

**Flexible Mixed-Effect Modeling of Functional Data,
with Applications to Process Monitoring**

by

Sofia A. Mosesova

A thesis
presented to the University of Waterloo
in fulfilment of the thesis requirement
for the degree of


Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2007

©Sofia A. Mosesova, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

A handwritten signature in black ink, appearing to read 'Sofia Mosesova', with a stylized, cursive script.

Sofia A. Mosesova

Abstract

High levels of automation in manufacturing industries are leading to data sets of increasing size and dimension. The challenge facing statisticians and field professionals is to develop methodology to help meet this demand.

Functional data is one example of high-dimensional data characterized by observations recorded as a function of some continuous measure, such as time. An application considered in this thesis comes from the automotive industry. It involves a production process in which valve seats are force-fitted by a ram into cylinder heads of automobile engines. For each insertion, the force exerted by the ram is automatically recorded every fraction of a second for about two and a half seconds, generating a force profile. We can think of these profiles as individual functions of time summarized into collections of curves.

The focus of this thesis is the analysis of functional process data such as the valve seat insertion example. A number of techniques are set forth. In the first part, two ways to model a single curve are considered: a b-spline fit via linear regression, and a nonlinear model based on differential equations. Each of these approaches is incorporated into a mixed effects model for multiple curves, and multivariate process monitoring techniques are applied to the predicted random effects in order to identify anomalous curves. In the second part, a Bayesian hierarchical model is used to cluster low-dimensional summaries of the curves into meaningful groups. The belief is that the clusters correspond to distinct types of processes (e.g. various types of “good” or “faulty” assembly). New observations can be assigned to one of these by calculating the probabilities of belonging to each cluster. Mahalanobis distances are used to identify new observations not belonging to any of the existing clusters. Synthetic and real data are used to validate the results.

Acknowledgements

I would like to thank Professor Hugh Chipman for introducing me to the topic as well as for his valuable input, patience and generous funding. My sincere gratitude also extends to Professors Jock MacKay, Stefan Steiner and Jim Ramsay for contributing their time and ideas to the project, and last but not least to my caring family and friends for their tremendous support, encouragement and inspiration.

Dedication

I dedicate the work in this thesis to my high school councilor, Mrs. Olson, who advised me that mathematics is a difficult subject for girls, and that I would have a much easier time taking home economics.

Contents

1	Introduction	1
1.1	Overview of Functional Data Analysis	1
1.2	Valve Seat Insertion Example	3
1.3	Goals and Techniques	7
1.4	Past Work	9
1.5	Software	12
2	Preprocessing and Exploratory Analysis	13
2.1	Curve Registration	13
2.1.1	Literature Review	14
2.1.2	Algorithm	14
2.2	Other Preliminaries	19
2.3	Data Visualization	21
3	Modeling a Single Curve	26
3.1	Smoothing Splines	27
3.2	Nonlinear Modeling	31
3.2.1	Motivation	34
3.2.2	Second-order DE Model	35
3.3	Discussion	37
4	Modeling Collections of Curves	38
4.1	Linear Mixed Effects	40
4.2	Nonlinear Mixed Effects	42
4.3	Parameter Estimation	43
4.3.1	Linear Model	43
4.3.2	Nonlinear Model	47
4.4	Distributional Properties of the Predicted Random Effects	48

5	Profile Monitoring	51
5.1	Chart for Individual Curves	52
5.2	Chart for Subgrouped Curves	54
5.3	Example	55
5.3.1	Model Specification	56
5.3.2	Phase I Analysis	56
5.3.3	Phase II Analysis	57
5.4	Discussion	61
5.4.1	Choosing Random Effects	61
5.4.2	Choice of Phase I Data	62
5.4.3	Sequential Charts	62
5.4.4	Charts Using Both Fixed and Random Effects	64
5.4.5	Charts for Multiple Valves	64
6	Model-Based Clustering	65
6.1	Background	66
6.2	Dimension Reduction	67
6.3	A Mixture Model for Multivariate Data	68
6.3.1	Covariance Structure	70
6.3.2	Model-Selection	72
6.4	Extension to Functional Data	74
6.5	Example	75
7	A Bayesian Approach to Clustering	83
7.1	Model	84
7.2	Priors	88
7.3	Markov chain Monte Carlo	94
7.3.1	Review	94
7.3.2	Gibbs Sampling Algorithm	98
7.3.3	Birth-Death MCMC	101
8	Bayesian Random Effects Clustering	107
8.1	Model	107
8.2	Priors	109
8.3	Estimation	111
8.4	Applications	113
8.4.1	Mode Detection	113
8.4.2	Bayesian Prediction	114

8.4.3	Profile Monitoring	115
8.5	Examples	117
8.5.1	Synthetic Data	117
8.5.2	Force Exertion Data	123
8.6	Discussion	131
9	Summary and Conclusions	135
A	Commonly Used Distributions	138
B	Full Conditionals	140
B.1	Bayesian Clustering Model	140
B.2	Bayesian Random Effects Clustering	144

List of Tables

6.1	MBC choices of covariance structure.	71
6.2	Tabulation of cluster membership against valve labels.	78
8.1	Synthetic data: true vs. predicted cluster labels.	120
8.2	Synthetic data: estimated and true parameter values.	120

List of Figures

1.1	A sample of five observed force curves.	2
1.2	Sample cylinder head.	4
1.3	Cartoon of a cylinder head.	5
1.4	Examples of observed valve-insertion data.	6
2.1	Observed force and first difference of force plotted against time.	16
2.2	Observed distance and first difference of distance plotted against time. . .	17
2.3	Curve registration algorithm by example.	20
2.4	Number of parts produced plotted by day.	22
2.5	Using PCA to explore variability in the force exertion data.	24
2.6	Interpretation of the first two principal components.	25
3.1	An example of b-spline smoothing.	29
3.2	Average R^2 values for a range of smoother d.f.	30
3.3	Second order DE fit to force exertion data.	32
4.1	Hierarchical structure of mixed-effects models for curve data.	39
4.2	Force curve fitted using a linear mixed-effects model.	41
5.1	Phase I analysis of force exertion data.	58
5.2	Phase II analysis of force exertion data.	60
5.3	Phase II MEWMA chart under the b-spline model.	63
6.1	BIC criterion for model selection.	77
6.2	BIC plot after subtracting the valve effects.	80
6.3	MBC results for force exertion data.	81
7.1	DAG of the Bayesian curve-clustering model.	86
7.2	A toy example of BDMCMC.	103

8.1	DAG of the Bayesian random effects clustering model.	110
8.2	Synthetic force exertion data by cluster label.	119
8.3	Synthetic data: inference for number of clusters.	121
8.4	Synthetic data: estimated vs. true parameters.	122
8.5	Trace plots of MCMC output for synthetic data.	124
8.6	Trace plots of MCMC output for February data.	125
8.7	Prediction of cluster membership probabilities for February data.	126
8.8	Curve means by cluster for February data.	127
8.9	Distance charts for force exertion data.	129
8.10	Prediction of cluster membership probabilities for January data.	130

Chapter 1

Introduction

1.1 Overview of Functional Data Analysis

Functional data analysis (FDA) is a branch of statistics involving the study of curve data. The field is continuously evolving, with the technological and computational advances of the past two decades offering new ways to collect and analyze such data.

Functional data are characterized by observations that are functions of some continuous measurement. Figure 1.1 provides an example. Here force exertions are plotted against time for a set of five valve seats inserted by a ram into the top of a cylinder head as part of assembling automobile engines. Other examples of functional data range from height and weight measurements taken at different ages for the same subject, to sound patterns, to weather conditions, to tumor size and brain activity measurements, all of which can be plotted as curves as they are recorded over time.

The focus of this thesis is on the analysis of functional data that arise from a process that generates curves over the span of weeks, months or years. Our goals are to monitor the process over time, and identify possible drifts and anomalous observations. That is, we wish to develop methodology for dealing with process curve data - ways to screen for changes in the process, and detect patterns and/or possible outliers.

In an effort to investigate some of these ideas we consider the above-mentioned application in automotive manufacturing. We focus on the insertion of eight valve seats into the cylinder heads of automobile engines. A number of aspects of this particular data set

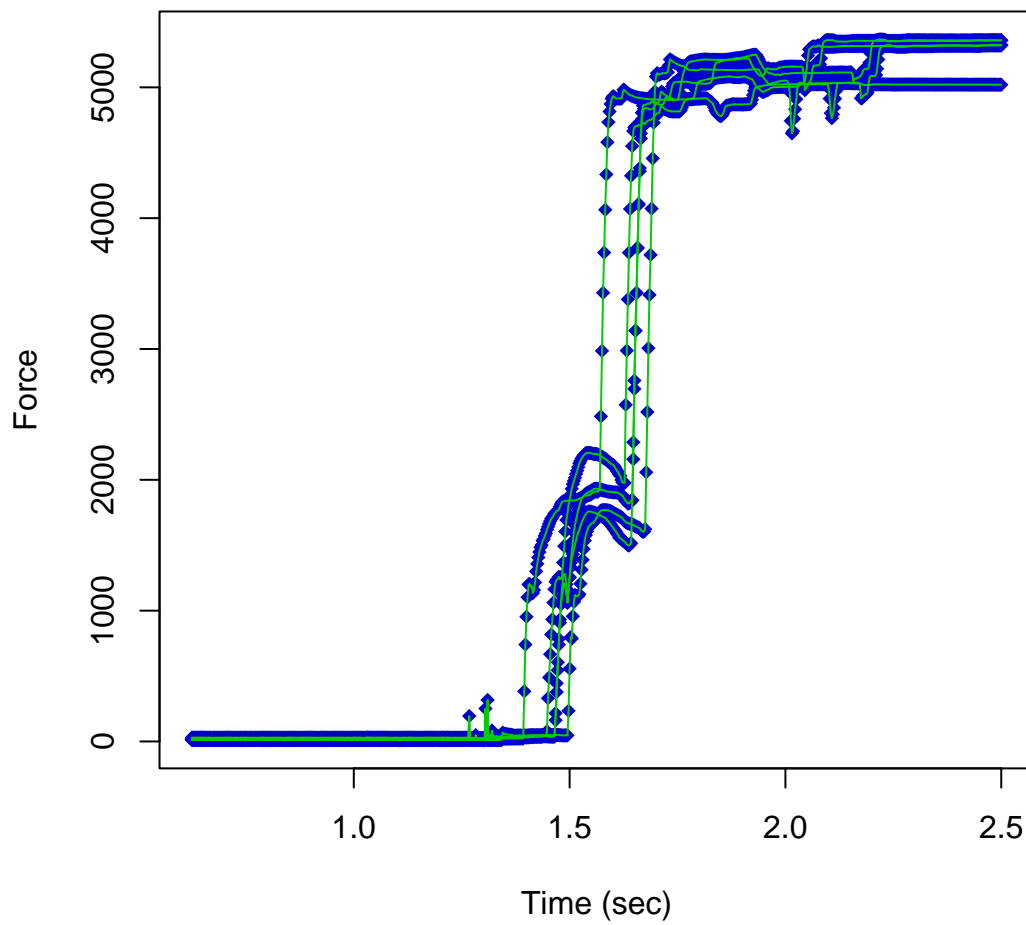


Figure 1.1: Five observed force curves for the valve seat insertion example. For each curve, observed force values are recorded by the machine at discrete time points. The smooth lines through these points help emphasize the fact that the curves can be thought of as functions of time.

make the analysis challenging. A key issue complicating the problem is the absence of any knowledge as to which of the insertions are faulty. That is, the observed functional response is the only information available about a particular insertion. Furthermore, the curves, observed on a fine time grid, are not aligned to ensure that the same pattern can be seen at the same time point. Thus some form of preprocessing of the raw data must take place before they can be analyzed. The total number of recorded observations, over 6,000, is also large spanning 41 days in two different months with eight different recordings for eight different valve seats. This is only a small fraction of the data generated over several years worth of production, a typical situation for highly automated processes such as this one.

1.2 Valve Seat Insertion Example

As the valve seat insertion data possess most of the qualities we wished to study, this is the primary example used to illustrate most of the ideas in this thesis. To gain a better understanding of these data, we briefly outline the basic workings of an eight-cylinder automobile engine.

The particular engines considered in our data set are V8, which simply means that they are V-shaped with four cylinders on each side. Fixed on top of the engine are the two cylinder heads. The eight valves, which are located on top of each cylinder head, are primarily made of steel. The two types of valves (four of each kind) are intake valves that allow fuel to enter the engine chamber and exhaust valves that release combustion gases from it. Valve seats, also made of steel, are inserted into the relatively soft aluminum head to protect it from the valves. Since the valves continuously open and close as fuel and gases travel in and out of the engine, it is absolutely crucial that the valve seats are securely attached to the cylinder heads so as to maintain a tight seal. If the valve seat is not securely locked into place during the engine assembly process, the part can eventually fall into the engine or become loose enough to cause leakage around the seat. Both are catastrophic failures.

Figures 1.2 and 1.3 help visualize the process of valve seat insertion. The four valve seats for each type of valve (intake and exhaust) are inserted simultaneously. For each



Figure 1.2: Cylinder head of an eight-valve engine as viewed from the top (right) and the bottom (left).

insertion, a machine automatically records the distance traveled by the ram as well as its exerted force every 400^{th} of a second for about 2.5 seconds.

In this example, data were collected for a total of 6,000 eight-cylinder engines over the span of 41 days throughout January and February 2000. We can think of the data as a series of 6,000 curves (or functions) of either force against time or distance against time. Figure 1.4 provides a closer look at one such observation (run #000740) recorded on February 21st. The data in the plot have been truncated at both ends of the insertion process, removing the flat and generally uninformative tails of the curves.

In Figure 1.4, the distance curve is much less informative than the force curve. From the force curve we can see that the point of impact between the ram and the valve seat must have occurred roughly around 1.49 seconds, at which time the exerted force drastically increased to adjust for resistance. In the distance curve, the same can be inferred from a small bump in the curve at the same time. The distance curve is predominantly a steadily increasing straight line. This is because the ram is designed to move at a constant velocity with the force exertion adjusted accordingly.

A possible alternative to analyzing the force-time curve would be to consider the first difference of force against time, however in that case our results become considerably more difficult to interpret.

