	2.75	-2.98	10.72	6.50
8	3.76	-2.83	9.41	5.00	4.15	-3.47	9.84	3.50

Table 6.27: Summary table of lLR's obtained using multivariate normal random-effects model with non constant within-group covariance model for comparing sets of size $n_s = 2$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

When n_s is increased from 1 to 2, the magnitudes of S and D increase slightly and both FP and FN rates decreased so this is outperforming $n_s = 1$. Comparing to DR-S, there is an overall decrease in FN with a slight increase in FP so this model is capturing the within-group variations present in the data.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
2	-	-	-	-	0.53	-0.83	34.56	11.67
3	-	-	-	-	1.08	-1.59	23.80	11.67
4	1.49	-1.87	15.96	8.33	1.60	-2.26	13.86	9.17
5	2.13	-2.24	12.92	6.67	2.21	-2.34	12.34	7.50
6	2.93	-3.04	9.06	6.67	2.94	-3.07	10.23	7.50
7	3.80	-3.48	8.60	7.50	3.85	-4.20	8.19	6.67
8	5.07	-4.11	6.96	5.83	4.91	-5.18	7.43	6.67

Table 6.28: Summary table of lLR's obtained using multivariate normal random-effects model with non constant within-group covariance model for comparing sets of size $n_s = 3$ for different choice of basis and number of basis functions (B).

We see some decrease in both FN and FP rates when comparing with DR-S for $n_s > 1$. This model is able to decrease FN rates to single digits but overall it might not worth the computation time as there is no visible advantage over DR-S. Moreover, when B-spline basis functions are used, the best setup is again $B = 6$.

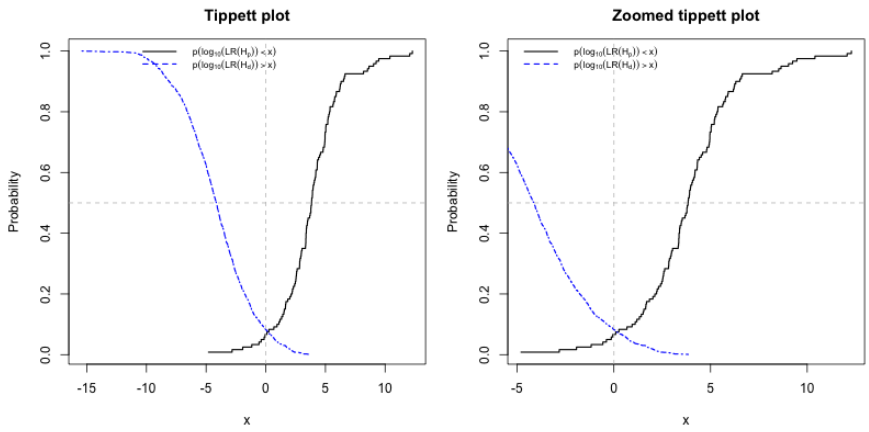


Figure 6.30: Tippett plot for cotton data with setup $n_s = 3$, $B = 7$ under model DR-C when eigenfunctions obtained from fPCA are used.

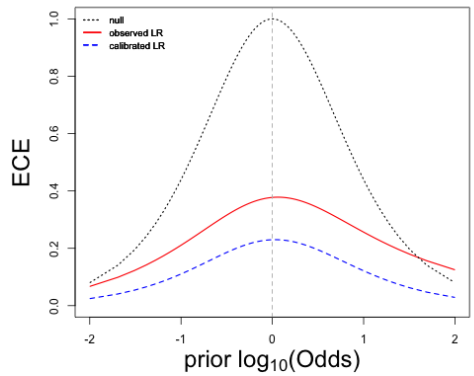


Figure 6.31: ECE plot for cotton data with setup $n_s = 3$, $B = 7$ under model DR-C when eigenfunctions obtained from fPCA are used.

This model performs similar to CA-ar according to the ECE plots but with much smaller FP and FN rates.

6.5.7 Conclusion

The best model for cotton data is multivariate normal random-effects model with non constant within-group covariance but not much better than dimension reduced multivariate random-effects model according to the summary tables. However, according to ECE CA-ar and DR-C perform best. Both of these models take into account some within-group variabilities.

6.6 Conclusion

DR-C is the best choice for all datasets although this is not clear by just looking at the summary tables. It can be seen that the most complicated model is still the best. It takes into account both between- and within-group variabilities. However, computationally CA-ar or DR-S can be used instead for ink and wool data.

Chapter 7

Sensitivity analysis

7.1 Introduction

In this chapter, we are mainly interested in the effect of hyperparameter estimates in the evaluation of likelihood ratios. Forensic cases are often unique in nature with different relevant populations so it is hard to find out the exact distributions that are useful for each case. Moreover, the size of the samples available for analysis is not always large enough for proper inference so a number of adjustments on the estimates are made and tested for their effects on likelihood ratios evaluated. Here we pick a few models that are introduced in Chapter 3 and make adjustments mainly with within-group variance or covariance matrices since the estimation of the means are more straight-forward. The models picked are simplified multivariate normal random-effects model (CA-S), constant within-group variance model (CA-const.), multivariate normal random-effects with autoregressive within-group covariance model (CA-ar) and dimension reduced multivariate normal random-effects model (DR-S), in the same order. Results will be presented under the relevant subsection for the given model under the section for each dataset. These include a set of summary tables each with different choice of n_s. Selection of set-ups that give the largest and smallest lLR's for that model will be picked and sets of $\{\boldsymbol{Y}_c, \boldsymbol{Y}_r\}$ will be chosen to showcase the sets of curves that are worst distinguished by $lLRs$ In other words, curves within the same groups that gives the smallest lLR and curves from different groups that give the largest lLR. Each of these sets will be drawn and compared visually and numerically

(by lLR).

Sections 7.1.1 to 7.1.4 describe how estimations of parameters will be manipulated under each model.

7.1.1 Simplified multivariate normal random-effects model

Under this model we have enormous FN rates for small intervals int and big difference in results for different choice of basis functions. It might be worthwhile to check the effect of using different estimates of the variance ($\hat{\sigma}^2$) on lLRs obtained. For this model, the estimates of the variance ($\hat{\sigma}^2$) will be both increased and decreased by 20%. This amount is set by trial and error to show that it is able to make some differences in FP or FN to give us an idea of the effect of different estimations on the lLR's.

7.1.2 Constant within-group variance model

Under contant within-group variance model, the hyperparameters associated with within-group variances are γ and δ. Four cases are considered for sensitivity analysis; they are

- A: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- B: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$
- C: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- D: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$

The results are presented for the same setups, that are, combinations of n_s and int as in Chapter 6. Results presented in Chapter 6 are also displayed under the Case Original.

7.1.3 Multivariate normal random-effects with autoregressive within-group covariance model

Recall that multivariate normal random-effects with autoregressive within-group covariance model differs from contant within-group variance model only by the assumption of an autocorrelated within-group variance-covariance matrix. The hyperparame-

ters associated with within-group variances are again γ and δ. Four cases are considered for sensitivity analysis; they are again

- A: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- B: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$
- C: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- D: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$

The results are presented for the same setups, that are, combinations of n_s and int as in Chapter 6. Results presented in Chapter 6 are also displayed under the Case Original.

7.1.4 Dimension reduced multivariate normal random-effects model

For the dimension reduced multivariate normal random-effects model, $\boldsymbol{U}$, the within-group variance-covariance is assumed to be constant for all groups, which can be over simplistic. We will consider four cases with varying $\boldsymbol{U}$.

- A: $\boldsymbol{U}_{new} = 0.5 \times \hat{\boldsymbol{U}}$
- B: $\boldsymbol{U}_{new} = 2.5 \times \hat{\boldsymbol{U}}$
- C: $\boldsymbol{U}_{new} = \hat{\boldsymbol{U}} - 0.2diag(\hat{\boldsymbol{U}})$
- D: $\boldsymbol{U}_{new} = \hat{\boldsymbol{U}} + 0.2diag(\hat{\boldsymbol{U}})$

The results are presented for the same setups, that are, combinations of n_s and int as in Chapter 6. However, only one selection of B is used due to the similarity of their performances. Results presented in Chapter 6 are also displayed under the Case Original.

7.1.5 Visual comparison and likelihood ratios

Although $lLRs$ are not meant to be classifiers, we do expect to see larger values of lLR either in the positive or negative direction for support of one proposition over another. For the purpose of performance evaluation, when a pair of sets of curves that are within

the same group but a negative lLR is obtained or vice versa using a given model, we claim that they are failed to be distinguished by lLR. After numerical results are presented for sensitivity analyses given a model, two sets of curves that failed to be distinguished by lLR's are drawn. They represent two setups selected using the table of lLR introduced in Figure 6.1 where setup is a combination of n_s and int (and B in some cases). There will be two sets of figures drawn to illustrate these cases. The first set of figures are curves from within the same group but negative lLR's are obtained using that model under the selected setup. The second set of figures show curves from different groups but positive lLR's are obtained under another selected setup. They are plotted together and separately for a total of 3 figures per set. The lLR's along with their setup will also be presented.

7.2 Ink data

Recall from Chapter 6 the best performing model for original ink data is DR-C according to the summary tables and ECE plots. Most of the ECE plots produced from results obtained using component additive models suggest too much loss in information although low FP and FN rates can be obtained when the right setup is chosen.

7.2.1 CA-S Simplified multivariate normal random-effects model - ink data

For this model, the estimates of the variance (σ^2) will be both increased and decreased by 20% and results will be presented in the same way as in Chapter 6.

Adjustment of	int	S	D	FP	FN
Subtract 20%	1	-59.34	-1069.62	0.03	74.86
Original	1	-45.41	-853.43	0.04	72.09
Plus 20%	1	-36.20	-709.37	0.05	70.00
Subtract 20%	5	-5.07	-204.48	0.45	44.86
Original	5	-2.63	-161.96	0.66	40.18
Plus 20%	5	-1.07	-133.68	0.88	35.45
Subtract 20%	15	2.35	-62.29	2.16	19.41
Original	15	2.88	-48.64	2.93	15.45
Plus 20%	15	3.17	-39.60	3.79	12.82

Table 7.1: Summary table of lLR's obtained using simplified multivariate normal random-effects model with varying estimation of σ^2 for different intervals (int) where 9 B-spline basis functions of order 3 are used for $n_s = 1$.

Adjustment of	int	S	D	FP	FN
Subtract 20%	1	-56.17	-3076.91	0.00	45.42
Original	1	-42.50	-2458.89	0.00	44.17
Plus 20%	1	-33.46	-2046.94	0.00	42.92
Subtract 20%	5	-3.08	-600.66	0.14	26.67
Original	5	-0.65	-478.51	0.19	22.08
Plus 20%	5	0.90	-397.14	0.27	20.42
Subtract 20%	15	4.15	-190.63	0.58	13.75
Original	15	4.71	-150.92	0.83	12.08
Plus 20%	15	5.02	-124.50	1.10	11.67

Table 7.2: Summary table of lLR's obtained using simplified multivariate normal random-effects model with varying estimation of σ^2 for different intervals (int) where 9 B-spline basis functions of order 3 are used for $n_s = 3$.

There are always trade-offs; a change in an estimate increases one of FP and FN and decreases the other one. We can see our original estimates do not always give the lowest FP or FN among all the adjustments given the same settings (int and n_s) but overall, they generally give the most balanced results in terms of FP and FN rates, which is optimal. For this model there are higher FN rates so the trade-offs are not balanced in magnitude; take $n_s = 1, int = 5$ as example, a drop of 9% for FN (from 44.86 to 40.18) resulted in an increase of almost 50% for FP (from 0.45 to 0.66) but magnitude-wise FN has much greater change so when FN rates are large, 'Plus 20%' or larger estimates of σ^2 gives better results which indicates a higher probability of

Adjustment of	int	S	D	FP	FN
Subtract 20%	1	-66.01	-5126.29	0.00	32.50
Original	1	-50.18	-4098.20	0.00	32.50
Plus 20%	1	-39.70	-3412.87	0.00	31.67
Subtract 20%	5	-4.37	-1006.63	0.00	24.17
Original	5	-1.50	-803.10	0.00	22.50
Plus 20%	5	0.35	-667.47	0.03	21.67
Subtract 20%	15	4.28	-323.40	0.29	12.50
Original	15	5.01	-256.94	0.45	12.50
Plus 20%	15	5.42	-212.69	0.48	7.50

Table 7.3: Summary table of lLR's obtained using simplified multivariate normal random-effects model with varying estimation of σ^2 for different intervals (int) where 9 B-spline basis functions of order 3 are used for $n_s = 5$.

getting a positive lLR when the tolerance for within-group error is increased. In other cases where both FP and FN are small, the estimates do not matter as much as either adjustments perform well.

The interval int selected is 1 with number of curves $n_s = 3$ in a set in a comparison. The lLR obtained is -1426.58.

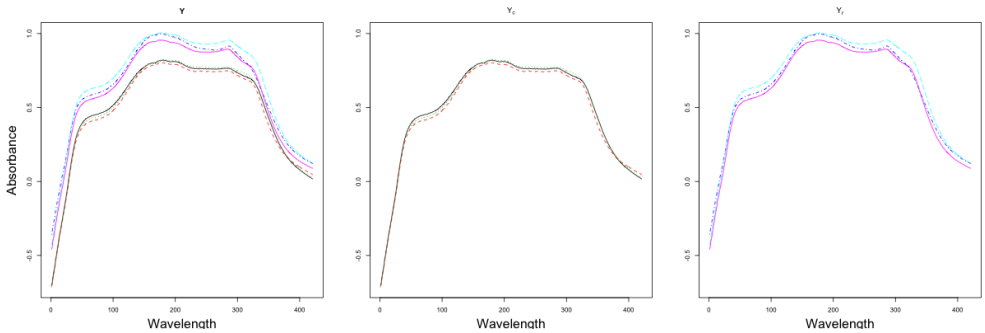


Figure 7.1: Curves from within the group 5 yet negative lLR is obtained under model CA-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The interval int selected is 15 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is 7.79.

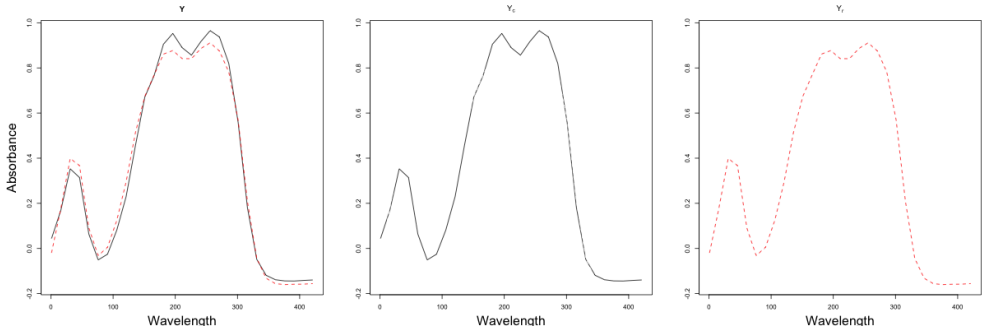


Figure 7.2: Curves from groups 17 and 1 yet positive lLR is obtained under model CA-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

From these two sets of figures we can tell the model fails when there are vertical separations between curves. However, when the curves are reasonably close, a positive lLR will be obtained.

7.2.2 CA-const. Constant within-group variance model - ink data

Under contant within-group variance model, the hyperparameters associated with within-group variances are γ and δ. Four cases are considered for sensitivity analysis; they are

- A: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- B: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$
- C: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- D: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$

The results are presented the same way as in Chapter 6 alongside results from Chapter 6 reproduced here under the Case Original.

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		-101.19	-472.61	0.01	78.45					
1	A	-102.43	-473.78	0.01	78.55	B	-98.08	-467.57	0.01	78.09
1	C	-104.00	-477.38	0.01	78.59	D	-100.00	-471.50	0.01	78.23
5		-8.14	-80.36	0.27	57.23					
5	A	-8.50	-80.52	0.26	57.82	B	-6.65	-77.25	0.30	54.95
5	C	-9.43	-83.32	0.26	59.05	D	-7.82	-80.25	0.27	57.00
15		1.65	-21.98	1.06	30.32					
15	A	1.60	-21.77	1.03	30.86	B	2.21	-20.15	1.39	26.59
15	C	1.24	-23.74	0.88	33.55	D	1.69	-22.21	1.09	29.95

Table 7.4: Summary table of lLR's obtained using constant within group covariance model and manipulating estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 9 and 3 for B-spline basis functions for $n_s = 1$. Refer to Section 7.1.2 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		-61.52	-1024.49	0.00	47.08					
1	A	-61.61	-1024.24	0.00	47.08	B	-60.71	-1022.26	0.00	47.08
1	C	-61.91	-1026.31	0.00	47.08	D	-61.41	-1024.73	0.00	47.08
5		-3.42	-196.49	0.14	34.17					
5	A	-3.47	-196.20	0.14	34.17	B	-2.77	-194.48	0.14	32.50
5	C	-3.69	-198.14	0.16	33.75	D	-3.35	-196.78	0.14	34.17
15		3.79	-61.35	0.47	17.08					
15	A	3.77	-61.03	0.47	17.08	B	4.22	-59.65	0.50	15.83
15	C	3.68	-62.76	0.46	17.50	D	3.81	-61.67	0.47	17.08

Table 7.5: Summary table of lLR's obtained using constant within group covariance model and manipulating estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 9 and 3 for B-spline basis functions for $n_s = 3$.

Based on the results in Tables 7.4 to 7.6, it can be seen that when the modek (constant and independent within-group variance for all points on the curve for a given group with a common centre curve) fails to account for characteristics that are essential for distinguishing between curves, the resulting likelihood ratios are not sensitive to the estimation of hyperparameters.

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		-50.84	-1579.49	0.00	33.33					
1	A	-50.86	-1579.13	0.00	33.33	B	-50.34	-1577.69	0.00	33.33
1	C	-50.92	-1580.87	0.00	33.33	D	-50.80	-1579.84	0.00	33.33
5		-1.43	-309.11	0.06	27.50					
5	A	-1.44	-308.75	0.06	27.50	B	-0.99	-307.39	0.06	27.50
5	C	-1.47	-310.44	0.06	27.50	D	-1.40	-309.46	0.06	27.50
15		4.73	-99.65	0.26	15.00					
15	A	4.73	-99.29	0.26	15.83	B	5.08	-98.07	0.29	15.00
15	C	4.75	-100.88	0.29	15.83	D	4.75	-100.00	0.29	15.00

Table 7.6: Summary table of lLR's obtained using constant within group covariance model and manipulating estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 9 and 3 for B-spline basis functions for $n_s = 5$. Refer to Section 7.1.2 for cases (adjustments).

The interval int selected is 1 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is -594.15.

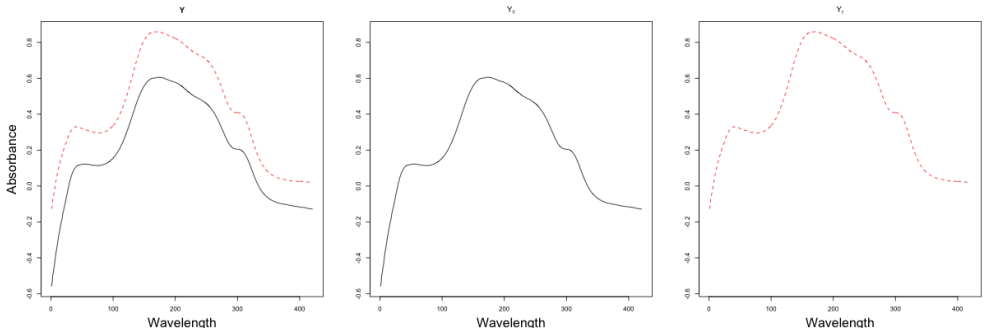


Figure 7.3: Curves from within the group 7 yet negative lLR is obtained under model CA-const.. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The interval int selected is 15 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is 9.91.

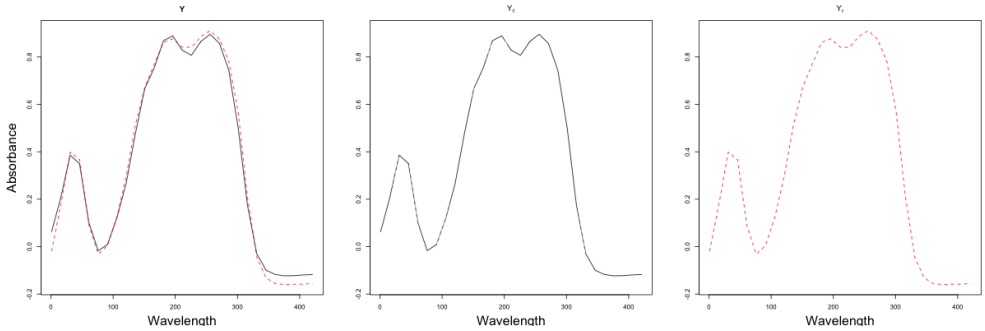


Figure 7.4: Curves from groups 17 and 1 yet positive lLR is obtained under model CA-const.. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

Again this model is sensitive to the distance (separation) between the curves; curves closer together have a higher probability of getting lLR that is greater than 1 and vice versa.

7.2.3 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model - ink data

Four cases are considered for sensitivity analysis. They are the same as CA-const..

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		0.63	-86.98	0.50	32.64					
1	A	0.77	-86.81	0.51	31.82	B	1.06	-86.32	0.54	30.59
1	C	0.82	-87.01	0.52	32.05	D	0.57	-87.06	0.49	32.86
5		3.39	-24.51	1.39	13.82					
5	A	3.44	-24.39	1.42	13.73	B	4.03	-23.25	1.83	10.55
5	C	3.76	-24.77	1.46	14.14	D	3.40	-24.57	1.40	13.82
15		3.26	-9.48	3.86	8.18					
15	A	3.25	-9.24	3.90	8.32	B	3.81	-7.77	6.83	4.23
15	C	4.00	-9.99	4.41	9.14	D	3.29	-9.68	3.89	8.18

Table 7.7: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model with varying estimation of δ and γ for different intervals (int) where 9 B-spline basis functions of order 3 are used for $n_s = 1$. Refer to Section 7.1.3 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		1.45	-260.01	0.14	26.67					
1	A	1.60	-259.84	0.14	26.67	B	1.86	-259.38	0.14	26.67
1	C	1.67	-260.01	0.17	26.67	D	1.38	-260.10	0.13	26.67
5		5.16	-79.21	0.56	14.17					
5	A	5.24	-79.05	0.56	13.33	B	5.80	-77.93	0.57	13.33
5	C	5.59	-79.41	0.60	13.75	D	5.15	-79.29	0.56	14.17
15		5.12	-33.32	0.98	7.92					
15	A	5.15	-33.03	0.97	7.92	B	5.86	-31.30	1.30	6.67
15	C	6.00	-33.76	1.15	7.50	D	5.15	-33.55	0.94	7.92

Table 7.8: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model with varying estimation of δ and γ for different intervals (int) where 9 B-spline basis functions of order 3 are used for $n_s = 3$. Refer to Section 7.1.3 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		0.76	-437.12	0.06	22.50					
1	A	0.91	-436.95	0.06	22.50	B	1.16	-436.5	0.06	21.67
1	C	0.98	-437.11	0.06	22.50	D	0.69	-437.21	0.06	22.50
5		5.57	-135.80	0.35	15.83					
5	A	5.65	-135.64	0.32	15.83	B	6.2	-134.53	0.35	15.00
5	C	6.03	-135.99	0.38	15.00	D	5.56	-135.89	0.35	15.83
15		5.88	-58.18	0.58	10.83					
15	A	5.91	-57.88	0.61	10.83	B	6.66	-56.10	0.77	8.33
15	C	6.81	-58.58	0.67	10.83	D	5.92	-58.42	0.61	10.83

Table 7.9: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model with varying estimation of δ and γ for different intervals (int) where 9 B-spline basis functions of order 3 are used for $n_s = 5$. Refer to Section 7.1.3 for cases (adjustments).

This model is also not sensitive to the choice of hyperparameters (γ and δ) although there are exceptional cases ($int = 15$ under Case B).

The interval int selected is 1 with number of curves $n_s = 3$ in a set in a comparison. The lLR obtained is -105.92.

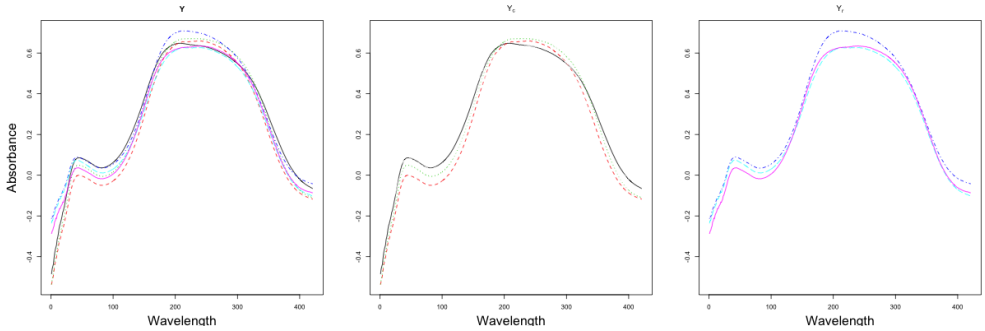


Figure 7.5: Curves from within the group 22 yet negative lLR is obtained under model CA-ar. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The reason why this pair of curves gives a negative lLR is unclear from the plots.

The interval int selected is 15 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is 7.42.

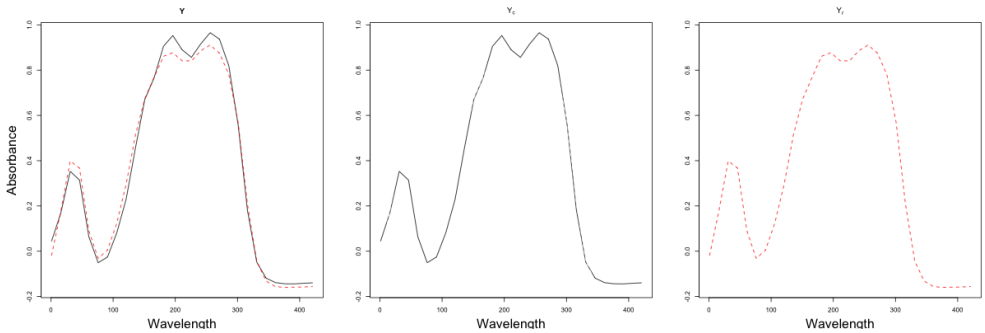


Figure 7.6: Curves from groups 17 and 1 yet positive lLR is obtained under model CA-ar. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The curves are of similar shape; with local minima and maxima close to the other set's.

7.2.4 DR-S Dimension reduced multivariate normal random-effects model - ink data

For the dimension reduced multivariate normal random-effects model, $\boldsymbol{U}$, the within-group variance-covariance is assumed to be constant for all groups. For the sensitivity analyses we will consider four cases with varying $\boldsymbol{U}$.

- A: $\boldsymbol{U}_{new} = 0.5 \times \hat{\boldsymbol{U}}$
- B: $\boldsymbol{U}_{new} = 2.5 \times \hat{\boldsymbol{U}}$
- C: $\boldsymbol{U}_{new} = \hat{\boldsymbol{U}} - 0.2diag(\hat{\boldsymbol{U}})$
- D: $\boldsymbol{U}_{new} = \hat{\boldsymbol{U}} + 0.2diag(\hat{\boldsymbol{U}})$

The results are presented for the same setups, that are, combinations of n_s and int as in Chapter 6. However, only selections of B will be considered used due to the similarity of their performances. Results are presented as in Chapter 6 with results from Chapter 6 reproduced here under the Case Original.

B	Case	S	D	FP	FN	Case	S	D	FP	FN
5	A	2.49	-60.06	2.98	10.00	C	2.37	-29.61	34.20	12.77
5	Original	2.63	-28.34	6.25	2.73	Original	2.63	-28.34	6.25	2.73
5	B	2.17	-9.90	12.62	0.09	D	2.43	-11.90	9.43	0.73
7	A	3.68	-104.50	1.85	8.05	C	4.97	22.29	69.81	1.64
7	Original	3.86	-49.76	4.07	0.95	Original	3.86	-49.76	4.07	0.95
7	B	3.21	-17.76	9.41	0.00	D	3.38	-17.79	7.62	0.32
9	A	4.73	-134.21	1.38	7.09	C	0.45	87.59	86.43	37.27
9	Original	4.95	-63.87	3.21	0.73	Original	4.95	-63.87	3.21	0.73
9	B	4.09	-22.79	8.93	0.00	D	3.93	-21.35	7.25	0.23

Table 7.10: Summary table of lLR's obtained using dimension reduced multivariate normal random-effects model with varying estimates for $\boldsymbol{U}$ for different number (B) of B-spline basis functions of order 3 for $n_s = 1$.

B	Case	S	D	FP	FN	Case	S	D	FP	FN
5	A	3.75	-184.37	1.03	7.50	C	3.70	-90.46	31.11	9.17
5	Original	3.83	-89.89	2.34	3.33	Original	3.83	-89.89	2.34	3.33
5	B	3.32	-33.79	6.08	0.42	D	3.60	-40.35	4.09	2.50
7	A	5.58	-320.08	0.68	7.08	C	6.42	57.57	65.44	1.25
7	Original	5.63	-156.68	1.44	1.67	Original	5.63	-156.68	1.44	1.67
7	B	4.84	-59.50	3.93	0.00	D	5.04	-59.92	2.89	0.42
9	A	7.43	-410.70	0.48	5.00	C	9.11	268.17	90.28	0.00
9	Original	7.38	-200.96	1.11	1.25	Original	7.38	-200.96	1.11	1.25
9	B	6.28	-76.25	3.02	0.00	D	6.09	-71.70	2.48	0.42

Table 7.11: Summary table of lLR's obtained using dimension reduced multivariate normal random-effects model with varying estimates for $\boldsymbol{U}$ for different number (B) of B-spline basis functions of order 3 for $n_s = 3$.

B	Case	S	D	FP	FN	Case	S	D	FP	FN
5	A	4.29	-311.07	0.61	10.00	C	4.00	-158.44	29.87	9.17
5	Original	4.39	-152.95	1.47	5.00	Original	4.39	-152.95	1.47	5.00
5	B	3.88	-58.66	3.69	0.00	D	4.14	-69.99	2.60	3.33
7	A	6.21	-538.67	0.42	8.33	C	6.89	89.12	63.56	5.00
7	Original	6.33	-265.58	0.87	3.33	Original	6.33	-265.58	0.87	3.33
7	B	5.60	-102.56	2.40	0.00	D	5.77	-103.63	1.70	2.50
9	A	8.27	-691.86	0.22	5.83	C	9.39	438.22	87.72	0.00
9	Original	8.31	-341.00	0.64	2.50	Original	8.31	-341.00	0.64	2.50
9	B	7.28	-131.60	1.86	0.00	D	7.07	-123.90	1.63	0.80

Table 7.12: Summary table of lLR's obtained using dimension reduced multivariate normal random-effects model with varying estimates for $\boldsymbol{U}$ for different number (B) of B-spline basis functions of order 3 for $n_s = 5$.

Overall, our original estimates are robust in terms of sums of FN and FP.

The number of basis functions B selected is 9 with number of curves $n_s = 3$ in a set in a comparison. The lLR obtained is -11.04.