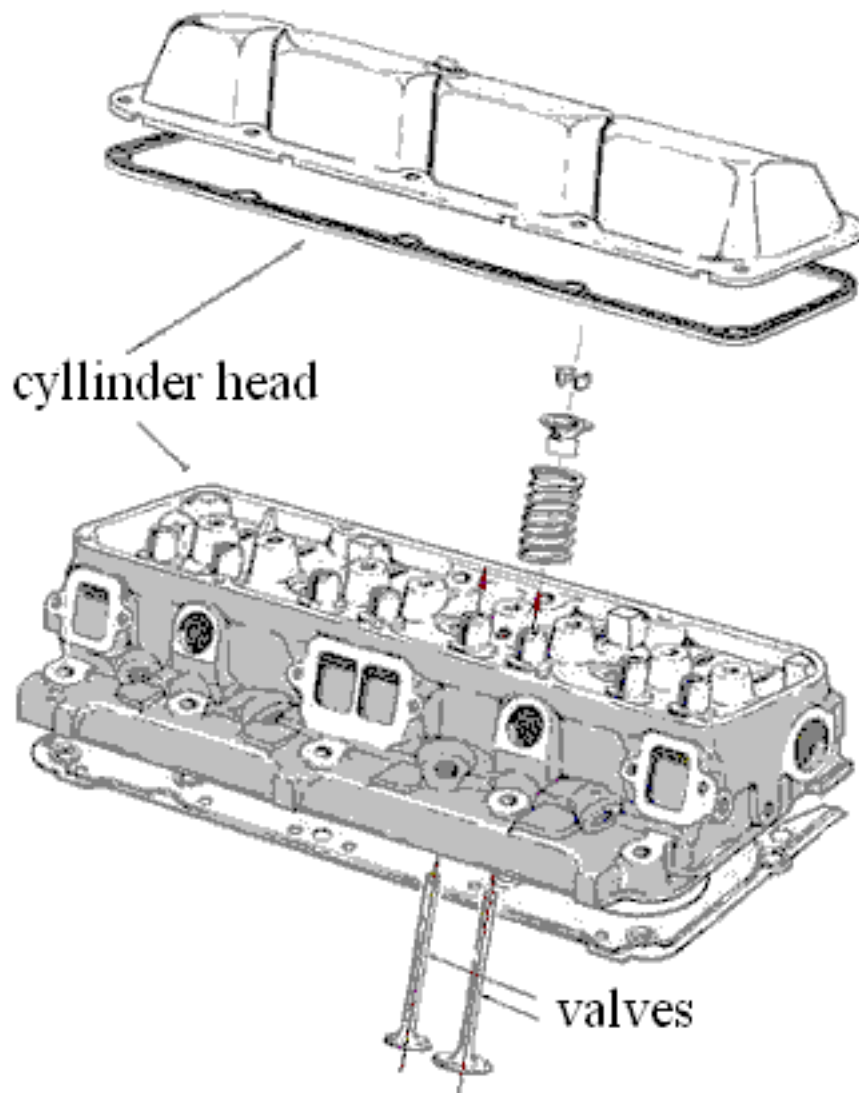


Figure 1.3: Cylinder head of an eight valve engine as viewed from the top.

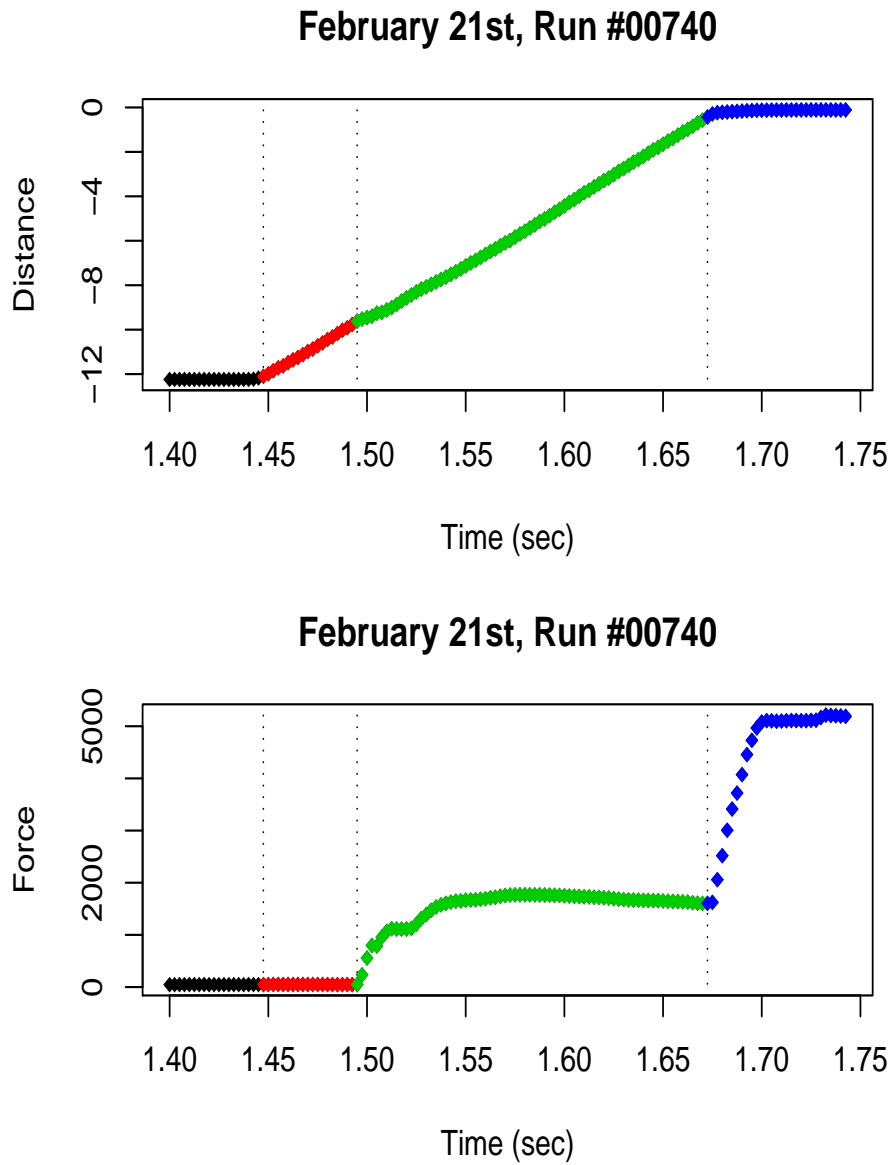


Figure 1.4: A recording from a single insertion in terms of distance against time (top) and force against time (bottom). The data were observed on February 21st (run #00740).

1.3 Goals and Techniques

In the valve seat insertion example, we seek ways to differentiate between curves that are “in-control” in a sense that the valve is inserted properly and all the possible “out-of-control” curves, where the insertion is somehow unusual. We wish to find anomalous features in the curves, which would indicate which valve seats have not been inserted properly. We also hope to screen for possible changes in the insertion process as production progresses over time. In general however, our goal is to develop statistical methodology for dealing with functional process data - ways of extracting low-dimensional summaries from it, monitoring the process, detecting patterns, changes and possible outliers, and incorporating information about potentially multiple sources of systematic effects (e.g., eight valve seats inserted on the same head).

This thesis develops a number of techniques to achieve these goals. These can be divided into three general categories listed below.

- **Preliminaries.**

A unique attribute of functional observations is the need to align them prior to the analysis. This preprocessing step is called *curve registration*, and it ensures that important parts of the curve (e.g., the start and end of the insertion process) occur at roughly the same time, making the curves comparable. In this thesis, we describe a curve registration algorithm for force exertion data. Other general issues pertaining to functional data, such as exploratory analysis, are also considered.

- **Profile monitoring.**

Profile monitoring is a functional extension of process monitoring, a statistical tool for detecting anomalous observations during an ongoing process. Conventional methods for monitoring multivariate data involve calculating Hotelling T^2 statistics as measures of discrepancy between the data and their expected value, scaled by the covariance. Thus, observations that correspond to significantly high values of this test statistic are flagged as unusual.

One functional alternative to the multivariate approach is to monitor low-dimensional summaries of the curves, such as estimated b-spline coefficients, obtained from smoothing the data. We extend this idea by monitoring predicted random effects from a

mixed effects model fitted to the functional data. Two different monitoring techniques are developed for linear and nonlinear mixed effects models.

The novelty in monitoring the random coefficients is that this enables us to detect small departures from the basic shape of the profiles, after controlling for the overall trend common to all of the data. By using considerably fewer random effects than time points on a curve, we are also able to monitor in a lower dimension than the original data, which increases the power of our monitoring procedure, assuming that the model is correct.

- **Curve clustering.**

A way to identify more than one possible change in a production process is to separate curves into groups based on differences in their observed values. Curve clustering is an approach used to explore this idea in this thesis. A new model is proposed, in which the data are fit using a sum of three components: fixed effects that control for any known differences amongst the curves (e.g., differences between the eight valves), random effects that take into account curve-specific departures from the fixed effects, and an error term that represents the amount of roughness/wiggliness in the observed curves. Within the context of the model, the random effects are clustered and a Bayesian approach is used to estimate all of the unknown parameters.

The advantages of clustering random coefficients parallel those of monitoring them, while also providing a flexible and interpretable representation of the random effect distribution. The Bayesian implementation also enables us to estimate the number of clusters by treating this quantity as one of the unknown parameters in the model, and to assess uncertainty with respect to this and other model parameters.

There are many different applications of curve clustering. In this thesis, we apply it to the monitoring problem in two ways. First, if clusters correspond to interesting changes in the production process, new observations can be classified as following one of these processes by calculating probabilities of belonging to corresponding clusters. Second, we develop a monitoring tool that defines outliers to be observations that do not belong to any of the previously identified clusters.

The last two topics make up the primary focus of our investigation, whereas the first one

is of secondary interest, but nonetheless important in assuring that subsequent analyses are reliable. The remainder of the thesis is organized as follows: after a brief literature review, we end this chapter with a short description of the software used to implement the new models and analyze the data. This is followed by a summary of preprocessing and functional data exploration techniques in Chapter 2. Modeling curves individually and collectively is addressed in Chapters 3 and 4. Profile monitoring techniques for detecting outlying curves using mixed effects models are developed in Chapter 5. In Chapters 6 and 7, curve clustering is reviewed, and a new model is proposed in Chapter 8. The thesis is concluded with a summary in Chapter 9.

1.4 Past Work

FDA is an exciting area of statistics which has received considerable attention in the past few years. The types of problems that arise in FDA depend heavily on particular aspects of the data being considered, and thus many of the papers published on the topic are application-specific. Some examples involving industrial process data that are functional include antenna manufacturing data (Jeong and Lu, 2004), density profile data for wooden boards (Walker and Wright, 2002), nano-machining in semi-conductor manufacturing (Ganesan and Das, 2002), tonnage stamping in the automotive industry (Jin and Shi, 2001) and many more. In this section we review several publications that have been useful in our research.

FDA Fundamentals

The best starting points for any type of FDA are two books by Ramsay & Silverman (1997 and 2001). The first book, entitled “Functional Data Analysis,” provides a detailed introduction to FDA, with an abundance of examples to ease the reader through the concepts. An underlying assumption is that each functional observation is intrinsically smooth, and in many cases smoothing techniques are used to recover the functional form of the data. With this in mind, the authors adopt such fundamental statistical concepts as principal components analysis, discriminant analysis, and linear modeling within the functional context. Curve registration - a preprocessing step of stretching and aligning the

curves in order to facilitate pointwise comparison between them - is another topic discussed in considerable detail in the book.

The second book, entitled “Applied Functional Data Analysis,” is exactly that - an applied complement to the first text, comprised entirely of case studies in FDA. Each example is carefully chosen to introduce a new concept. As in the first book, the authors rely heavily on the use of derivatives and differential equations to solve problems ranging from curve registration to modeling to discriminant analysis of functional data. Much of the material covered in the second chapter of this thesis is based on the ideas developed in these books.

Process Monitoring

The wide range of names for functional data in the process-monitoring community include “signals”, “waveform signals”, “signatures”, “profiles” and “curves”. We will use the last two terms interchangeably. Previous work in this area is well summarized by Woodall et. al. (2004), who observed that simple linear regression is the most common tool for summarizing curves. Kang & Albin (2000), Kim et. al. (2003), and Mahmoud & Woodall (2004) propose numerous ways to use multivariate T^2 -charts to monitor curves that are well summarized by a slope and an intercept.

With respect to process monitoring of nonlinear profiles, Jin & Shi (2001) describe an adaptive feature-extracting technique for data with limited or no prior “in-control” information. They use wavelet transforms to extract a reduced feature set of the data, then apply an iterative procedure to compare the Hotelling T^2 criteria for each new observation to the base set of “normal” observations. At each step, an observation that is deemed unusual is either classified into a known cluster of outliers or becomes a part of a new outlying cluster.

Walker & Wright (2002) use generalized additive models to compare collections of curves. They model the functional component of the data with smoothers, and the remaining variability is assumed to come from independent identically distributed (*iid*) normal errors. Unusual curves are flagged by conducting an appropriate F -test. Williams et. al. (2005) extend this idea by using a multivariate T^2 control chart to monitor the curves.

These approaches differ from ours in that we monitor the random coefficients obtained

from fitting a mixed effects model to the data. Abramovich & Angelini (2006), Guo (2002), Antoniadis & Sapatinas (2004), Morris & Carroll (2006) and Morris, Arroyo et. al. (2006) are just a few of the papers that discuss fitting random effects models to functional data; however this work is constrained to linear models only. The two papers by Morris and co-authors are closely related to our b-spline model, with the exception that wavelet basis functions are used in place of the b-spines. One disadvantage of wavelets for the purposes of process-monitoring is the fact that they require fitting many more coefficients than b-splines, consequently leading to a large number of random effects to be predicted and monitored.

Curve Clustering

Another approach to detecting outlying curves is to systematically separate the profiles into meaningful groups. Clustering is one such technique. In the case of force exertion, each cluster may correspond to a different type of insertion process, some of which may be “good” and others “bad”. Examining cluster means can provide insight into how the curves differ from one another.

A number of different clustering algorithms have been proposed for functional data. Ones that have a probabilistic context include James & Sugar (2003), Gafney & Smyth (2003), Chudova et. al. (2004), and Zhou & Wakefield (2005). An idea common to all of these papers is to cluster compact summaries of the curves rather than the observed data.

James & Sugar (2003) extend Banfield and Raftery’s (1993) model-based technique to cluster b-spline coefficients in a smoothing model that regresses observed data onto a basis of spline functions. Their method is tailored to sparsely observed functional data, with few values recorded for each curve. The use of smoothing splines has a dual advantage of preserving the functional form of the data and achieving dimension reduction (if the number of b-spline coefficients being used is fewer than the number of time points at which curves values are observed). Zhou & Wakefield (2005) implement the same regression model using a fully Bayesian framework, while Gafney & Smyth (2003) maximize the posterior (as opposed to the likelihood) to obtain point estimates of all parameters. Chudova et. al. (2004) apply Gafney & Smyth’s approach to clustering gene-expression curves.

A detailed overview of the b-spline clustering model that unifies the four papers is

presented in Chapters 6 and 7, as a prelude to the more general Bayesian random effects clustering (BREC) model developed in Chapter 8. This model differs from its predecessors in that it takes into account the common basic shape of the curves by including corresponding fixed effect terms in the clustering model.

1.5 Software

All of the techniques described in this thesis were implemented using R version 2.4.0, an open-source implementation of the S programming language (R Development Core Team, 2006).

Another resource for the computational exploration of functional data is a package written to analyze some of the examples in Ramsay & Silverman (2005). These functions are coded in R/S-PLUS and MATLAB and are available on Professor Jim Ramsay's website¹. A potential alternative is an FDA module developed by MathSoftTM, a commercial distributor of S-PLUS. The product implements some of the exploratory FDA tools developed by Ramsay & Silverman (2005) in an efficient and easy to use package.

¹<ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/>

Chapter 2

Preprocessing and Exploratory Analysis

Data preprocessing is the practice of transforming raw data into a suitable format for analysis. Although this is not the primary focus of our investigation, being able to preprocess data properly is a crucial step in ensuring computational ease as well as accuracy and reliability of our modeling procedures. An important preprocessing tool for functional data is curve registration, which is designed to reduce phase variation in the curves. An algorithm for registering valve-seat insertion data is presented in §2.1. Other preprocessing steps taken prior to modeling the data are summarized in §2.2.

An equally valuable preliminary step is to explore the underlying structure in the data before doing any modeling. Section 2.3 is dedicated to the use of principal component analysis to identify (and visualize) the extent of variation in the overall shape of the curves. For functional data, in particular, any knowledge gained about common sources of variability in the curves can be utilized in modeling them.

2.1 Curve Registration

A distinguishing characteristic of functional observations is the need to align the curves prior to analyzing the data in order to facilitate a comparison of common features between them. This is referred to as curve registration, a process that involves registering landmark

features in the curves and aligning them to one another across all other curves. In this section we describe an algorithm developed for registering the valve seat insertion data.

2.1.1 Literature Review

Ramsay & Silverman (2005, Chapter 7) identify two approaches to registering functional data - shift registration and landmark (or feature) registration. A unifying idea behind the two methods is to reduce phase variability, defined as the variation in the timings of prominent features of the curves. This is achieved by replacing the existing time axis \mathbf{t} for each curve i with some function $h_i(\mathbf{t})$.

The first approach involves shifting the time scale for each curve individually to some value that minimizes phase variability. This is generally achieved by using an iterative algorithm (e.g., Newton-Raphson) to shift the curves in a way that minimizes a measure of discrepancy, such as the sum of squared differences, between the observed curves and some target curve, such as their mean. Methods of this sort are referred to as shift methods, because the curves are shifted (i.e. $h_i(\mathbf{t}) = \mathbf{t} + c_i$ for some constants c_i) in order to minimize the criterion.

Another option is to identify specific features that are unique to all of the curves in the dataset, and shift, scale or warp each of the curves along the time axes so that the landmarks occur at the same time points. The features are called registration points, which are quantities such as curve maxima or curve minima. The approach is called landmark registration, because the curves are shifted, scaled ($h_i(\mathbf{t}) = c_i \cdot \mathbf{t}$) or warped ($h_i(\mathbf{t})$ are some nonlinear functions of \mathbf{t}) to ensure that all registration points are aligned.

The algorithm developed in this section is an adaptation of the landmark and shift registration techniques.

2.1.2 Algorithm

A registration algorithm developed specifically for the valve seat insertion data consists of four steps:

STEP 1: Identify landmark features as initial values for registration points.

For each individual curve, pick two registration points corresponding to the beginning and the end of the insertion process. We use first differences of force and distance in order to identify these landmark features. Although this was not the only possibility considered, the use of differences has proven to work quite effectively for the valve seat insertion data. Descriptions of the two landmarks and how these were identified are provided below.

- *Start of the process.*

If we define the beginning of the process as the initial point of impact between the ram and the valve seat, then this point can be determined as the time when the force differences are first at their local maximum. This is because a spike in the force is needed to push the seat forward upon impact. The resulting point is marked by a dashed dark blue line in Figure 2.1. By comparing the top and the bottom panels of this plot, we see that the start point in the force curve at the top corresponds to a local maximum in the force differences curve at the bottom. As an added precaution, the starting point is chosen 20 units to the left of the initial estimate of the starting value, in order to include some observed values just before the ram hits the valve seat.

- *End of the process.*

The second landmark feature represents the completion of the insertion process, which we assume to be the time at which the ram has stopped moving. Using first differences of distance to estimate speed, the end of the process is taken to be the last time point at which the speed is above zero. Due to the roughness of the first differences, it is sometimes difficult to know when the speed is zero. As an alternative, we used the average of all speed values over the entire span of the insertion process as a measure of close to zero. The average speed is marked by a horizontal line in Figure 2.2. A vertical dashed line identifies the last time when speed is greater than the average line, which estimates the end of the process. As was the case with the starting point, the final end point is chosen 20 time units to the right of its original location to allow some room for error. The final end point is represented by a solid light blue vertical line in Figures 2.2 and 2.1.

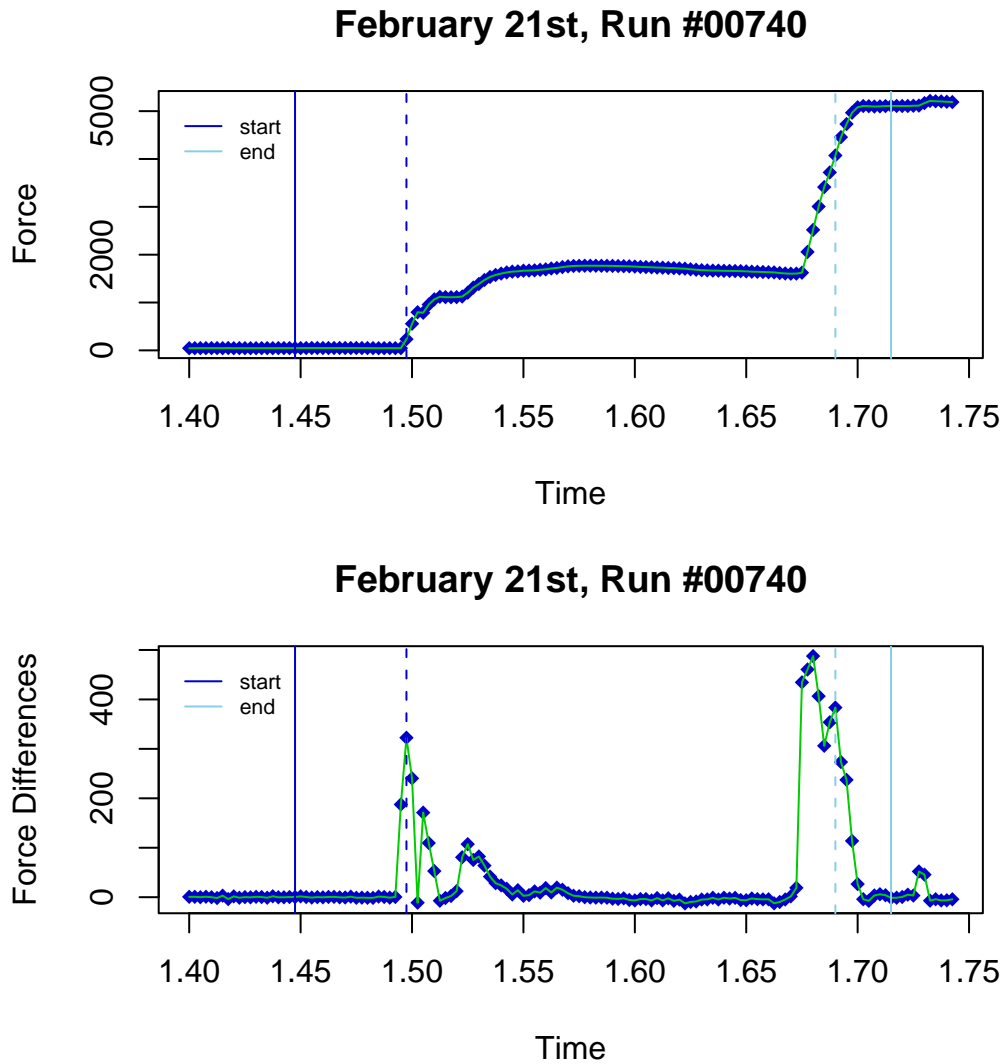


Figure 2.1: Plots of force against time (top) and first difference of force against time (bottom) for a single observation. The dashed line in each curve correspond to the points marking the beginning and end of the insertion process as summarized in Step 1 of the registration algorithm. The solid lines mark the final registration points after the minimization procedure described in Step 2.

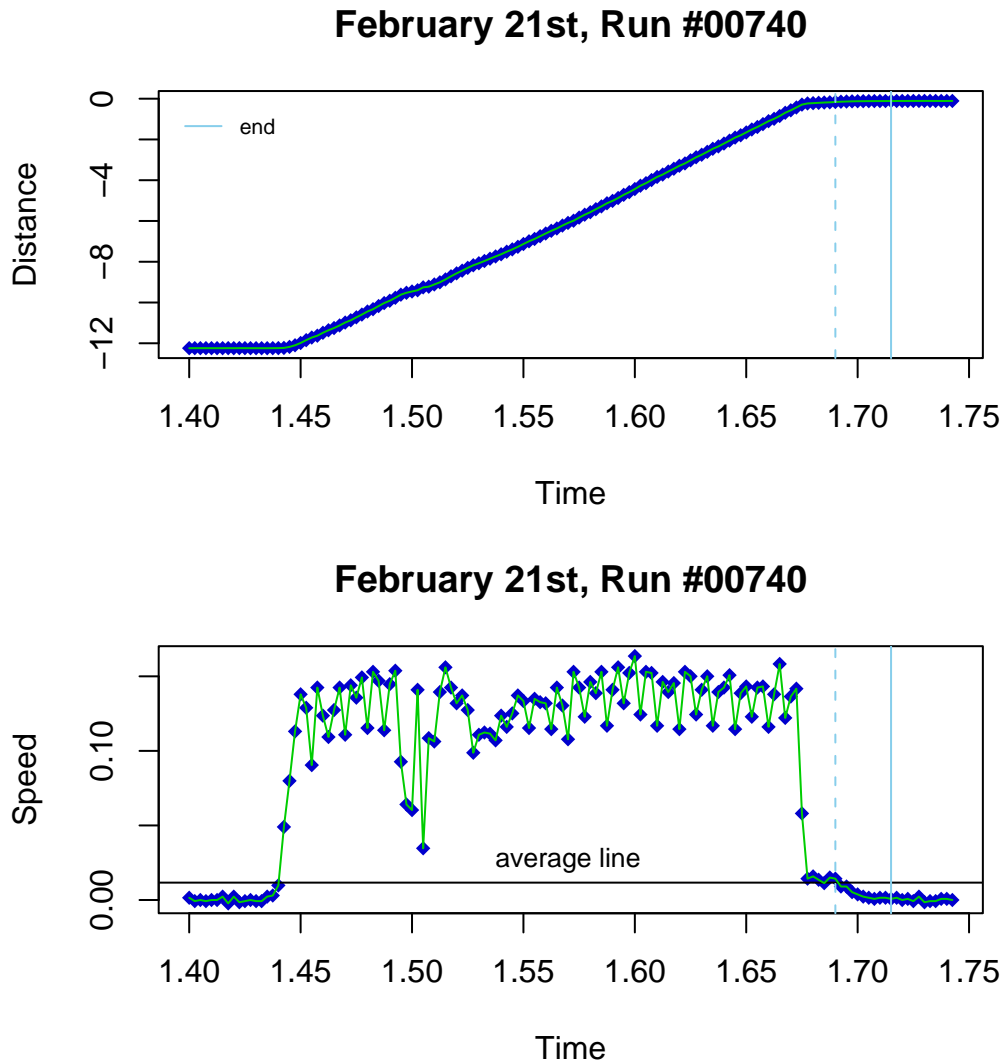


Figure 2.2: Plots of distance against time (top) and first difference of distance against time (bottom) for a single observation. The dashed lines in each curve correspond to the point marking the end of the insertion process as summarized in Step 1 of the registration algorithm. The solid line marks the final end point after the minimization procedure described in Step 2.

STEP 2: Refine the registration points.

To ensure that true locations of the landmark features are identified, final registration points are refined to minimize the sum of squared differences between the individual curves and a target curve $\mu(\mathbf{t})$, where $\mathbf{t} = [1, 2, \dots, 100]^T$ is a standardized time scale. If we let $y_i(\mathbf{t}; s_i, e_i)$ represent the i^{th} observed curve cropped at the registration points (s_i, e_i) and interpolated to time grid \mathbf{t} , then using registration points obtained in step 1 as starting values, the refined start point s_i and end point e_i are found by minimizing

$$\sum_{t=1}^{100} \{y_i(t; s_i, e_i) - \mu(t)\}^2. \quad (2.1)$$

The purpose of (2.1) is to shift the observed data $\mathbf{y}_i(\mathbf{t}^*)$ along the time scale towards a target value $\mu(\mathbf{t})$, such that the start and the end of the process are observed at roughly the same time points across all curves. The minimization problem is solved iteratively using a Newton-type algorithm offered by the `nlm` nonlinear optimization routine in `R`.

An initial estimate of the target curve $\mu(\mathbf{t})$ is obtained from step 1 as the mean of the first twenty curves, cropped at the registration points and interpolated to the same time scale $\mathbf{t} = [1, \dots, 100]^T$. Then, in the current step 2, the target is continuously updated by using the new registration points to crop and interpolate each curve, and setting $\mu(\mathbf{t})$ to be the average the registered data. An alternative is to specify the target as the cross-sectional mean of the raw curves interpolated to time grid \mathbf{t} , but even from the five curves displayed in Figure 1.1, it is clear that the cross-sectional mean of the raw data describes the curves quite poorly.

STEP 3: Crop and interpolate the data.

The next step is to crop each curve at the registration points, focusing only on the portion of the process from the very beginning of the insertion to the very end. This is justified by the fact that there is very little activity in the curves at the flat tail-ends of the registration points (see Figure 1.1).

Once cropped, each force curve is assigned a new time scale of 1 through 100 and the force values are interpolated with respect to these 100 grid points accordingly. In doing so,

we force each curve to be on the same time scale, with the same starting and end points across all of the curves. The particular choice of 100 dimensions for the new time scale was made because the cropped data contained roughly that many values for each curve.

Figure 2.3 helps visualize the registration process for a subset of the last five insertions made on February 21st. The top panel of the plot shows the raw data, with the two registration points marked as filled circles and squares along the curves (Steps 1 and 2). The middle panel shows the cropped curves before they have been aligned (Step 3), and the bottom panel illustrates the curves after the data has been registered, cropped and interpolated.

STEP 4: Re-register the data (optional).

As a final step, it may be helpful to iterate through steps 2 and 3 one more time in order to stabilize landmark feature selection. Re-registering would reduce dependence on the initial set of curves chosen as a baseline.

Registration of multiple valves was achieved by sweeping over each valve individually and registering all of the curves within that valve first before moving on to the next. This means that all of valve 1 curves were registered first, then valve 2 curves, valve 3 curves, and so forth.

2.2 Other Preliminaries

Although observed data on both the distance and the force at which the valves were inserted are available, we restrict our attention to force curves. This is done because unlike the distance curves, which are mostly linear, the force curves seem to contain more information about the dynamic behaviour of the process (see Figure 1.4).