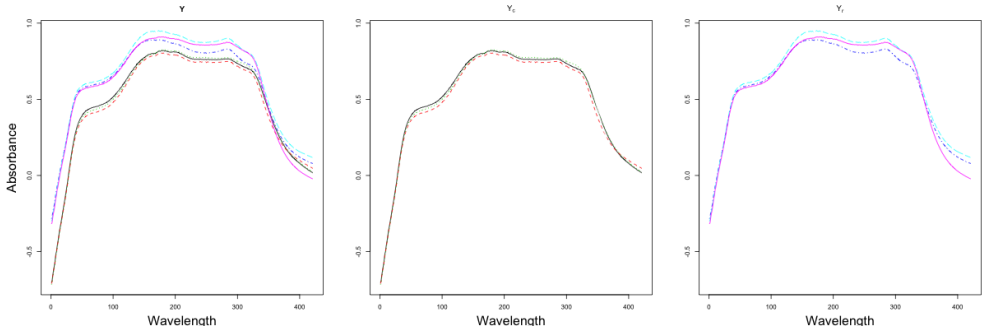


Figure 7.7: Curves from within the group 5 yet negative lLR is obtained under model DR-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The number of basis functions B selected is 9 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is 29.43.

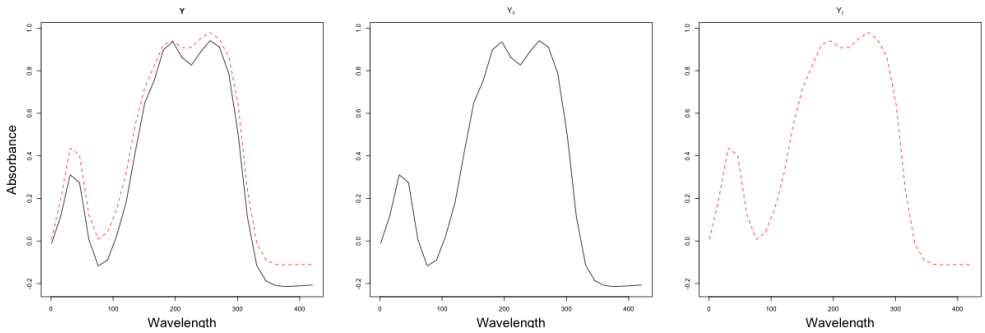


Figure 7.8: Curves from groups 17 and 1 yet positive lLR is obtained under model DR-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The same groups (17 and 1) are picked up by all component-wise additive models.

7.2.5 Conclusion

If a model fails to account for characteristics that are essential for distinguishing between curves, the resulting likelihood ratios are not sensitive to the estimation of hyperparameters; however, if a model fits well, the performance is optimum even for a slight change in the estimation.

7.3 Wool data

Recall from Chapter 6 wool data had the worst result among all datasets. Some models are able to obtain low FP and FN rates but the ECE plots said otherwise. The only acceptable models are CA-ar and DR-C under certain setups.

7.3.1 CA-S Simplified multivariate normal random-effects model - wool data

For this model, the estimates of the variance (σ^2) will be both increased and decreased by 20% and results will be presented in the same way as in Chapter 6.

Adjustment of	int	S	D	FP	FN
Subtract 20%	1	-3.23	-42.38	4.10	44.44
Original	1	-1.90	-33.09	5.42	40.33
Plus 20%	1	-1.06	-26.95	6.64	36.56
Subtract 20%	2	-0.05	-19.01	9.22	30.67
Original	2	0.47	-14.58	11.56	27.78
Plus 20%	2	0.78	-11.67	13.65	24.78
Subtract 20%	3	0.68	-12.44	12.92	25.33
Original	3	0.97	-9.42	16.09	21.22
Plus 20%	3	1.12	-7.45	18.93	19.22

Table 7.13: Summary table of lLR's obtained using simplified multivariate normal random-effects model with varying estimation of σ^2 for different intervals (int) where 6 B-spline basis functions of order 3 are used for $n_s = 1$.

Adjustment of	int	S	D	FP	FN
Subtract 20%	1	-1.95	-77.26	1.97	34.50
Original	1	-0.77	-60.87	2.60	31.00
Plus 20%	1	-0.02	-50.00	3.42	26.00
Subtract 20%	2	0.87	-35.87	4.84	22.00
Original	2	1.32	-27.96	6.51	20.50
Plus 20%	2	1.57	-22.72	7.66	17.50
Subtract 20%	3	1.48	-24.09	7.27	18.50
Original	3	1.72	-18.62	9.24	15.00
Plus 20%	3	1.84	-15.02	10.79	10.50

Table 7.14: Summary table of lLR's obtained using simplified multivariate normal random-effects model with varying estimation of σ^2 for different intervals (int) where 6 B-spline basis functions of order 3 are used for $n_s = 2$.

Based on the results above, it can be seen that when the assumption is wrong, the resulting likelihood ratios are not sensitive to the estimation of variance.

Adjustment of	int	S	D	FP	FN
Subtract 20%	1	-4.53	-124.23	1.23	35.00
Original	1	-2.70	-98.31	1.75	33.33
Plus 20%	1	-1.54	-81.09	2.05	28.33
Subtract 20%	2	-0.10	-58.67	2.98	25.00
Original	2	0.65	-46.07	4.15	21.67
Plus 20%	2	1.11	-37.71	5.32	20.00
Subtract 20%	3	0.94	-39.88	4.80	20.00
Original	3	1.38	-31.14	6.26	19.17
Plus 20%	3	1.64	-25.36	7.49	18.33

Table 7.15: Summary table of lLR's obtained using simplified multivariate normal random-effects model with varying estimation of σ^2 for different intervals (int) where 6 B-spline basis functions of order 3 are used for $n_s = 3$.

The interval int selected is 1 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is -95.09.

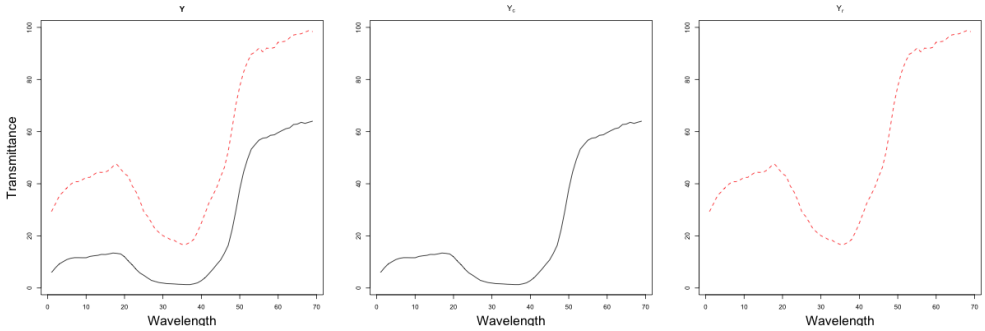


Figure 7.9: Curves from within the group 14 yet negative lLR is obtained under model CA-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The interval int selected is 3 with number of curves $n_s = 3$ in a set in a comparison. The lLR obtained is 4.87.

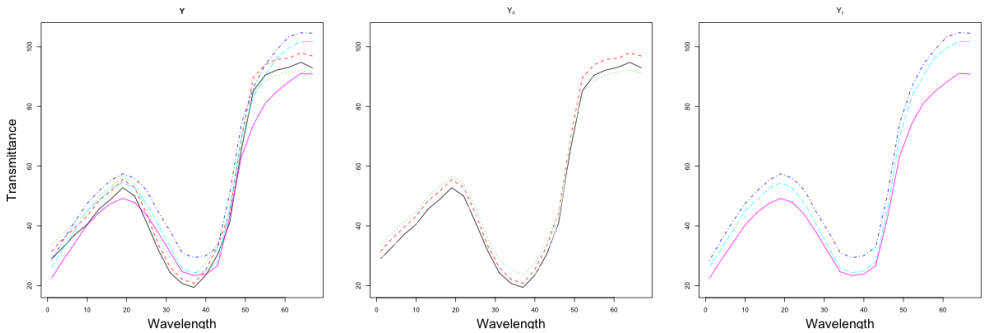


Figure 7.10: Curves from groups 12 and 7 yet positive lLR is obtained under model CA-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

Note that these two sets of curves have means very close to each other, which might be the reason for a positive lLR obtained.

7.3.2 CA-const. Constant within group covariance model - wool data

Under contant within-group variance model, the hyperparameters associated with within-group variances are γ and δ. Four cases are considered for sensitivity analysis; they are

- A: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- B: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$
- C: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- D: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$

The results are presented the same way as in Chapter 6 alongside results from Chapter 6 reproduced here under the Case Original.

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		-3.02	-26.74	2.83	47.44					
1	A	-3.21	-26.78	2.63	48.22	B	0.21	-20.82	7.25	34.56
1	C	-3.63	-30.41	3.01	47.78	D	-2.89	-26.78	2.95	47.22
2		-0.13	-11.55	6.89	35.56					
2	A	-0.12	-11.33	6.73	35.89	B	1.74	-7.56	16.87	21.67
2	C	-0.02	-14.04	7.87	35.67	D	-0.14	-11.76	7.04	35.11
3		0.46	-7.31	10.52	29.56					
3	A	0.51	-7.02	10.45	29.44	B	1.72	-4.17	23.93	15.89
3	C	0.73	-9.46	11.10	29.78	D	0.42	-7.56	10.64	29.67

Table 7.16: Summary table of lLR's obtained using constant within group covariance model with varying estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 6 and 3 for B-spline basis functions for $n_s = 1$. Refer to Section 7.1.2 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		-1.76	-47.21	1.32	40.00					
1	A	-1.90	-47.16	1.32	39.50	B	1.20	-41.44	3.12	27.00
1	C	-1.02	-49.46	2.01	34.00	D	-1.62	-47.28	1.38	39.50
2		0.83	-21.36	3.95	26.50					
2	A	0.81	-21.15	3.85	27.00	B	2.89	-16.88	8.03	17.00
2	C	1.87	-23.03	5.86	23.00	D	0.85	-21.56	3.95	26.50
3		1.33	-13.83	5.89	21.00					
3	A	1.36	-13.54	5.72	21.00	B	2.91	-9.99	12.40	10.50
3	C	2.40	-15.40	8.49	19.50	D	1.32	-14.09	5.86	21.00

Table 7.17: Summary table of lLR's obtained using constant within group covariance model with varying estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 6 and 3 for B-spline basis functions for $n_s = 2$. Refer to Section 7.1.2 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		-3.87	-72.28	0.64	40.00					
1	A	-3.99	-72.12	0.64	40.00	B	-0.77	-66.33	1.87	33.33
1	C	-2.91	-74.28	1.11	37.50	D	-3.74	-72.43	0.64	39.17
2		0.02	-33.56	2.22	30.83					
2	A	0.01	-33.30	2.16	31.67	B	2.40	-28.63	5.26	21.67
2	C	1.20	-35.14	3.86	29.17	D	0.06	-33.81	2.28	29.17
3		0.91	-22.05	3.74	25.00					
3	A	0.93	-21.71	3.68	25.00	B	2.88	-17.63	8.48	17.50
3	C	2.09	-23.61	5.85	24.17	D	0.91	-22.35	3.74	25.00

Table 7.18: Summary table of lLR's obtained using constant within group covariance model with varying estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 6 and 3 for B-spline basis functions for $n_s = 3$. Refer to Section 7.1.2 for cases (adjustments).

Based on the tables above, the results are not sensitive to the estimations of the hyperparameters.

The interval int selected is 1 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is -58.74.

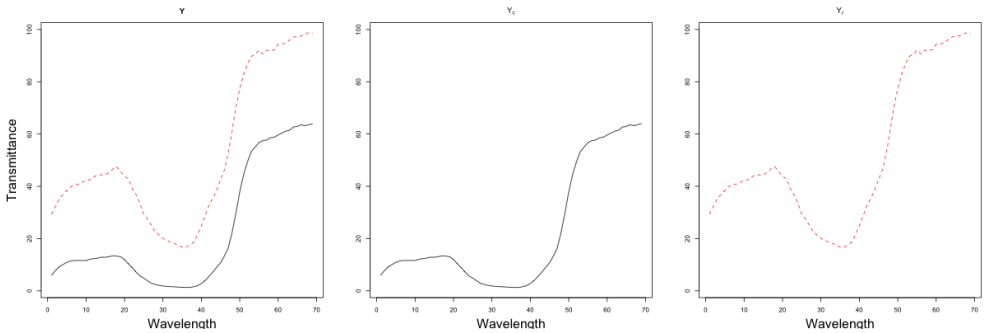


Figure 7.11: Curves from within the group 14 yet negative lLR is obtained under model CA-const.. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The interval int selected is 3 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is 5.24.

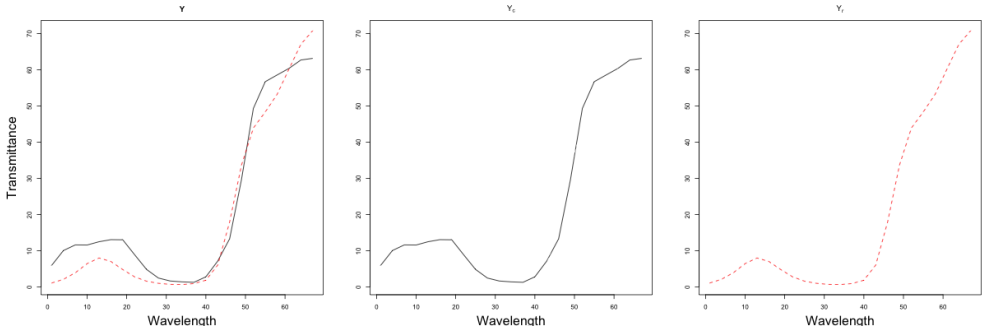


Figure 7.12: Curves from groups 14 and 7 yet positive lLR is obtained under model CA-const.. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

These two curves have different shapes; however, it is also hard to distinguish by eyes.

7.3.3 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model - wool data

Four cases are considered for sensitivity analysis. They are the same as CA-const..

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		1.96	-1.50	31.51	2.78					
1	A	2.02	-1.49	32.29	2.78	B	2.97	-0.2	53.74	1.33
1	C	3.01	-0.76	46.08	2.56	D	1.94	-1.47	31.82	2.89
2		1.30	-0.87	34.35	3.44					
2	A	1.35	-0.87	34.78	3.44	B	2.13	0.37	65.71	0.00
2	C	2.80	0.12	56.41	2.44	D	1.28	-0.85	34.81	3.33
3		1.06	-1.09	33.43	6.22					
3	A	1.11	-1.05	33.87	6.00	B	1.67	0.05	59.71	0.89
3	C	2.48	-0.41	50.47	4.33	D	1.04	-1.11	33.57	6.44

Table 7.19: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model with varying estimation of δ and γ for different intervals (int) where 6 B-spline basis functions of order 3 are used for $n_s = 1$. Refer to Section 7.1.3 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		3.01	-4.65	16.64	3.00					
1	A	3.06	-4.68	16.94	3.00	B	4.25	-3.07	30.82	1.50
1	C	4.06	-3.96	25.36	3.00	D	3.01	-4.58	16.78	3.00
2		2.18	-2.68	19.77	3.00					
2	A	2.22	-2.72	19.84	3.00	B	3.42	-0.89	43.91	1.00
2	C	3.88	-1.60	36.38	3.00	D	2.17	-2.62	20.49	3.00
3		1.84	-2.85	19.97	4.00					
3	A	1.90	-2.83	20.30	4.00	B	2.86	-1.08	42.83	2.00
3	C	3.63	-1.95	37.01	3.00	D	1.82	-2.84	20.26	4.00

Table 7.20: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model with varying estimation of δ and γ for different intervals (int) where 6 B-spline basis functions of order 3 are used for $n_s = 2$. Refer to Section 7.1.3 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		3.68	-8.45	9.47	3.33					
1	A	3.72	-8.50	9.53	3.33	B	5.05	-6.69	17.78	1.67
1	C	4.67	-7.84	15.91	3.33	D	3.70	-8.35	10.00	3.33
2		2.72	-5.15	12.81	3.33					
2	A	2.75	-5.22	13.10	3.33	B	4.23	-2.99	29.30	2.50
2	C	4.42	-4.14	25.38	3.33	D	2.73	-5.05	13.04	3.33
3		2.23	-5.32	12.81	7.50					
3	A	2.28	-5.32	13.10	7.50	B	3.56	-3.11	29.47	2.50
3	C	4.09	-4.46	27.31	4.17	D	2.22	-5.30	12.87	7.50

Table 7.21: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model with varying estimation of δ and γ for different intervals (int) where 6 B-spline basis functions of order 3 are used for $n_s = 3$. Refer to Section 7.1.3 for cases (adjustments).

Our original estimates are robust in terms of lowest FP or FN.

The interval int selected is 3 with number of curves $n_s = 3$ in a set in a comparison. The lLR obtained is -5.85.

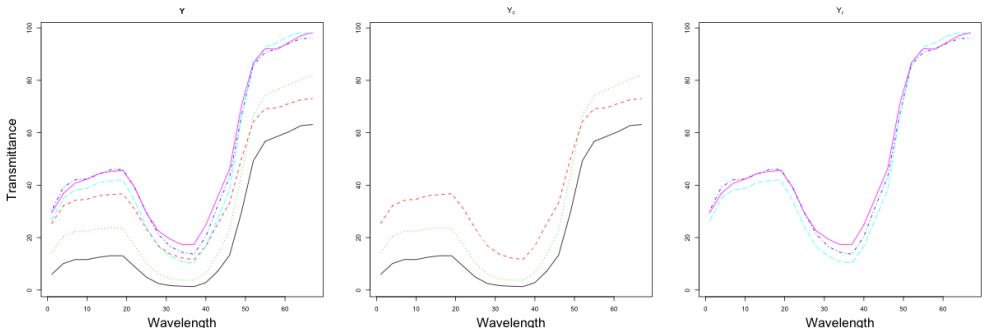


Figure 7.13: Curves from within the group 14 yet negative lLR is obtained under model CA-ar. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

Curves from the same group (14) have been picked up for all component-wise additive models to give negative lLRs.

The interval int selected is 2 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is 3.55.

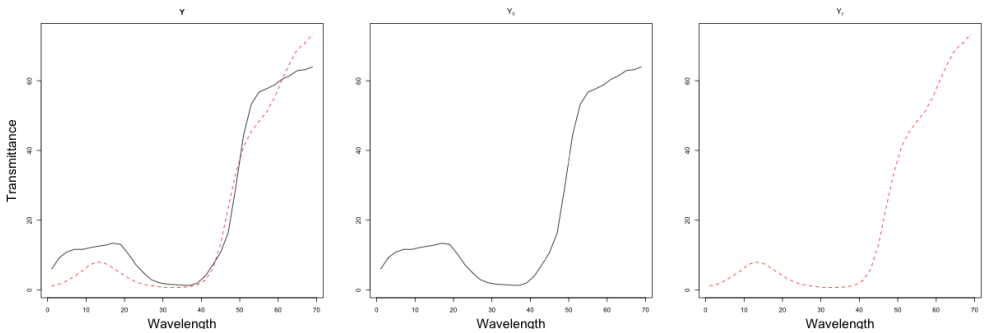


Figure 7.14: Curves from groups 14 and 7 yet positive lLR is obtained under model CA-ar. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

These two curves have different shapes; however, it is also hard to distinguish by eyes when looking at them separately.

7.3.4 DR-S Dimension reduced multivariate normal random-effects model - wool data

For the dimension reduced multivariate normal random-effects model, $\boldsymbol{U}$, the within-group variance-covariance is assumed to be constant for all groups. For the sensitivity analyses we will consider four cases with varying $\boldsymbol{U}$.

- A: $\boldsymbol{U}_{new} = 0.5 \times \hat{\boldsymbol{U}}$
- B: $\boldsymbol{U}_{new} = 2.5 \times \hat{\boldsymbol{U}}$
- C: $\boldsymbol{U}_{new} = \hat{\boldsymbol{U}} - 0.2diag(\hat{\boldsymbol{U}})$
- D: $\boldsymbol{U}_{new} = \hat{\boldsymbol{U}} + 0.2diag(\hat{\boldsymbol{U}})$

The results are presented the same way as in Chapter 6 alongside results from Chapter 6 reproduced here under the Case Original. However, only selections of B will be considered due to the similarity of their performances.

B	Case	S	D	FP	FN	Case	S	D	FP	FN
4	A	1.28	-15.90	5.93	15.00	C	1.98	5.33	90.22	0.78
4	Original	1.39	-6.97	10.70	7.67	Original	1.39	-6.97	10.70	7.67
4	B	1.09	-2.05	23.47	3.00	D	1.05	-1.44	24.83	3.00
6	A	0.72	-23.40	3.00	16.00	C	-0.79	23.26	87.90	33.44
6	Original	1.48	-10.21	6.17	9.33	Original	1.48	-10.21	6.17	9.33
6	B	1.33	-3.01	17.96	3.89	D	1.43	-1.89	19.21	4.44
8	A	1.06	-36.84	1.63	17.00	C	2.34	-463.97	68.60	37.11
8	Original	2.10	-16.24	3.72	9.33	Original	2.10	-16.24	3.72	9.33
8	B	1.88	-4.93	12.25	4.67	D	1.63	-2.24	17.72	4.56

Table 7.22: Summary table of lLR's obtained using dimension reduced multivariate normal random-effects model with varying estimates for $\boldsymbol{U}$ for different number (B) of B-spline basis functions of order 3 for $n_s = 1$. Refer to Section 7.1.4 for cases (adjustments).

B	Case	S	D	FP	FN	Case	S	D	FP	FN
4	A	1.61	-33.33	3.49	13.00	C	2.38	8.72	85.33	1.50
4	Original	1.82	-15.39	6.64	7.50	Original	1.82	-15.39	6.64	7.50
4	B	1.53	-5.10	13.29	4.50	D	1.54	-3.77	14.74	3.00
6	A	0.41	-48.75	1.51	16.50	C	8.60	53.81	89.08	6.50
6	Original	1.74	-22.44	3.03	9.50	Original	1.74	-22.44	3.03	9.50
6	B	1.87	-7.39	8.26	4.00	D	2.08	-5.07	10.26	4.00
8	A	0.81	-77.25	0.82	18.00	C	7.06	19.84	73.03	8.50
8	Original	2.65	-35.64	1.55	8.50	Original	2.65	-35.64	1.55	8.50
8	B	2.71	-11.91	5.36	3.50	D	2.46	-5.96	9.24	4.00

Table 7.23: Summary table of lLR's obtained using dimension reduced multivariate normal random-effects model with varying estimates for $\boldsymbol{U}$ for different number (B) of B-spline basis functions of order 3 for $n_s = 2$. Refer to Section 7.1.4 for cases (adjustments).

B	Case	S	D	FP	FN	Case	S	D	FP	FN
4	A	1.36	-51.50	2.16	15.83	C	2.69	11.54	80.53	1.67
4	Original	1.89	-24.26	4.56	10.83	Original	1.89	-24.26	4.56	10.83
4	B	1.75	-8.43	9.42	6.67	D	1.79	-6.52	10.12	5.00
6	A	-1.10	-76.07	0.88	20.83	C	10.34	69.99	91.52	2.50
6	Original	1.28	-35.78	1.93	13.33	Original	1.28	-35.78	1.93	13.33
6	B	2.00	-12.36	5.73	5.83	D	2.37	-8.79	6.20	5.00
8	A	-1.61	-119.90	0.53	19.17	C	11.15	29.73	77.37	4.17
8	Original	1.92	-56.41	1.05	14.17	Original	1.92	-56.41	1.05	14.17
8	B	2.90	-19.63	2.98	5.83	D	2.87	-10.24	5.44	5.83

Table 7.24: Summary table of lLR's obtained using dimension reduced multivariate normal random-effects model with varying estimates for $\boldsymbol{U}$ for different number (B) of B-spline basis functions of order 3 for $n_s = 3$. Refer to Section 7.1.4 for cases (adjustments).

Based on sums of FN and FP alone, larger variance (covariance) as represented by cases B and D sometimes perform better than original estimates.

The number of basis functions B selected is 6 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is -50.17.

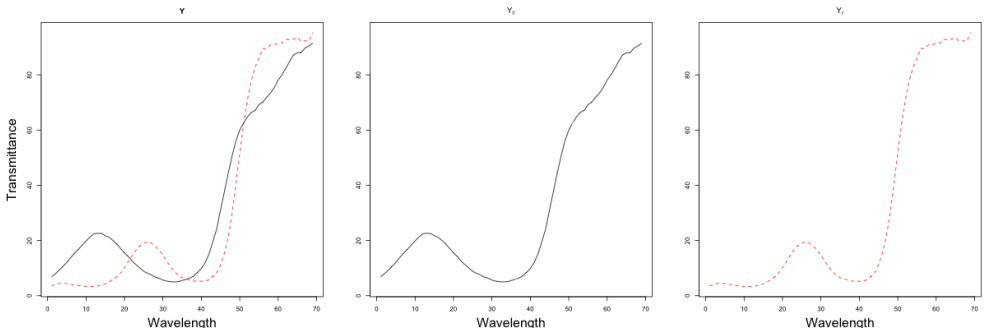


Figure 7.15: Curves from within the group 7 yet negative lLR is obtained under model DR-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

It makes sense to get a negative lLR for this pair of curves as they look completely different and the within-group variation is higher than all other groups.

The number of basis functions B selected is 6 with number of curves $n_s = 3$ in a set in a comparison. The lLR obtained is 4.34.

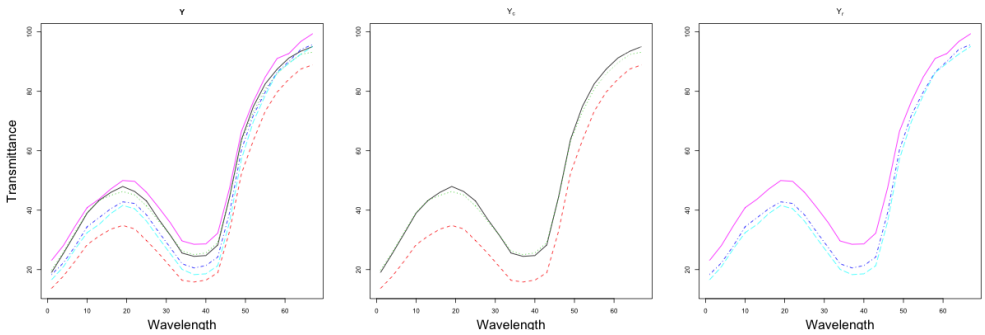


Figure 7.16: Curves from groups 10 and 5 yet positive lLR is obtained under model DR-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

These sets of curves are hard to tell apart by eye.

7.3.5 Conclusion

We can always tell from the plots why lLR's of the wrong sign as expected are obtained as these groups are exceptionally hard to distinguish, even by eyes. However, the

selected cases are the most extreme ones so it might be easier than those not drawn here. Overall, the models are performing as expected.

7.4 Cotton data

Recall from Chapter 6 that CA-ar and DR-C are the best performing models for original cotton data. This might be due to the fact that within-group variation is considered or modelled correctly.

7.4.1 CA-S Simplified multivariate normal random-effects model - cotton data

For this model, the estimates of the variance (σ^2) will be both increased and decreased by 20% and results will be presented in the same way as in Chapter 6.

Adjustment of	int	S	D	FP	FN
Subtract 20%	1	-8.46	-36.50	10.40	54.56
Original	1	-6.08	-28.38	13.13	51.44
Plus 20%	1	-4.54	-23.02	15.38	48.44
Subtract 20%	2	-2.71	-16.40	19.09	45.11
Original %	2	-1.68	-12.51	22.16	41.11
Plus 20%	2	-1.03	-9.97	24.98	39.00
Subtract 20%	3	-1.03	-9.98	25.13	39.11
Original	3	-0.44	-7.49	28.60	35.00
Plus 20%	3	-0.08	-5.87	31.25	31.56

Table 7.25: Summary table of lLR's obtained using simplified multivariate normal random-effects model with varying estimation of σ^2 for different intervals (int) where 6 B-spline basis functions of order 3 are used for $n_s = 1$.

The trade-off among FP and FN suggests there is limitations as how well the model can perform and the estimate of parameter does not help too much if the model does not fit well.

Adjustment of	int	S	D	FP	FN
Subtract 20%	1	-7.48	-64.35	5.86	45.50
Original	1	-5.23	-50.58	7.93	43.50
Plus 20%	1	-3.77	-41.45	9.70	41.00
Subtract 20%	2	-2.04	-30.10	13.19	37.50
Original	2	-1.06	-23.37	16.64	35.50
Plus 20%	2	-0.44	-18.93	19.61	33.50
Subtract 20%	3	-0.44	-18.97	19.70	33.50
Original	3	0.12	-14.58	23.32	33.00
Plus 20%	3	0.46	-11.69	25.59	31.50

Table 7.26: Summary table of lLR's obtained using simplified multivariate normal random-effects model with varying estimation of σ^2 for different intervals (int) where 6 B-spline basis functions of order 3 are used for $n_s = 2$.

Adjustment of	int	S	D	FP	FN
Subtract 20%	1	-7.16	-92.46	3.86	37.50
Original	1	-4.92	-72.99	4.85	31.67
Plus 20%	1	-3.46	-60.07	6.37	28.33
Subtract 20%	2	-1.74	-43.97	8.83	28.33
Original	2	-0.75	-34.40	10.99	28.33
Plus 20%	2	-0.14	-28.06	13.92	27.50
Subtract 20%	3	-0.15	-28.13	14.44	27.50
Original	3	0.42	-21.83	18.36	23.33
Plus 20%	3	0.76	-17.67	20.82	22.50

Table 7.27: Summary table of lLR's obtained using simplified multivariate normal random-effects model with varying estimation of σ^2 for different intervals (int) where 6 B-spline basis functions of order 3 are used for $n_s = 3$.

The interval int selected is 1 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is -93.99.

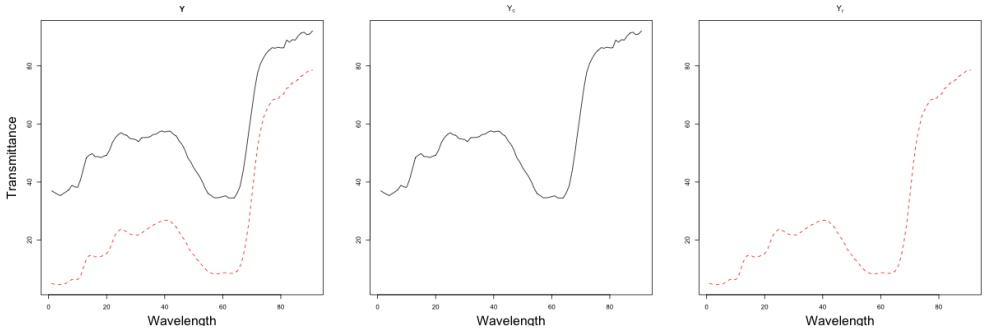


Figure 7.17: Curves from within the group 5 yet negative lLR is obtained under model CA-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

Even though the curves being compared have the same shape as represented by local minimum and maximums, the separation or vertical distance might be the cause of getting a negative lLR.

The interval int selected is 3 with number of curves $n_s = 3$ in a set in a comparison. The lLR obtained is 6.23.