We further restrict our attention to insertions made on a subset of 3,776 engine heads. These include all of the insertions made in January and 1,008 insertions from the last six days in February. Note that January was the start of production for this process. This choice of data is made to facilitate demonstration of process monitoring methods later in §5.3 and §8.5.2. In particular, the 1,008 insertions from February will represent the process “in control” and the January insertions will represent a process that is possibly

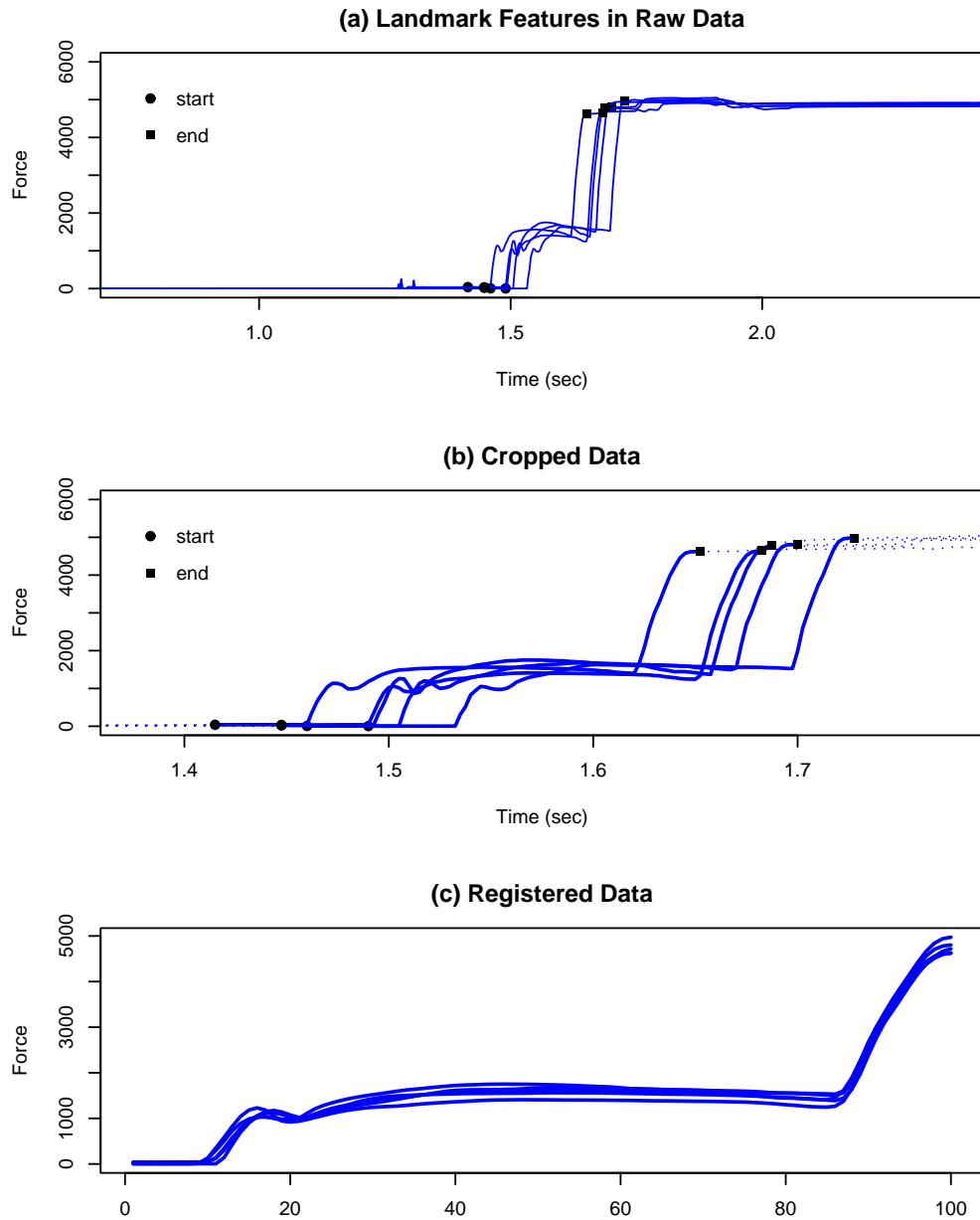


Figure 2.3: Force against time curves for the last five runs of February 21, 2000.

wandering. Figure 2.4 motivates this logic. The plot displays the number of parts produced per day within each month. We see that very few insertions were made at the beginning of January. The mean number of parts produced that month is 126 (standard deviation = 69), compared to 168 (standard deviation = 60) in February. The fact that the number of insertions made in January is fewer and more variable than in February supports our suspicion that production is more stable at the end of February.

Due to a suspected malfunction in the recordings made for valve six, only seven of the eight valves are considered for a grand total of 26,432 curves. For some of these curves, the insertion process ended with inconceivably large force values. These were scaled back to the last largest recorded value prior to the start of the erratic behaviour in the curve in order to be able to proceed with the analysis.

2.3 Data Visualization

One useful approach to assessing key sources of variability in multivariate data is principal components analysis (PCA). To do this, PCA creates a new orthogonal co-ordinate system, formed by taking linear combinations of the original n -dimensional data, called principal components (Ramsay & Silverman 2005, Chapter 8). This is accomplished in such a way that the first principal component (PC) accounts for the most variability in the data, the second one captures the most of whatever variability is left over, and so forth for the remaining $n-2$ PCs.

Based on the analysis of February valve seat insertion data, the first two PCs explain 91% of the overall variation among the curves. By examining these PCs, we may be able to better understand the structure of the data. We restrict our attention to February curves in order to understand variability when the system is likely to be under control. In Chapter 5, we utilize our findings for these data in detecting unusual behaviour amongst the January insertions. For illustration purposes here, we ignore valve labels and explore variation in all of the curves. Unscaled data are analyzed because all of the curves are observed on the same scale.

We use an approach proposed by Jones & Rice (1996) to visualize the structure contained in the first few PCs. The method involves projecting the data onto the first few

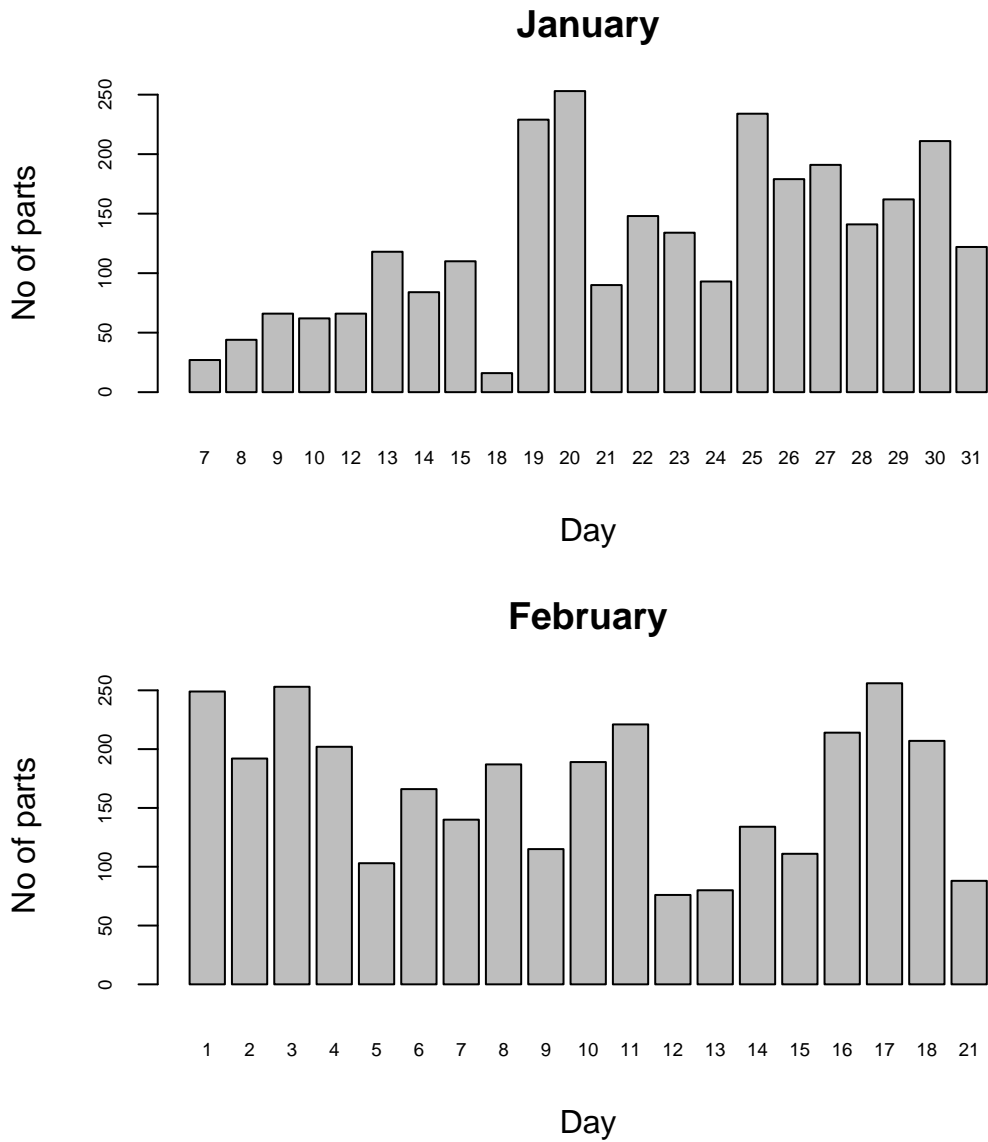


Figure 2.4: Number of parts produced per day plotted against day.

principal components and plotting the curves corresponding to the smallest and largest projected values of each component. Figure 2.5 demonstrates this idea for the first two PCs of the force exertion profiles. Curves with extreme values in the first principal component (see top panel) differ the most in the long flat section spanning time points 25 through 75. Variation is also evident in both the time points at which sharp vertical increases occur and in the extent of the sinusoidal loop in the curves between time points 10 and 20. Sinusoidal variation is further emphasized by looking at the bottom panel describing variability contained in the second PC. Variation in the second PC also stems from the flat section of the curve between time points 85 through 100, and the amount of decay in the long flat segment of the curve at times 25 through 75.

Hastie, Tibshirani & Friedman (2002, page 490) offer an alternative method to visualizing the structure described by the first two PCs. Their approach is demonstrated on the force data in Figure 2.6. The first two PC scores are plotted in the top panel. We can think of each point in this plot as a single curve reduced to 2D. One way to understand the extent of profile-to-profile variation as represented by the first two principal components is to examine changes in the original curves as we move from one extreme part of the space to another. For example, curves corresponding to the points highlighted in black in the top panel of Figure 2.6 are plotted in the same order in the bottom panel. The black points correspond to the 5th, 25th, 50th, 75th and 95th quantiles of the two principal components.

Looking from left to right in the bottom plot of Figure 2.6 (across the range of variation in the first principal component), it appears that the profiles differ primarily in the amount of vertical shift in the long flat middle section. This suggests that the majority of the variability in the data occurs during the middle of the insertion process. Similarly, looking from the bottom to the top, the amount of curvature in the first spike of the curves is increasingly more pronounced. The end flat segment of the curve also seems to be changing in its level. Combined, the extent of curvature in the first spike of the curve and the amount of force applied at the end of the insertion appear to be secondary sources of variation in the profiles, accounting for 6% of the variability in the data. These conclusions are in agreement with the findings based on the Jones & Rice (1996) technique.

Interpreting variability described by higher order PCs is more challenging. As these tend to explain less variation, it is harder to visually detect differences between the curves.

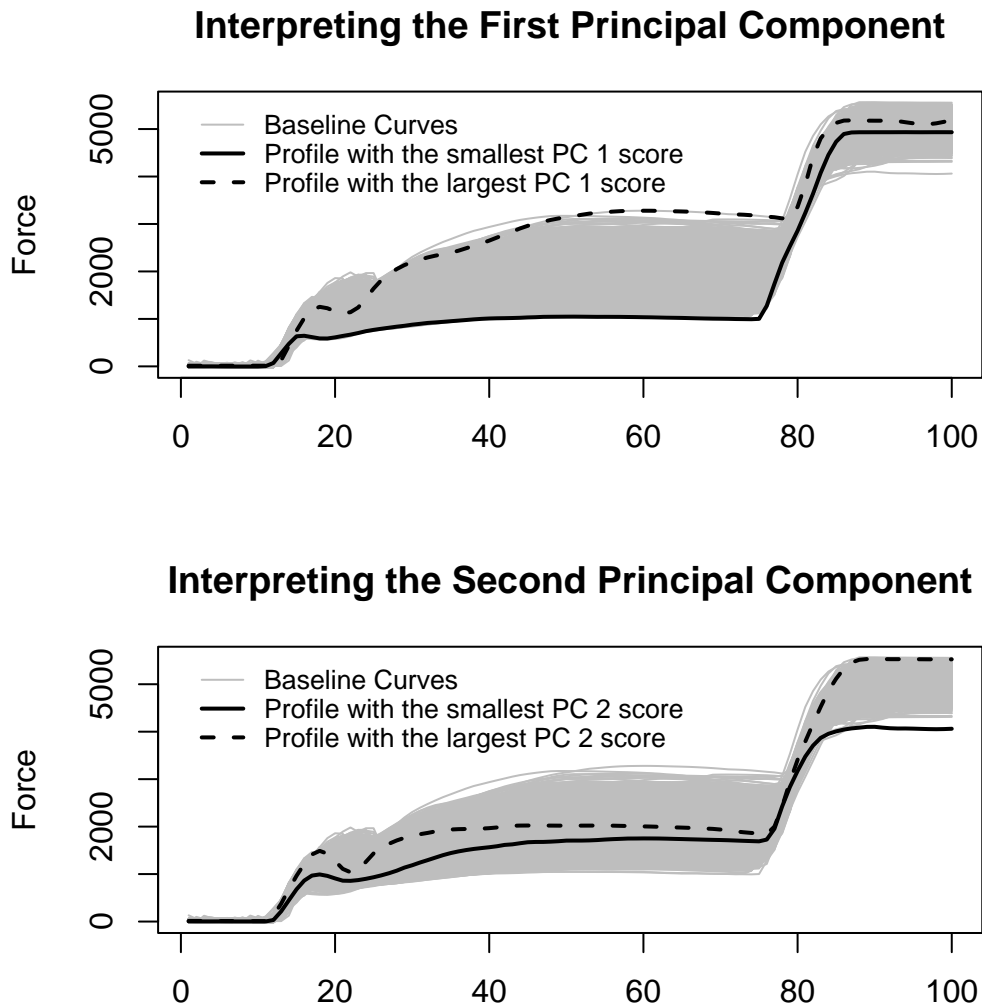


Figure 2.5: Exploring variability in the February force exertion data as summarized by the first two principal components (PCs). The data are plotted in light grey, and curves corresponding to the smallest and largest PC scores are highlighted in black.

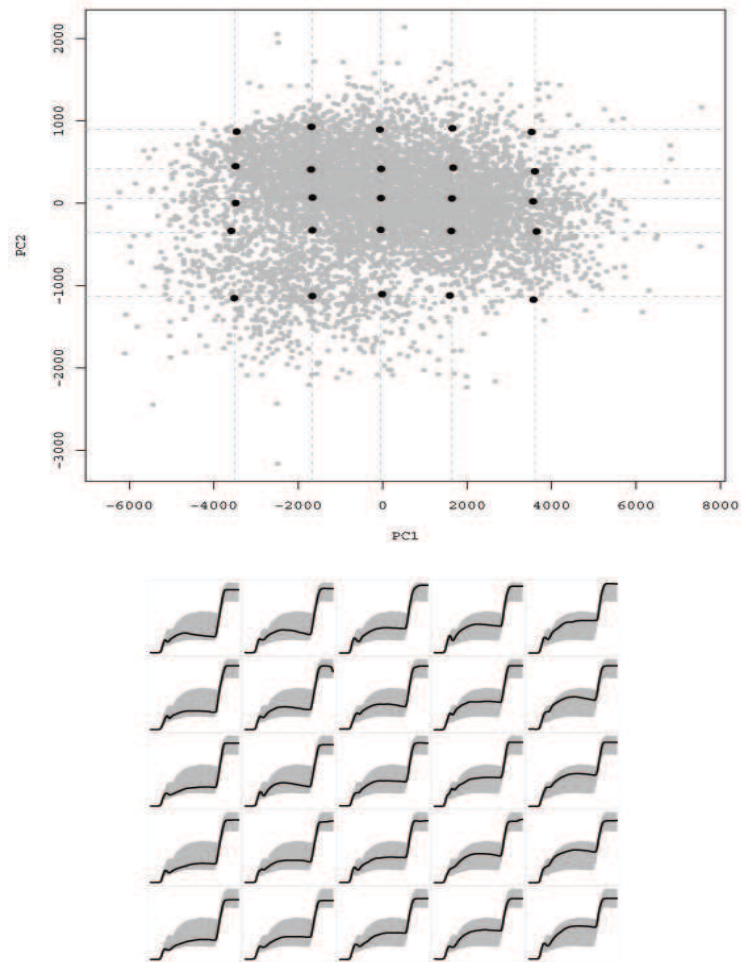


Figure 2.6: The top panel is a plot of the data projected onto the first two principal components for the valve seat insertion example. Each point is an insertion for one valve. All observations are plotted in grey in both panels. The dashed grid lines mark five quantiles of the two principal components, with the points highlighted in black chosen within the closest proximity of the intersections on the grid. Each of these points corresponds to a single functional observation plotted in the bottom panel.

Chapter 3

Modeling a Single Curve

By definition, functional data are characterized by the presence of a systematic pattern in the data points. Looking back at the force curve in Figure 1.4, for example, it is clear that force has a “stair-step” pattern with two significant spikes occurring at 1.49 and 1.67 seconds (marked by the last two vertical dashed lines). The first one occurs when the ram reached the valve seat, which requires an increase in the force in order to push the seat forward. The second spike likely corresponds to the time when the valve seat is fully inserted, but the force exerted by the ram increases to counteract the resistance from the cylinder head. A less significant sinusoidal-looking spike occurs at 1.52 seconds, and corresponds to some other (less obvious) oscillation in the force values. It is this systematic pattern that makes this a functional observation.

In this chapter we describe smoothing and nonlinear modeling as two ways to characterize the functional behavior of a single curve. A fundamental goal of both approaches is to provide a low-dimensional representation of the data. Smoothing is an off-the-shelf technique that can be applied to any curve, whereas nonlinear modeling utilizes functional information about the dynamics of the system and is tailored specifically to the force profiles presented here. In both cases, dimension reduction is achieved by using considerably fewer coefficients than time points on a curve to model each profile.

3.1 Smoothing Splines

A well-known approach for capturing the functional form of a curve from observed data is smoothing. It assumes that a vector of response values $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$ and discrete times $\mathbf{t} = [t_1, t_2, \dots, t_n]^T$ at which they are observed have a functional relationship of the form $f(\mathbf{t})$. That is, let

$$\mathbf{y} = f(\mathbf{t}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (3.1)$$

where $f(\mathbf{t}) = [f(t_1), \dots, f(t_n)]^T$ is a univariate function f evaluated at components of \mathbf{t} .

Although the exact form of f is unknown, one possibility is to assume that it is a linear combination of p known basis functions, such that

$$f(\mathbf{t}) = \sum_{k=1}^p \theta_k b_k(\mathbf{t}) = B\boldsymbol{\theta}, \quad (3.2)$$

where $b_1(\mathbf{t}), \dots, b_p(\mathbf{t})$ are called basis functions because (3.2) is a basis-expansion of $f(\mathbf{t})$. Using compact notation, $B = \{b_j(\mathbf{t})\}_{j=1, \dots, p}$ represents an $n \times p$ matrix whose columns contain the p basis functions observed at n time points. Then the vector of unknown coefficients $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_p]^T$ can be estimated by minimizing the least-squares criterion

$$SSE = [\mathbf{y} - B\boldsymbol{\theta}]^T [\mathbf{y} - B\boldsymbol{\theta}]$$

with respect to $\boldsymbol{\theta}$.

A special case of this idea is polynomial smoothing. Here the unknown functional form of a curve is estimated using a p -degree Taylor series expansion of $f(\mathbf{t})$ about some point $\mathbf{t} = \mathbf{a}$, which is given by

$$f(\mathbf{t}) \approx f(\mathbf{a}) + f'(\mathbf{a})\mathbf{t} + \dots + f^{(p-1)}(\mathbf{a})\mathbf{t}^{p-1} = \sum_{k=0}^{p-1} f^{(k)}(\mathbf{a})\mathbf{t}^k.$$

Letting $\boldsymbol{\theta} = [f(\mathbf{a}), f'(\mathbf{a}), \dots, f^{(p-1)}(\mathbf{a})]^T$ and $B = [\mathbf{1}, \mathbf{t}^2, \dots, \mathbf{t}^{p-1}]$, we get the expression in (3.2). Since the true forms of f and subsequently its derivatives $f^{(k)}$ are unknown, coefficients $\boldsymbol{\theta}$ cannot be evaluated directly. Instead we can estimate them by regressing

observed data \mathbf{y} on B , a matrix whose columns correspond to the polynomial (basis) functions evaluated at the vector of observed times \mathbf{t} .

A flexible generalization of the polynomial fit widely used in practice for modeling nonperiodic data is spline smoothing (Ramsay, 2005, pg. 46). Following James and Sugar (2003), we use a special type of spline functions called b-splines, described by de Boor (2001) and implemented in the R `splines` library (R Development Core Team, 2006).

In b-spline smoothing, the basis functions $b_1(\mathbf{t}), \dots, b_p(\mathbf{t})$ are piecewise polynomials constrained to ensure that a linear combination of these functions is smooth at a set of points, called interior knots. The order of the polynomials, the number and the placement of the knots determine the level of desired smoothness. If first-degree polynomials (line segments) are used, the resulting function is continuous but not differentiable. For quadratic polynomials, the smoothed function is differentiable, and for cubic polynomials it is twice-differentiable.

The total number of b-spline coefficients (p) used to fit a single curve is called the smoother degrees of freedom (d.f.). For curves that have a zero intercept, the d.f. is equal to the degree of the polynomials being used plus the number of interior knots. To see this, suppose that we split a curve into 18 regions at 17 equally-spaced interior knot points and smooth within each region using piecewise cubic polynomials (see Figure 3.1(a)). A cubic polynomial has four terms and thus 4×18 parameters must be estimated. However 3×17 of these are constrained to be the same in order to ensure that the estimated function and its first two derivatives are continuous at the knots. Thus, a total of $(4 \times 18) - (3 \times 17) = 21$ coefficients need to be estimated. If we exclude the intercept, this number drops down to 20. Thus $p = 3 + 17 = 20$.

Figure 3.1(b) provides an example of a b-spline fit to one of the curves from the force exertion process. We let $p = 20$, generating 20 piecewise cubic basis functions to smooth the data at 17 interior knots. The basis functions are plotted in heavier lines below the original data (circled points) and the fitted smooth (solid black line). The fitted line was obtained by regressing the observed data onto the matrix of basis function values.

Due to the inherent relationship between the number of knots and the degrees of freedom, smoothness can be controlled by adjusting the number of basis functions used to fit the data. One way to choose the number of basis functions needed to provide adequate fit

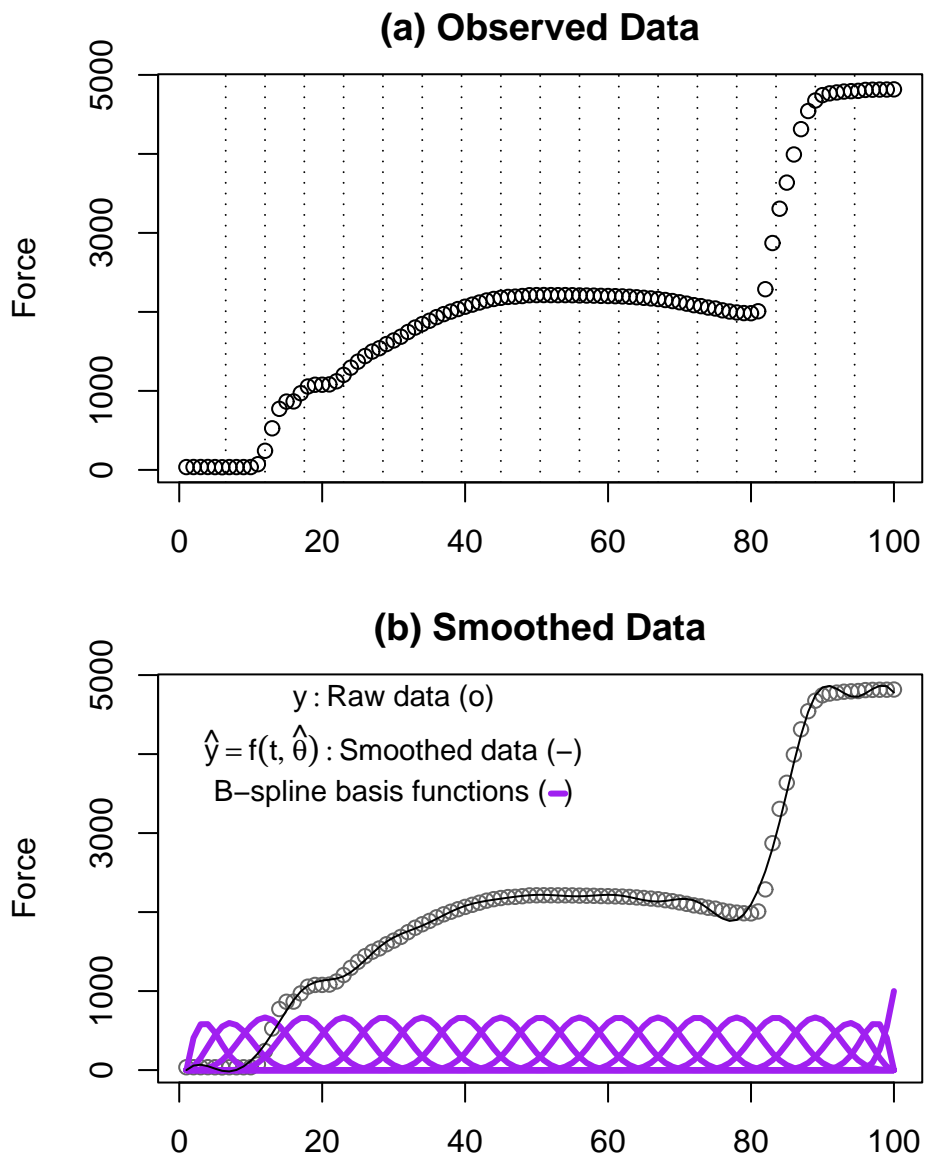


Figure 3.1: (a) A single force exertion curve (points) split into 18 regions at 17 equally spaced boundary points (dashed lines). (b) A b-spline fit $\hat{y} = f(t; \hat{\theta})$ to the observed data. Heavier curves at the bottom are the $p = 20$ basis functions $b_1(t), \dots, b_{20}(t)$ used to smooth over the data.

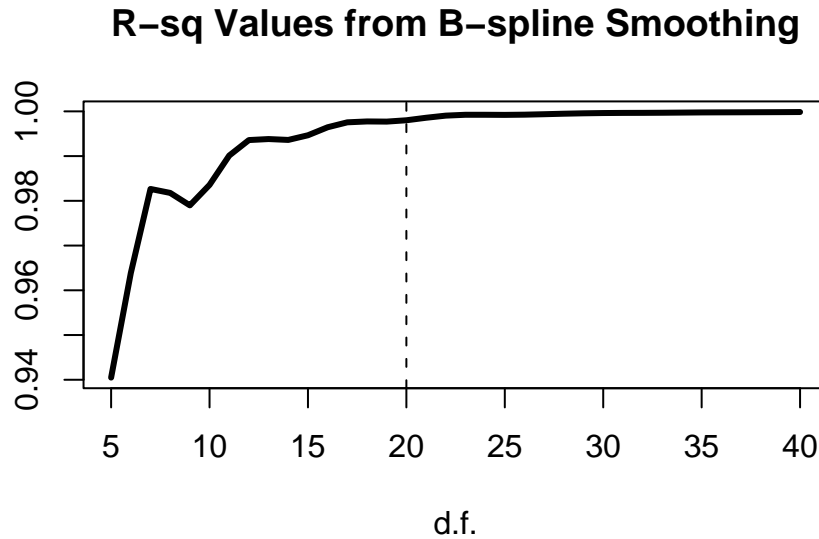


Figure 3.2: R^2 goodness-of-fit measures averaged over all of the observations and plotted against the number of basis functions (d.f.) used to smooth over the force exertion data.

is to examine goodness of fit measures for varying degrees of freedom, then pick a value of p that provides the desired amount improvement in the data fit.

Figure 3.2 illustrates this idea for the force exertion data. Letting $\hat{\mathbf{y}}_i = [\hat{y}_{i1}, \dots, \hat{y}_{in}]^T$ denote a b-spline smooth to the i^{th} observed curve ($i = 1, \dots, m$) recorded at n time points, we define

$$R^2 = 1 - \frac{\sum_{i=1}^m \sum_{j=1}^n (y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^m \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2}.$$

The solid black line in Figure 3.2 connects R^2 values calculated by fitting b-spline smoothers to all of the available curves at varying degrees of freedom. It is easy to see that 20 basis functions provide an excellent fit to the data (R^2 is close to 1), and very little improvement is achieved by using more basis functions.