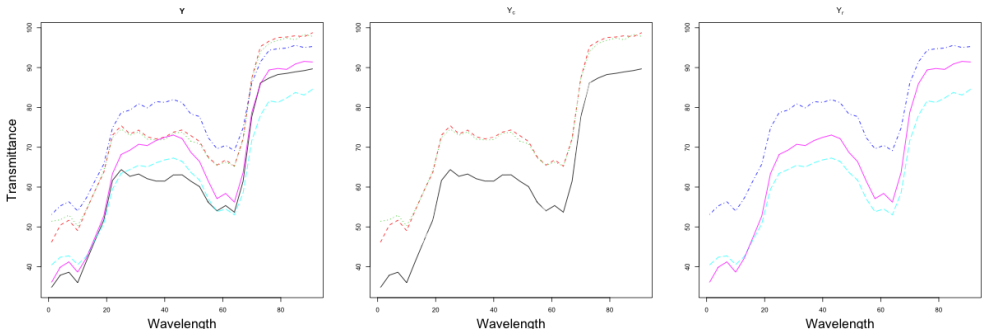


Figure 7.18: Curves from groups 5 and 3 yet positive lLR is obtained under model CA-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

We can see some differences in the shapes among the sets of curves being compared; however, a positive lLR is obtained, which suggests more weight is put on the distance than the shape.

7.4.2 CA-const. Constant within group covariance model - cotton data

Under contant within-group variance model, the hyperparameters associated with within-group variances are γ and δ. Four cases are considered for sensitivity analysis; they are

- A: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- B: $\gamma_{new} = 0.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$
- C: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 0.5 \times \hat{\delta}$
- D: $\gamma_{new} = 1.5 \times \hat{\gamma}, \delta_{new} = 1.5 \times \hat{\delta}$

The results are presented the same way as in Chapter 6 alongside results from Chapter 6 reproduced here under the Case Original.

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		-9.21	-28.54	5.00	61.89					
1	A	-9.52	-28.93	4.51	62.67	B	-6.25	-24.48	11.84	53.11
1	C	-10.39	-30.89	3.98	64.00	D	-8.95	-28.22	5.53	60.78
2		-3.28	-12.41	11.43	53.00					
2	A	-3.37	-12.53	10.69	53.67	B	-1.15	-9.37	24.39	42.00
2	C	-3.93	-14.08	9.53	57.33	D	-3.20	-12.32	12.12	52.22
3		-1.69	-7.58	16.86	48.78					
3	A	-1.71	-7.60	16.08	49.33	B	0.00	-5.11	31.83	35.11
3	C	-2.10	-8.92	14.74	51.67	D	-1.67	-7.59	17.49	48.67

Table 7.28: Summary table of lLR's obtained using constant within group covariance model and manipulating estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 6 and 3 for B-spline basis functions for $n_s = 1$. Refer to Section 7.1.2 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		-8.93	-46.48	1.81	49.00					
1	A	-9.10	-46.67	1.84	49.50	B	-6.70	-43.19	4.14	48.00
1	C	-8.91	-47.56	2.47	48.50	D	-8.74	-46.28	2.04	49.00
2		-3.05	-21.28	5.16	47.50					
2	A	-3.12	-21.35	5.00	47.50	B	-1.21	-18.49	11.51	41.50
2	C	-2.83	-22.08	6.91	45.50	D	-2.95	-21.19	5.43	47.50
3		-1.42	-13.36	8.62	44.00					
3	A	-1.45	-13.37	8.49	44.50	B	0.16	-10.89	18.36	36.50
3	C	-1.10	-14.00	11.15	42.00	D	-1.36	-13.33	8.91	44.00

Table 7.29: Summary table of lLR's obtained using constant within group covariance model and manipulating estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 6 and 3 for B-spline basis functions for $n_s = 2$. Refer to Section 7.1.2 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		-9.31	-64.38	1.70	46.67					
1	A	-9.43	-64.49	1.70	46.67	B	-7.31	-61.34	2.63	42.50
1	C	-8.93	-65.06	2.34	45.83	D	-9.14	-64.22	1.75	45.83
2		-3.20	-30.17	3.45	40.83					
2	A	-3.27	-30.22	3.16	40.00	B	-1.46	-27.46	7.49	35.83
2	C	-2.72	-30.68	4.91	37.50	D	-3.10	-30.09	3.57	40.83
3		-1.47	-19.20	5.73	37.50					
3	A	-1.50	-19.21	5.85	37.50	B	0.10	-16.73	12.87	33.33
3	C	-0.91	-19.61	8.54	34.17	D	-1.40	-19.17	6.02	37.50

Table 7.30: Summary table of lLR's obtained using constant within group covariance model and manipulating estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 6 and 3 for B-spline basis functions for $n_s = 3$. Refer to Section 7.1.2 for cases (adjustments).

The interval int selected is 2 with number of curves $n_s = 1$ in a set in a comparison.

The lLR obtained is -35.42.

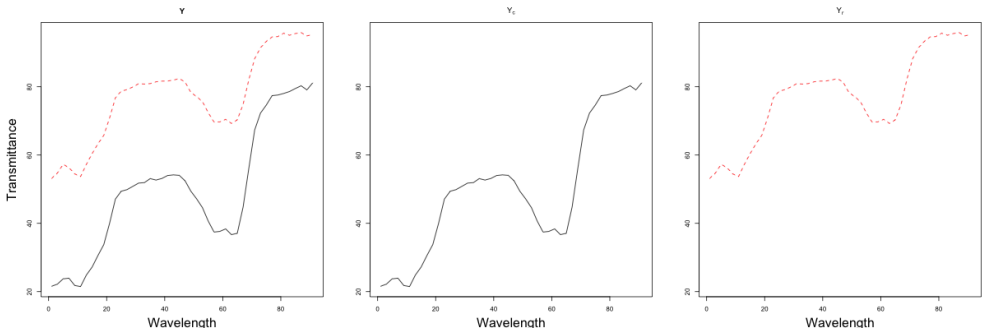


Figure 7.19: Curves from within the group 6 yet negative lLR is obtained under model CA-const.. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The interval int selected is 3 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is 4.55.

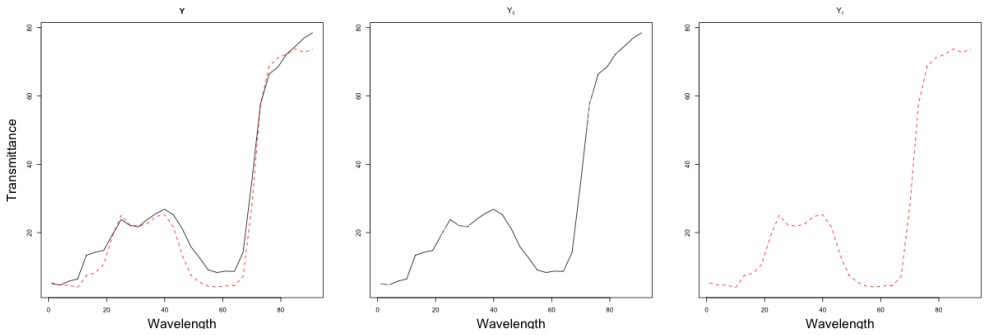


Figure 7.20: Curves from groups 5 and 3 yet positive lLR is obtained under model CA-const.. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

Based on these figures, it can be seen that distance is again the main reason for the lLR's of wrong signs than expected are obtained

7.4.3 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model - cotton data

Four cases are considered for sensitivity analysis. They are the same as CA-const..

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		0.69	-1.66	29.69	15.78					
1	A	0.72	-1.67	31.32	15.33	B	1.91	-0.23	58.10	3.67
1	C	2.22	-0.35	55.78	6.00	D	0.70	-1.61	29.59	16.11
2		0.56	-0.94	35.04	13.89					
2	A	0.61	-0.94	36.92	12.56	B	1.48	0.17	66.69	2.11
2	C	1.84	0.10	62.16	5.11	D	0.56	-0.91	34.78	14.44
3		0.46	-0.80	36.86	15.89					
3	A	0.52	-0.78	38.84	14.44	B	1.22	0.16	68.01	5.44
3	C	1.57	0.05	62.14	7.11	D	0.45	-0.80	36.71	16.44

Table 7.31: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model and manipulating estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 6 and 3 for B-spline basis functions for $n_s = 1$. Refer to Section 7.1.3 for cases (adjustments).

Our original estimates are robust in terms of lowest FP or FN.

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		1.23	-3.78	20.95	8.00					
1	A	1.28	-3.78	22.17	8.00	B	2.58	-2.20	40.56	1.50
1	C	2.93	-2.31	42.99	1.50	D	1.24	-3.72	21.02	8.50
2		0.96	-2.27	27.24	8.50					
2	A	1.01	-2.26	28.36	5.00	B	2.05	-0.94	50.30	1.50
2	C	2.45	-1.04	49.31	1.50	D	0.96	-2.24	26.81	8.50
3		0.79	-1.91	28.88	9.00					
3	A	0.85	-1.88	30.10	8.00	B	1.75	-0.70	53.32	2.50
3	C	2.16	-0.83	51.28	3.00	D	0.78	-1.89	28.98	10.00

Table 7.32: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model and manipulating estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 6 and 3 for B-spline basis functions for $n_s = 2$. Refer to Section 7.1.3 for cases (adjustments).

int	Case	S	D	FP	FN	Case	S	D	FP	FN
1		1.47	-6.26	14.74	9.17					
1	A	1.52	-6.27	15.85	10.00	B	2.90	-4.59	30.94	1.67
1	C	3.20	-4.77	33.63	4.17	D	1.49	-6.19	15.03	9.17
2		1.11	-3.87	20.88	12.50					
2	A	1.16	-3.87	21.87	10.00	B	2.32	-2.41	38.95	3.33
2	C	2.64	-2.61	39.88	2.50	D	1.12	-3.82	20.82	12.50
3		0.92	-3.23	23.39	10.83					
3	A	0.98	-3.21	24.44	10.83	B	2.00	-1.87	42.92	4.17
3	C	2.35	-2.10	41.75	5.83	D	0.92	-3.20	23.51	10.00

Table 7.33: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model and manipulating estimation of δ and γ for different intervals (int) where number of basis (B) and order of basis used are 6 and 3 for B-spline basis functions for $n_s = 3$. Refer to Section 7.1.3 for cases (adjustments).

The interval int selected is 3 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is -3.42.

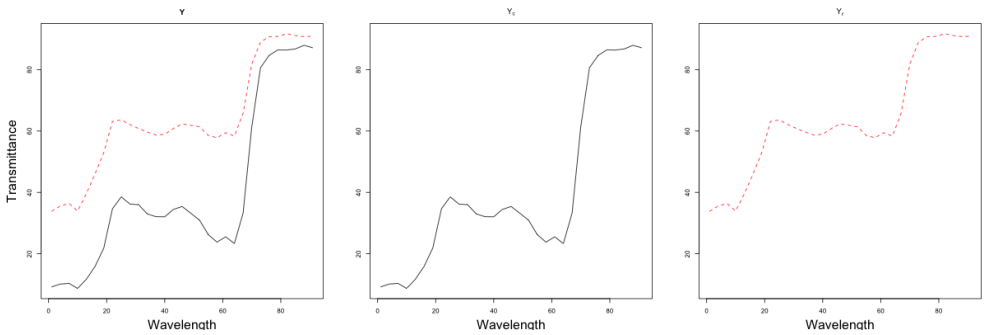


Figure 7.21: Curves from within the group 19 yet negative lLR is obtained under model CA-ar. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The interval int selected is 3 with number of curves $n_s = 2$ in a set in a comparison. The lLR obtained is 1.90.

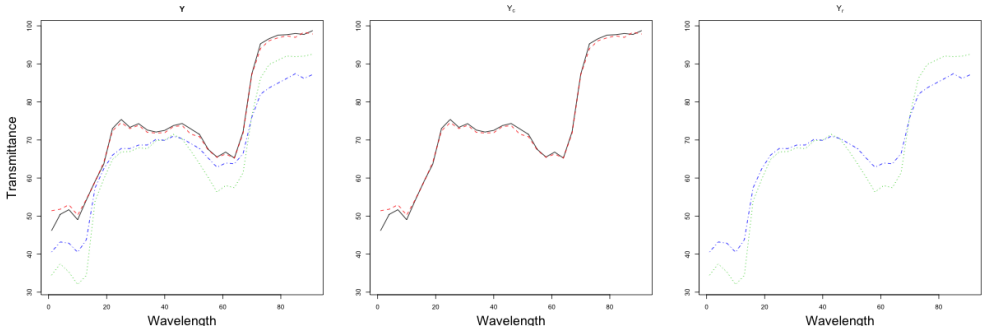


Figure 7.22: Curves from groups 15 and 11 yet positive lLR is obtained under model CA-ar. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

Similarly to CA-const., distance is more important than shape.

7.4.4 DR-S Dimension reduced multivariate normal random-effects model - cotton data

For the dimension reduced multivariate normal random-effects model, $\boldsymbol{U}$, the within-group variance-covariance is assumed to be constant for all groups. For the sensitivity analyses we will consider four cases with varying $\boldsymbol{U}$.

- A: $\boldsymbol{U}_{new} = 0.5 \times \hat{\boldsymbol{U}}$
- B: $\boldsymbol{U}_{new} = 2.5 \times \hat{\boldsymbol{U}}$
- C: $\boldsymbol{U}_{new} = \hat{\boldsymbol{U}} - 0.2diag(\hat{\boldsymbol{U}})$
- D: $\boldsymbol{U}_{new} = \hat{\boldsymbol{U}} + 0.2diag(\hat{\boldsymbol{U}})$

The results are presented the same way as in Chapter 6 alongside results from Chapter 6 reproduced here under the Case Original. However, only selections of B will be considered due to the similarity of their performances.

B	Case	S	D	FP	FN	Case	S	D	FP	FN
4	A	0.40	-2.64	21.49	21.33	C	5.93	2.57	58.93	44.33
4	Original	0.48	-0.90	34.89	12.33	Original	0.48	-0.90	34.89	12.33
4	B	0.30	-0.14	54.11	5.78	D	0.34	-0.29	49.59	10.89
6	A	0.33	-6.34	11.32	17.22	C	1.57	47.90	74.66	40.44
6	Original	0.71	-2.37	20.80	10.78	Original	0.71	-2.37	20.80	10.78
6	B	0.54	-0.49	41.47	5.89	D	0.44	-0.25	48.41	9.22
8	A	0.80	-14.66	6.57	15.33	C	-1.96	-37.10	54.92	34.44
8	Original	1.41	-5.87	13.22	10.33	Original	1.41	-5.87	13.22	10.33
8	B	1.12	-1.42	29.71	6.11	D	0.60	-0.30	45.76	7.00

Table 7.34: Summary table of lLR's obtained using dimension reduced multivariate normal random-effects model with varying estimates for $\boldsymbol{U}$ for different number (B) of B-spline basis functions of order 3 for $n_s = 1$. Refer to Section 7.1.4 for cases (adjustments).

B	Case	S	D	FP	FN	Case	S	D	FP	FN
4	A	0.78	-5.69	13.39	17.50	C	-9.04	-1.23	61.48	32.50
4	Original	0.83	-2.22	23.59	10.00	Original	0.83	-2.22	23.59	10.00
4	B	0.57	-0.49	43.39	4.00	D	0.61	-0.83	41.18	6.00
6	A	0.66	-13.15	6.05	16.50	C	4.02	23.43	89.80	5.00
6	Original	1.17	-5.41	12.24	10.00	Original	1.17	-5.41	12.24	10.00
6	B	0.97	-1.38	27.47	5.50	D	0.83	-0.80	37.37	4.00
8	A	1.18	-30.64	3.16	15.00	C	6.34	-26.73	32.96	60.50
8	Original	2.14	-13.23	6.84	12.50	Original	2.14	-13.23	6.84	12.50
8	B	1.85	-3.79	17.37	7.50	D	1.13	-0.98	33.42	3.00

Table 7.35: Summary table of lLR's obtained using dimension reduced multivariate normal random-effects model with varying estimates for $\boldsymbol{U}$ for different number (B) of B-spline basis functions of order 3 for $n_s = 2$. Refer to Section 7.1.4 for cases (adjustments).

Our original estimates are robust in terms of lowest FP or FN.

B	Case	S	D	FP	FN	Case	S	D	FP	FN
4	A	0.94	-8.96	9.01	20.83	C	3.24	5.13	86.90	2.50
4	Original	1.02	-3.71	17.08	11.67	Original	1.02	-3.71	17.08	11.67
4	B	0.75	-0.95	35.26	1.67	D	0.78	-1.50	34.15	5.00
6	A	0.99	-20.37	4.21	18.33	C	7.95	11.50	85.79	21.67
6	Original	1.56	-8.77	8.54	9.17	Original	1.56	-8.77	8.54	9.17
6	B	1.31	-2.49	21.29	3.33	D	1.10	-1.53	29.06	3.33
8	A	1.53	-46.90	1.93	20.00	C	-16.25	-32.17	32.87	64.17
8	Original	2.83	-20.84	4.50	12.50	Original	2.83	-20.84	4.50	12.50
8	B	2.46	-6.42	12.40	4.17	D	1.51	-1.89	24.39	3.33

Table 7.36: Summary table of lLR's obtained using dimension reduced multivariate normal random-effects model with varying estimates for $\boldsymbol{U}$ for different number (B) of B-spline basis functions of order 3 for $n_s = 3$. Refer to Section 7.1.4 for cases (adjustments).

The number of basis functions B selected is 6 with number of curves $n_s = 1$ in a set in a comparison. The lLR obtained is -21.28.

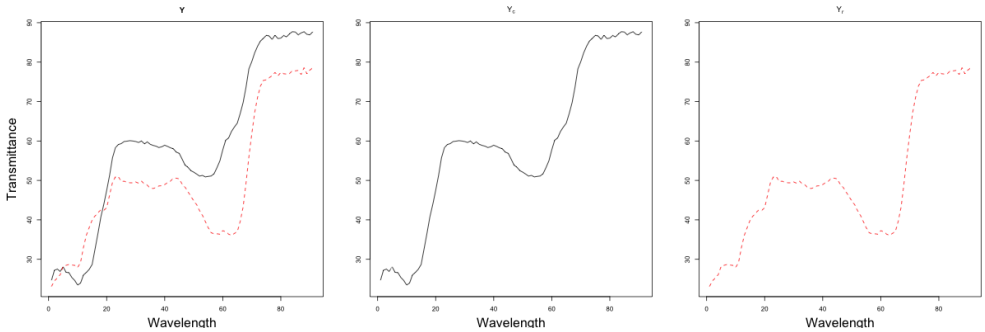


Figure 7.23: Curves from within the group 9 yet negative lLR is obtained under model DR-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The within-group variation among curves within group 9 makes it hard to find them from the same group.

The number of basis functions B selected is 6 with number of curves $n_s = 3$ in a set in a comparison. The lLR obtained is 3.51.

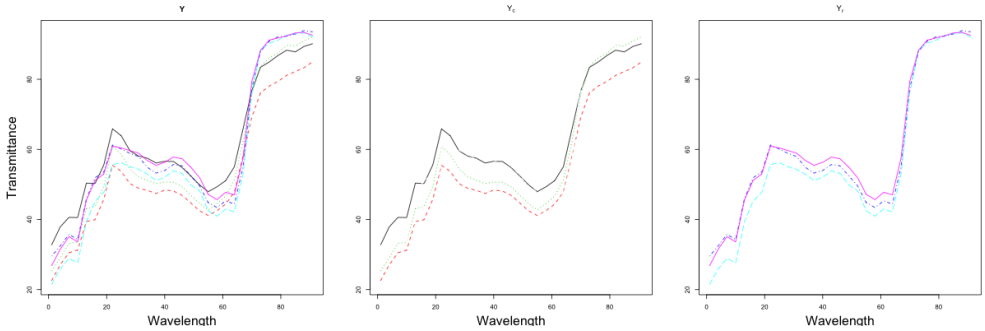


Figure 7.24: Curves from groups 13 and 7 yet positive lLR is obtained under model DR-S. The second panel shows the first set of curves, the third panel shows the second set of curves, the first panel shows all of these plotted together.

The shape of these sets of curves are somewhat different; however, they have local minima and maxima very close to the other set's so this can probably explain the positive lLR obtained. Moreover, it is also hard to tell apart from eye when they are plotted together.

7.4.5 Conclusion

The estimates are only sensitive if the model fits.

7.5 Conclusion

We do not need to worry about what estimates to use for variance as a small change has no effect on the overall performance of a model if it is a good fit to our data.

Chapter 8

More results - data preprocessing

8.1 Introduction

Recall in Chapter 4 and Chapter 7 that our models fail to model the separation between curves within-groups. We would like to manually process the data before modelling. The process we choose is taking differences since the general shapes of the curves are very similar but the separation of curves can depend on the shape at each point.

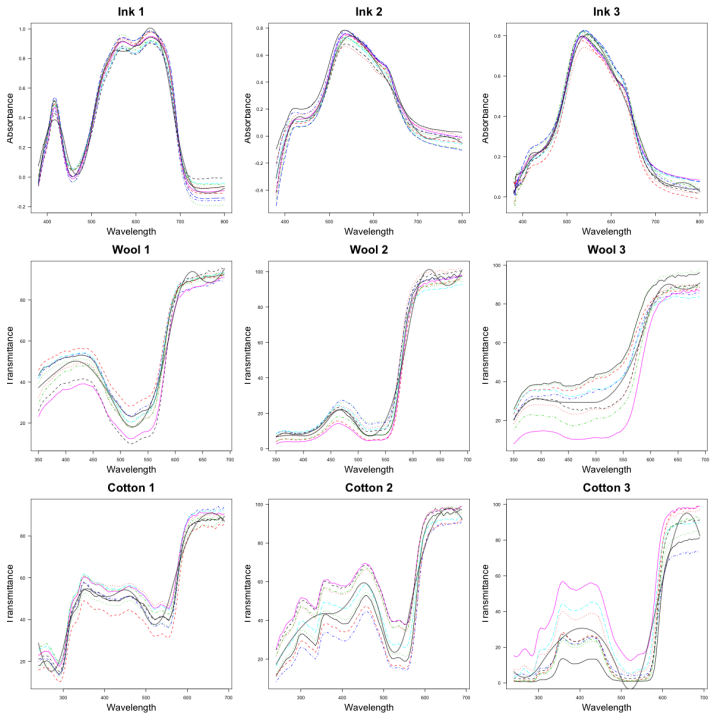


Figure 8.1: Plots of original data that show separation of curves within groups along with fitted mean curves using B-spline basis functions of order 3.

We can see from Figure 8.1 that even though B-spline basis functions can be used to approximate the means quite well, most of our models assume curves are centred at the means, which is not the case and therefore give mediocre results. The separation we see in all 3 types of ink as well as wool 1 and cotton 1 from Figure 8.1 are plain vertical separations but wool 3, cotton 2 and 3 show differences in curvatures as well as vertical separations. This can be seen easily by only looking at purple and black curves in wool 3 and cotton 3. However, cotton 2 is more complicated as there seems to be horizontal shifts between curves as well. We hope to resolve most of these problems by taking the differences of consecutive points on the curve. The new points are calculated as

$$\boldsymbol{y}^*_{ki} = [\frac{y_{ki2} - y_{ki1}}{2}, \frac{y_{ki3} - y_{ki1}}{2}, \cdots, \frac{y_{kim} - y_{ki,m-2}}{2}, \frac{y_{kim} - y_{ki,m-1}}{2}] = \frac{\boldsymbol{y}^{d,0}_{ki} + \boldsymbol{y}^{0,d}_{ki}}{2}$$

where

$$\begin{aligned}\boldsymbol{y}^{d,0}_{ki} &= [y_{ki2}, y_{ki3}, \cdots, y_{kim}, y_{kim}] - \boldsymbol{y}_{ki} \\ \boldsymbol{y}^{0,d}_{ki} &= \boldsymbol{y}_{ki} - [y_{ki1}, y_{ki1}, \cdots, y_{ki,m-2}, y_{ki,m-1}]\end{aligned}$$

for all ith curve in all group k. The rest of the chapter is organised as follows. Each section includes results for one dataset. There will be a short summary of the dataset along with some graphs to compare the new dataset with the original. Four of our five models proposed in Chapter 3 will be used to evaluate likelihood ratios using the same procedure described in Section 6.1 and reported in the same way. The same choices of parameters will be used for the ease of comparison so no more selection of basis functions will be done like in Chapter 5. The results will be compared with those in Chapter 6.

8.2 Ink data

Sample of ink data consists of $K = 40$ groups of $n = n_k = 10$ MSP measurements of absorbance $\boldsymbol{y}_{ki}$ versus wavelength for $1 \leq i \leq n$ for all k. Absorbance are measured at wavelengths ranging from 380-800 nm with intervals of 1nm so using all the points,

that is, taking interval or $int = 1$, the total number of points, the dimension of our data, is $m = 421$.

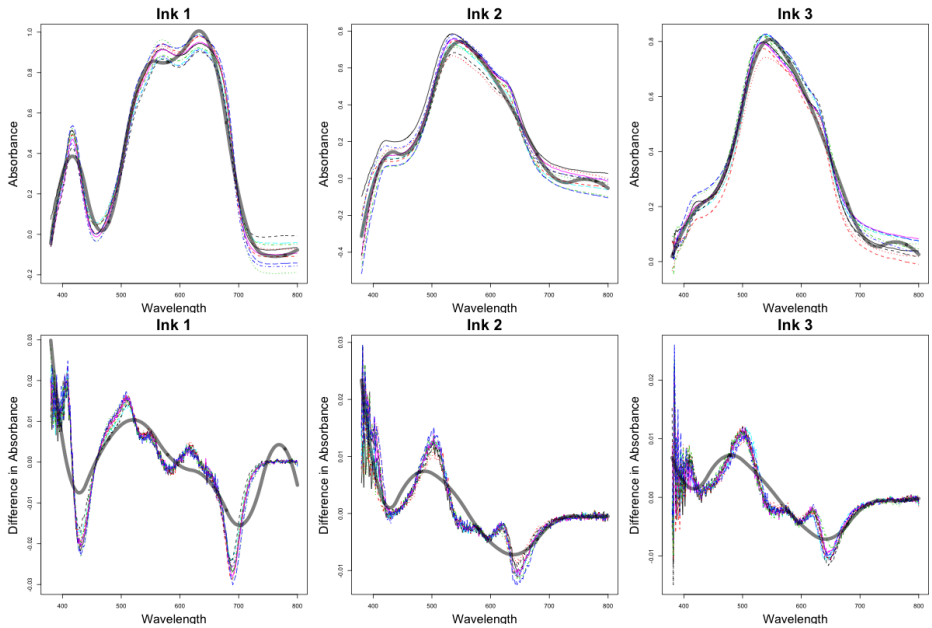


Figure 8.2: Fitting original and the first differences of three types of ink using 9 B-spline basis functions of order 3.

Using the same number of B-spline basis functions gives worse fit by the look but the scales of these plots are very different so residuals are a bit smaller. It is harder to differentiate between Ink 2 and Ink 3 compared to either one and Ink 1 due to the locations of their minima and maxima. However, curves from within the same groups are much closer now.

8.2.1 Summary table for preprocessed ink data

For each model, three tables of results will be reported for ink data for 3 distinct values of n_s. The three values are 1, 3 and 5. Since we have 10 measurements of one sample for each of the 40 different types of ink, there are $10 \times 11 \div 2 = 55$ within-group and $10 \times 10 = 100$ between-group lLR's for comparisons between 40 and $40 \times 39 \div 2 = 780$ pairs of groups for $n_s = 1$. For $n_s = 3$, lLR's are obtained for comparing a sets of $n_s = 3$ measurements with another (mutually exclusive) set of $n_s = 3$ measurements so there are $\lfloor \frac{10}{3} \rfloor \times (\lfloor \frac{10}{3} \rfloor + 1) \div 2 = 6$ within group and $\lfloor \frac{10}{3} \rfloor \times \lfloor \frac{10}{3} \rfloor = 9$ between group

lLR's for comparisons between 40 and $40 \times 39 \div 2 = 780$ pairs of groups. For $n_s = 5$ there are $\lfloor \frac{10}{5} \rfloor \times (\lfloor \frac{10}{5} \rfloor + 1) \div 2 = 3$ within-group and $\lfloor \frac{10}{5} \rfloor \times \lfloor \frac{10}{5} \rfloor = 4$ between-group lLR's for comparisons between 40 and $40 \times 39 \div 2 = 780$ pairs of groups.

8.2.2 CA-S Simplified multivariate normal random-effects model - ink data

Log likelihood ratios calculated using the simplified multivariate normal random-effects model for preprocessed ink data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	2.81	-113.85	0.97	21.86	-14.44	-581.25	0.04	66.09
1	5	3.37	-18.44	6.33	7.05	6.95	-6.11	18.78	2.77
3	1	4.04	-343.78	0.28	19.17	-14.03	-1696.85	0.00	43.33
3	5	5.38	-60.78	2.17	7.08	16.53	-35.68	7.81	2.50
5	1	4.02	-584.43	0.10	18.33	-16.97	-2856.73	0.00	31.67
5	5	6.20	-106.71	1.28	8.33	22.47	-73.52	5.03	2.50

Table 8.1: Summary table of lLR's obtained using simplified multivariate normal random-effects model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 9 and order of basis used is 3 for B-spline basis functions.

The result is improved compared to those obtained using the original data as shown in Table 6.2 in terms of smaller S and D in magnitudes and huge drop in FN rates from over 70% for $n_s = 1,\ int = 1$ when B-spline basis functions are used, to a reasonable 22%. The best setup $n_s = 5,\ int = 5$ also has much lowered FN although this is offset by an increase in FP. Using eigenfunctions, on the other hand, does not result in significant improvement but some improvements can be seen. Comparing to results in Table 6.2, the magnitudes of S and D are halved for $int = 1$. It is surprising to see the effects of taking differences to the use of different basis functions. Under the setup $n_s = 3,\ int = 5$, the performance of using B-spline basis functions is comparable to that of using eigenfunctions from fPCA in terms of sums of FP and FN in contrast to

the much larger FN in Table 6.2 when B-spline basis functions are used.

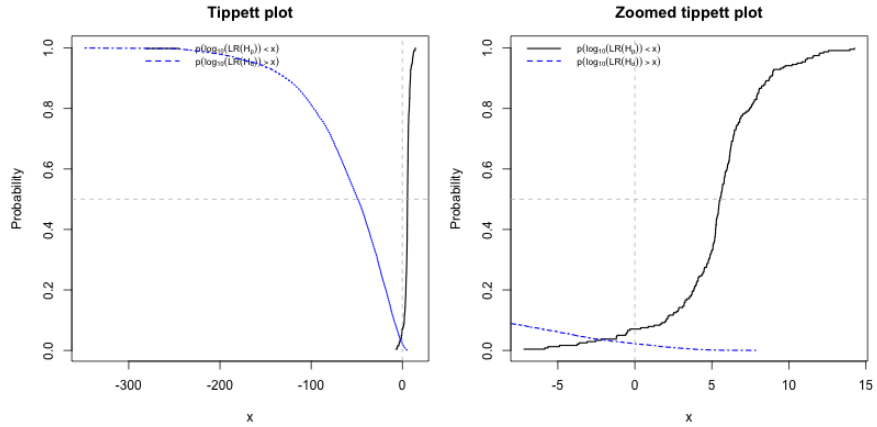


Figure 8.3: Tippett plot for ink data with setup $n_s = 3$, $int = 5$ under model CA-S.

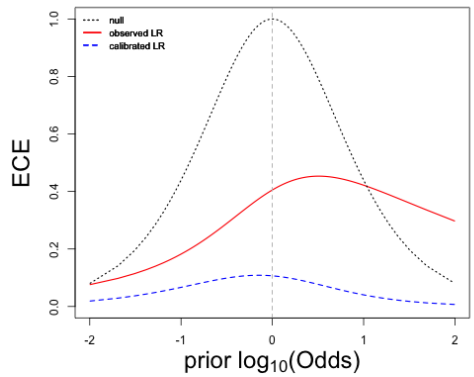


Figure 8.4: ECE for ink data with setup $n_s = 3$, $int = 5$ under model CA-S.

The ECE shows that using first order differences results in better fit by the model as indicated by much smaller loss of information.

8.2.3 CA-const. Constant within-group variance model - ink data

Log likelihood ratios calculated using the constant within-group variance model for preprocessed ink data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	0.81	-100.02	0.49	33.05	-14.10	-225.42	0.15	65.09
1	5	3.08	-16.28	3.48	10.64	7.76	-11.68	10.48	5.23
3	1	2.39	-299.17	0.16	25.83	-10.90	-616.01	0.07	42.50
3	5	4.99	-54.13	1.14	9.17	22.96	-59.46	3.97	3.33
5	1	1.92	-501.31	0.03	22.50	-12.01	-1006.22	0.00	29.17
5	5	5.66	-93.08	0.45	12.50	34.02	-119.63	2.37	3.33

Table 8.2: Summary table of lLR's obtained using constant within-group variance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 9 and order of basis used is 3 for B-spline basis functions.

Again, the results as shown in Table 8.2 are improved compared to those obtained using the original data as shown in Table 6.3, especially when B-spline basis functions are used; FN rates are halved for smaller n_s. The best setup is still $n_s = 5,\ int = 5$ when eigenfunctions from fPCA are used. The signs for S are corrected (positive now) when B-spline basis functions are used and D almost a third in magnitude. Overall, larger n_s and int gives better results.

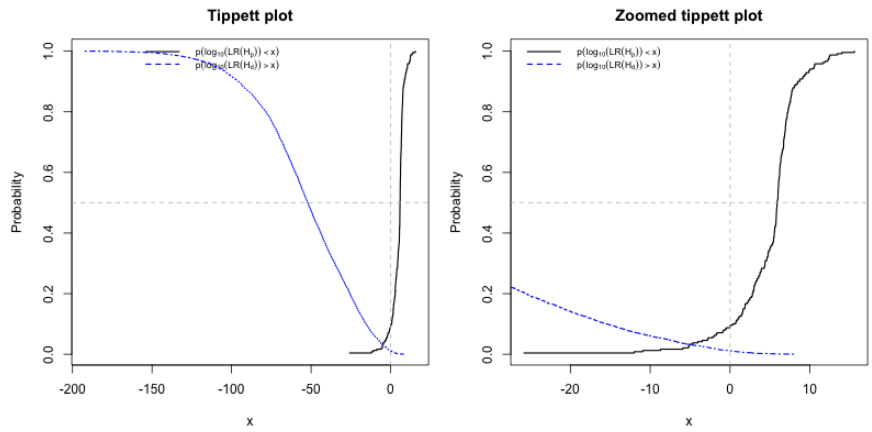


Figure 8.5: Tippett plot for ink data with setup $n_s = 3, int = 5$ under model CA-const..

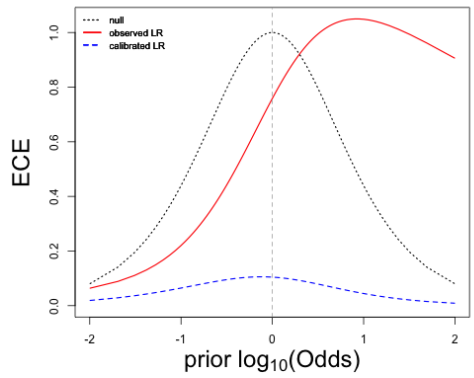


Figure 8.6: ECE for ink data with setup $n_s = 3$, $int = 5$ under model CA-const..

Similarly to CA-S, the ECE shows that using first order differences results in better fit by the model as indicated by much smaller loss of information comparing to using original data.

8.2.4 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model - ink data

Log likelihood ratios calculated using the multivariate normal random-effects with autoregressive within-group covariance model for preprocessed ink data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	0.48	-19.51	3.14	29.41	-5.51	-104.19	0.18	59.59
1	5	1.49	-7.17	7.11	18.41	2.86	-6.52	13.28	20.23
3	1	2.64	-58.12	0.90	18.33	-1.73	-285.52	0.07	37.92
3	5	3.16	-24.05	2.35	12.50	10.87	-29.03	4.25	11.67
5	1	3.60	-99.92	0.45	18.33	-0.66	-471.63	0.00	30.00
5	5	3.60	-41.99	1.31	10.83	16.87	-57.33	1.60	8.33

Table 8.3: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 9 and order of basis used is 3 for B-spline basis functions.

This is the first model for ink data that performs similarly for differences as for original data. However, the magnitudes of D are reduced at least when B-spline basis functions are used. Overall, the results obtained when eigenfunctions from fPCA are used are worse than those from original data but still outperform those when B-spline basis functions are used for this model.

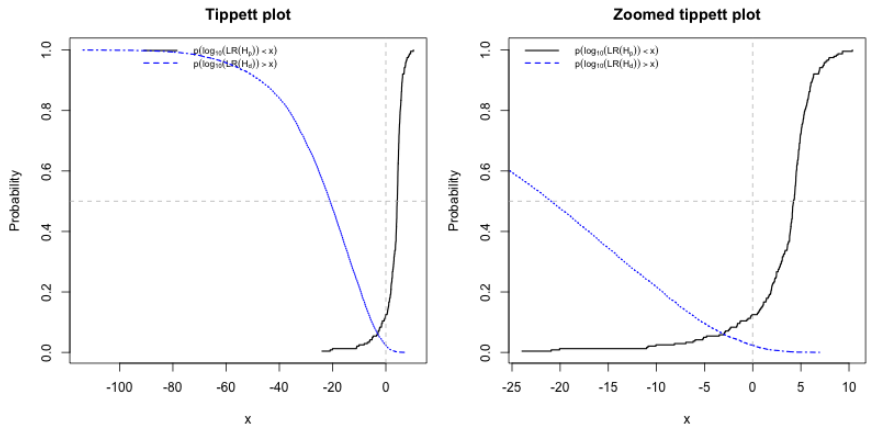


Figure 8.7: Tippett plot for ink data with setup $n_s = 3$, $int = 5$ under model CA-ar.

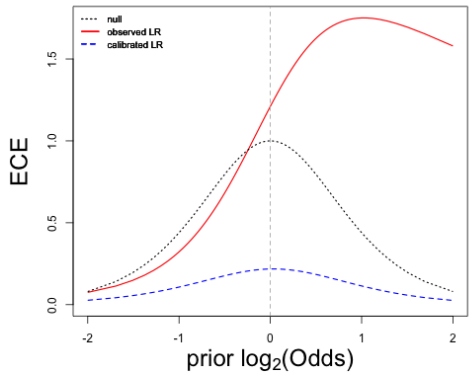


Figure 8.8: ECE for ink data with setup $n_s = 3$, $int = 5$ under model CA-ar.

There is no difference in the shape of the ECE obtained using first order differences under CA-ar compared to using original data just like the summary tables suggest.

8.2.5 DR-S Dimension reduced multivariate random-effects model - ink data

Log likelihood ratios calculated using the dimension reduced multivariate random-effects model for preprocessed ink data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.55	-2.89	31.64	6.50
2	-	-	-	-	1.25	-12.05	13.46	3.55
3	-	-	-	-	2.09	-27.14	9.04	2.14
4	2.09	-19.83	7.00	2.82	2.89	-42.78	5.56	1.18
5	2.80	-36.63	5.65	1.91	3.45	-47.80	4.69	0.82
6	3.42	-38.97	4.61	1.59	3.90	-54.51	4.31	0.68
7	4.01	-49.22	4.34	0.59	4.26	-62.94	4.06	0.73
8	4.78	-56.05	4.09	0.59	4.73	-65.79	3.39	0.68
9	5.52	-61.06	3.48	0.32	4.91	-68.84	3.22	0.59

Table 8.4: Summary table of lLR's for comparing sets of size $n_s = 1$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.70	-9.63	21.14	7.92
2	-	-	-	-	1.69	-38.34	6.51	2.92
3	-	-	-	-	2.84	-84.72	3.52	2.92
4	3.00	-63.18	3.02	3.75	3.93	-133.21	2.49	1.25
5	4.02	-115.16	2.26	2.50	4.81	-149.42	2.39	0.00
6	4.95	-123.22	1.67	0.42	5.53	-170.49	2.14	0.00
7	5.87	-155.38	1.52	0.42	6.12	-196.25	1.94	0.42
8	7.05	-177.32	1.47	0.42	6.94	-205.58	1.57	0.00
9	8.20	-194.16	1.44	0.42	7.29	-214.95	1.45	0.42

Table 8.5: Summary table of lLR's for comparing sets of size $n_s = 3$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

For $n_s = 1$, taking the difference makes no difference to the performance; however, for $n_s > 1$, both FP and FN rates dropped even to 0 in many cases for FN. This suggests that FN is caused by the separation of curves in the original data.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.73	-16.71	16.96	10.00
2	-	-	-	-	1.83	-65.40	4.20	7.50
3	-	-	-	-	3.11	-144.39	2.08	5.00
4	3.46	-108.13	1.89	5.83	4.32	-225.90	1.31	2.50
5	4.57	-195.18	1.60	4.17	5.32	-253.04	1.38	0.00
6	5.62	-209.56	1.06	2.50	6.17	-288.54	1.03	0.00
7	6.68	-263.53	0.96	0.83	6.88	-332.10	1.06	0.00
8	8.00	-301.32	0.67	0.00	7.83	-348.40	0.64	0.00
9	9.29	-329.79	0.64	0.00	8.30	-364.27	0.54	0.00

Table 8.6: Summary table of lLR's for comparing sets of size $n_s = 5$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

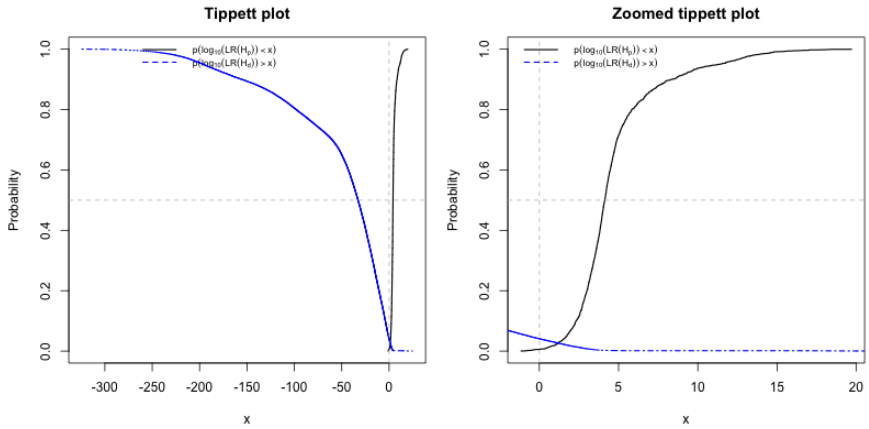


Figure 8.9: Tippett plot for ink data with setup $n_s = 1$, $B = 8$ under model DR-S.

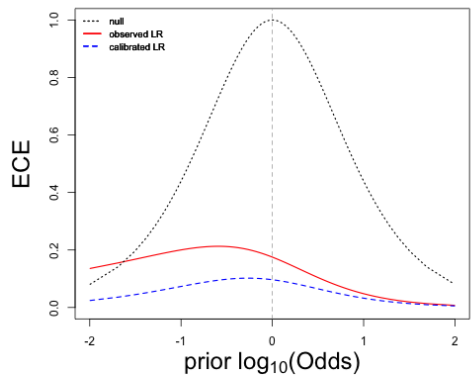


Figure 8.10: ECE for ink data with setup $n_s = 1$, $B = 8$ under model DR-S.

There is almost no loss of information for $\log_{10}(Odds)$ large.

8.2.6 Conclusion

Taking differences makes component-wise additive models perform better as indicated by Tippett and ECE plots. However, there is still room for improvement as indicated by high ECE at some $\log_{10}(Odd)$ values.

8.3 Wool data

Sample of ink data consists of $K = 20$ groups of $n = n_k = 9$ MSP measurements of transmittance $\boldsymbol{y}_{ki}$ versus wavelength for $1 \leq i \leq n$ for all k. Transmittance are measured at wavelengths ranging from 350-690 nm with intervals of 5 nm so using all the points, that is, taking interval or $int = 5$, the total number of points, the dimension of our data, is $m = 69$.

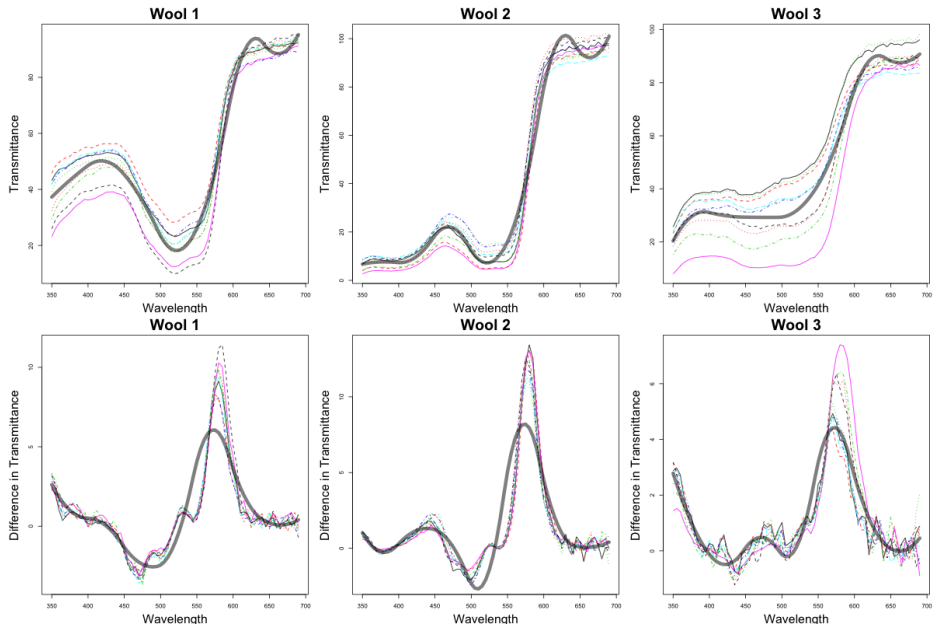


Figure 8.11: Fitting original and the first derivative of three types of wool using 6 B-spline basis functions of order 3.

Taking differences gets rid of some separations as we can see from Wool 2 in Figure 8.11 but like we mentioned previously in Chapter 4 there are more variations that

are dependent of the location or magnitudes as shown around 600 nm of Wool 3, which are still there and they represent within-group variations for our new data.

8.3.1 Summary tables for preprocessed wool data

For each model, three tables of results will be reported for wool data for 3 distinct values of n_s. The three values are 1, 2 and 3. Since we have 9 measurements of one sample for each of the 20 different types of wool fibres, there are $9 \times 10 \div 2 = 45$ within-group and $9 \times 9 = 81$ between-group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups for $n_s = 1$. For $n_s = 2$, lLR's are obtained for comparing sets of $n_s = 2$ measurements with another (mutually exclusive) set of $n_s = 2$ measurements so there are $\lfloor \frac{9}{2} \rfloor \times (\lfloor \frac{9}{2} \rfloor + 1) \div 2 = 10$ within group and $\lfloor \frac{9}{2} \rfloor \times \lfloor \frac{9}{2} \rfloor = 16$ between group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups. For $n_s = 3$ there are $\lfloor \frac{9}{3} \rfloor \times (\lfloor \frac{9}{3} \rfloor + 1) \div 2 = 6$ within-group and $\lfloor \frac{9}{3} \rfloor \times \lfloor \frac{9}{3} \rfloor = 9$ between-group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups.

8.3.2 CA-S Simplified multivariate normal random-effects model - wool data

Log likelihood ratios calculated using the simplified multivariate normal random-effects model for preprocessed wool data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	1.28	-0.93	28.21	3.22	-2.33	-73.63	0.71	31.33
1	2	0.73	-0.17	43.74	1.78	2.65	-11.56	8.01	7.22
2	1	2.02	-2.90	16.97	4.00	-3.42	-147.35	0.46	26.00
2	2	1.32	-0.84	28.72	2.00	3.74	-26.69	3.19	6.00
3	1	2.43	-5.29	10.29	5.00	-6.94	-224.33	0.41	27.50
3	2	1.73	-1.78	21.40	3.33	3.84	-43.26	2.28	7.50

Table 8.7: Summary table of lLR's obtained using simplified multivariate normal random-effects model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

Using B-spline basis functions performs worse on differences than original data in terms of enormous increases in FP rates. However, we see a slight improvement when eigenfunctions from fPCA are used. Overall, the setup $n_s = 3$, $int = 2$ gives reasonable result.

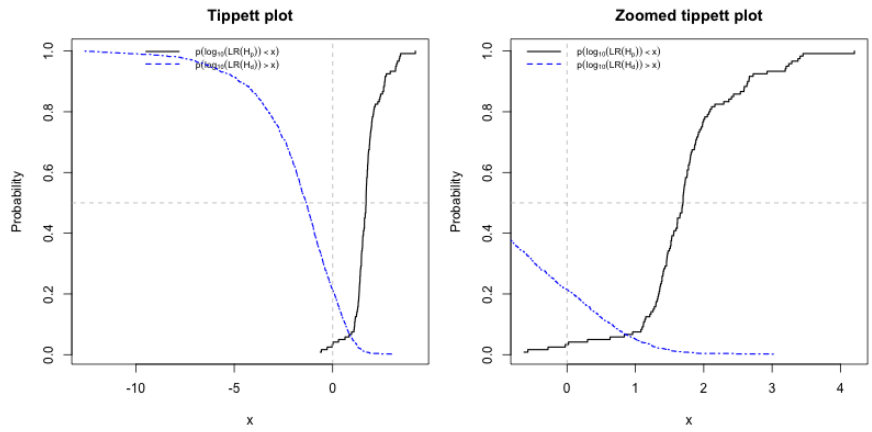


Figure 8.12: Tippett plot for wool data with setup $n_s = 3$, $int = 2$ under model CA-S.

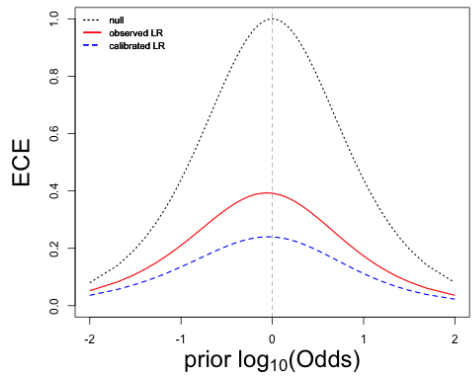


Figure 8.13: ECE for wool data with setup $n_s = 3$, $int = 2$ under model CA-S.

A slight improvement in performance suggested by the summary tables is actually a great one as ECE plots as shown in Figure 8.13 suggest. Using differences makes the model perform much better compared to using original data.

8.3.3 CA-const. Constant within-group variance model - wool data

Log likelihood ratios calculated using the constant within-group variance model for preprocessed wool data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	1.85	-4.58	17.67	5.44	-0.67	-35.87	0.83	33.56
1	2	1.37	-1.56	30.22	3.33	3.11	-8.70	5.54	8.78
2	1	2.66	-10.76	9.74	5.00	-0.53	-68.07	0.49	27.50
2	2	2.08	-4.20	19.31	3.00	4.94	-21.58	2.40	8.00
3	1	2.95	-18.05	5.91	8.33	-2.22	-102.56	0.41	25.00
3	2	2.46	-7.48	13.33	3.33	5.60	-36.49	1.46	10.83

Table 8.8: Summary table of lLR's obtained using constant within-group variance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

Compared to results obtained using original data, there is an general decline of FN especially when B-spline basis functions are used but this is offset by some increase in

FP. Overall the performance is improved and the best setup is $n_s = 2$, $int = 2$ when eigenfunctions from fPCA are used.

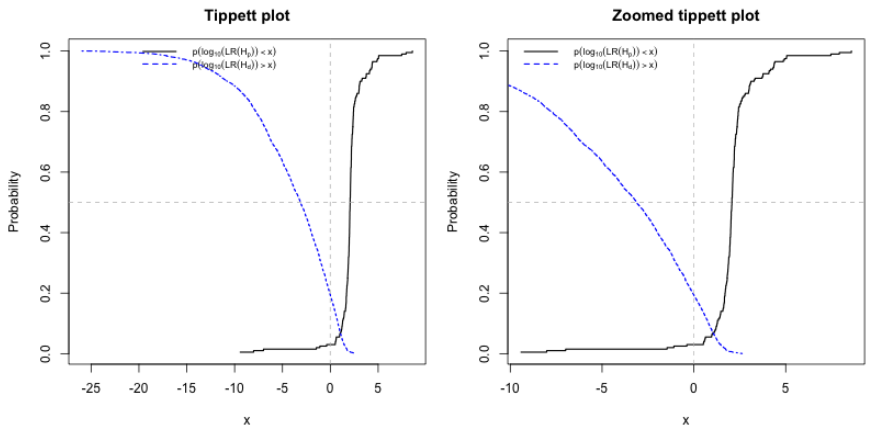


Figure 8.14: Tippett plot for wool data with setup $n_s = 2$, $int = 2$ under model CA-const..

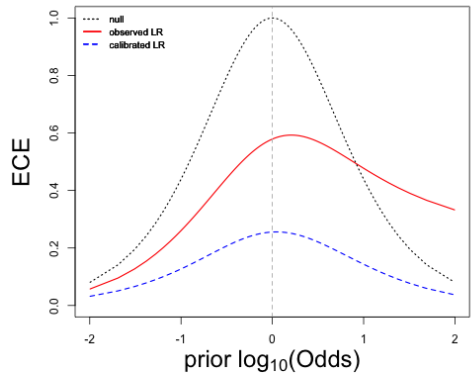


Figure 8.15: ECE for wool data with setup $n_s = 2$, $int = 2$ under model CA-const..

Similar to CA-S, this model performs much better on differences than on original data according to ECE plots.

8.3.4 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model - wool data

Log likelihood ratios calculated using the multivariate normal random-effects with autoregressive within-group covariance model for preprocessed wool data are sum-

marised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	0.89	-1.80	36.02	4.89	1.51	-16.27	3.70	16.00
1	2	0.82	-0.92	40.82	3.11	2.40	-6.67	9.85	7.78
2	1	1.29	-4.21	26.81	5.50	2.03	-32.66	1.48	13.50
2	2	1.27	-2.40	30.82	2.50	3.71	-16.14	4.54	6.50
3	1	1.44	-7.16	18.83	5.83	1.46	-50.83	1.05	14.17
3	2	1.47	-4.26	22.75	3.33	4.18	-27.15	3.10	7.50

Table 8.9: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

Based on Table 8.9 the performance is worse than those obtained using original data when B-spline basis functions are used. There is very little improvement for when eigenfunctions from fPCA are used. Compared to using original data, there is no improvement for $n_s < 3$, especially when B-spline basis functions are used. However, $n_s = 3,\ int = 2$ is acceptable in terms of FP and FN.

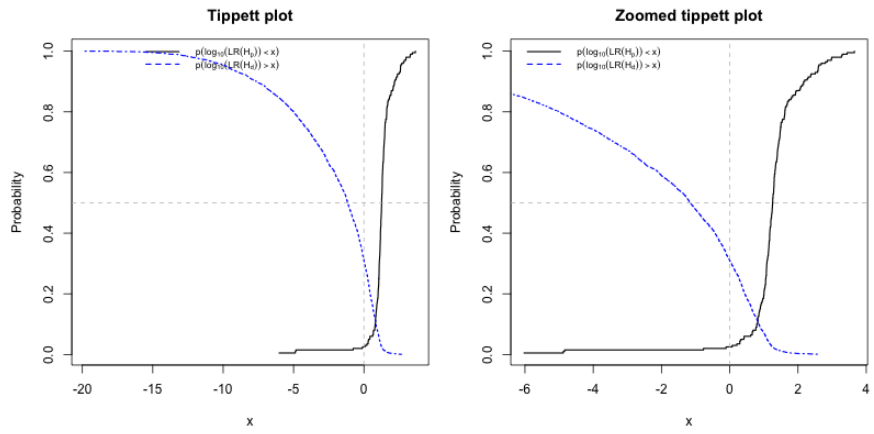


Figure 8.16: Tippett plot for wool data with setup $n_s = 2$, $int = 2$ under model CA-ar.

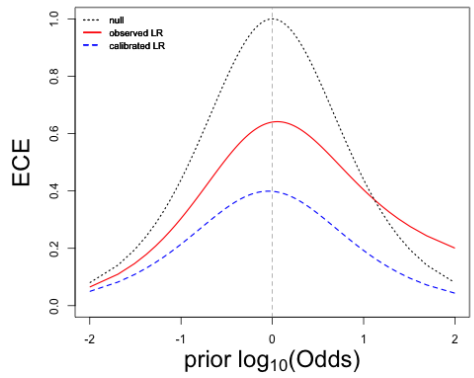


Figure 8.17: ECE for wool data with setup $n_s = 2$, $int = 2$ under model CA-ar.

The improvement made by using differences is more easily seen when comparing the ECE plots.

8.3.5 DR-S Dimension reduced multivariate random-effects model - wool data

Log likelihood ratios calculated using the dimension reduced multivariate random-effects model for preprocessed wool data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.30	-1.09	42.07	4.67
2	-	-	-	-	0.71	-2.69	23.31	7.33
3	-	-	-	-	1.11	-4.65	14.42	9.33
4	1.42	-4.66	11.33	8.22	1.56	-6.30	10.86	9.67
5	1.63	-8.40	6.99	8.67	1.80	-9.87	6.43	7.33
6	2.19	-11.68	5.11	7.89	2.36	-15.72	2.90	6.78
7	2.25	-14.24	3.90	9.11	2.47	-16.84	2.53	7.11
8	2.45	-17.05	3.29	8.78	2.80	-17.91	2.25	6.89
9	2.54	-20.94	2.82	8.67	2.89	-19.13	2.20	6.89

Table 8.10: Summary table of lLR's for comparing sets of size $n_s = 1$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

Under this model, the results obtained from using original data was already quite good. There is no improvement when differences are used.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.34	-2.56	34.38	4.50
2	-	-	-	-	0.88	-6.18	15.23	6.50
3	-	-	-	-	1.39	-10.60	7.57	9.00
4	1.92	-10.73	5.30	9.00	1.94	-14.37	5.46	9.00
5	2.03	-18.77	2.96	7.50	2.18	-22.03	2.24	8.50
6	2.89	-26.05	2.60	7.50	2.87	-34.23	1.25	8.50
7	2.90	-31.42	1.61	7.00	2.96	-36.85	0.99	8.00
8	3.15	-37.34	1.58	6.00	3.46	-39.30	1.15	8.00
9	3.18	-45.54	1.12	7.00	3.62	-41.82	1.15	8.00

Table 8.11: Summary table of lLR's for comparing sets of size $n_s = 2$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

When n_s increases from 1 to 2, FP rates generally decreased.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.27	-4.18	30.70	5.00
2	-	-	-	-	0.80	-9.89	11.93	5.00
3	-	-	-	-	1.35	-16.75	5.56	9.17
4	1.99	-17.34	3.22	13.33	1.89	-22.67	3.22	11.67
5	1.96	-29.68	1.58	10.83	1.97	-34.46	1.17	12.50
6	2.99	-40.52	1.58	10.00	2.67	-53.42	0.88	11.67
7	2.83	-48.96	0.88	9.17	2.62	-57.53	0.70	12.50
8	2.89	-58.69	0.76	9.17	3.10	-61.43	0.70	12.50
9	2.46	-72.03	0.64	10.83	3.26	-65.41	0.70	13.33

Table 8.12: Summary table of lLR's for comparing sets of size $n_s = 3$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

FN rates goes up for larger n_s and B.

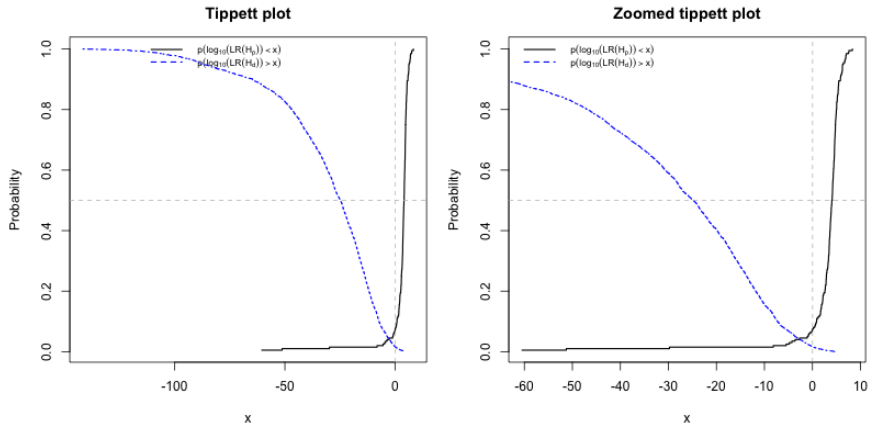


Figure 8.18: Tippett plot for wool data with setup $n_s = 2$, $B = 7$ under model DR-S.

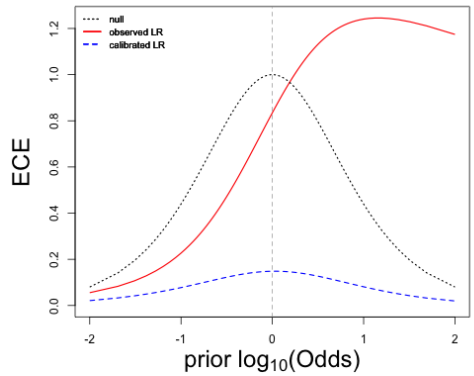


Figure 8.19: ECE plot for wool data with setup $n_s = 2$, $B = 7$ under model DR-S.

There is no visible improvement according to ECE plot as shown in Figure 8.19 compared to use of original data for model DR-S.

8.3.6 Conclusion

The improvements of performances of models when differences are used instead of original data is easy to see for all component-wise additive models for wool data.

8.4 Cotton data

Sample of ink data consists of $K = 20$ groups of $n = n_k = 9$ MSP measurements of transmittance $\boldsymbol{y}_{ki}$ versus wavelength for $1 \leq i \leq n$ for all k. Transmittance are measured at wavelengths ranging from 240-690 nm with intervals of 5 nm so using all the points, that is, taking interval or $int = 5$, the total number of points, the dimension of our data, is $m = 91$.

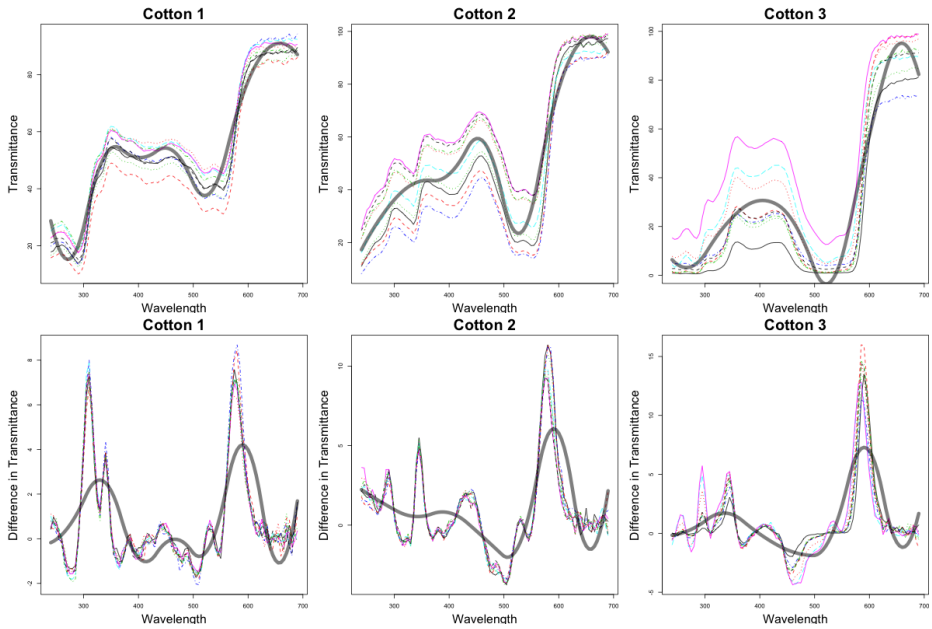


Figure 8.20: Fitting original and the first differences of three types of cotton using 6 B-spline basis functions of order 3.