Other ways to adjust smoothing include changing the degree of the piecewise polynomials and/or the placement of the knots. A common choice and the default in R is to use

cubic b-splines and space the knots at the quantiles of the time scale. These specifications are used throughout this thesis. One possible alternative is to place the knots in such a way that the most interesting aspects of the curve are weighted more heavily than the remainder of the curve. For example, for the force against time curves like the one in Figure 3.1, we might want to de-emphasize the long flat segment of the curve and focus on the peaks instead. We can do this by selectively placing more knots at the peaks.

Advantages of smoothing are numerous, the key benefit being that the b-spline model is linear in the parameters, meaning that model coefficients are easy (and fast) to estimate and have well-established properties. One drawback is that a large number of these parameters are often needed to adequately represent the functional form of a curve. For valve seat insertion, the number of b-spline coefficients that are used (20) is nearly double the number of parameters in some of the nonlinear models, which we introduce next.

3.2 Nonlinear Modeling

While linear models simplify computation, a broader range of nonlinear functions will often provide a better fit to the data. For the force exertion example, such models are generated by observing the dynamics of the process, which are reflected in the general shape and prominent features of the curve. For example, referring back to the force curve fitted using smoothing splines in Figure 3.1 and replotted in Figure 3.3(a), at least three important events appear to occur during the insertion process. These are marked by the three significant spikes in the force values around 10, 20 and 80 time units, which are emphasized by vertical dashed lines in the plot. Within each region, the force functions appear to be exponentially increasing towards force values marked by the horizontal dashed lines.

Combining this information, we propose

$$f(\mathbf{t}) = \sum_{k=1}^3 \{ \alpha_k (1 - e^{-\beta_k(\mathbf{t}-\delta_k)}) \cdot I(\mathbf{t} > \delta_k) \} \quad (3.3)$$

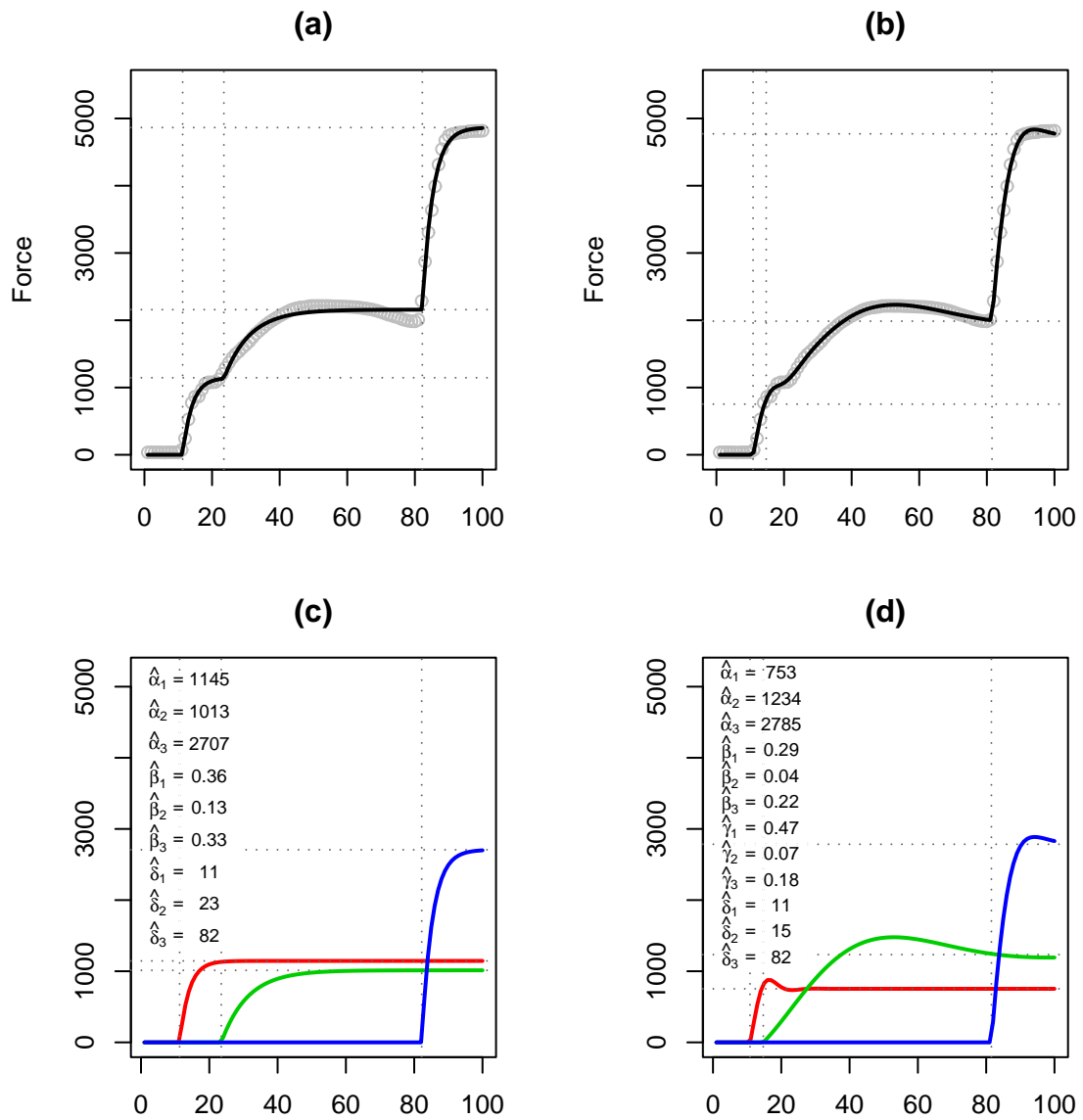


Figure 3.3: Nonlinear least-squares fits of the (a) first-order and (b) second-order DE models for one realization of the force exertion data (points). Three piecewise functions plotted in (c) and (d) are summed to obtain fits in (a) and (b) respectively.

as a nonlinear model for the curve. Here

$$I(t > \delta_k) = \begin{cases} 1 & \text{if } t > \delta_k \\ 0 & \text{otherwise} \end{cases}$$

constrains each of the three piecewise solutions to specific regions of the curve for which $t > \delta_k$. Thus, δ_1 , δ_2 and δ_3 represent the times at which each of the three events initiate. An engineering term for these quantities is reaction times, because they measure the amount of time it takes for the system to react to the initiated events.

The proposed model is the solution to a first-order system of differential equations, and will be motivated in §3.2.1. An important characteristic of (3.3) is its exponential form, which reflects the fact that at each stage of the insertion process force exertion monotonically increases, eventually reaching a steady-state of

$$\lim_{t \rightarrow \infty} \alpha_k (1 - e^{-\beta_k(t-\delta_k)}) = \alpha_k$$

for positive values of β_k .

The terms α_k are called gain parameters, because they quantify an increase in force exertion going from one part of insertion to the next. By time $t \approx \frac{4}{\beta_k} + \delta_k$, the process is nearly stabilized at level $\alpha_1 + \dots + \alpha_k$, since

$$\alpha_k (1 - e^{-\beta_k(t-\delta_k)})|_{t=\frac{4}{\beta_k}+\delta_k} = 0.98\alpha_k \approx \alpha_k.$$

Thus β_k are called response speed parameters.

Following Ramsay (2000), the unknown parameters in (3.3) are found using nonlinear least squares. That is, $\boldsymbol{\theta} = [\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \delta_1, \delta_2, \delta_3]^T$ are estimated by minimizing

$$SSE = \sum_{j=1}^n \{y_j - f(t_j; \boldsymbol{\theta})\}^2 = [\mathbf{y} - f(\mathbf{t}; \boldsymbol{\theta})]^T [\mathbf{y} - f(\mathbf{t}; \boldsymbol{\theta})]$$

with respect to $\boldsymbol{\theta}$. As before, $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$ represents a vector of force values observed at times $\mathbf{t} = [t_1, t_2, \dots, t_n]^T$. We use the `nls` function in **R/S-PLUS** (R Development Core Team, 2006) to accomplish the minimization; another good option is the `lsqcurvefit`

function in MATLAB's Optimization Toolbox.

A fitted curve (3.3) is displayed as a solid black line in Figure 3.3(a). The three distinct regions of the curve are modeled using three different nonlinear functions, which are plotted in Figure 3.3(c). We assume that these correspond to different events in the valve-seat insertion process.

As expected, $\hat{\delta}_1 = 11$, $\hat{\delta}_2 = 23$, and $\hat{\delta}_3 = 82$, marked by dashed vertical lines, correspond to the times at which each of the three events were initiated. Force amounts accumulated during each part of the process are estimated to be $\hat{\alpha}_1 = 1145$, $\hat{\alpha}_2 = 1013$, and $\hat{\alpha}_3 = 2707$. Cumulative gain parameters ($\hat{\alpha}_1$, $\hat{\alpha}_1 + \hat{\alpha}_2$, and $\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3$) are marked by horizontal dashed lines in Figures 3.3(a) and (c).

Estimated response speed parameters are $\hat{\beta}_1 = 0.36$, $\hat{\beta}_2 = 0.13$ and $\hat{\beta}_3 = 0.33$. These quantify how quickly changes occur. The higher value of $\hat{\beta}_3$ than $\hat{\beta}_2$, for example, indicates that force is increasing more rapidly at the end of the insertion process, a fact that is clearly evident from the shape of the force curve.

As can be seen from Figure 3.3(a), the fitted model has clear deficiencies in describing some of the curvature in the observed data. A more flexible nonlinear model is considered in §3.2.2. Both models are motivated as solutions of differential equations, and the rationale behind this is presented next.

3.2.1 Motivation

A key assumption behind modeling force data is that force changes continuously with time via some mapping function $f(t)$. This implies the existence of a first and possibly higher order derivatives of $f(t)$, which we denote by $D^{(n)}f(t) = \frac{d^{(n)}f(t)}{dt^{(n)}}$ using Euler notation. Our goal is to incorporate any information we may have about the derivatives, i.e. the dynamics of the process, in modeling the curves. That is, we seek to use models based on differential equations (DEs).

Ramsay (2000) introduced these ideas to functional data analysis by using data to estimate parameters of interest in a DE model. Recent developments by Ramsay, Hooker et. al. (2006) extend this methodology to allow its application to just about any dynamic process.

For valve seat insertion data, the process is characterised by three critically damped

spikes in force exertion. These are well characterized by the first order DE

$$Df(\mathbf{t}) = \sum_{k=1}^3 \{\alpha_k \beta_k I(\mathbf{t} \geq \delta_k) - \beta_k f(\mathbf{t})\}. \quad (3.4)$$

The function provided in (3.3) is a solution to this equation. The result was obtained by combining solutions to the three first-order DEs inside the sum of (3.4), which can be found in introductory texts on DEs, e.g. Ogunnaike and Ray (1994).

3.2.2 Second-order DE Model

Although the first order model is able to describe an overall pattern in the process, it appears to be inadequate at summarizing local features in the shape of the curve (refer to Figure 3.3(a)). One concern is that it has difficulty representing the oscillations between $20 < t < 80$. The three terms being summed in (3.3) to obtain the fit (see Figure 3.3(a)) are monotonically increasing functions that flatten out without oscillating. The fit is also abrupt in transitioning to the piecewise solutions at times 10, 20 and 80.

An extension of the first-order relationship in (3.4) which allows for oscillations is the solution to a second-order DE is given by:

$$f(\mathbf{t}) = \sum_{k=1}^3 \{\alpha_k (1 - e^{-\beta_k(\mathbf{t}-\delta_k)} \cos(\gamma_k(\mathbf{t} - \delta_k))) \cdot I(\mathbf{t} \geq \delta_k)\}. \quad (3.5)$$

With the exception of the new sinusoidal parameters, unknown quantities in (3.5) have the same interpretation as for the first order model. This follows from the fact that if all $\gamma_k = 0$, (3.5) simplifies to (3.3). The new parameters γ_1, γ_2 and γ_3 measure oscillation frequency in the force values. Higher absolute values of these parameters correspond to rapidly oscillating parts of the force curve. One caution is that $\cos(x) = \cos(-x)$, and therefore γ_k and $-\gamma_k$ generate the same curve estimates. To alleviate any confusion, we recommend either constraining these parameters to be positive during the nonlinear estimation procedure or taking absolute values of the final estimates.

A fitted solution to the model

$$\mathbf{y} = f(\mathbf{t}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

where $f(\mathbf{t})$ has nonlinear form (3.5) is displayed as a solid black line in Figure 3.3(b). The three piecewise solutions that sum to the fitted line are displayed in Figure 3.3(d). Dampening oscillations in the shapes of these functions provide the desired flexibility in fitting the data.

The improvement in the fit is encouraging. The middle part of the curve and the smoothness at $t = 20$ are clearly better fit by the second-order model. Time parameters $\hat{\boldsymbol{\delta}} = (11, 15, 82)$ are in close correspondence with our earlier estimates of reaction times. The only exception is $\hat{\delta}_2$, which is now closer to $\hat{\delta}_1$, implying that the process may be dominated by two rather than three major events or that the first event lasts a shorter time. Cumulative gain parameters $\sum \hat{\boldsymbol{\alpha}}_k = (753, 1987, 4772)$ are also similar to the first-order estimates (1145, 2158, 4865), with the amount of force accumulated during the first part of the insertion process ($\hat{\alpha}_1$) adjusted for initial oscillation.

Estimated response speed parameters $\hat{\boldsymbol{\beta}} = (.29, .04, .22)$ are close to the earlier estimates, with the exception of $\hat{\beta}_2$, which is considerably lower than the previous value, implying that it takes longer to reach a steady state. This is evident from examining the shape of the function for $15 < t < 82$, which curves upward before stabilizing to a constant value. Such curvature is modeled by the oscillation frequencies, which are estimated to be $\hat{\boldsymbol{\gamma}} = (0.47, 0.07, 0.18)$.

In summary, fitting results for the two nonlinear DE models support the following interpretation of the parameters:

- i) Difference in reaction times, $\delta_{k+1} - \delta_k$, indicates the duration of the k^{th} part of the process;
- ii) α_k equals the gain, or the amount of force accumulated during event k . The overall gain of the system is given by $\sum_k \alpha_k$;
- iii) β_k measures the decay, or the speed with which force stabilizes;
- iv) $|\gamma_k|$ represents oscillation frequency in the force values.

We can think of decay and oscillation frequency as shape parameters, and gain and reaction time as scale parameters. That is, decay and oscillation frequency carry information about the twists and turns in the curve, whereas gain and reaction time simply shift parts of the plot toward higher or lower values.

Points (i)-(iv) highlight the biggest advantage of using DEs to model functional data - interpretability of the parameters. Using only a handful of parameters, we are able to specify a model that makes sense in the context of our problem. A disadvantage of this approach is the fact that the rather complicated nature of the solutions involve a nonlinear model, and an iterative algorithm is needed to estimate the parameters at hefty computational costs. Many problems may also involve higher-order DEs which are less easy to solve, and may not have closed-form solutions.