Taking differences almost get rid of the vertical separations completely except at some small ranges of wavelengths. Following the simulations in Chapter 5 and results from Chapter 7, we expect our models to perform better on this new set of data.

8.4.1 Summary tables for preprocessed cotton data

For each model, results will be reported for cotton data for 3 distinct values of n_s. The three values are 1, 2 and 3. Since we have 9 measurements of one sample for each of the 20 different types of cotton fibres, there are $9 \times 10 \div 2 = 45$ within-group and $9 \times 9 = 81$ between-group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$

pairs of groups for $n_s = 1$. For $n_s = 2$, lLR's are obtained for comparing sets of $n_s = 2$ measurements with another (mutually exclusive) set of $n_s = 2$ measurements so there are $\lfloor \frac{9}{2} \rfloor \times (\lfloor \frac{9}{2} \rfloor + 1) \div 2 = 10$ within group and $\lfloor \frac{9}{2} \rfloor \times \lfloor \frac{9}{2} \rfloor = 16$ between group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups. For $n_s = 3$ there are $\lfloor \frac{9}{3} \rfloor \times (\lfloor \frac{9}{3} \rfloor + 1) \div 2 = 6$ within-group and $\lfloor \frac{9}{3} \rfloor \times \lfloor \frac{9}{3} \rfloor = 9$ between-group lLR's for comparisons between 20 and $20 \times 19 \div 2 = 190$ pairs of groups.

8.4.2 CA-S Simplified multivariate normal random-effects model - cotton data

Log likelihood ratios calculated using the simplified multivariate normal random-effects model for preprocessed cotton data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	0.56	-0.21	50.28	5.00	-4.03	-57.32	2.51	41.78
1	2	0.24	-0.02	59.95	2.78	2.16	-11.15	12.39	18.89
2	1	1.09	-0.76	39.61	2.00	-3.10	-110.30	1.41	36.50
2	2	0.53	-0.15	53.06	0.00	3.65	-24.86	6.18	16.00
3	1	1.49	-1.53	31.81	0.83	-2.51	-163.82	0.88	34.17
3	2	0.79	-0.37	47.13	0.00	4.58	-39.42	3.98	17.50

Table 8.13: Summary table of lLR's obtained using simplified multivariate normal random-effects model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

Like the effect of taking differences on wool data, high FN on original data became high FP for when B-spline basis functions are used. This can possibly suggest that separation causes high FN rates. However, a different pattern is seen when eigenfunctions from fPCA are used; both FP and FN decline.

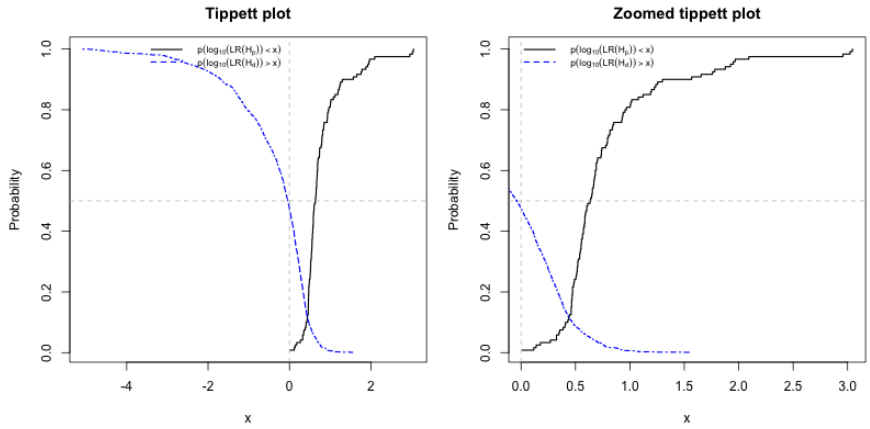


Figure 8.21: Tippett plot for cotton data with setup $n_s = 3$, $int = 2$ under model CA-S.

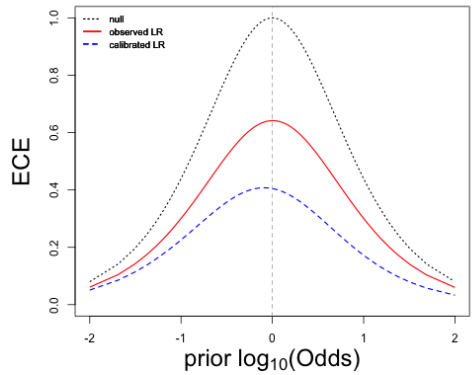


Figure 8.22: ECE plot for cotton data with setup $n_s = 3$, $int = 2$ under model CA-S.

Despite the high FP rate, this model is much better calibrated compared to using original data based on the ECE plots.

8.4.3 CA-const. Constant within-group variance model - cotton data

Log likelihood ratios calculated using the constant within-group variance model for preprocessed cotton data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	0.71	-1.46	36.30	15.33	-3.49	-35.68	1.41	42.33
1	2	0.52	-0.44	47.39	11.78	2.11	-9.69	10.23	19.44
2	1	1.32	-3.34	27.66	7.00	-2.34	-63.48	0.49	39.50
2	2	0.94	-1.17	39.64	4.00	3.63	-21.46	4.51	19.50
3	1	1.55	-5.60	20.99	8.33	-2.38	-91.81	0.23	35.83
3	2	1.12	-2.12	33.57	3.33	4.42	-34.15	2.57	19.17

Table 8.14: Summary table of lLR's obtained using constant within-group variance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

Based on Table 8.14 the pattern of results obtained from this model looks very similar to that obtained from CA-S. However, FP rates generally decrease but this is offset by increase in FN so overall, this model performs similar to CA-S as well. Using eigenfunctions from fPCA resulted in much lowered FN for use of differences compared to the use of original data.

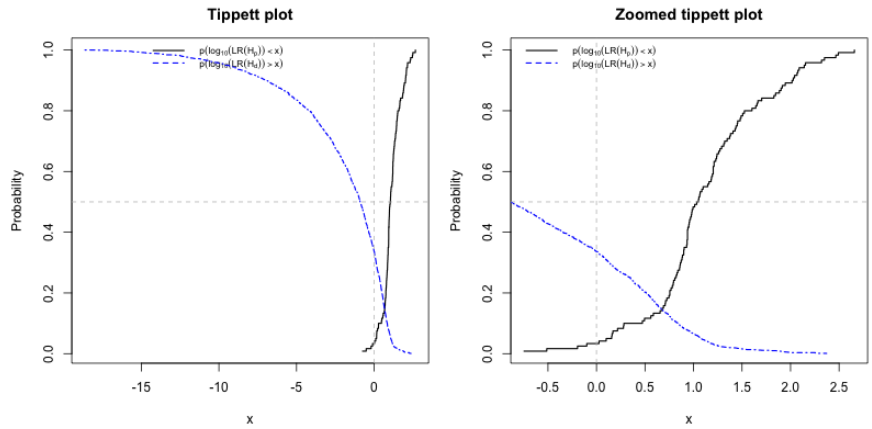


Figure 8.23: Tippett plot for cotton data with setup $n_s = 3$, $int = 2$ under model CA-const..

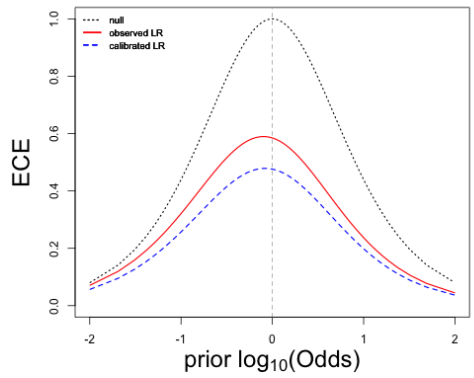


Figure 8.24: ECE plot for cotton data with setup $n_s = 3$, $int = 2$ under model CA-const..

Compared to original data, this model has much smaller loss of information when differences are used.

8.4.4 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model - cotton data

Log likelihood ratios calculated using the multivariate normal random-effects with autoregressive within-group covariance model for preprocessed cotton data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

		B-spline				fPCA			
n_s	int	S	D	FP	FN	S	D	FP	FN
1	1	0.36	-0.47	49.74	15.56	0.69	-17.02	5.27	28.00
1	2	0.32	-0.24	51.88	11.89	1.98	-7.71	13.00	16.33
2	1	0.60	-1.15	41.68	12.50	1.82	-33.13	2.89	22.00
2	2	0.55	-0.68	45.03	7.50	3.26	-17.16	6.38	14.00
3	1	0.17	-2.47	22.63	33.33	2.56	-48.83	2.22	24.17
3	2	0.44	-1.41	33.10	19.17	4.06	-27.18	4.27	15.83

Table 8.15: Summary table of lLR's obtained using multivariate normal random-effects with autoregressive within-group covariance model for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different n_s and intervals (int) where number of basis (B) used is 6 and order of basis used is 3 for B-spline basis functions.

According to Table 8.15 the performance improves when differences are used instead of original data (see Table 6.22) when eigenfunctions from fPCA are used. However, when B-spline basis functions are used, FP goes up compared to original data used. When B-spline basis functions are used, CA-ar performs worse than CA-const. in terms of sums of FP and FN when differences are used whereas CA-ar outperforms CA-const. when original data is used.

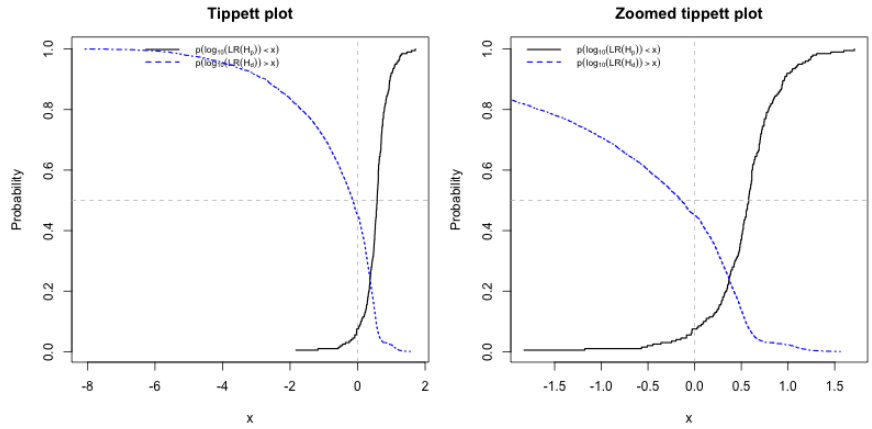


Figure 8.25: Tippett plot for cotton data with setup $n_s = 2$, $int = 2$ under model CA-ar.

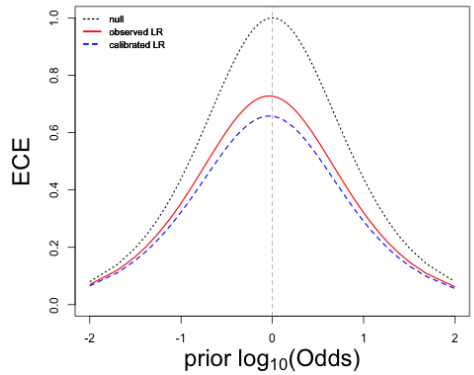


Figure 8.26: ECE plot for cotton data with setup $n_s = 2$, $int = 2$ under model CA-ar.

There is a slightly larger loss of information compared to when original data is used. However, it is still close to calibrated LR using the PAV algorithm so is still acceptable although the FP is quite high.

8.4.5 DR-S Dimension reduced multivariate random-effects model - cotton data

Log likelihood ratios calculated using the dimension reduced multivariate random-effects model for preprocessed cotton data are summarised in tables and plots for assessing the performance are drawn for one selection of setups.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.32	-0.54	46.54	9.44
2	-	-	-	-	0.65	-1.93	27.75	8.33
3	-	-	-	-	0.96	-2.23	27.84	9.11
4	0.71	-1.24	27.80	10.89	0.94	-3.01	25.54	8.22
5	0.83	-1.61	25.61	10.56	0.92	-4.83	19.45	8.67
6	1.12	-2.05	22.94	10.33	1.26	-9.23	13.85	9.22
7	1.29	-3.87	15.86	9.11	1.58	-11.19	12.46	8.44
8	1.68	-5.58	17.58	7.78	1.78	-12.08	12.05	8.11
9	1.61	-7.40	14.91	9.44	1.74	-12.87	12.09	9.78

Table 8.16: Summary table of lLR's for comparing sets of size $n_s = 1$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.48	-1.32	38.91	7.00
2	-	-	-	-	0.97	-4.37	21.91	7.00
3	-	-	-	-	1.41	-5.29	20.59	6.50
4	1.19	-3.09	18.78	7.50	1.35	-6.86	18.03	9.50
5	1.39	-3.98	16.22	6.50	1.29	-10.56	12.99	10.00
6	1.90	-5.09	12.47	7.50	1.75	-19.77	7.83	11.00
7	2.10	-8.87	8.36	9.00	2.23	-23.97	7.34	7.50
8	2.81	-12.49	10.00	9.50	2.54	-25.90	7.37	8.00
9	2.69	-16.30	7.80	10.00	2.65	-27.41	7.73	8.00

Table 8.17: Summary table of lLR's for comparing sets of size $n_s = 2$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

There are slight improvements of performance when differences are used instead of original data in terms of lowered FN and FP for smaller B. However, based on summary tables alone, using differences does not outperform use of original data. The only setups that give improvements are those with $n_s = 3$ and large B.

	B-spline				fPCA			
B	S	D	FP	FN	S	D	FP	FN
1	-	-	-	-	0.52	-2.13	34.04	6.67
2	-	-	-	-	1.15	-6.93	17.54	5.83
3	-	-	-	-	1.77	-8.41	16.14	5.00
4	1.47	-5.13	14.85	9.17	1.80	-10.80	13.68	6.67
5	1.78	-6.45	10.47	7.50	2.02	-16.42	9.06	9.17
6	2.47	-8.43	8.83	8.33	2.59	-30.20	5.67	8.33
7	2.69	-14.44	5.79	9.17	3.12	-36.68	5.50	8.33
8	3.72	-20.45	6.67	7.50	3.53	-39.66	5.20	7.50
9	3.80	-25.61	5.03	10.00	3.74	-41.88	5.50	7.50

Table 8.18: Summary table of lLR's for comparing sets of size $n_s = 3$ for different choices of basis (B-splines and eigenfunctions obtained using functional principal component analysis) for different numbers B of basis functions.

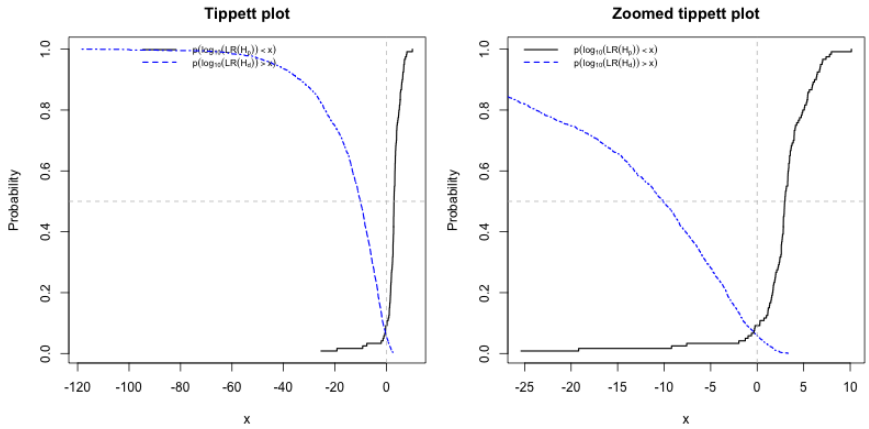


Figure 8.27: Tippett plot for cotton data with setup $n_s = 3$, $B = 7$ under model DR-S.

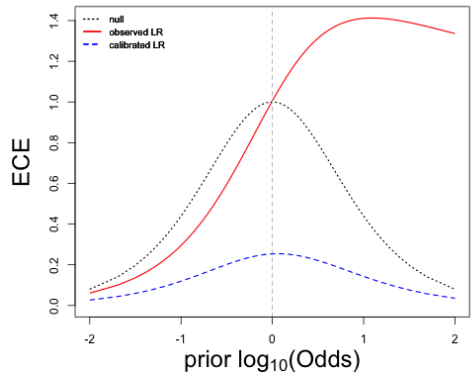


Figure 8.28: ECE plot for cotton data with setup $n_s = 3$, $B = 7$ under model DR-S.

Based on ECE plots, there is no improvement made by this model when differences are used instead of original data.

8.4.6 Conclusion

Some improvements can be seen for all four models, especially as indicated by ECE plots for CA-S and CA-const..

8.5 Conclusion

There are improvements for all models for all datasets. Even though some are small when looking at the summary tables alone, ECE plots show otherwise. Overall, wool data shows the most improvements among all datasets when differences are used in replacement of original data.

Chapter 9

Conclusion, recommendations and future direction

9.1 Summary

We have developed models to evaluate evidence in the form of functional data by the use of likelihood ratios with applications in microspectrophotometry data. Previous work on the evaluation of evidence in the form of functional data used either a score approach for likelihood ratio calculation or visual comparisons. We developed models for the calculation of likelihood ratios in a probabilistic approach to take into account different levels of variation together.

Overall, two types of models are developed, one is based on fundamental functional data analaysis that decomposes the curves as a sum of two components, a smooth underlying curve and some error, and one analyses the dimension reduced representation of the data. In either one, two types of basis functions are used: B-spline basis functions and eigenfunctions obtained from using functional principal component analysis. We can tell from the model fittings and simulations in Chapter 5 what variations exist in the data and what characteristics each model is able to capture. Based on these, we found results presented as summary tables, Tippett plots and ECE plots in Chapter 6 to be consistent with our findings in Chapter 5 about each model and further gained insights about the models through sensitivity analysis in Chapter 7. Finally, we improved the performances of the models through preprocessing of our data with results

presented in Chapter 8.

Although we only worked on microspectrophotometry data, the same techniques we employed in analysing functional data can be readily applied on any data of similar type. Overall, our approach provides an objective and innovative way of calculating likelihood ratios for the evaluation of evidence in the form of functional data.

9.2 Recommendations

When measurements are obtained, the first thing to do is to examine the variations among them, both within- and between-groups. Moreover, they should be compared visually to pick up the characteristics that can be used to distinguish whether two sets of evidence are from the same origin or not. Based on the properties of each dataset, an appropriate basis must be chosen first for the purpose of dimension reduction if the dimension of the original (input) data is large, meaning larger than samples available. After that, analyses can be done to distinguish the main features that can help to differentiate among different evidence. Variable selection might be necessary for further dimension reduction. Based on the complexity of the data, whether it has greater between- or within-group variations, appropriate models can be chosen to accommodate that by the use of a hierarchical model with different covariance structures. The number of levels required depends on the number of levels present in the data that are essential for distinguishing between groups. The exact structures of the parameters such as variance-covariance matrices, can be modified based on that of the data,i.e., independent or conditional independent.

9.3 Future research directions

The results of our proposed models for evaluating evidence using likelihood ratios by comparing evidence in the form of functional data is excellent when the right model and setups are chosen. However, the following can still be addressed for generalisation purposes and possibly better results.

- Assumption of normality for between-group distribution

Only normality is assumed for different levels of our data, which might not always be the case as we can see in Chapter 5. Kernel densities can be used instead if one wishes to consider the complexity.

- Consider different basis and datasets

  So far, B-spline basis functions and eigenfunctions are used for all datasets because of them being microspectrophotometry data; however, the same methodology can be easily generalised to other types of (functional) data, thus different choices of basis functions might be needed. For example, it might be better to use wavelet transform for fourier transformed infrared spectra (FTIR) based on the properties (shapes) of the data. After the transformation, it might be worthwhile to select the variables to pick up the most prominent ones.

- Model and data complexity

  We would be interested to analyse data that are more complicated: either contaminated or collected at different times as it might add another level of variability. Moreover, models we developed so far are yet able to capture the separations among curves from within the same groups although this problem is solved by pre-processing the data as we can see from the results presented in Chapter 8. However, it is always preferred if a model is able to capture all characteristics that are useful in distinguishing between evidence.

Appendix A

Distributions

Distributions used throughout the thesis have probability density function specified here.

A univariate random variable $X \in \mathbb{R}$ following a normal distribution, denoted as $X \sim N(\mu, \sigma^2)$ has probability density function

$$f(x; \mu, \sigma) = (2\pi\sigma^2)^{-1/2}\text{exp}\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}.$$

A multivariate random variable $\boldsymbol{X} \in \mathbb{R}^m$ following a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, denoted as $\boldsymbol{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has probability density function

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-m/2}|\boldsymbol{\Sigma}|^{-1/2}\text{exp}\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \quad \boldsymbol{x} \in \mathbb{R}^m.$$

A univariate random variable $X \in (0, \infty)$ following a gamma distribution with shape parameter $\gamma > 0$ and scale parameter $\delta > 0$, denoted as $X \sim \Gamma(\gamma, \delta)$ has probability density function

$$f(x; \gamma, \delta) = \frac{\delta^\gamma}{\Gamma(\gamma)}x^{\gamma-1}\text{exp}\left(-\delta x\right), \quad x \in \mathbb{R}^+.$$

If $X \sim \Gamma(\gamma, \delta)$, its inverse X^{-1} follows an inverse gamma distribution. A univariate random variable $X \in (0, \infty)$ following an inverse gamma distribution with shape parameter $\gamma > 0$ and scale parameter $\delta > 0$, denoted as $X \sim \mathbf{Inv-Gamma}(\gamma, \delta)$

has probability density function

$$f(x;\gamma,\delta) = \frac{\delta^{\gamma}}{\Gamma(\gamma)} x^{-\gamma-1} \text{exp}\left(-\frac{\delta}{x}\right), \quad x \in \mathbb{R}^{+}.$$

An positive semi-definite random variable $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ following an inverse Wishart distribution with scale matrix $\mathbf{\Omega}$ and degrees of freedom parameter $\nu > p-1$, denoted as $\mathbf{\Sigma} \sim \mathcal{W}^{-1}(\mathbf{\Omega}, \nu)$ has probability density function

$$f(\Sigma;\mathbf{\Omega},\nu) = \frac{|\mathbf{\Omega}|^{\frac{\nu}{2}}}{2^{\nu p/2}\Gamma_p(\nu/2)} |\Sigma|^{-\frac{\nu+p+1}{2}} \text{exp}\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Omega}\mathbf{X}^{-1}\right)\right\}$$

where tr$(\cdot)$ is the trace function.

Appendix B

Systems of B-spline basis function used throughout

Two examples of obtaining basis functions are to be described. They are the systems of B-spline basis functions used in our analysis. The first example is the set of basis functions where there are $N = 3$ equidistant interior knots with boundary knots at (0,4), and the order of the splines is $o = 3$. The augmented knot sequence used to construct the B-spline is $\boldsymbol{\tau}^* = (\tau_0, \ldots, \tau_{N+2o-1}) = (0, 0, 0, 1, 2, 3, 4, 4, 4)$. The domains used here are for illustration only. The number of order $o = 3$ basis functions is $B = 6 = N + o$. The exact formula for the bases are derived below. These are obtained using Equations (2.1) and (2.2) recursively as laid out in Figure 2.3. Every basis of order o is a linear combination of bases of degree $o - 1$. Only non-zero B's are shown.

$$B_{2,1}(x) = \begin{cases} 1 & \text{if } \tau_2 \leq x < \tau_3 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Likewise,

$$B_{3,1}(x) = \begin{cases} 1 & \text{if } 1 \leq x < 2 \\ 0 & \text{otherwise,} \end{cases} \quad B_{4,1}(x) = \begin{cases} 1 & \text{if } 2 \leq x < 3 \\ 0 & \text{otherwise,} \end{cases}$$

$$B_{5,1}(x) = \begin{cases} 1 & \text{if } 3 \leq x < 4 \\ 0 & \text{otherwise.} \end{cases}$$

For order 2, we have

$$
\begin{aligned}
B_{1,2}(x) &= \alpha_{1,2}B_{1,1}(x) + \left(1 - \frac{x-\tau_2}{\tau_3-\tau_2}\right) B_{2,1}(x) \\
&= (1-x)B_{2,1}(x) = \begin{cases} 1-x & \text{if } 0 \le x < 1 \\ 0 & \text{otherwise,} \end{cases} \\
B_{2,2}(x) &= \frac{x-\tau_2}{\tau_3-\tau_2}B_{2,1}(x) + \left(1 - \frac{x-\tau_3}{\tau_4-\tau_3}\right) B_{3,1}(x) \\
&= xB_{2,1}(x) + (1-(x-1))B_{3,1}(x) = \begin{cases} x & \text{if } 0 \le x < 1 \\ 2-x & \text{if } 1 \le x < 2 \\ 0 & \text{otherwise,} \end{cases} \\
B_{3,2}(x) &= \frac{x-\tau_3}{\tau_4-\tau_3}B_{3,1}(x) + \left(1 - \frac{x-\tau_4}{\tau_5-\tau_4}\right) B_{4,1}(x) \\
&= (x-1)B_{3,1}(x) + (1-(x-2))B_{4,1}(x) = \begin{cases} x-1 & \text{if } 1 \le x < 2 \\ 3-x & \text{if } 2 \le x < 3 \\ 0 & \text{otherwise,} \end{cases} \\
B_{4,2}(x) &= \frac{x-\tau_4}{\tau_5-\tau_4}B_{4,1}(x) + \left(1 - \frac{x-\tau_5}{\tau_6-\tau_5}\right) B_{5,1}(x) \\
&= (x-2)B_{4,1}(x) + (1-(x-3))B_{5,1}(x) = \begin{cases} x-2 & \text{if } 2 \le x < 3 \\ 4-x & \text{if } 3 \le x < 4 \\ 0 & \text{otherwise,} \end{cases} \\
B_{5,2}(x) &= \frac{x-\tau_5}{\tau_6-\tau_5}B_{5,1}(x) + \left(1 - \frac{x-\tau_6}{\tau_7-\tau_6}\right) B_{6,1}(x) \\
&= (x-3)B_{5,1}(x) + (1-(x-4))B_{6,1}(x) = \begin{cases} x-3 & \text{if } 3 \le x < 4 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

For order 3, we have

$$
\begin{aligned}
B_{0,3}(x) &= \alpha_{0,3}B_{0,2}(x) + \left(1 - \frac{x-\tau_1}{\tau_3-\tau_1}\right) B_{1,2}(x) = (1-x)B_{1,2}(x) \\
&= \begin{cases} (1-x)^2 & \text{if } 0 \le x < 1 \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}
$$

$$B_{1,3}(x) = \frac{x-\tau_1}{\tau_3-\tau_1}B_{1,2}(x) + \left(1-\frac{x-\tau_2}{\tau_4-\tau_2}\right)B_{2,2}(x) = xB_{1,2}(x) + (1-(x/2))B_{2,2}(x)$$
$$= \begin{cases} x(1-x)+\frac{(2-x)x}{2} & \text{if } 0 \le x < 1 \\ \frac{(2-x)^2}{2} & \text{if } 1 \le x < 2 \\ 0 & \text{otherwise,} \end{cases}$$
$$B_{2,3}(x) = \frac{x-\tau_2}{\tau_4-\tau_2}B_{2,2}(x) + \left(1-\frac{x-\tau_3}{\tau_5-\tau_3}\right)B_{3,2}(x) = \frac{x}{2}B_{2,2}(x) + (1-\frac{x-1}{2})B_{3,2}(x)$$
$$= \frac{x}{2}[xB_{2,1}(x) + (2-x)B_{3,1}(x)] + \left(1-\frac{x-1}{2}\right)[(x-1)B_{3,1}(x) + (3-x)B_{4,1}(x)]$$
$$= \begin{cases} \frac{x^2}{2} & \text{if } 0 \le x < 1 \\ \frac{x(2-x)+(3-x)(x-1)}{2} & \text{if } 1 \le x < 2 \\ \frac{(3-x)^2}{2} & \text{if } 2 \le x < 3 \\ 0 & \text{otherwise,} \end{cases}$$
$$B_{3,3}(x) = \frac{x-\tau_3}{\tau_5-\tau_3}B_{3,2}(x) + \left(1-\frac{x-\tau_4}{\tau_6-\tau_4}\right)B_{4,2}(x) = \frac{x-1}{2}B_{3,2}(x) + \left(1-\frac{x-2}{2}\right)B_{4,2}(x)$$
$$= \frac{x-1}{2}[(x-1)B_{3,1}(x) + (3-x)B_{4,1}(x)]$$
$$+ \left(1-\frac{x-2}{2}\right)[(x-2)B_{4,1}(x) + (4-x)B_{5,1}(x)]$$
$$= \begin{cases} \frac{(x-1)^2}{2} & \text{if } 1 \le x < 2 \\ \frac{(x-1)(3-x)+(4-x)(x-2)}{2} & \text{if } 2 \le x < 3 \\ \frac{(4-x)^2}{2} & \text{if } 3 \le x < 4 \\ 0 & \text{otherwise,} \end{cases}$$
$$B_{4,3}(x) = \frac{x-\tau_4}{\tau_6-\tau_4}B_{4,2}(x) + \left(1-\frac{x-\tau_5}{\tau_7-\tau_5}\right)B_{5,2}(x) = \frac{x-2}{2}B_{4,2}(x) + \left(1-\frac{x-3}{2}\right)B_{5,2}(x)$$
$$= \frac{x-2}{2}[(x-2)B_{4,1}(x) + (4-x)B_{5,1}(x)] + (4-x)(x-3)B_{5,1}(x)$$
$$= \begin{cases} \frac{(x-2)^2}{2} & \text{if } 2 \le x < 3 \\ \frac{(x-2)(4-x)}{2} + (4-x)(x-3) & \text{if } 3 \le x < 4 \\ 0 & \text{otherwise,} \end{cases}$$
$$B_{5,3}(x) = \frac{x-\tau_5}{\tau_7-\tau_5}B_{5,2}(x) + (1-\alpha_{6,3})B_{6,2}(x) = (x-3)B_{5,2}(x)$$
$$= \begin{cases} (x-3)^2 & \text{if } 3 \le x < 4 \\ 0 & \text{otherwise.} \end{cases}$$

The second example is the set of basis functions used for fibre data where there are $N = 6$ equidistant interior knots with boundary knots at (0,7), and the order of the splines is $o = 3$. The augmented knot sequence used to construct the B-spline is $(0, 0, 0, 1, 2, 3, 4, 5, 6, 7, 7, 7)$. Again, the boundary knots at (0,7) is for demonstration only. The number of basis functions (at order $o = 3$) is $B = 9 = N + o$. The exact formula for the bases are derived below. These are obtained using Equations 2.1, 2.2 and 2.3 recursively like that in Figure 2.3. Every basis of order o is a linear combination of bases of degree $o - 1$. Only non-zero B's are laid out.