3.3 Discussion

Several factors are worth noting when deciding whether the b-spline or the nonlinear model is best suited for the problem. Advantages of b-splines are that they are linear in the parameters and generic enough to be applied to any data without the need for understanding the dynamics of the process. This also means that they are easier and faster to fit computationally. In contrast, the nonlinear model involves parameters that are easy to interpret in the context of the problem. As with any situation where there is a choice between two different modeling techniques, there is no “best” model, but there might be one that is more appropriate than the other in terms of the trade-offs between computation and interpretability.

For the force data, we also believe that the nonlinear model is better at explaining key sources of variability because it is application-specific. The model uses derivatives to characterize observable changes in the process. Specifically, each of the three sections of the force curves can be represented as a solution to a second-order differential equation (DE). We chose to use a second-order DEs as it allows more flexibility than the simpler first-order DE in modeling curvature at the δ_k . For more information on the use of DEs to model functional data, see Ramsay & Silverman (2005).

Chapter 4

Modeling Collections of Curves

Let us now consider a collection of m independent profiles: $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$. One possible extension of (3.1) to describing the curve \mathbf{y}_i is to allow the structural parameters to vary from curve to curve, leading to the model:

$$\mathbf{y}_i = f(\mathbf{t}; \boldsymbol{\theta}_i) + \boldsymbol{\epsilon}_i; \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (4.1)$$

where $\boldsymbol{\theta}_i = [\theta_{i1}, \dots, \theta_{ip}]^T$ is a vector of p unknown model parameters specific to curve i , and $\boldsymbol{\epsilon}_i$ is a random component measuring pointwise roughness in the curve. To emphasize our interest in the structural parameters, we write $f(\boldsymbol{\theta}_i)$ instead of $f(\mathbf{t}; \boldsymbol{\theta}_i)$, suppressing the time variable. For b-splines $f(\boldsymbol{\theta}_i)$ is the linear function $f(\boldsymbol{\theta}_i) = B\boldsymbol{\theta}_i$, and for the DE model, it is the nonlinear function of form (3.5) with coefficients $\boldsymbol{\theta}_i = [\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \beta_{i1}, \beta_{i2}, \beta_{i3}, \gamma_{i1}, \gamma_{i2}, \gamma_{i3}, \delta_{i1}, \delta_{i2}, \delta_{i3}]^T$.

A recurrent idea in this thesis is to utilize information contained in the estimated model parameters rather than the observed data in monitoring the insertion process, identifying general trends, process drifts and/or outliers. In this context, (4.1) has far too many parameters to be useful. An extension of (4.1) is the mixed effects model, which employs a set of fixed parameters ($\boldsymbol{\mu}$) that is the same across all curves and characterizes their common shape, and random parameters ($\boldsymbol{\eta}_i$) that describe curve-specific variation from the overall shape. That is, we let $\boldsymbol{\theta}_i = (\boldsymbol{\mu}, \boldsymbol{\eta}_i)$, where $\boldsymbol{\mu}$ is a vector that does not vary with i , and $\boldsymbol{\eta}_i$ is a random effect that varies across curves $i = 1, \dots, m$. The manner in which

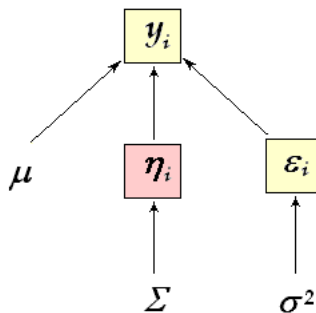


Figure 4.1: Hierarchical dependencies in the mixed-effects models (4.2) and (4.3). Boxed and unboxed quantities correspond to random and fixed quantities respectively. Here, $Cov(\boldsymbol{\eta}_i) = \Sigma$ and $Cov(\boldsymbol{\epsilon}_i) = \sigma^2 \mathbf{I}_n$

$\boldsymbol{\mu}$ and $\boldsymbol{\eta}_i$ are combined to represent f will depend on whether a b-spline or a nonlinear model is used. Details are provided in the next section.

By including both fixed ($\boldsymbol{\mu}$) and random ($\boldsymbol{\eta}_i$) parameters, each curve is modeled individually, while at the same time borrowing strength from similarities amongst all profiles. This is evident from Figure 4.1, which illustrates hierarchical dependencies among the data and the parameters. The diagram highlights two sources of variability associated with each curve: within curve variability (σ^2) and variability in the structural parameters (Σ) contributing to differences among the individual curves. That is, by treating some of the structural parameters as random effects, we allow the profiles to vary in both roughness ($\boldsymbol{\epsilon}_i$) and shape ($\boldsymbol{\eta}_i$). While it is possible to place restrictions on the form of Σ (e.g., forcing it to be diagonal), to allow flexibility we do not explore this idea here.

The number of random effects used to model each curve can vary. For parsimony and to ease computation, we suggest using fewer random effects than the number of fixed effects. That is, even though a potentially large number of parameters may be needed to model an overall trend in the curves, few parameters can generally be used to describe curve-specific departures away from the overall trend.

In the remainder of this chapter, we consider two possible applications of the mixed effects to b-spline and nonlinear models.

4.1 Linear Mixed Effects

A straightforward application of mixed effects to b-splines is the Laird & Ware (1982) model:

$$\mathbf{y}_i = B_1 \boldsymbol{\mu} + B_2 \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad (4.2)$$

$$\boldsymbol{\eta}_i \sim \mathcal{N}_q(\mathbf{0}, \Sigma) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

Here B_1 is an $n \times p$ matrix whose columns contain p known basis functions evaluated at times \mathbf{t} , and $\boldsymbol{\mu}$ is a fixed vector of corresponding b-spline coefficients. B_2 is a matrix of q basis functions and $\boldsymbol{\eta}_i$ s are the random coefficients for each curve.

An example of a single curve from the February data, fit using (4.2) with $p = 20$ and $q = 4$, is displayed as a heavy black line in Figure 4.2(a) alongside the raw data. A reconstruction using only the fixed effect is marked on the same plot as a thin red line. This represents the fitted mean of all February curves under consideration. The plot stresses the role played by the random effects in modeling departures of a specific curve from the overall average (red line). For this particular observation, the contribution of the random effects appears to be in the flat middle and end segments of the curve. The 24 basis functions used to model the data are plotted below, with the 20 columns of B_1 and four columns of B_2 shown in Figures 4.2(b) and 4.2(c) respectively.

According to model (4.2), the shape of each profile is dictated by an overall mean curve $B_1 \boldsymbol{\mu}$ plus a curve-specific departure $B_2 \boldsymbol{\eta}_i$. We focus on the latter. The hope is that predictions of the random effects $\boldsymbol{\eta}_i$ will capture enough key information about unique attributes of each profile to allow monitoring the process by observing changes in $\hat{\boldsymbol{\eta}}_i$ rather than the original data \mathbf{y}_i . This is a form of dimension reduction in that the $\hat{\boldsymbol{\eta}}_i$ s serve as compact low-dimensional summaries of the differences among the curves. In an effort to summarize curve-specific variation in as few parameters as possible, we use a large number of b-splines ($p = 20$) to model the mean profile, but only a few ($q = 4$) to describe curve-specific departures from the overall mean.

A conceptual equivalent of the b-spline mixed-effects model involves subtracting the functional average of all the profiles ($\bar{\mathbf{y}}$) from each curve, and modeling the residual curves

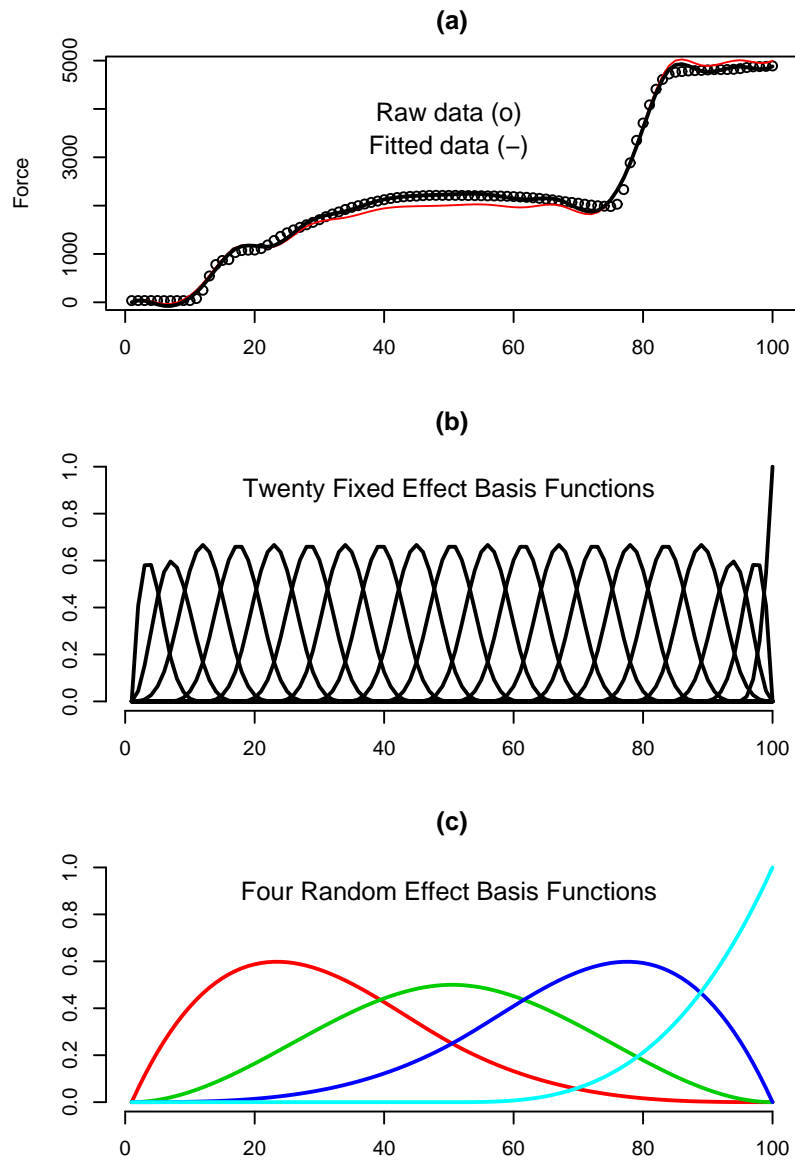


Figure 4.2: Linear mixed-effects model fitted to a single observation in the force exertion data (a), and 24 basis function used to obtain the fit ((b) and (c)).

using

$$\begin{aligned} \mathbf{y}_i - \bar{\mathbf{y}} &= B_2 \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \\ \boldsymbol{\eta}_i &\sim \mathcal{N}_q(\mathbf{0}, \Sigma) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n). \end{aligned}$$

While this leads to similar results, an advantage of (4.2) is that it generalizes easily to more complicated scenarios. For example, rather than fitting a common fixed effect, we may wish to control for the valve effect (recall there are eight in total) and the ram effect (two rams simultaneously insert four valve seats at a time) by including corresponding fixed terms in the model. Another advantage of fitting (4.2) is that there is no need to assume that all curves are observed on the same time scale (eliminating the need for curve registration), whereas subtracting an average level requires this to be true.

4.2 Nonlinear Mixed Effects

An extension of the Laird & Ware (1982) mixed model to the nonlinear case is discussed in Lindstrom & Bates (1990). We adopt a special case by letting

$$\mathbf{y}_i = f\left(\underbrace{\boldsymbol{\mu}}_{p \times 1} + \underbrace{Z}_{p \times q} \underbrace{\boldsymbol{\eta}_i}_{q \times 1}\right) + \underbrace{\boldsymbol{\epsilon}_i}_{n \times 1}, \quad (4.3)$$

$$\boldsymbol{\eta}_i \sim \mathcal{N}_q(\mathbf{0}, \Sigma) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

for $i = 1, \dots, m$ and $q \leq p$. Here $\boldsymbol{\theta}_i = \boldsymbol{\mu} + Z\boldsymbol{\eta}_i$, where Z is an indicator matrix dictating which of the p parameters have a random component.

In the application, the nonlinear form of $f(\boldsymbol{\theta}_i)$ is given by

$$f(\boldsymbol{\theta}_i) = \sum_{k=1}^3 \left\{ \alpha_{ik} \left(1 - e^{-\beta_{ik}(t-\delta_{ik})} \cos(\gamma_{ik}(t-\delta_{ik})) \right) \cdot I(t > \delta_{ik}) \right\}.$$

The structural parameters are $\boldsymbol{\theta}_i = [\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \beta_{i1}, \beta_{i2}, \beta_{i3}, \gamma_{i1}, \gamma_{i2}, \gamma_{i3}, \delta_{i1}, \delta_{i2}, \delta_{i3}]^T$, and their interpretation is discussed in §3.2. The formulation of our model in (4.3) allows all 12 of these parameters to differ for each curve (i.e. $Z = I_p$ or equivalently $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \boldsymbol{\eta}_i$).

However it seems likely that only a handful of the parameters capture significant changes in the curves. It is only these parameters that we wish to have random components, with the remaining parameters fixed across all of the curves.

A key distinction between (4.3) and the linear model (4.2) is that in (4.3) there is an additive relationship between the fixed and random effects (i.e. $\boldsymbol{\theta}_i = \boldsymbol{\mu} + Z\boldsymbol{\eta}_i$), which implies that some (or all) of the fixed effects have associated random components. This is different from the linear case, where we use fixed effects to fit the mean curve ($B_1\boldsymbol{\mu}$) and random effects ($B_2\boldsymbol{\eta}_i$) to fit the residual curves after subtracting off the mean profile. In the linear case, all we have to decide is how many parameters we wish to make random; for the nonlinear model, we must decide which of the fixed effects will have associated random components. An example of how to make this choice for the force exertion data is presented in §4.3.2 and implemented in §5.3.

4.3 Parameter Estimation

The first step in monitoring the structural parameters is to estimate them and establish the distributional properties. In this section we summarize some of the theoretical results for mixed-effect models. Basic ideas are introduced in the context of b-spline and DE models, though these generalize directly to all other mixed-effect models in the respective linear and nonlinear domains.

Resources used in this section include results and discussions by Laird & Ware (1982), Lindstrom & Bates (1988 and 1990), Demidenko (2004) and Pinheiro & Bates (2001). The last two books in particular provide an excellent overview of linear and nonlinear mixed-effects models.

4.3.1 Linear Model

A unique attribute of model (4.2) and mixed-effects models in general is the fact that they combine both frequentist and Bayesian ideas. Demidenko (2004) describes it best in the summary of his book:

“The mixed model technique is a child of the marriage of the frequentist and Bayesian approaches. Similar to the Bayesian approach, a mixed model specifies the model in a hierarchical fashion (see Figure 4.1), assuming that parameters are random. However, unlike the Bayesian approach, hyperparameters are estimated from the data as in the frequentist approach. As in the Bayesian approach, one has to make a decision as to the prior distribution, but that distribution may contain unknown parameters that are estimated from the data, as in the frequentist approach”.