$$B_{2,1}(x) = \begin{cases} 1 & \text{if } \tau_2 \leq x < \tau_3 \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Likewise,

$$B_{3,1}(x) = \begin{cases} 1 & \text{if } 1 \leq x < 2 \\ 0 & \text{otherwise.} \end{cases} \quad B_{4,1}(x) = \begin{cases} 1 & \text{if } 2 \leq x < 3 \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{5,1}(x) = \begin{cases} 1 & \text{if } 3 \leq x < 4 \\ 0 & \text{otherwise.} \end{cases} \quad B_{6,1}(x) = \begin{cases} 1 & \text{if } 4 \leq x < 5 \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{7,1}(x) = \begin{cases} 1 & \text{if } 5 \leq x < 6 \\ 0 & \text{otherwise.} \end{cases} \quad B_{8,1}(x) = \begin{cases} 1 & \text{if } 6 \leq x < 7 \\ 0 & \text{otherwise.} \end{cases}$$

For order 2, we have

$$\begin{aligned} B_{1,2}(x) &= \alpha_{1,2} B_{1,1}(x) + \left(1 - \frac{x - \tau_2}{\tau_3 - \tau_2}\right) B_{2,1}(x) \\ &= (1 - x) B_{2,1}(x) = \begin{cases} 1 - x & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \\ B_{2,2}(x) &= \frac{x - \tau_2}{\tau_3 - \tau_2} B_{2,1}(x) + \left(1 - \frac{x - \tau_3}{\tau_4 - \tau_3}\right) B_{3,1}(x) \\ &= x B_{2,1}(x) + (1 - (x - 1)) B_{3,1}(x) = \begin{cases} x & \text{if } 0 \leq x < 1 \\ 2 - x & \text{if } 1 \leq x < 2 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

$$B_{3,2}(x) = \frac{x-\tau_3}{\tau_4-\tau_3}B_{3,1}(x) + \left(1 - \frac{x-\tau_4}{\tau_5-\tau_4}\right) B_{4,1}(x)$$

$$= (x-1)B_{3,1}(x) + (1-(x-2))B_{4,1}(x) = \begin{cases} x-1 & \text{if } 1 \le x < 2 \\ 3-x & \text{if } 2 \le x < 3 \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{4,2}(x) = \frac{x-\tau_4}{\tau_5-\tau_4}B_{4,1}(x) + \left(1 - \frac{x-\tau_5}{\tau_6-\tau_5}\right) B_{5,1}(x)$$

$$= (x-2)B_{4,1}(x) + (1-(x-3))B_{5,1}(x) = \begin{cases} x-2 & \text{if } 2 \le x < 3 \\ 4-x & \text{if } 3 \le x < 4 \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{5,2}(x) = \frac{x-\tau_5}{\tau_6-\tau_5}B_{5,1}(x) + \left(1 - \frac{x-\tau_6}{\tau_7-\tau_6}\right) B_{6,1}(x)$$

$$= (x-3)B_{5,1}(x) + (1-(x-4))B_{6,1}(x) = \begin{cases} x-3 & \text{if } 3 \le x < 4 \\ 5-x & \text{if } 4 \le x < 5 \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{6,2}(x) = \frac{x-\tau_6}{\tau_7-\tau_6}B_{6,1}(x) + \left(1 - \frac{x-\tau_7}{\tau_8-\tau_7}\right) B_{7,1}(x)$$

$$= (x-4)B_{6,1}(x) + (1-(x-5))B_{7,1}(x) = \begin{cases} x-4 & \text{if } 4 \le x < 5 \\ 6-x & \text{if } 5 \le x < 6 \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{7,2}(x) = \frac{x-\tau_7}{\tau_8-\tau_7}B_{7,1}(x) + \left(1 - \frac{x-\tau_8}{\tau_9-\tau_8}\right) B_{8,1}(x)$$

$$= (x-5)B_{7,1}(x) + (1-(x-6))B_{8,1}(x) = \begin{cases} x-5 & \text{if } 5 \le x < 6 \\ 7-x & \text{if } 6 \le x < 7 \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{8,2}(x) = \frac{x-\tau_8}{\tau_9-\tau_8}B_{8,1}(x) + \left(1 - \frac{x-\tau_9}{\tau_{10}-\tau_9}\right) B_{9,1}(x)$$

$$= (x-6)B_{8,1}(x) + (1-(x-7))B_{9,1}(x) = \begin{cases} x-6 & \text{if } 6 \le x < 7 \\ 0 & \text{otherwise.} \end{cases}$$

For order 3, we have

$$
\begin{aligned}
B_{0,3}(x) &= \alpha_{0,3}B_{0,2}(x) + \left(1 - \frac{x-\tau_1}{\tau_3-\tau_1}\right) B_{1,2}(x) = (1-x)B_{1,2}(x) \\
&= \begin{cases} (1-x)^2 & \text{if } 0 \le x < 1 \\ 0 & \text{otherwise,} \end{cases} \\
B_{1,3}(x) &= \frac{x-\tau_1}{\tau_3-\tau_1}B_{1,2}(x) + \left(1 - \frac{x-\tau_2}{\tau_4-\tau_2}\right) B_{2,2}(x) = xB_{1,2}(x) + (1-(x/2))B_{2,2}(x) \\
&= \begin{cases} x(1-x) + \frac{(2-x)x}{2} & \text{if } 0 \le x < 1 \\ \frac{(2-x)^2}{2} & \text{if } 1 \le x < 2 \\ 0 & \text{otherwise,} \end{cases} \\
B_{2,3}(x) &= \frac{x-\tau_2}{\tau_4-\tau_2}B_{2,2}(x) + \left(1 - \frac{x-\tau_3}{\tau_5-\tau_3}\right) B_{3,2}(x) = \frac{x}{2}B_{2,2}(x) + \left(1 - \frac{x-1}{2}\right) B_{3,2}(x) \\
&= \frac{x}{2}[xB_{2,1}(x) + (2-x)B_{3,1}(x)] + \left(1 - \frac{x-1}{2}\right)[(x-1)B_{3,1}(x) + (3-x)B_{4,1}(x)] \\
&= \begin{cases} \frac{x^2}{2} & \text{if } 0 \le x < 1 \\ \frac{x(2-x)+(3-x)(x-1)}{2} & \text{if } 1 \le x < 2 \\ \frac{(3-x)^2}{2} & \text{if } 2 \le x < 3 \\ 0 & \text{otherwise,} \end{cases} \\
B_{3,3}(x) &= \frac{x-\tau_3}{\tau_5-\tau_3}B_{3,2}(x) + \left(1 - \frac{x-\tau_4}{\tau_6-\tau_4}\right) B_{4,2}(x) = \frac{x-1}{2}B_{3,2}(x) + \left(1 - \frac{x-2}{2}\right) B_{4,2}(x) \\
&= \frac{x-1}{2}[(x-1)B_{3,1}(x) + (3-x)B_{4,1}(x)] \\
&\quad + \left(1 - \frac{x-2}{2}\right)[(x-2)B_{4,1}(x) + (4-x)B_{5,1}(x)] \\
&= \begin{cases} \frac{(x-1)^2}{2} & \text{if } 1 \le x < 2 \\ \frac{(x-1)(3-x)+(4-x)(x-2)}{2} & \text{if } 2 \le x < 3 \\ \frac{(4-x)^2}{2} & \text{if } 3 \le x < 4 \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
B_{4,3}(x) &= \frac{x-\tau_4}{\tau_6-\tau_4}B_{4,2}(x) + \left(1-\frac{x-\tau_5}{\tau_7-\tau_5}\right)B_{5,2}(x) = \frac{x-2}{2}B_{4,2}(x) + \left(1-\frac{x-3}{2}\right)B_{5,2}(x) \\
&= \frac{x-2}{2}[(x-2)B_{4,1}(x) + (4-x)B_{5,1}(x)] + \frac{5-x}{2}[(x-3)B_{5,1}(x) + (5-x)B_{6,1}(x)] \\
&= \begin{cases} \frac{(x-2)^2}{2} & \text{if } 2 \le x < 3 \\ \frac{(x-2)(4-x)}{2} + \frac{(5-x)(x-3)}{2} & \text{if } 3 \le x < 4 \\ \frac{(5-x)^2}{2} & \text{if } 4 \le x < 5 \\ 0 & \text{otherwise,} \end{cases} \\
B_{5,3}(x) &= \frac{x-\tau_5}{\tau_7-\tau_5}B_{5,2}(x) + \left(1-\frac{x-\tau_6}{\tau_8-\tau_6}\right)B_{6,2}(x) = \frac{x-3}{2}B_{5,2}(x) + (1-\frac{x-4}{2})B_{6,2}(x) \\
&= \frac{x-3}{2}[(x-3)B_{5,1}(x) + (5-x)B_{6,1}(x)] \\
&\quad + \left(1-\frac{x-4}{2}\right)[(x-4)B_{6,1}(x) + (6-x)B_{7,1}(x)] \\
&= \begin{cases} \frac{(x-3)^2}{2} & \text{if } \ 3 \le x < 4 \\ \frac{(x-3)(5-x)}{2} + \frac{(6-x)(x-4)}{2} & \text{if } \ 4 \le x < 5 \\ \frac{(6-x)^2}{2} & \text{if } \ 5 \le x < 6 \\ 0 & \text{otherwise,} \end{cases} \\
B_{6,3}(x) &= \frac{x-\tau_6}{\tau_8-\tau_6}B_{6,2}(x) + \left(1-\frac{x-\tau_7}{\tau_9-\tau_7}\right)B_{7,2}(x) = \frac{x-4}{2}B_{6,2}(x) + \left(1-\frac{x-5}{2}\right)B_{7,2}(x) \\
&= \frac{x-4}{2}[(x-4)B_{6,1}(x) + (6-x)B_{7,1}(x)] \\
&\quad + \left(1-\frac{x-5}{2}\right)[(x-5)B_{7,1}(x) + (7-x)B_{8,1}(x)] \\
&= \begin{cases} \frac{(x-4)^2}{2} & \text{if } 4 \le x < 5 \\ \frac{(x-4)(6-x)}{2} + \frac{(7-x)(x-5)}{2} & \text{if } 5 \le x < 6 \\ \frac{(7-x)^2}{2} & \text{if } 6 \le x < 7 \\ 0 & \text{otherwise,} \end{cases} \\
B_{7,3}(x) &= \frac{x-\tau_7}{\tau_9-\tau_7}B_{7,2}(x) + \left(1-\frac{x-\tau_8}{\tau_{10}-\tau_8}\right)B_{8,2}(x) = \frac{x-5}{2}B_{7,2}(x) + (1-(x-6))B_{8,2}(x) \\
&= \frac{x-5}{2}[(x-5)B_{7,1}(x) + (7-x)B_{8,1}(x)] + (7-x)(x-6)B_{8,1}(x) \\
&= \begin{cases} \frac{(x-5)^2}{2} & \text{if } 5 \le x < 6 \\ \frac{(x-5)(7-x)}{2} + (7-x)(x-6) & \text{if } 6 \le x < 7 \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}
$$

$$B_{8,3}(x) = \frac{x-\tau_8}{\tau_{10}-\tau_8}B_{8,2}(x) + (1-\alpha_{9,3})B_{9,2}(x) = (x-6)B_{8,2}(x)$$
$$= \begin{cases} (x-6)^2 & \text{if } 6 \le x < 7 \\ 0 & \text{otherwise.} \end{cases}$$

Appendix C

Derivation of likelihood ratios

C.1 CA-S Simplified multivariate normal random-effects model

$$LR = \frac{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_p)}{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_d)} = \frac{\int_{\boldsymbol{\theta}} \prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) \prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{D}) d\boldsymbol{\theta}}{\prod_{q\in\{c,r\}} \int_{\boldsymbol{\theta}} \prod_{i=1}^{n_q} f(\boldsymbol{y}_{qi}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{D}) d\boldsymbol{\theta}}$$

C.1.1 Likelihood ratio evaluation under prosecution proposition

The numerator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and the control curve come from the same origin, or $\boldsymbol{\theta}_r = \boldsymbol{\theta}_c$.

$$\begin{aligned}
&\int \prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) \prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi\theta}, \sigma^2\boldsymbol{I}_m) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{D}) d\boldsymbol{\theta} \\
&= \int (2\pi)^{-(n_c+n_r)m/2} \sigma^{-(n_c+n_r)m} |2\pi\boldsymbol{D}|^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_c} (\boldsymbol{y}_{ci} - \boldsymbol{\Phi\theta})^T (\boldsymbol{y}_{ci} - \boldsymbol{\Phi\theta}) \right\} \\
&\quad \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_r} (\boldsymbol{y}_{ri} - \boldsymbol{\Phi\theta})^T (\boldsymbol{y}_{ri} - \boldsymbol{\Phi\theta}) \right\} \exp\left\{ -\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\eta})^T \boldsymbol{D}^{-1} (\boldsymbol{\theta} - \boldsymbol{\eta}) \right\} d\boldsymbol{\theta} \\
&= \int (2\pi)^{-(n_c+n_r)m/2} \sigma^{-(n_c+n_r)m} |2\pi\boldsymbol{D}|^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_c} \boldsymbol{y}_{ci}^T \boldsymbol{y}_{ci} + \sum_{i=1}^{n_r} \boldsymbol{y}_{ri}^T \boldsymbol{y}_{ri} \right) - \frac{1}{2} \boldsymbol{\eta}^T \boldsymbol{D}^{-1} \boldsymbol{\eta} \right\} \\
&\quad \exp\left\{ -\frac{1}{2} \left[\frac{n_c + n_r}{\sigma^2} \boldsymbol{\theta}^T \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \boldsymbol{\theta}^T \boldsymbol{D}^{-1} - \left(\frac{2}{\sigma^2} \left(\sum_{i=1}^{n_c} \boldsymbol{y}_{ci}^T + \sum_{i=1}^{n_r} \boldsymbol{y}_{ri}^T \right) \boldsymbol{\Phi} + 2\boldsymbol{\eta}^T \boldsymbol{D}^{-1} \right) \right] \boldsymbol{\theta} \right\} d\boldsymbol{\theta}
\end{aligned}$$

$$=(2\pi)^{-(n_c+n_r)m/2}\sigma^{-(n_c+n_r)m}|\boldsymbol{D}|^{-1/2}|\boldsymbol{\Sigma}_n^*|^{1/2}$$
$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{y}_{ci}-\frac{1}{2\sigma^2}\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{y}_{ri}-\frac{1}{2}\boldsymbol{\eta}^T\boldsymbol{D}^{-1}\boldsymbol{\eta}+\frac{1}{2}\boldsymbol{\mu}_n^{*T}\boldsymbol{\Sigma}_n^{*-1}\boldsymbol{\mu}_n^*\right\}$$

where

$$\boldsymbol{\Sigma}_n^{*-1}=\frac{n_c+n_r}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{D}^{-1}$$
$$\boldsymbol{\mu}_n^*=\left(\frac{n_c+n_r}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{D}^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{\Phi}^T\left(\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}+\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}\right)+\boldsymbol{D}^{-1}\boldsymbol{\eta}\right)$$

C.1.2 Likelihood ratio evaluation under alternative proposition

The denominator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and the control curve come from different origins, independently.

$$\prod_{q\in\{c,r\}}\int\prod_{i=1}^{n_q}f(\boldsymbol{y}_{qi}|\boldsymbol{\Phi}\boldsymbol{\theta}_q,\sigma^2\boldsymbol{I})f(\boldsymbol{\theta}_q|\boldsymbol{\eta},\boldsymbol{D})d\boldsymbol{\theta}_q$$

The two terms are similar so a general case is derived.

$$\int\prod_{i=1}^{n_q}f(\boldsymbol{y}_{qi}|\boldsymbol{\Phi}\boldsymbol{\theta}_q,\sigma^2 I)f(\boldsymbol{\theta}_q|\boldsymbol{\eta},\boldsymbol{D})d\boldsymbol{\theta}_q$$
$$=(2\pi)^{-n_qm/2}\sigma^{-n_qm}|\boldsymbol{D}|^{-1/2}|\boldsymbol{\Sigma}_q^*|^{1/2}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_q}\boldsymbol{y}_{qi}^T\boldsymbol{y}_{qi}-\frac{1}{2}\boldsymbol{\eta}^T\boldsymbol{D}^{-1}\boldsymbol{\eta}+\frac{1}{2}\boldsymbol{\mu}_q^{*T}\boldsymbol{\Sigma}_q^{*-1}\boldsymbol{\mu}_q^*\}$$

where

$$\boldsymbol{\Sigma}_q^{*-1}=\frac{n_q}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{D}^{-1}$$
$$\boldsymbol{\mu}_q^*=\left(\frac{n_q}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{D}^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n_q}\boldsymbol{\Phi}^T\boldsymbol{y}_{qi}+\boldsymbol{D}^{-1}\boldsymbol{\eta}\right).$$

C.1.3 Likelihood ratio

Putting the numerator and denominator together we get

$$LR = \frac{|\boldsymbol{\Sigma}_n^*|^{1/2}\text{exp}\{\frac{1}{2}\boldsymbol{\mu}_n^{*T}\boldsymbol{\Sigma}_n^{*-1}\boldsymbol{\mu}_n^*\}}{|\boldsymbol{\Sigma}_c^*|^{1/2}|\boldsymbol{\Sigma}_r^*|^{1/2}|\boldsymbol{D}|^{-1/2}\text{exp}\left\{-\frac{1}{2}\boldsymbol{\eta}^T\boldsymbol{D}^{-1}\boldsymbol{\eta}+\frac{1}{2}\boldsymbol{\mu}_c^{*T}\boldsymbol{\Sigma}_c^{*-1}\boldsymbol{\mu}_c^*+\frac{1}{2}\boldsymbol{\mu}_r^{*T}\boldsymbol{\Sigma}_r^{*-1}\boldsymbol{\mu}_r^*\right\}}.$$

C.1.4 Estimate of hyperparameters using relevant population

$\boldsymbol{y}_{ki} \sim N_m(\boldsymbol{\Phi}\boldsymbol{\theta}_k, \sigma^2\boldsymbol{I}_m), \quad k = 1, \ldots, K, \quad i = 1, \ldots, n_k$
$\boldsymbol{\theta}_q \sim N_B(\boldsymbol{\eta}, \boldsymbol{D})$, $\boldsymbol{D}$ diagonal ($\boldsymbol{D}_{ii} = \omega_i^2$)

$$\hat{\sigma}^2 = \frac{1}{K(mn-B)}\sum_{k=1}^{K}\sum_{i=1}^{n_q}\left\|\boldsymbol{y}_{ki} - \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}_k\right\|^2,$$

$$\hat{\boldsymbol{\eta}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_k,$$

$$\hat{D}_{bb} = \hat{\omega}_b = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\theta}_b^{(k)} - \hat{\eta}_b\right)^2 - \frac{\hat{\sigma^2}}{n_k}\boldsymbol{\Phi}^T\boldsymbol{\Phi}.$$

C.1.5 Simulation

Under this model, datasets will be generated by first simulate group means $\boldsymbol{\theta}_k \sim N(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{D}})$ then $\boldsymbol{y}_{ki} \sim N(\boldsymbol{\Phi}\boldsymbol{\theta}_k, \hat{\sigma}^2\boldsymbol{I}_m)$ for $1 \leq k \leq K$ and $1 \leq i \leq n_k$. Descriptions of this model can be found in Section 3.2.2.

C.2 CA-const. Constant within-group variance model

$$\begin{aligned} LR &= \frac{f(\boldsymbol{Y}_c, \mathbf{Y}_r|H_p)}{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_d)} \\ &= \frac{\int_{\sigma^2}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi}\boldsymbol{\theta}, \sigma^2\boldsymbol{I}_m)\prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi}\boldsymbol{\theta}, \sigma^2\boldsymbol{I}_m)f(\boldsymbol{\theta}|\boldsymbol{\eta}, \sigma^2\boldsymbol{C})f(\sigma^2|\gamma, \delta)d\boldsymbol{\theta}d\sigma^2}{\prod_{q\in\{c,r\}}\int_{\sigma^2}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_q} f(\boldsymbol{y}_{qi}|\boldsymbol{\Phi}\boldsymbol{\theta}, \sigma^2\boldsymbol{I}_m)f(\boldsymbol{\theta}|\boldsymbol{\eta}, \sigma^2\boldsymbol{C})f(\sigma^2|\gamma, \delta)d\boldsymbol{\theta}d\sigma^2} \end{aligned}$$

C.2.1 Likelihood ratio evaluation under prosecution proposition

The numerator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and from the control curve come from the same origin, or $\boldsymbol{\theta}_r$ =

$\boldsymbol{\theta}_c$.

$$
\begin{aligned}
&\int\int \prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi\theta}, \boldsymbol{I}/\lambda) \prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi\theta}, \boldsymbol{I}/\lambda) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}/\lambda) d\boldsymbol{\theta} f(\lambda|\gamma, \delta) d\lambda \\
=&\int f(\lambda|\gamma, \delta) \int \prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi\theta}, \boldsymbol{I}/\lambda) \prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi\theta}, \boldsymbol{I}/\lambda) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}/\lambda) d\boldsymbol{\theta} d\lambda \\
=&\int f(\lambda|\gamma, \delta) \int \left(\frac{\lambda}{2\pi}\right)^{n_c m/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n_c}(\boldsymbol{y}_{ci} - \boldsymbol{\Phi\theta})^T(\boldsymbol{y}_{ci} - \boldsymbol{\Phi\theta})\right\} \\
&\left(\frac{\lambda}{2\pi}\right)^{n_r m/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n_r}(\boldsymbol{y}_{ri} - \boldsymbol{\Phi\theta})^T(\boldsymbol{y}_{ri} - \boldsymbol{\Phi\theta})\right\} \\
&\left(\frac{\lambda}{2\pi}\right)^{B/2} |\boldsymbol{C}|^{-1/2} \exp\left\{-\frac{\lambda}{2}(\boldsymbol{\theta} - \boldsymbol{\eta})^T \boldsymbol{C}^{-1}(\boldsymbol{\theta} - \boldsymbol{\eta})\right\} d\boldsymbol{\theta} d\lambda \\
=&\int f(\lambda|\gamma, \delta) \int \left(\frac{\lambda}{2\pi}\right)^{\frac{(n_c+n_r)m+B}{2}} |\boldsymbol{C}|^{-1/2} \exp\left\{-\frac{\lambda}{2}\left[\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{y}_{ci} + \sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{y}_{ri} + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}\right]\right\} \\
&\exp\left\{-\frac{\lambda}{2}\left[(n_c + n_r)\boldsymbol{\theta}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi\theta} + \boldsymbol{\theta}^T\boldsymbol{C}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\theta} - 2\left(\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T + \sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\right)\boldsymbol{\Phi\theta}\right]\right\} d\boldsymbol{\theta} d\lambda
\end{aligned}
$$

Let

$$
\begin{aligned}
\boldsymbol{C}_n^{*-1} &= (n_c + n_r)\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1} \\
\boldsymbol{\mu}_n^* &= \left((n_c + n_r)\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}\right)^{-1}\left(\boldsymbol{\eta}^T\boldsymbol{C}^{-1} + \left(\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T + \sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\right)\boldsymbol{\Phi}\right)^T
\end{aligned}
$$

Overall, the integral then becomes

$$
\begin{aligned}
&\int \frac{\delta^\gamma}{\Gamma(\gamma)}\lambda^{\gamma-1}\exp\{-\delta\lambda\}\left(\frac{\lambda}{2\pi}\right)^{\frac{(n_c+n_r)m}{2}} \left(|\boldsymbol{C}^{*-1}||\boldsymbol{C}|\right)^{-1/2} \\
&\qquad\qquad \exp\left\{-\frac{\lambda}{2}\left[\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{y}_{ci} + \sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{y}_{ri} + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta} - \boldsymbol{\mu}_n^{*T}\boldsymbol{C}_n^{*-1}\boldsymbol{\mu}_n^*\right]\right\} d\lambda \\
=&\frac{\delta^\gamma}{\Gamma(\gamma)}\left(\frac{1}{2\pi}\right)^{\frac{(n_c+n_r)m}{2}} \left|\sum_{i=1}^{n_c}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \sum_{i=1}^{n_r}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}\right|^{-1/2} |\boldsymbol{C}|^{-1/2} \frac{\Gamma(\gamma^*)}{\delta^{*\gamma^*}}
\end{aligned}
$$

With

$$\gamma^* = \gamma + \frac{(n_c + n_r)m}{2}$$
$$\delta^* = \delta + \frac{1}{2}\left(\sum_{i=1}^{n_c} \boldsymbol{y}_{ci}^T \boldsymbol{y}_{ci} + \sum_{i=1}^{n_r} \boldsymbol{y}_{ri}^T \boldsymbol{y}_{ri} + \boldsymbol{\eta}^T \boldsymbol{C}^{-1} \boldsymbol{\eta} - \boldsymbol{\mu}_n^{*T} \boldsymbol{C}_n^{*-1} \boldsymbol{\mu}_n^*\right)$$

C.2.2 Likelihood ratio evaluation under alternative proposition

The denominator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and from the control curve have different origins, or $\boldsymbol{\theta}_r \neq \boldsymbol{\theta}_c$.

$$\prod_{q\in\{c,r\}} \int_\lambda \int_{\boldsymbol{\theta}} \prod_{i=1}^{n_q} f(\boldsymbol{y}_{qi}|\boldsymbol{\Phi}\boldsymbol{\theta}, \boldsymbol{I}/\lambda) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}/\lambda) d\boldsymbol{\theta} f(\lambda|\gamma, \delta) d\lambda$$
$$= \prod_{q\in\{c,r\}} \frac{\delta^\gamma}{\Gamma(\gamma)} \left(\frac{1}{2\pi}\right)^{\frac{n_q m}{2}} \left|\sum_{i=1}^{n_c} \boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}\right|^{-1/2} |\boldsymbol{C}|^{-1/2} \frac{\Gamma(\gamma_q^*)}{\delta_q^{*\gamma_q^*}}$$

where

$$\gamma_q^* = \gamma + \frac{n_q m}{2}$$
$$\delta_q^* = \delta + \frac{1}{2}\left[\sum_{i=1}^{n_q} \boldsymbol{y}_{qi}^T \boldsymbol{y}_{qi} + \boldsymbol{\eta}^T \boldsymbol{C}^{-1} \boldsymbol{\eta} - \left(\boldsymbol{\eta}^T \boldsymbol{C}^{-1} + \sum_{i=1}^{n_q} \boldsymbol{y}_{qi}^T \boldsymbol{\Phi}\right) \left(n_q \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \boldsymbol{C}^{-1}\right)^{-1} \left(\boldsymbol{\eta}^T \boldsymbol{C}^{-1} + \boldsymbol{\Phi}^T \sum_{i=1}^{n_q} \boldsymbol{y}_{qi}\right)^T\right]$$

C.2.3 Likelihood ratio

Putting the numerator and denominator together we get

$$LR = \frac{\Gamma(\gamma)|\boldsymbol{C}|^{1/2}}{\delta^\gamma} \frac{\Gamma(\gamma^*)}{\Gamma(\gamma_c^*)\Gamma(\gamma_r^*)} \frac{\delta_c^{*\gamma_c^*} \delta_r^{*\gamma_r^*}}{\delta^{*\gamma^*}} \frac{|(n_c + n_r)\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}|^{-1/2}}{|n_c\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}|^{-1/2} |n_r\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{C}^{-1}|^{-1/2}}.$$

C.2.4 Estimation of hyperparameters using relevant population

$\boldsymbol{y}_{ki} \sim N_m(\boldsymbol{\Phi\theta}_k, \boldsymbol{I}_m/\lambda_k), \quad k = 1, \ldots, K$

$$\begin{aligned}
\hat{\lambda}_k &= \frac{1}{RSS_k/(mn_k - B)} \\
\hat{\lambda} &= \frac{1}{K}\sum_k \hat{\lambda}_k = \gamma/\delta \\
s^2_\lambda &= \frac{1}{K-1}\sum_k (\hat{\lambda}_k - \hat{\lambda})^2 = \gamma/\delta^2 \\
\hat{\delta} &= \hat{\lambda}/s^2_\lambda, \quad \hat{\gamma} = \hat{\lambda}^2/s^2_\lambda
\end{aligned}$$

$\boldsymbol{\theta}_k \sim N_B(\boldsymbol{\eta}, \boldsymbol{C}/\lambda_k), \quad k = 1, \ldots, K$

$$\begin{aligned}
\hat{\boldsymbol{\eta}} &= \frac{1}{K}\sum_k \hat{\boldsymbol{\theta}}_k \\
\hat{\boldsymbol{\theta}}_k &= \frac{1}{n_k}\sum_i \hat{\boldsymbol{\theta}}_{ki} \\
\hat{\boldsymbol{C}} &= \frac{1}{K}\sum_k \hat{\text{var}}(\boldsymbol{\theta}_k)\hat{\lambda} \\
&= \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\eta}})^2 / \frac{\sum_{k=1}^K \hat{\sigma}^2_k}{K} - \frac{1}{n}\boldsymbol{\Phi}^T\boldsymbol{\Phi}
\end{aligned}$$

C.2.5 Simulation

Under this model, datasets will be generated by first simulate group variances $\sigma^2_k \sim Inv{-}Gam(\hat{\gamma}, \hat{\delta})$ then group means $\boldsymbol{\theta}_k \sim N(\hat{\boldsymbol{\eta}}, \sigma^2_k\hat{\boldsymbol{C}})$ and finally, $\boldsymbol{y}_{ki} \sim N(\boldsymbol{\Phi\theta}_k, \sigma^2_k\boldsymbol{I}_m)$ for $1 \leq k \leq K$ and $1 \leq i \leq n_k$. Descriptions of this model can be found in Section 3.2.3.

C.3 CA-ar Multivariate normal random-effects with autoregressive within-group covariance model

$$
\begin{aligned}
&\frac{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_p)}{f(\boldsymbol{Y}_c, \boldsymbol{Y}_r|H_d)} \\
=&\frac{\int_{\sigma^2}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\theta}, \sigma^2\boldsymbol{P})\prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\theta}, \sigma^2\boldsymbol{P})f(\boldsymbol{\theta}|\boldsymbol{\eta}, \sigma^2\boldsymbol{C})f(\sigma^2|\gamma, \delta)d\boldsymbol{\theta}d\sigma^2}{\prod_{q\in\{c,r\}}\int_{\sigma^2}\int_{\boldsymbol{\theta}}\prod_{i=1}^{n_q} f(\boldsymbol{y}_{ci}|\boldsymbol{\theta}, \sigma^2\boldsymbol{P})f(\boldsymbol{\theta}|\boldsymbol{\eta}, \sigma^2\boldsymbol{C})f(\sigma^2|\gamma, \delta)d\boldsymbol{\theta}d\sigma^2}
\end{aligned}
$$

C.3.1 Likelihood ratio evaluation under prosecution proposition

The numerator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and the control curve come from the same origin, or $\boldsymbol{\theta}_r = \boldsymbol{\theta}_c$.

$$
\begin{aligned}
&\int\int\prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi\theta}, \boldsymbol{P}/\lambda)\prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi\theta}, \boldsymbol{P}/\lambda)f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}/\lambda)d\boldsymbol{\theta}f(\lambda|\gamma, \delta)d\lambda \\
=&\int f(\lambda|\gamma, \delta)\int\prod_{i=1}^{n_c} f(\boldsymbol{y}_{ci}|\boldsymbol{\Phi\theta}, \boldsymbol{P}/\lambda)\prod_{i=1}^{n_r} f(\boldsymbol{y}_{ri}|\boldsymbol{\Phi\theta}, \boldsymbol{P}/\lambda)f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}/\lambda)d\boldsymbol{\theta}d\lambda \\
=&\int f(\lambda|\gamma, \delta)\int\left(\frac{\lambda}{2\pi}\right)^{n_c m/2}|\boldsymbol{P}|^{-n_c/2}\text{exp}\left\{-\frac{\lambda}{2}\sum_{i=1}^{n_c}(\boldsymbol{y}_{ci}-\boldsymbol{\Phi\theta})^T\boldsymbol{P}^{-1}(\boldsymbol{y}_{ci}-\boldsymbol{\Phi\theta})\right\} \\
&\left(\frac{\lambda}{2\pi}\right)^{n_r m/2}|\boldsymbol{P}|^{-n_r/2}\text{exp}\left\{-\frac{\lambda}{2}\sum_{i=1}^{n_r}(\boldsymbol{y}_{ri}-\boldsymbol{\Phi\theta})^T\boldsymbol{P}^{-1}(\boldsymbol{y}_{ri}-\boldsymbol{\Phi\theta})\right\} \\
&\left(\frac{\lambda}{2\pi}\right)^{B/2}|\boldsymbol{C}|^{-1/2}\text{exp}\left\{-\frac{\lambda}{2}(\boldsymbol{\theta}-\boldsymbol{\eta})^T\boldsymbol{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\eta})\right\}d\boldsymbol{\theta}d\lambda \\
=&\int f(\lambda|\gamma, \delta)\int\left(\frac{\lambda}{2\pi}\right)^{\frac{(n_c+n_r)m+B}{2}}|\boldsymbol{P}|^{-\frac{n_c+n_r}{2}}|\boldsymbol{C}|^{-1/2} \\
&\text{exp}\left\{-\frac{\lambda}{2}\left(\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ci}+\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ri}+\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}\right)\right\} \\
&\text{exp}\left\{-\frac{\lambda}{2}\left(\sum_{i=1}^{n_c}\boldsymbol{\theta}^T\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi\theta}+\sum_{i=1}^{n_r}\boldsymbol{\theta}^T\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi\theta}+\boldsymbol{\theta}^T\boldsymbol{C}^{-1}\boldsymbol{\theta}\right)\right\} \\
&\text{exp}\left\{-\frac{\lambda}{2}\left(-2\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\theta}-2\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi\theta}-2\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi\theta}\right)\right\}d\boldsymbol{\theta}d\lambda
\end{aligned}
$$

Let

$$\boldsymbol{P}_n^{*-1} = (n_c+n_r)\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\boldsymbol{C}^{-1}$$

$$\boldsymbol{\mu}_n^* = \left((n_c+n_r)\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\boldsymbol{C}^{-1}\right)^{-1}\left(\boldsymbol{\eta}^T\boldsymbol{C}^{-1}+\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}\right)^T$$

Overall, the integral then becomes

$$\begin{aligned}
&\int \frac{\delta^\gamma}{\Gamma(\gamma)}\lambda^{\gamma-1}\exp\{-\delta\lambda\}\left(\frac{\lambda}{2\pi}\right)^{\frac{(n_c+n_r)m}{2}} |\boldsymbol{P}_n^*|^{1/2}|\boldsymbol{P}|^{-\frac{n_c+n_r}{2}}|\boldsymbol{C}|^{-1/2}\\
&\qquad\exp\left\{-\frac{\lambda}{2}(\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ci}+\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ri}+\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}-\boldsymbol{\mu}_n^{*T}\boldsymbol{P}_n^{*-1}\boldsymbol{\mu}_n^*)\right\}d\lambda\\
=&\frac{\delta^\gamma}{\Gamma(\gamma)}\left(\frac{1}{2\pi}\right)^{\frac{(n_c+n_r)m}{2}}\left|(n_c+n_r)\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\boldsymbol{C}^{-1}\right|^{-1/2}|\boldsymbol{P}|^{-\frac{n_c+n_r}{2}}|\boldsymbol{C}|^{-1/2}\frac{\Gamma(\gamma^*)}{\delta^{*\gamma^*}}
\end{aligned}$$

With

$$\begin{aligned}
\gamma^* =&\gamma+\frac{(n_c+n_r)m}{2}\\
\delta^* =&\delta+\frac{1}{2}\left(\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ci}+\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ri}+\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}-\boldsymbol{\mu}_n^{*T}\boldsymbol{P}_n^{*-1}\boldsymbol{\mu}_n^*\right)\\
=&\delta+\frac{1}{2}\left\{\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ci}+\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{ri}+\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}\right.\\
&-\left(\boldsymbol{\eta}^T\boldsymbol{C}^{-1}+\left(\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T+\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}^T\right)\boldsymbol{P}^{-1}\boldsymbol{\Phi}\right)\\
&\left.((n_c+n_r)\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}+\boldsymbol{C}^{-1})^{-1}\left(\boldsymbol{C}^{-1}\boldsymbol{\eta}+\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\left(\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}+\sum_{i=1}^{n_r}\boldsymbol{y}_{ri}\right)\right)\right\}
\end{aligned}$$

C.3.2 Likelihood ratio evaluation under prosecution proposition

The denominator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and the control curve come from different origins, independently.

$$\prod_{q\in\{c,r\}}\int\int\prod_{i=1}^{n_q}f(\boldsymbol{y}_{qi}|\boldsymbol{\Phi}\boldsymbol{\theta},\boldsymbol{P}/\lambda)f(\boldsymbol{\theta}|\boldsymbol{\eta},\boldsymbol{C}/\lambda)d\boldsymbol{\theta}f(\lambda|\gamma,\delta)d\lambda$$

The two terms are similar so a general case is derived.

$$
\begin{aligned}
&\int\int \prod_{i=1}^{n_q} f(\boldsymbol{y}_{qi}|\boldsymbol{\Phi\theta}, \boldsymbol{P}/\lambda_q) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}/\lambda_q) d\boldsymbol{\theta} f(\lambda|\gamma,\delta) d\lambda_q \\
&= \int f(\lambda|\gamma,\delta) \int \left(\frac{\lambda}{2\pi}\right)^{\frac{n_q m}{2}} |\boldsymbol{P}|^{-n_q/2} \text{exp}\left\{-\frac{\lambda}{2}\sum_{i=1}^{n_q}(\boldsymbol{y}_{qi}-\boldsymbol{\Phi\theta})^T\boldsymbol{P}^{-1}(\boldsymbol{y}_{qi}-\boldsymbol{\Phi\theta})\right\} \\
&\qquad \left(\frac{\lambda}{2\pi}\right)^{B/2} |\boldsymbol{C}|^{-1/2} \text{exp}\left\{-\frac{\lambda}{2}(\boldsymbol{\theta}-\boldsymbol{\eta})^T\boldsymbol{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\eta})\right\} d\boldsymbol{\theta} d\lambda \\
&= \int f(\lambda|\gamma,\delta) \int \left(\frac{\lambda}{2\pi}\right)^{\frac{n_q m+B}{2}} |\boldsymbol{P}|^{-n_q/2} |\boldsymbol{C}|^{-1/2} \text{exp}\left\{-\frac{\lambda}{2}\left(\sum_{i=1}^{n_q}\boldsymbol{y}_{qi}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{qi} + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}\right)\right\} \\
&\qquad \text{exp}\left\{-\frac{\lambda}{2}\left(\sum_{i=1}^{n_q}\boldsymbol{\theta}^T\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi\theta} + \boldsymbol{\theta}^T\boldsymbol{C}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\theta} - 2\sum_{i=1}^{n_c}\boldsymbol{y}_{ci}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi\theta}\right)\right\} d\boldsymbol{\theta} d\lambda
\end{aligned}
$$

Let

$$
\begin{aligned}
\boldsymbol{P}_q^{*-1} &= \sum_{i=1}^{n_q} \boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi} + \boldsymbol{C}^{-1} \\
\boldsymbol{\mu}_q^* &= \left(\sum_{i=1}^{n_q} \boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi} + \boldsymbol{C}^{-1}\right)^{-1} \left(\boldsymbol{\eta}^T\boldsymbol{C}^{-1} + \sum_{i=1}^{n_q}\boldsymbol{y}_{ci}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}\right)^T
\end{aligned}
$$

Overall, the integral then becomes

$$
\begin{aligned}
&\int \frac{\delta^\gamma}{\Gamma(\gamma)}\lambda^{\gamma-1}\text{exp}\{-\delta\lambda\}\left(\frac{\lambda}{2\pi}\right)^{\frac{n_q m}{2}} |\boldsymbol{P}_q^*|^{1/2}|\boldsymbol{P}|^{-n_q/2}|\boldsymbol{C}|^{-1/2} \\
&\qquad\qquad \text{exp}\left\{-\frac{\lambda}{2}\left(\sum_{i=1}^{n_q}\boldsymbol{y}_{qi}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{qi} + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta} - \boldsymbol{\mu}_q^{*T}\boldsymbol{P}_q^{*-1}\boldsymbol{\mu}_q^*\right)\right\} d\lambda \\
&= \frac{\delta^\gamma}{\Gamma(\gamma)}\left(\frac{1}{2\pi}\right)^{\frac{n_q m}{2}} \left|\sum_{i=1}^{n_q}\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi} + \boldsymbol{C}^{-1}\right|^{-1/2} |\boldsymbol{P}|^{-n_q/2}|\boldsymbol{C}|^{-1/2}\frac{\Gamma(\gamma_q^*)}{\delta_q^{*\gamma_q^*}}
\end{aligned}
$$

With

$$
\begin{aligned}
\gamma_q^* &= \gamma + \frac{n_q m}{2} \\
\delta_q^* &= \delta + \frac{1}{2}\left(\sum_{i=1}^{n_q}\boldsymbol{y}_{qi}^T\boldsymbol{P}^{-1}\boldsymbol{y}_{qi} + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta} - \boldsymbol{\mu}_q^{*T}\boldsymbol{P}_q^{*-1}\boldsymbol{\mu}_q^*\right)
\end{aligned}
$$

$$= \delta + \frac{1}{2}\left[\sum_{i=1}^{n_q} \boldsymbol{y}_{qi}^T \boldsymbol{P}^{-1}\boldsymbol{y}_{qi} + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta} - \left(\boldsymbol{\eta}^T\boldsymbol{C}^{-1} + \sum_{i=1}^{n_q}\boldsymbol{y}_{qi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi}\right)(n_q\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi} + \boldsymbol{C}^{-1})^{-1}\left(\boldsymbol{C}^{-1}\boldsymbol{\eta} + \boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\sum_{i=1}^{n_q}\boldsymbol{y}_{qi}\right)\right]$$

C.3.3 Likelihood ratio

Putting the numerator and denominator together we get

$$LR = \frac{\Gamma(\gamma)|\boldsymbol{C}|^{1/2}}{\delta^{\gamma}}\frac{\Gamma(\gamma^*)}{\Gamma(\gamma_c^*)\Gamma(\gamma_r^*)}\frac{\delta_c^{*\gamma_c^*}\delta_r^{*\gamma_r^*}}{\delta^{*\gamma^*}}\frac{|(n_c+n_r)\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi} + \boldsymbol{C}^{-1}|^{-1/2}}{|n_c\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi} + \boldsymbol{C}^{-1}|^{-1/2}|n_r\boldsymbol{\Phi}^T\boldsymbol{P}^{-1}\boldsymbol{\Phi} + \boldsymbol{C}^{-1}|^{-1/2}}.$$

C.3.4 Estimation of hyperparameters from relevant population

$Y_{kij} = f_k(j) + r_{kij} = f_k(j) + \sigma_k\epsilon_{kij}$

$r_{kij} = \sigma_k\epsilon_{kij} = \sigma_k(\psi_1\epsilon_{ki,j-1} + \psi_2\epsilon_{ki,j-2} + ... + \psi_p\epsilon_{ki,j-p}) + \omega_{kij}$

Preliminary analysis suggests p=1. So

$$r_{kij} = \psi r_{ki,j-1} + \omega_{kij}$$
$$\sigma_k\epsilon_{kij} = \psi(\sigma_k\epsilon_{ki,j-1}) + \omega_{kij}.$$

Equating the variances of both side gives

$$Var(\sigma_k\epsilon_{kij}) = Var(\psi(\sigma_k\epsilon_{ki,j-1}) + \omega_{kij})$$
$$\sigma_k^2 = \psi_{(k)}^2\sigma_k^2 + \tau_k^2$$

variables with subscript (k) are transitional (for estimation only)

$$\hat{\sigma}_k^2 = \frac{\hat{\tau}_k^2}{1-\hat{\psi}_{(k)}^2} = \frac{1}{\hat{\lambda}_k}$$
$$\text{where } \hat{\tau}_k^2 = \frac{\sum_i\sum_{j=2}^m(r_{kij} - \hat{\psi}_{(k)}r_{ki,j-1})^2}{n(m-1)-1}$$
$$\hat{\lambda} = \frac{1}{K}\sum_k\hat{\lambda}_k = \gamma/\delta$$
$$s_\lambda^2 = \frac{1}{K-1}\sum_k(\hat{\lambda}_k - \hat{\lambda})^2 = \gamma/\delta^2$$
$$\hat{\delta} = \hat{\lambda}/s_\lambda^2, \quad \hat{\gamma} = \hat{\lambda}^2/s_\lambda^2$$

$$\hat{\boldsymbol{P}}_{(k)} = \begin{pmatrix} 1 & \hat{\psi}_{(k)} & \hat{\psi}^2_{(k)} & \cdots & \hat{\psi}^{m-1}_{(k)} \\ \hat{\psi}_{(k)} & 1 & \hat{\psi}_{(k)} & & \\ \hat{\psi}^2_{(k)} & \hat{\psi}_{(k)} & 1 & & \\ \vdots & & & \ddots & \vdots \\ \hat{\psi}^{m-1}_{(k)} & & & \cdots & 1 \end{pmatrix}$$

$$\hat{\boldsymbol{P}} = \frac{\sum_{k=1}^{K} \hat{\boldsymbol{P}}_{(k)}}{K}$$

$\boldsymbol{\theta} \sim N_B(\boldsymbol{\eta}, \boldsymbol{C}/\lambda), \quad k = 1, \ldots, K$

$$\begin{aligned} \hat{\boldsymbol{\eta}} &= \frac{1}{K} \sum_k \hat{\boldsymbol{\theta}}_k \\ \hat{\boldsymbol{\theta}}_k &= \frac{1}{n_k} \sum_i \hat{\boldsymbol{\theta}}_{ki} \\ \hat{\boldsymbol{C}} &= \frac{1}{K} \sum_k \hat{\text{var}}(\boldsymbol{\theta}_k) \hat{\lambda} \\ &= \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\eta}})^2 / \frac{\sum_{k=1}^{K} \hat{\sigma}_k^2}{K} - \frac{1}{n} \boldsymbol{\Phi}^T \boldsymbol{\Phi} \end{aligned}$$

C.3.5 Simulation

Under this model, datasets will be generated by first simulate group variances $\sigma_k^2 \sim Inv{-}Gam(\hat{\gamma}, \hat{\delta})$ then group means $\boldsymbol{\theta}_k \sim N(\hat{\boldsymbol{\eta}}, \sigma_k^2 \hat{\boldsymbol{C}})$ and finally, $\boldsymbol{y}_{ki} \sim N(\boldsymbol{\Phi}\boldsymbol{\theta}_k, \sigma_k^2 \hat{\boldsymbol{P}})$ for $1 \leq k \leq K$ and $1 \leq i \leq n_k$. Descriptions of this model can be found in Section 3.2.4.

C.4 DR-S Dimension reduced multivariate normal random-effects model

C.4.1 Likelihood ratio evaluation under prosecution proposition

The numerator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and the control curve come from the same origin, or $\boldsymbol{\theta}_r = \boldsymbol{\theta}_c$.

$$
\begin{aligned}
&\int \prod_{i=1}^{n_c} f(\boldsymbol{z}_{ci}|\boldsymbol{\theta},\boldsymbol{U}) \prod_{i=1}^{n_r} f(\boldsymbol{z}_{ri}|\boldsymbol{\theta},\boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta},\boldsymbol{C}) d\boldsymbol{\theta} \\
&= \int |2\pi\boldsymbol{U}|^{-n_c/2} |2\pi\boldsymbol{U}|^{-n_r/2} |2\pi\boldsymbol{C}|^{-1/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_c}(\boldsymbol{z}_{ci}-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ci}-\boldsymbol{\theta})\right\} \\
&\qquad \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_r}(\boldsymbol{z}_{ri}-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ri}-\boldsymbol{\theta})\right\} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\eta})^T C^{-1}(\boldsymbol{\theta}-\boldsymbol{\eta})\right\} d\boldsymbol{\theta} \\
&= \int |2\pi\boldsymbol{U}|^{-n_c/2} |2\pi\boldsymbol{U}|^{-n_r/2} |2\pi\boldsymbol{C}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\eta})^T C^{-1}(\boldsymbol{\theta}-\boldsymbol{\eta})\right\} \\
&\qquad \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_c}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c) - \frac{1}{2}\sum_{i=1}^{n_c}(\bar{\boldsymbol{z}}_c-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\bar{\boldsymbol{z}}_c-\boldsymbol{\theta})\right\} \\
&\qquad \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_r}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r) - \frac{1}{2}\sum_{i=1}^{n_r}(\bar{\boldsymbol{z}}_r-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\bar{\boldsymbol{z}}_r-\boldsymbol{\theta})\right\} d\boldsymbol{\theta} \\
&= \int |2\pi\boldsymbol{U}|^{-(n_c+n_r)/2} |2\pi\boldsymbol{C}|^{-1/2} \\
&\qquad \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_c}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c) - \frac{1}{2}\sum_{i=1}^{n_r}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)\right\} \\
&\qquad \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n_c}\bar{\boldsymbol{z}}_c^T\boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_c + \sum_{i=1}^{n_r}\bar{\boldsymbol{z}}_r^T\boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_r + \boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}\right)\right\} \\
&\qquad \exp\left\{-\frac{1}{2}(\sum_{i=1}^{n_c}\boldsymbol{\theta}^T\boldsymbol{U}^{-1}\boldsymbol{\theta} + \sum_{i=1}^{n_r}\boldsymbol{\theta}^T\boldsymbol{U}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T\boldsymbol{C}^{-1}\boldsymbol{\theta})\right\} \\
&\qquad \exp\left\{-\frac{1}{2}(-2\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\theta} - 2\sum_{i=1}^{n_c}\bar{\boldsymbol{z}}_c^T\boldsymbol{U}^{-1}\boldsymbol{\theta} - 2\sum_{i=1}^{n_r}\bar{\boldsymbol{z}}_r^T\boldsymbol{U}^{-1}\boldsymbol{\theta})\right\} d\boldsymbol{\theta}
\end{aligned}
$$

Let

$$
\begin{aligned}
\boldsymbol{\Sigma}_n^{*-1} &= \sum_{i=1}^{n_c}\boldsymbol{U}^{-1} + \sum_{i=1}^{n_r}\boldsymbol{U}^{-1} + \boldsymbol{C}^{-1} = (n_c+n_r)\boldsymbol{U}^{-1} + \boldsymbol{C}^{-1} \\
\boldsymbol{\eta}_n^* &= \left((n_c+n_r)\boldsymbol{U}^{-1} + \boldsymbol{C}^{-1}\right)^{-1}\left(\boldsymbol{C}^{-1}\boldsymbol{\eta} + \boldsymbol{U}^{-1}n_c\bar{\boldsymbol{z}}_c + \boldsymbol{U}^{-1}n_r\bar{\boldsymbol{z}}_r\right)
\end{aligned}
$$

Overall, the integral becomes

$$
\begin{aligned}
&|2\pi\boldsymbol{U}|^{-(n_c+n_r)/2}\,|2\pi\boldsymbol{C}|^{-1/2} \\
&\qquad \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_c}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)-\frac{1}{2}\sum_{i=1}^{n_r}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)\right\} \\
&\qquad \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n_c}\bar{\boldsymbol{z}}_c^T\boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_c+\sum_{i=1}^{n_r}\bar{\boldsymbol{z}}_r^T\boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_r+\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}-\boldsymbol{\eta}_n^{*T}\boldsymbol{\Sigma}_n^{*-1}\boldsymbol{\eta}_n^*\right)\right\}
\end{aligned}
$$

which can be shown to simply to

$$
|2\pi\boldsymbol{U}|^{-(n_c+n_r)/2}\,|2\pi\boldsymbol{C}|^{-1/2}\left|2\pi((n_c+n_r)\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1})^{-1}\right|^{1/2}\exp\left\{-\frac{1}{2}(h_1+h_2+h_3)\right\}
$$

where

$$
\begin{aligned}
h_1 &= \sum_{i=1}^{n_c}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)+\sum_{i=1}^{n_r}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r) \\
&= \text{tr}\left(\sum_{i=1}^{n_c}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)\right)+\text{tr}\left(\sum_{i=1}^{n_r}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)\right) \\
&= \text{tr}\left(\sum_{i=1}^{n_c}(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)(\boldsymbol{z}_{ci}-\bar{\boldsymbol{z}}_c)^T\boldsymbol{U}^{-1}\right)+\text{tr}\left(\sum_{i=1}^{n_r}(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)(\boldsymbol{z}_{ri}-\bar{\boldsymbol{z}}_r)^T\boldsymbol{U}^{-1}\right) \\
&= \text{tr}\left(\boldsymbol{S}_c\boldsymbol{U}^{-1}\right)+\text{tr}(\boldsymbol{S}_r\boldsymbol{U}^{-1}) \\
h_2 &= (\boldsymbol{z}^*-\boldsymbol{\eta})^T\left(\frac{\boldsymbol{U}}{n_c+n_r}+\boldsymbol{C}\right)^{-1}(\boldsymbol{z}^*-\boldsymbol{\eta}) \\
h_3 &= (\bar{\boldsymbol{z}}_c-\bar{\boldsymbol{z}}_r)^T\left(\frac{\boldsymbol{U}}{n_c}+\frac{\boldsymbol{U}}{n_r}\right)^{-1}(\bar{\boldsymbol{z}}_c-\bar{\boldsymbol{z}}_r)
\end{aligned}
$$

with

$$
\boldsymbol{z}^*=\frac{n_c\bar{\boldsymbol{z}}_c+n_r\bar{\boldsymbol{z}}_r}{n_c+n_r}.
$$

C.4.2 Likelihood ratio evaluation under alternative proposition

The denominator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and the control curve come from different origins, indepen-

dently.

$$\int \prod_{i=1}^{n_c} f(\boldsymbol{z}_{ci}|\boldsymbol{\theta}_c, \boldsymbol{U}) f(\boldsymbol{\theta}_c|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta}_c \int \prod_{i=1}^{n_r} f(\boldsymbol{z}_{ri}|\boldsymbol{\theta}_r, \boldsymbol{U}) f(\boldsymbol{\theta}_r|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta}_r$$

These two parts are similar, so the calculation is showed for the general case.

$$\begin{aligned}
&\int \prod_{i=1}^{n_q} f(\boldsymbol{z}_{qi}|\boldsymbol{\theta}_q, \boldsymbol{U}) f(\boldsymbol{\theta}_q|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta}_q \\
&= \int |2\pi \boldsymbol{U}|^{-n_q/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta}_q)^T \boldsymbol{U}^{-1}(\boldsymbol{z}_{qi}-\boldsymbol{\theta}_q)\right\} |2\pi \boldsymbol{C}|^{-1/2} \\
&\qquad \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_q-\boldsymbol{\eta})^T \boldsymbol{C}^{-1}(\boldsymbol{\theta}_q-\boldsymbol{\eta})\right\} d\boldsymbol{\theta}_q \\
&= |2\pi \boldsymbol{U}|^{-n_q/2} |2\pi \boldsymbol{C}|^{-1/2} \\
&\qquad \int \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\bar{\boldsymbol{z}}_q)^T \boldsymbol{U}^{-1}(\boldsymbol{z}_{qi}-\bar{\boldsymbol{z}}_q) - \frac{1}{2}\sum_{i=1}^{n_q}(\bar{\boldsymbol{z}}_q-\boldsymbol{\theta}_q)^T \boldsymbol{U}^{-1}(\bar{\boldsymbol{z}}_q-\boldsymbol{\theta}_q)\right\} \\
&\qquad \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_q-\boldsymbol{\eta})^T \boldsymbol{C}^{-1}(\boldsymbol{\theta}_q-\boldsymbol{\eta})\right\} d\boldsymbol{\theta}_q \\
&= |2\pi \boldsymbol{U}|^{-n_q/2} |2\pi \boldsymbol{C}|^{-1/2} \\
&\qquad \int \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\bar{\boldsymbol{z}}_q)^T \boldsymbol{U}^{-1}(\boldsymbol{z}_{qi}-\bar{\boldsymbol{z}}_q) - \frac{1}{2}\left(\sum_{i=1}^{n_q}\bar{\boldsymbol{z}}_q^T \boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_q + \boldsymbol{\eta}^T \boldsymbol{C}^{-1}\boldsymbol{\eta}\right)\right\} \\
&\qquad \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n_q}\boldsymbol{\theta}_q^T \boldsymbol{U}^{-1}\boldsymbol{\theta}_q + \boldsymbol{\theta}_q^T \boldsymbol{C}^{-1}\boldsymbol{\theta}_q - 2\boldsymbol{\eta}^T \boldsymbol{C}^{-1}\boldsymbol{\theta}_q - 2\sum_{i=1}^{n_q}\bar{\boldsymbol{z}}_q^T \boldsymbol{U}^{-1}\boldsymbol{\theta}_q\right)\right\} d\boldsymbol{\theta}_q
\end{aligned}$$

Let

$$\begin{aligned}
\boldsymbol{\Sigma}_q^{*-1} &= \sum_{i=1}^{n_q} \boldsymbol{U}^{-1} + \boldsymbol{C}^{-1} \\
\boldsymbol{\eta}_q^* &= \left(\sum_{i=1}^{n_q} \boldsymbol{U}^{-1} + \boldsymbol{C}^{-1}\right)^{-1} \left(C^{-1}\boldsymbol{\eta} + \sum_{i=1}^{n_q} \boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_q\right)
\end{aligned}$$

which integrates to

$$\begin{aligned}
&|2\pi \boldsymbol{U}\ |^{-n_q/2} |2\pi \boldsymbol{C}|^{-1/2} \left|2\pi \boldsymbol{\Sigma}_q^*\right|^{1/2} \\
&\qquad \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\bar{\boldsymbol{z}}_q)^T \boldsymbol{U}^{-1}(\boldsymbol{z}_{qi}-\bar{\boldsymbol{z}}_q) - \sum_{i=1}^{n_q}\bar{\boldsymbol{z}}_q^T \boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_q + \boldsymbol{\eta}^T \boldsymbol{C}^{-1}\boldsymbol{\eta} - \boldsymbol{\eta}_q^{*T} \boldsymbol{\Sigma}_q^{*-1}\boldsymbol{\eta}_q^*\right)\right\}
\end{aligned}$$

Using the relation

$$(\boldsymbol{A}+\boldsymbol{B})^{-1}=\boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right)^{-1}\boldsymbol{B}^{-1}=\boldsymbol{B}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right)^{-1}\boldsymbol{A}^{-1}$$

we get

$$\begin{aligned}\boldsymbol{A}^{-1}&=(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1})\\ \boldsymbol{B}^{-1}&=(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{A}\left(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right)\end{aligned}$$

hence

$$\begin{aligned}n_q\boldsymbol{U}^{-1}&=\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\boldsymbol{C}(n_q\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1})\\ \boldsymbol{C}^{-1}&=\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\frac{\boldsymbol{U}}{n_q}\left(n_q\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1}\right)\\ \left(n_q\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}&=\frac{\boldsymbol{U}}{n_q}\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\boldsymbol{C}=\boldsymbol{C}\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\frac{\boldsymbol{U}}{n_q}\end{aligned}$$

So

$$\begin{aligned}&\sum_{i=1}^{n_q}\bar{\boldsymbol{z}}_q^T\boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_q+\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}-\boldsymbol{\eta}_q^{*T}\boldsymbol{\Sigma}_q^{*-1}\boldsymbol{\eta}_q^*\\ =&\sum_{i=1}^{n_q}\bar{\boldsymbol{z}}_q^T\boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_q+\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{\eta}\\ &\quad-\left(\boldsymbol{\eta}^T\boldsymbol{C}^{-1}+\sum_{i=1}^{n_q}\bar{\boldsymbol{z}}_q^T\boldsymbol{U}^{-1}\right)\left(\sum_{i=1}^{n_q}\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\left(\boldsymbol{C}^{-1}\boldsymbol{\eta}+\sum_{i=1}^{n_q}\boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_q\right)\\ =&\bar{\boldsymbol{z}}_q^T\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\boldsymbol{C}(n_q\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1})\bar{\boldsymbol{z}}_q-\bar{\boldsymbol{z}}_q^T\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\boldsymbol{C}n_q\boldsymbol{U}^{-1}\bar{\boldsymbol{z}}_q\\ &\quad+\boldsymbol{\eta}^T\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\frac{\boldsymbol{U}}{n_q}(n_q\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1})\boldsymbol{\eta}-\boldsymbol{\eta}^T\boldsymbol{C}^{-1}\boldsymbol{C}\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\frac{\boldsymbol{U}}{n_q}\boldsymbol{C}^{-1}\boldsymbol{\eta}\\ &\quad-2\bar{\boldsymbol{z}}_q^Tn_q\boldsymbol{U}^{-1}\frac{\boldsymbol{U}}{n_q}\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{\eta}\\ =&\bar{\boldsymbol{z}}_q^T\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\bar{\boldsymbol{z}}_q+\boldsymbol{\eta}^T\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\boldsymbol{\eta}-2\bar{\boldsymbol{z}}_q^TT\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\boldsymbol{\eta}\\ =&(\bar{\boldsymbol{z}}_q-\boldsymbol{\eta})^T\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}(\bar{\boldsymbol{z}}_q-\boldsymbol{\eta}).\end{aligned}$$

The integral is then evaluated as

$$|2\pi \boldsymbol{U}|^{-n_q/2}|2\pi \boldsymbol{C}|^{-1/2}|2\pi \boldsymbol{\Sigma}_q^*|^{1/2}\text{exp}\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{S}_q\boldsymbol{U}^{-1}\right)-\frac{1}{2}\left(\bar{\boldsymbol{z}}_q-\boldsymbol{\eta}\right)^T\left(\frac{\boldsymbol{U}}{n_q}+\boldsymbol{C}\right)^{-1}\left(\bar{\boldsymbol{z}}_q-\boldsymbol{\eta}\right)\right\}$$

where

$$\boldsymbol{S}_q=\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\bar{\boldsymbol{z}}_q)(\boldsymbol{z}_{qi}-\bar{\boldsymbol{z}}_q)^T.$$

Putting these together with the numerator gives

$$\begin{aligned}LR=&\frac{|2\pi \boldsymbol{U}|^{-(n_c+n_r)/2}\,|2\pi \boldsymbol{C}|^{-1/2}\left|2\pi\left((n_c+n_r)\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2}\text{exp}\left\{-\frac{1}{2}\left(h_1+h_2+h_3\right)\right\}}{\prod_{q\in\{c,r\}}|2\pi \boldsymbol{U}|^{-n_q/2}\,|2\pi \boldsymbol{C}|^{-1/2}\left|2\pi\left(n_q\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2}\text{exp}\left\{-\frac{1}{2}\left(h_{1q}+h_{4q}\right)\right\}}\\ =&\frac{\left|\left((n_c+n_r)\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2}\text{exp}\left\{-\frac{1}{2}\left(h_2+h_3\right)\right\}}{|\boldsymbol{C}|^{-1/2}\left|\left(n_c\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2}\left|\left(n_r\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\right|^{1/2}\text{exp}\left\{-\frac{1}{2}\left(h_{4c}+h_{4r}\right)\right\}}.\end{aligned}$$

C.4.3 Estimation of hyperparameters using relevant population

$\boldsymbol{z}_{ki}\sim N_B(\boldsymbol{\theta}_k,\boldsymbol{U}),\quad k=1,\ldots,K,\quad i=1,\ldots,n_k$
$\boldsymbol{\theta}_k\sim N_B(\boldsymbol{\eta},\boldsymbol{C}),\quad k=1,\ldots,K$

$$\begin{aligned}\hat{\boldsymbol{\eta}}&=\frac{1}{K}\sum_k\hat{\boldsymbol{\theta}}_k\\ \hat{\boldsymbol{\theta}}_k&=\frac{1}{n_k}\sum_i\boldsymbol{z}_{ki}\\ \hat{\boldsymbol{C}}&=\frac{1}{K-1}\sum_k\hat{Var}(\boldsymbol{\theta}_k)=\frac{1}{K-1}\sum_k\left(\hat{\boldsymbol{\theta}}_k-\hat{\boldsymbol{\eta}}\right)\left(\hat{\boldsymbol{\theta}}_k-\hat{\boldsymbol{\eta}}\right)^T-\frac{\hat{\boldsymbol{U}}}{n_k}.\end{aligned}$$

C.4.4 Simulation

Under this model, datasets will be generated by first simulate group means $\boldsymbol{\theta}_k\sim N(\hat{\boldsymbol{\eta}},\hat{\boldsymbol{C}})$ then $\boldsymbol{z}_{ki}\sim N(\boldsymbol{\theta}_k,\hat{\boldsymbol{U}})$ for $1\leq k\leq K$ and $1\leq i\leq n_k$. To compare with original data we will reconstruct $\hat{\boldsymbol{y}}_{ki}$ as $\boldsymbol{\Phi}\boldsymbol{z}_{ki}$. Descriptions of this model can be found in Section 3.3.1.

C.5 DR-C Multivariate normal random-effects with non constant within-group covariance

C.5.1 Likelihood ratio evaluation under prosecution proposition

The numerator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and the control curve come from the same origin, or $\boldsymbol{\theta}_r = \boldsymbol{\theta}_c$ and $\boldsymbol{U}_r = \boldsymbol{U}_c$.

We are interested in

$$m(\boldsymbol{Z}|H_p) = \int\int \prod_{i=1}^{n_c} f(\boldsymbol{z}_{ci}|\boldsymbol{\theta}, \boldsymbol{U}) \prod_{i=1}^{n_r} f(\boldsymbol{z}_{ri}|\boldsymbol{\theta}, \boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta} f(\boldsymbol{U}|\boldsymbol{\Omega}, \nu) d\boldsymbol{U}$$

which is difficult to evaluate analytically. However, using Bayes' Theorem as in Chib (1995), the marginal likelihood can be written as

$$m(\boldsymbol{Z}|H_p) = \frac{f(\boldsymbol{Z}|\boldsymbol{\Psi}, H_p)\pi(\boldsymbol{\Psi}|H_p)}{\pi(\boldsymbol{\Psi}|\boldsymbol{Z}, H_p)}$$

where $\boldsymbol{\Psi} = (\boldsymbol{\theta}, \boldsymbol{U})$. Denoting the maximum likelihood estimate as $\boldsymbol{\Psi}^*$, the estimate of the marginal density on logarithmic scale is

$$\log\left[\hat{m}(\boldsymbol{Z}|H_p)\right] = \log\left[f(\boldsymbol{Z}|\boldsymbol{\Psi}^*, H_p)\right] + \log\left[\pi(\boldsymbol{\Psi}^*|H_p)\right] - \log\left[\hat{\pi}(\boldsymbol{\Psi}^*|\boldsymbol{Z}, H_p)\right] \quad \text{(C.1)}$$

where $\hat{\pi}(\boldsymbol{\Psi}^*|\boldsymbol{z}, H_p)$ can be estimated using Gibbs sampling algorithm described in Bozza et al (2008).

The density function of the observation is given by

$$\begin{aligned} f(\boldsymbol{Z}|\boldsymbol{\Psi}, H_p) &= \prod_{i=1}^{n_c} f(\boldsymbol{z}_{ci}|\boldsymbol{\theta}, \boldsymbol{U}) \prod_{i=1}^{n_r} f(\boldsymbol{z}_{ri}|\boldsymbol{\theta}, \boldsymbol{U}) \\ &= \prod_{q\in\{c,r\}} \prod_{i=1}^{n_l} (2\pi)^{-p/2} |\boldsymbol{U}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{z}_{qi} - \boldsymbol{\theta})^T \boldsymbol{U}^{-1} (\boldsymbol{z}_{qi} - \boldsymbol{\theta})\right\} \end{aligned}$$

The prior density for $\boldsymbol{\Psi}$ is given by

$$
\begin{aligned}
&f(\boldsymbol{\Psi}|H_p) \\
&= (2\pi)^{-p/2}|\boldsymbol{C}|^{-1/2}\text{exp}\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\eta})^T\boldsymbol{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\eta})\right\}\frac{|\boldsymbol{\Omega}|^{\frac{\nu}{2}}|\boldsymbol{U}|^{-\frac{\nu+p+1}{2}}}{2^{\nu p/2}\Gamma_p(\nu/2)}\text{exp}\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}\boldsymbol{U}^{-1}\right)\right\}.
\end{aligned}
$$

The complete conditional density of $\boldsymbol{\theta}$ is then

$$
\begin{aligned}
f(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{U}) &= \frac{f(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{U})f(\boldsymbol{\theta}|\boldsymbol{\eta},\boldsymbol{C})}{\int f(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{U})f(\boldsymbol{\theta}|\boldsymbol{\eta},\boldsymbol{C})d\boldsymbol{\theta}} \\
&\propto \text{exp}\left\{-\frac{1}{2}\left[\sum_{l=1}^{2}\sum_{i=1}^{n_l}(\boldsymbol{z}_{li}-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{li}-\boldsymbol{\theta})+(\boldsymbol{\theta}-\boldsymbol{\eta})^T\boldsymbol{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\eta})\right]\right\}
\end{aligned}
$$

which can be shown to be still of type normal with parameters $(\boldsymbol{\eta}^*, \boldsymbol{C}^*)$, where

$$
\begin{aligned}
\boldsymbol{C}^* &= \left(\sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}\boldsymbol{U}^{-1}+\boldsymbol{C}^{-1}\right)^{-1} \\
\boldsymbol{\eta}^* &= \boldsymbol{C}^*\left(\boldsymbol{C}^{-1}\boldsymbol{\eta}+\sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}\boldsymbol{U}^{-1}\boldsymbol{z}_{qi}\right).
\end{aligned}
$$

And the complete conditional density of $\boldsymbol{U}$ would be

$$
\begin{aligned}
&f(\boldsymbol{U}|\boldsymbol{z},\boldsymbol{\theta}) \\
&\propto |\boldsymbol{U}|^{-(n_c+n_r)/2}|\boldsymbol{U}|^{-(\nu+p+1)/2}\text{exp}\left\{-\frac{1}{2}\left[\sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}(\boldsymbol{z}_{li}-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\boldsymbol{z}_{li}-\boldsymbol{\theta})+\text{tr}\left(\boldsymbol{\Omega}\boldsymbol{U}^{-1}\right)\right]\right\} \\
&\propto |\boldsymbol{U}|^{-(n_c+n_r+\nu+p+1)/2}\text{exp}\left\{-\frac{1}{2}\left[\text{tr}\left(\sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}(\boldsymbol{z}_{li}-\boldsymbol{\theta})(\boldsymbol{y}_{li}-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}\right)+\text{tr}\left(\boldsymbol{\Omega}\boldsymbol{U}^{-1}\right)\right]\right\} \\
&\propto |\boldsymbol{U}|^{-(n_c+n_r+\nu+p+1)/2}\text{exp}\left\{-\frac{1}{2}\left[\text{tr}\left(\left(\boldsymbol{\Omega}+\sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta})(\boldsymbol{z}_{qi}-\boldsymbol{\theta})^T\right)\boldsymbol{U}^{-1}\right)\right]\right\}
\end{aligned}
$$

which can be shown to be still of type inverse Wishart with parameters $(\boldsymbol{\Omega}^*, \nu^*)$, where

$$
\begin{aligned}
\boldsymbol{\Omega}^* &= \boldsymbol{\Omega}+\sum_{q\in\{c,r\}}\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta})(\boldsymbol{z}_{qi}-\boldsymbol{\theta})^T \\
\nu^* &= \nu+n_c+n_r.
\end{aligned}
$$

The algorithm is then

1. Estimate $\hat{\boldsymbol{\eta}}$, $\hat{\boldsymbol{C}}$, and $\hat{\boldsymbol{\Omega}}$ from background
2. Sample pairs of $\boldsymbol{\theta}^g \sim N_B(\boldsymbol{\eta}^*, \boldsymbol{C}^*)$ and $\boldsymbol{U}^g \sim \mathcal{IW}(\boldsymbol{\Omega}^*, \nu^*)$, $g = 1, \ldots, G$
3. Obtain maximum likelihood approximation of $\boldsymbol{\Psi}^* = (\boldsymbol{\theta}^*, \boldsymbol{U}^*)$ as

$$\boldsymbol{\Psi}^* = max_{\boldsymbol{\Psi}^g} f(\boldsymbol{z}|\boldsymbol{\Psi}^g, H_p)$$

4. Compute

$$\hat{\pi}(\boldsymbol{U}^*|\boldsymbol{Z}) = \sum_{g=1}^{G} \frac{\pi(\boldsymbol{U}^*|\boldsymbol{Z}, \boldsymbol{\theta}^g)}{G}$$

5. Posterior is then given by $\hat{\pi}(\boldsymbol{\Psi}^*|\boldsymbol{Z}) = \pi(\boldsymbol{\theta}^*|\boldsymbol{U}^*, \boldsymbol{Z})\hat{\pi}(\boldsymbol{U}^*|\boldsymbol{Z})$.

The marginal likelihood (on logarithmic scale) can then be estimated using equation (C.1).

C.5.2 Likelihood ratio evaluation under alternative proposition

The denominator of the likelihood ratio is evaluated under the proposition that the data for the recovered curve and the control curve come from different origins.
We are interested in

$$m(\boldsymbol{Z}|H_p) = \prod_{q\in\{c,r\}} \int\int \prod_{i=1}^{n_q} f(\boldsymbol{z}_{qi}|\boldsymbol{\theta}, \boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta} f(\boldsymbol{U}|\boldsymbol{\Omega}, \nu) d\boldsymbol{U}$$

which can also be estimated using

$$m(\boldsymbol{Z}|H_d) = m(\boldsymbol{Z}_c)m(\boldsymbol{Z}_r) = \frac{f(\boldsymbol{Z}_c|\boldsymbol{\Psi})\pi(\boldsymbol{\Psi})}{\pi(\boldsymbol{\Psi}|\boldsymbol{Z}_c)} \frac{f(\boldsymbol{Z}_r|\boldsymbol{\Psi})\pi(\boldsymbol{\Psi})}{\pi(\boldsymbol{\Psi}|\boldsymbol{Z}_r)} \tag{C.2}$$

$$= \prod_{q\in\{c,r\}} \frac{f(\boldsymbol{Z}_q|\boldsymbol{\Psi})\pi(\boldsymbol{\Psi})}{\pi(\boldsymbol{\Psi}|\boldsymbol{Z}_q)} = \prod_{q\in\{c,r\}} m(\boldsymbol{Z}_q) \tag{C.3}$$

where $\boldsymbol{\Psi} = (\boldsymbol{\theta}, \boldsymbol{U})$. Denoting the maximum likelihood estimate as $\boldsymbol{\Psi}^*$, the estimate of the marginal density on logarithmic scale is

$$\log\left[\hat{m}(\boldsymbol{Z}_q)\right] = \log\left[f(\boldsymbol{Z}_q|\boldsymbol{\Psi}^*)\right] + \log\left[\pi(\boldsymbol{\Psi}^*)\right] - \log\left[\hat{\pi}(\boldsymbol{\Psi}^*|\boldsymbol{Z}_q)\right]$$

where $\hat{\pi}(\boldsymbol{\Psi}_q^*|\boldsymbol{Z}_q)$ can be estimated using Gibbs sampling algorithm described in Bozza et al. (2008).

The density function of the observation is given by

$$\begin{aligned} f(\boldsymbol{Z}_q|\boldsymbol{\Psi}) &= \prod_{i=1}^{n_q} f(\boldsymbol{z}_{qi}|\boldsymbol{\theta}, \boldsymbol{U}) \\ &= \prod_{i=1}^{n_q} (2\pi)^{-p/2} |\boldsymbol{U}|^{-1/2} \text{exp}\left\{-\frac{1}{2}(\boldsymbol{z}_{qi} - \boldsymbol{\theta})^T \boldsymbol{U}^{-1} (\boldsymbol{z}_{qi} - \boldsymbol{\theta})\right\}. \end{aligned}$$

The prior density for $\boldsymbol{\Psi}$ is given by

$$f(\boldsymbol{\Psi}|H_d) = (2\pi)^{-p/2}|\boldsymbol{C}|^{-1/2}\text{exp}\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\eta})^T \boldsymbol{C}^{-1}(\boldsymbol{\theta} - \boldsymbol{\eta})\} \frac{Const.|\boldsymbol{\Omega}|^{\frac{\nu}{2}}}{|\boldsymbol{U}|^{\frac{\nu+p+1}{2}}} \text{exp}\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}\boldsymbol{U}^{-1}\right)\}.$$

The complete conditional density of $\boldsymbol{\theta}$ is then

$$\begin{aligned} f(\boldsymbol{\theta}|\boldsymbol{Z}_q, \boldsymbol{U}) &= \frac{f(\boldsymbol{Z}_q|\boldsymbol{\theta}, \boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C})}{\int f(\boldsymbol{Z}_q|\boldsymbol{\theta}, \boldsymbol{U}) f(\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{C}) d\boldsymbol{\theta}} \\ &\propto \text{exp}\left\{-\frac{1}{2}\left[\sum_{i=1}^{n_q} (\boldsymbol{z}_{qi} - \boldsymbol{\theta})^T \boldsymbol{U}^{-1} (\boldsymbol{z}_{qi} - \boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\eta})^T \boldsymbol{C}^{-1} (\boldsymbol{\theta} - \boldsymbol{\eta})\right]\right\} \end{aligned}$$

which can be shown to be still of type normal with parameters $(\boldsymbol{\eta}^*, \boldsymbol{C}^*)$, where

$$\begin{aligned} \boldsymbol{C}^* &= (\sum_{i=1}^{n_q} \boldsymbol{U}^{-1} + \boldsymbol{C}^{-1})^{-1} \\ \boldsymbol{\eta}^* &= \boldsymbol{C}^*(\boldsymbol{C}^{-1}\boldsymbol{\eta} + \sum_{i=1}^{n_q} \boldsymbol{U}^{-1}\boldsymbol{z}_{qi}). \end{aligned}$$

And the complete conditional density of $\boldsymbol{U}$ would be

$$f(\boldsymbol{U}|\boldsymbol{Z}_q, \boldsymbol{\theta}) \propto |\boldsymbol{U}|^{-n_q/2} |\boldsymbol{U}|^{-(\nu+p+1)/2} \text{exp}\left\{-\frac{1}{2}\left[\sum_{i=1}^{n_q} (\boldsymbol{z}_{qi} - \boldsymbol{\theta})^T \boldsymbol{U}^{-1} (\boldsymbol{z}_{qi} - \boldsymbol{\theta}) + \text{tr}\left(\boldsymbol{\Omega}\boldsymbol{U}^{-1}\right)\right]\right\}$$

$$\propto |\boldsymbol{U}|^{-(n_q+\nu+p+1)/2}\text{exp}\left\{-\frac{1}{2}\left[\text{tr}\left(\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta})(\boldsymbol{z}_{qi}-\boldsymbol{\theta})^T\boldsymbol{U}^{-1}\right)+\text{tr}(\boldsymbol{\Omega}\boldsymbol{U}^{-1})\right]\right\}$$
$$\propto |\boldsymbol{U}|^{-(n_q+\nu+p+1)/2}\text{exp}\left\{-\frac{1}{2}\left[\text{tr}\left(\left(\boldsymbol{\Omega}+\sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta})(\boldsymbol{z}_{qi}-\boldsymbol{\theta})^T\right)\boldsymbol{U}^{-1}\right)\right]\right\}$$

which can be shown to be still of type inverse Wishart with parameters $(\boldsymbol{\Omega}^*, \nu^*)$, where

$$\boldsymbol{\Omega}^* = \boldsymbol{\Omega} + \sum_{i=1}^{n_q}(\boldsymbol{z}_{qi}-\boldsymbol{\theta})(\boldsymbol{z}_{qi}-\boldsymbol{\theta})^T$$
$$\nu^* = \nu + n_q.$$

The algorithm is then

1. Estimate $\hat{\boldsymbol{\eta}}$, $\hat{\boldsymbol{C}}$, and $\hat{\boldsymbol{\Omega}}$ from background
2. Sample $\boldsymbol{\theta}^g|\boldsymbol{U}^g, \boldsymbol{Z}_q \sim N_B(\boldsymbol{\eta}^*, \boldsymbol{C}^*)$ and update $\boldsymbol{\theta}^*$ to equal to $\boldsymbol{\theta}^g$ if $f(\boldsymbol{Z}_q|\boldsymbol{\theta}^g, \boldsymbol{U}^*)\pi(\boldsymbol{\theta}^g, \boldsymbol{U}^*) = max_{\boldsymbol{\theta}^g} f(\boldsymbol{Z}_q|\boldsymbol{\theta}^g, \boldsymbol{U}^*)\pi(\boldsymbol{\theta}^g, \boldsymbol{U}^*)$
3. Sample $\boldsymbol{U}^g|\boldsymbol{\theta}^g, \boldsymbol{Z}_q \sim \mathcal{IW}(\boldsymbol{\Omega}^*, \nu^*)$ and update $\boldsymbol{U}^*$ to equal to $\boldsymbol{U}^g$ if $f(\boldsymbol{Z}_q|\boldsymbol{\theta}^*, \boldsymbol{U}^g)\pi(\boldsymbol{\theta}^*, \boldsymbol{U}^g) = max_{\boldsymbol{U}^g} f(\boldsymbol{Z}_q|\boldsymbol{\theta}^*, \boldsymbol{U}^g)\pi(\boldsymbol{\theta}^*, \boldsymbol{U}^g)$
4. Compute
$$\hat{\pi}(\boldsymbol{U}^*|\boldsymbol{Z}_q) = \sum_{g=1}^{G}\frac{\pi(\boldsymbol{U}^*|\boldsymbol{Z}_q, \boldsymbol{\theta}^g)}{G}$$
5. Posterior is then given by $\hat{\pi}(\boldsymbol{\Psi}^*|\boldsymbol{Z}_q) = \pi(\boldsymbol{\theta}^*|\boldsymbol{U}^*, \boldsymbol{Z}_q)\hat{\pi}(\boldsymbol{U}^*|\boldsymbol{Z}_q)$.

The marginal likelihood on logarithmic scale ($\log \hat{m}(\boldsymbol{Z}_q|H_d)$) can then be estimated using equation (C.2).

C.5.3 Likelihood ratio

Putting the numerator and denominator together we get

$$\log(LR) = \log\left[\hat{m}(\boldsymbol{Z}|H_p)\right] - \log\left[\hat{m}(\boldsymbol{Z}_c|H_d)\right] - \log\left[\hat{m}(\boldsymbol{Z}_r|H_d)\right].$$

C.5.4 Estimation of hyperparameters using relevant population

$\boldsymbol{z}_{ki} \sim N_B(\boldsymbol{\theta}_k, \boldsymbol{U}_k), \quad k = 1, \ldots, K, \quad i = 1, \ldots, n_k$

$\boldsymbol{\theta}_k \sim N_B(\boldsymbol{\eta}, \boldsymbol{C}), \quad k = 1, \ldots, K$

$\boldsymbol{U}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Omega}, \nu).$

$$\hat{\boldsymbol{\eta}} = \frac{1}{K} \sum_k \hat{\boldsymbol{\theta}}_k$$

$$\hat{\boldsymbol{\theta}}_k = \frac{1}{n_k} \sum_i \boldsymbol{z}_{ki}$$

$$\hat{\boldsymbol{C}} = \frac{1}{K-1} \sum_k \hat{Var}(\boldsymbol{\theta}_k) = \frac{1}{K-1} \sum_k \left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\eta}}\right) \left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\eta}}\right)^T$$

$$\hat{\boldsymbol{\Omega}} = (\nu - B - 1) \sum_k \hat{\boldsymbol{U}}_k / K.$$

C.5.5 Simulation

Under this model, datasets will be generated by first simulate group means $\boldsymbol{\theta}_k \sim N(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{C}})$ then within-group covariance $\boldsymbol{U} \sim \mathcal{W}^{-1}(\hat{\boldsymbol{\Omega}}, \hat{\nu})$ and $\boldsymbol{z}_{ki} \sim N(\boldsymbol{\theta}_k, \hat{\boldsymbol{U}})$ for $1 \leq k \leq K$ and $1 \leq i \leq n_k$. To compare with original data we will reconstruct $\hat{\boldsymbol{y}}_{ki}$ as $\boldsymbol{\Phi}\boldsymbol{z}_{ki}$. Descriptions of this model can be found in Section 3.3.2.

Appendix D

Parameter and variance estimation

D.1 Generalised inverse and pseudo-determinant

When evaluating the probability density function of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, it often requires inverting $\boldsymbol{\Sigma}$ and raising its determinant to power of halves. This should not result in any problems provided $\boldsymbol{\Sigma}$ is a positive definite covariance matrix. However, when a covariance matrix $\boldsymbol{\Sigma}^*$ is estimated from data, the estimated matrix can be negative definite.

The problem arises when using multivariate analysis of variance to estimate between-group covariance matrix in a random-effects model with nested covariance structure. The estimation requires taking the difference between between-group and within-group mean squares and can produce an estimate that is negative definite. This can be an indication that the real between-group covariance is zero or the model is wrong since the estimate is unbiased (Searle, 1992). Solutions have been proposed to resolve this problem in Amemiya (1985) but it only guarantees nonnegative definite, i.e., it only concerns point estimate of $\boldsymbol{\Sigma}^*$ which can be singular. Therefore, we will use generalised inverse and pseudo-determinants when these problems arise.

A matrix $\boldsymbol{M}$ has to be non-singular and square to be invertible. However, there are cases where the inverse is needed when $\boldsymbol{M}$ is rectangular or not of full rank. For these matrices, the Moore–Penrose inverse can be obtained, i.e., for any m-by-n matrix $\boldsymbol{M}$, an n-by-m matrix $\boldsymbol{M}^*$ can be found so that $\boldsymbol{M}\boldsymbol{M}^*\boldsymbol{M} = \boldsymbol{M}$ and $\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}^* = \boldsymbol{M}^*$.

The determinant of a positive semi-definite matrix is non-negative since the de-

terminant of a matrix equals to the product of its eigenvalues and they should all be positive. However, when Σ is estimated using data, there might be eigenvalues that are negative. In such cases, pseudo-determinants calculated as the product of positive eigenvalues, is used.

Bibliography

Adam, C. D. (2008). In situ luminescence spectroscopy with multivariate analysis for the discrimination of black ballpoint pen ink-lines on paper. *Forensic Science International 182*(1-3), 27–34.

Adam, C. D., S. L. Sherratt, and V. L. Zholobenko (2008). Classification and individualisation of black ballpoint pen inks using principal component analysis of uv-vis absorption spectra. *Forensic Science International 174*(1), 16–25.

Aitken, C. G. G., Y.-T. Chang, P. Buzzini, G. Zadora, and G. Massonnet (2019). The evaluation of evidence for microspectrophotometry data using functional data analysis. *Forensic Science International 305*, 110007.

Aitken, C. G. G. and D. Lucy (2004, February). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics 53*(1), 109–122.

Aitken, C. G. G. and F. Taroni (2005). *Statistics and the Evaluation of Evidence for Forensic Scientists: Second Edition*. Wiley.

Aitken, C. G. G., G. Zadora, and D. Lucy (2007). A two-level model for evidence evaluation. *Journal of Forensic Sciences 52*(2), 412–419.

Amemiya, Y. (1985). What should be done when an estimated between-group covariance matrix is not nonnegative definite? *The American Statistician 39*(2), 112–117.

Banas, K., A. Banas, H. O. Moser, M. Bahou, W. Li, P. Yang, M. Cholewa, and S. K. Lim (2010). Multivariate analysis techniques in the forensics investigation of the postblast residues by means of fourier transform-infrared spectroscopy. *Analytical Chemistry 82*(7), 3038–3044.

Bojko, K., C. Roux, and B. J. Reedy (2008). An examination of the sequence of intersecting lines using attenuated total reflectance fourier transform infrared spectral imaging*. *Journal of Forensic Sciences 53*(6), 1458–1467.

Bozdogan, H. (1987, Sep). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika 52*(3), 345–370.

Bozza, S., F. Taroni, R. Marquis, and M. Schmittbuhl (2008, June). Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *57*(3), 329–341.

Braz, A., M. López-López, and C. García-Ruiz (2013). Raman spectroscopy for forensic analysis of inks in questioned documents. *Forensic Science International 232*(1-3), 206–212.

Burfield, R., C. Neumann, and C. P. Saunders (2015, December). Review and application of functional data analysis to chemical data—the example of the comparison, classification, and database search of forensic ink chromatograms. *Chemometrics and Intelligent Laboratory Systems 149*, 97–106.

Buzzini, P. and G. Massonnet (2015). The analysis of colored acrylic, cotton, and wool textile fibers using micro-raman spectroscopy. part 2: Comparison with the traditional methods of fiber examination. *Journal of Forensic Sciences 60*(3), 712–720.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research 1*(2), 245–276.

Chib, S. (1995, December). Marginal likelihood from the gibbs output. *90*(432), 1313–1321.

Cody, R. B., J. A. Laramée, and H. D. Durst (2005). Versatile new ion source for the analysis of materials in open air under ambient conditions. *Analytical Chemistry 77*(8), 2297–2302.

Cover, T. and J. Thomas (2005). *Elements of Information Theory*. John Wiley and Sons.

de Souza Lins Borba, F., R. Saldanha Honorato, and A. de Juan (2015). Use of raman spectroscopy and chemometrics to distinguish blue ballpoint pen inks. *Forensic Science International 249*, 73–82.

De Wael, K., K. Van Dijck, and F. Gason (2015). Discrimination of reactively-dyed cotton fibres with thin layer chromatography and uv microspectrophotometry. *Science & Justice 55*(6), 422–430.

Denman, J. A., W. M. Skinner, K. P. Kirkbride, and I. M. Kempson (2010). Organic and inorganic discrimination of ballpoint pen inks by tof-sims and multivariate statistics. *Applied Surface Science 256*(7), 2155–2163.

Frank, R. S. and S. P. Sobol (1990). Fibres and their examination in forensic science. In A. Maehly and R. Williams (Eds.), *Forensic Science Progress*, Berlin, Heidelberg, pp. 41–125. Springer Berlin Heidelberg.

Gonzalez-Rodriguez, J., A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech Language 20*(2-3), 331–355.

Hepler, A. B., C. P. Saunders, L. J. Davis, and J. Buscaglia (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Science International 219*(1-3), 129–140.

Kher, A., M. Mulholland, E. Green, and B. Reedy (2006). Forensic classification of ballpoint pen inks using high performance liquid chromatography and infrared spectroscopy with principal components analysis and linear discriminant analysis. *Vibrational Spectroscopy 40*(2), 270–277.

Konishi, S. and G. Kitagawa (1996). Generalised information criteria in model selection. *Biometrika 83*(4), 875–890.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika 64*(2), 207–213.

Marquis, R., F. Taroni, S. Bozza, and M. Schmittbuhl (2006). Quantitative characterization of morphological polymorphism of handwritten characters loops. *Forensic Science International 164*(2-3), 211–220.

Martyna, A., D. Lucy, G. Zadora, B. M. Trzcinska, D. Ramos, and A. Parczewski (2013). The evidential value of microspectrophotometry measurements made for pen inks. *Anal. Methods 5*, 6788–6795.

Martyna, A., A. Michalska, and G. Zadora (2015). Interpretation of ftir spectra of polymers and raman spectra of car paints by means of likelihood ratio approach supported by wavelet transform for reducing data dimensionality. *Analytical and Bioanalytical Chemistry 407*(12), 3357–3376.

Martyna, A., G. Zadora, T. Neocleous, A. Michalska, and N. Dean (2016). Hybrid approach combining chemometrics and likelihood ratio framework for reporting the evidential value of spectra. *Analytica Chimica Acta 931*, 34–46.

Massonnet, G., P. Buzzini, G. Jochem, M. Staube, T. Coyle, C. Roux, J. Thomas, H. Leijenhorst, Z. van Zanten, R. Griffin, K. Wiggins, and S. Chabli (2003). Evaluation of raman spectroscopy for the analysis of coloured fibres: A collaborative study. *Forensic Science International 136*, 124–124.

Neumann, C., R. Ramotowski, and T. Genessay (2011). Forensic examination of ink by high-performance thin layer chromatography - the united states secret service digital ink library. *Journal of Chromatography A 1218*(19), 2793–2811.

Pfefferli, P. W. (1983). Application of microspectrophotometry in document examination. *Forensic Science International 23*(2), 129–136.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis* (Second edition. ed.). Springer series in Statistics. New York: Springer.

Ray, P. (2016). The evaluation of fibre evidence in the investigation of serious crime.

Roux, C., M. Novotny, I. Evans, and C. Lennard (1999). A study to investigate the evidential value of blue and black ballpoint pen inks in australia. *Forensic Science International 101*(3), 167–176.

Searle, S. R. S. R. (1992). *Variance components*. Wiley series in probability and mathematical statistics. Applied probability and statistics. New York: Wiley.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics 8*(1), 147–164.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika 68*(1), 45–54.

Smalldon, K. and A. Moffat (1973). The calculation of discriminating power for a series of correlated attributes. *Journal of the Forensic Science Society 13*(4), 291–295.

Takáts, Z., J. M. Wiseman, B. Gologan, and R. G. Cooks (2004). Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science (New York, N.Y.) 306*(5695).

Thanasoulias, N. C., N. A. Parisis, and N. P. Evmiridis (2003). Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their vis spectra. *Forensic Science International 138*(1), 75–84.

Thanasoulias, N. C., E. T. Piliouris, M.-S. E. Kotti, and N. P. Evmiridis (2002). Application of multivariate chemometrics in forensic soil discrimination based on the uv-vis spectrum of the acid fraction of humus. *Forensic Science International 130*(2), 73–82.

Was-Gubala, J. and R. Starczak (2015). Uv-vis microspectrophotometry as a method of differentiation between cotton fibre evidence coloured with reactive dyes. *Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy 142*, 118–125.

White, P. (2004). *Crime scene to court : the essentials of forensic science* (Second edition.. ed.). Cambridge, UK: Royal Society of Chemistry.

Zadora, G., A. Martyna, D. Ramos, and C. G. G. Aitken (2013). *Statistical Analysis in Forensic Science : Evidential Values of Multivariate Physicochemical Data.* Wiley.

Zięba-Palus, J. and M. Kunicki (2006). Application of the micro-ftir spectroscopy, raman spectroscopy and xrf method examination of inks. *Forensic Science International 158*(2), 164–172.