Specifically, if we rewrite model (4.2) as

$$\mathbf{y}_i \sim \mathcal{N}_n(B_1 \boldsymbol{\mu}, V) \quad \text{where } V = B_2 \Sigma B_2^T + \sigma^2 \mathbf{I}_n, \quad (4.4)$$

from a frequentist perspective unknown parameters include $\boldsymbol{\mu}$, Σ , and σ^2 . These are generally estimated using maximum likelihood or least squares. However, the unobserved or latent $\boldsymbol{\eta}_i$ are also parameters of interest. If anything, these are the structural quantities that are the most interesting as they contain key profile-specific information. Conditional expectations of $\boldsymbol{\eta}_i$ based on sample data can be used to estimate these quantities. That is, we let $\hat{\boldsymbol{\eta}}_i = E(\boldsymbol{\eta}_i | \mathbf{y}_i)$, which is essentially a posterior mean.

Although this methodology has a Bayesian flavor, our interpretation remains frequentist. The use of expectations may give a misleading impression that $\hat{\boldsymbol{\eta}}_i$ is a fixed quantity. It is important to remember that this is not the case. This is because the expectation is taken with respect to $\boldsymbol{\eta}_i$ not the data. As such $\hat{\boldsymbol{\eta}}_i$ will still depend on the random \mathbf{y}_i and therefore will itself be random. Many authors like to distinguish between the two different ways in which the fixed and random effects are estimated by calling the former estimates and the latter predictions. For consistency, we adopt the same terminology here.

Estimating Fixed Effects

Common approaches to estimating $\boldsymbol{\mu}$, Σ , and σ^2 are maximum likelihood (ML) and its close relative restricted maximum likelihood (REML). The basic idea behind both is to determine parameter values that are best supported by the data. In ML this is achieved

by maximizing the log-likelihood function (Lindstrom and Bates, 1990):

$$l(\boldsymbol{\mu}, \sigma^2, \Sigma | \mathbf{y}) = -\frac{m}{2} \log |V^{-1}| - \frac{1}{2} \sum_{i=1}^m \{(\mathbf{y}_i - B_1 \boldsymbol{\mu})^T V^{-1} (\mathbf{y}_i - B_1 \boldsymbol{\mu})\}, \quad (4.5)$$

with respect to $\boldsymbol{\mu}$, Σ , and σ^2 . Here $V = B_2 \Sigma B_2^T + \sigma^2 \mathbf{I}_n$ denotes overall variability in the curves as given in (4.4).

A closer look at expression (4.4) reveals that this evaluation problem is equivalent to that of generalized least squares (GLS), with parameter estimates depending largely on the form of Σ , which dictates the form of V . If true values of σ^2, Σ and therefore V are available, then by letting $\mathbf{y}_i^* = V^{-1/2} \mathbf{y}_i$, we can re-write (4.4) into an OLS problem

$$\mathbf{y}_i^* \sim \mathcal{N}_n (V^{-1/2} B_1 \boldsymbol{\mu}, \mathbf{I}_n),$$

with the least squares estimate of $\boldsymbol{\mu}$ taking on familiar form

$$\hat{\boldsymbol{\mu}}_{GLS} = (B_1^T V^{-1} B_1)^{-1} B_1^T V^{-1} \bar{\mathbf{y}}.$$

For fixed V , $\hat{\boldsymbol{\mu}}_{GLS}$ is also the MLE of $\boldsymbol{\mu}$. The estimate also possesses a myriad of other desirable properties, the primary of which is the fact that it is the best linear unbiased estimator (BLUE) of $\boldsymbol{\mu}$ (Pinheiro & Bates, 2001). It is best in a sense that it has the lowest mean squared error among all estimators of its kind, linear with respect to the response, and unbiased because its expected value with respect to the data is equal to $\boldsymbol{\mu}$.

In practice, the true values of σ^2, Σ and subsequently V are unknown. In the absence of these quantities, the feasible generalized least squares (FGLS) estimate of $\boldsymbol{\mu}$ is given by

$$\hat{\boldsymbol{\mu}} = (B_1^T \hat{V}^{-1} B_1)^{-1} B_1^T \hat{V}^{-1} \bar{\mathbf{y}}, \quad (4.6)$$

where $\hat{V} = B_2 \hat{\Sigma} B_2^T + \hat{\sigma}^2 \mathbf{I}_n$. Although $\hat{\boldsymbol{\mu}}$ is not exactly BLUE, as sample size m increases towards infinity, $\hat{V} \approx V$ so that asymptotically $\hat{\boldsymbol{\mu}}$ is approximately BLUE (Harville, 1990).

As indicated earlier, variance estimates $\hat{\sigma}^2$ and $\hat{\Sigma}$ are generally obtained by maximizing the log-likelihood function (4.5). A REML alternative is to estimate the variance compo-

nents by maximizing

$$l_R(\hat{\boldsymbol{\mu}}, \sigma^2, \Sigma | \mathbf{y}) = -\frac{1}{2} \log |B_1^T V^{-1} B_1| + l(\hat{\boldsymbol{\mu}}_{GLS}, \sigma^2, \Sigma | \mathbf{y}),$$

where $l(\boldsymbol{\mu}, \sigma^2, \Sigma | \mathbf{y})$ is the log-likelihood function, as shown in (4.5). This approach is used in this thesis and is preferred to ML because it takes into account the degrees of freedom lost in estimating the fixed effects and therefore tends to produce variance estimates that are less biased. REML and ML estimates of $\boldsymbol{\mu}$ are equivalent. For a discussion of the REML approach, the reader is referred to Pinheiro & Bates (2001). Computational details of the iterative algorithms used to compute parameter estimates, such as the EM algorithm or the Newton-Raphson method, can be found in Laird & Ware (1982), Lindstrom and Bates (1988) and many other papers and textbooks written on the topic.

Predicting Random Effects

While $\boldsymbol{\mu}$ is fixed for all curves, (4.2) dictates that the random effects $\boldsymbol{\eta}_i$ vary. Thus, inference about these parameters is based on their estimated distribution, and predictions of the random effects are obtained using conditional expectation (Lange & Ryan, 1989). For σ^2 and Σ known, Harville (1990) showed that the best linear unbiased predictor (BLUP) of the random components is given by

$$E(\boldsymbol{\eta}_i | \mathbf{y}_i, \sigma^2, \Sigma) = \Sigma B_2^T V^{-1} (\mathbf{y}_i - B_1 \hat{\boldsymbol{\mu}}_{GLS}).$$

Then much like with the fixed effects, for large enough sample size m , approximate BLUPs of the random effects in the absence of true knowledge of the variance components are given by

$$\hat{\boldsymbol{\eta}}_i = \hat{\Sigma} B_2^T \hat{V}^{-1} (\mathbf{y}_i - B_1 \hat{\boldsymbol{\mu}}), \quad (4.7)$$

where $\hat{\boldsymbol{\mu}}$ is the approximate BLUE as stated in (4.6) and $\hat{\sigma}^2$ and $\hat{\Sigma}$ are estimated numerically from the data using the ML or REML framework.

4.3.2 Nonlinear Model

An alternative to b-splines is the nonlinear DE model:

$$\mathbf{y}_i = \sum_{k=1}^3 \left\{ \alpha_{ik} \left(1 - e^{-\beta_{ik}(\mathbf{t} - \delta_{ik})} \cos(\gamma_{ik}(\mathbf{t} - \delta_{ik})) \right) \cdot I(\mathbf{t} > \delta_{ik}) \right\} + \boldsymbol{\epsilon}_i, \quad (4.8)$$

$$\boldsymbol{\theta}_i \sim \mathcal{N}_q(\mathbf{0}, \Sigma) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

This model has 12 structural parameters $\boldsymbol{\theta}_i = [\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \beta_{i1}, \beta_{i2}, \beta_{i3}, \gamma_1, \gamma_2, \gamma_3, \delta_{i1}, \delta_{i2}, \delta_{i3}]^T$, and their interpretation is discussed in §3.2. In (4.8), we allow all 12 parameters to be different for each curve (each of the components of $\boldsymbol{\theta}_i$ is indexed by i). As in the b-spline model, estimating a large number of random effects is computationally challenging and not useful for monitoring purposes. Thus we seek only a subset of $\boldsymbol{\theta}_i$ as random effects.

The number and the choice of random effects should be set based on process knowledge, obtained from field experts. If such information is limited, we can use conclusions drawn from PCA. Recall from the PCA results in §2.3 that the most common causes of variability in the force profiles appear to be:

- vertical shift in the long flat segment at times 25 through 75;
- vertical shift in the flat end part at times 85 through 100;
- the times at which the three major parts of the curve begin to change;
- amount of decay in the long flat segment at times 25 through 75;
- extent of curvature in the profile at times 10 through 20.

For illustrative purposes, we focus on the first three points on this list, each corresponding to specific parameters in the nonlinear DE model. That is we have chosen five gain and reaction-time parameters α_{i2} , α_{i3} , δ_{i1} , δ_{i2} , and δ_{i3} to be random effects in our model. In an earlier discussion, we referred to these as the level parameters, because they determine the amount of vertical and horizontal shift in the force curves. In the context of model (4.3), this particular choice of random effects implies setting

$$\boldsymbol{\theta}_i = [\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3, \delta_1, \delta_2, \delta_3]^T + [0, \eta_{i1}, \eta_{i2}, 0, 0, 0, 0, 0, 0, \eta_{i3}, \eta_{i4}, \eta_{i5}]^T$$

or equivalently $\boldsymbol{\theta}_i = \boldsymbol{\mu} + Z\boldsymbol{\eta}_i$, where $\boldsymbol{\mu} = [\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3, \delta_1, \delta_2, \delta_3]^T$, Z is a 12×5 sparse matrix of zeros everywhere except for entries (2,1), (3,2), (10,3), (11,4) and

(12,5), which are equal to 1, and $\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \eta_{i3}, \eta_{i4}, \eta_{i5}]^T$.

Estimating these quantities is complicated by the fact that $f(\boldsymbol{\theta}_i)$ is nonlinear in terms of the structural parameters. Due to the complicated nature of the model, even when the variance components σ^2 and Σ are known, there are no closed-form solutions for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\eta}}_i$. Instead these must be calculated via numerical optimization. The usual approach is an adaptation of the Gauss-Newton technique, which works iteratively by linearizing $f(\boldsymbol{\theta}_i)$ using a Taylor series expansion about some initial value (as shown later in §4.4) and finding ML or REML estimates of the parameters under the approximation. These replace the initial values and the process is repeated in this fashion until convergence is attained. Detailed descriptions of the Gauss-Newton approach to nonlinear regression can be found in Gallant (1975), Bates & Watts (1988) and Greenwood (2004), with discussions in the context of mixed models developed in Lindstrom & Bates (1990), Wolfinger (1993) and Pinhero & Bates (2001).

4.4 Distributional Properties of the Predicted Random Effects

To summarize, for the linear model with known variance components in $V = B_2 \Sigma B_2^T + \sigma^2 \mathbf{I}_n$, the estimated fixed and the predicted random effects are

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= (B_1^T V^{-1} B_1)^{-1} B_1^T V^{-1} \bar{\mathbf{y}}; \\ \hat{\boldsymbol{\eta}}_i &= \Sigma B_2^T V^{-1} (\mathbf{y}_i - B_1 \hat{\boldsymbol{\mu}}) = \Sigma B_2^T V^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}).\end{aligned}$$

where $\mathbf{y}_i \sim \mathcal{N}_n(B_1 \boldsymbol{\mu}, V)$.

Thus the sampling distribution of the predicted random effects is as stated below.

Result 4.4.1 *Let $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ denote a collection of m independent profiles modeled by the linear mixed-effects model (4.2). Then for σ^2 and Σ known,*

- $\hat{\boldsymbol{\mu}} \sim \mathcal{N}_p(\boldsymbol{\mu}, (m B_1^T V^{-1} B_1)^{-1})$ and
- $\hat{\boldsymbol{\eta}}_i \sim \mathcal{N}_q(\mathbf{0}, \frac{m-1}{m} \Sigma B_2^T V^{-1} B_2 \Sigma)$ independently for $i = 1, \dots, m$.

The result is a direct consequence of the fact that every linear combination of multivariate normals is itself normal.

When variance components are not known, they can be replaced by their estimated values. Under such circumstances, Result 4.4.1 will be an approximation. Since none of the MLEs have closed-form expressions, it is not clear how to prove this mathematically. However in §5.3 we will show empirical results based on the analysis of real data that seem to support this claim.

Determining large-sample properties for the nonlinear DE model is even more challenging since there are no closed-form solutions for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\eta}}_i$ even when variances are known. Some progress can be made by linearizing $f(\boldsymbol{\theta}_i)$ and then applying Result 4.4.1 as an approximation. Using first-order Taylor series expansion of a nonlinear vector function $f(\boldsymbol{\theta}_i)$ about some fixed values of the structural parameters denoted by $\boldsymbol{\theta}_i^*$, we get

$$f(\boldsymbol{\theta}_i) \approx f(\boldsymbol{\theta}_i^*) + J(\boldsymbol{\theta}_i^*)^T(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)$$

where $J(\boldsymbol{\theta}_i^*) = \frac{df(\boldsymbol{\theta})}{d\boldsymbol{\theta}}|_{\boldsymbol{\theta}_i=\boldsymbol{\theta}_i^*}$ is a $p \times n$ matrix whose rows correspond to partial derivatives of $f(\boldsymbol{\theta}_i)$ with respect to $\boldsymbol{\theta}_i$ evaluated at $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^*$. It follows from (4.3) that

$$\mathbf{y}_i \approx f(\boldsymbol{\theta}_i^*) - J(\boldsymbol{\theta}_i^*)^T\boldsymbol{\theta}_i^* + J(\boldsymbol{\theta}_i^*)^T\boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i,$$

or equivalently

$$\mathbf{y}_i - f(\boldsymbol{\theta}_i^*) + J(\boldsymbol{\theta}_i^*)^T\boldsymbol{\theta}_i^* \approx J(\boldsymbol{\theta}_i^*)^T\boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i.$$

Setting $\mathbf{y}_i^* = \mathbf{y}_i - f(\boldsymbol{\theta}_i^*) + J(\boldsymbol{\theta}_i^*)^T\boldsymbol{\theta}_i^*$ and substituting $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \boldsymbol{\eta}_i$, we get the linearized mixed effects model similar to (4.2):

$$\mathbf{y}_i^* \approx J(\boldsymbol{\theta}_i^*)^T\boldsymbol{\mu} + J(\boldsymbol{\theta}_i^*)^TZ\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \text{for } i = 1, \dots, m$$

$$\boldsymbol{\eta}_i \sim \mathcal{N}_q(\mathbf{0}, \Sigma) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n).$$

All of the usual results from the linear case roughly extend here. For example, when variances are known and $\boldsymbol{\theta}_i^*$ is some fixed estimate of the structural parameters at the present stage of the iteration, approximate estimates of the fixed effects and predictions of

the random effects are given by

$$\hat{\boldsymbol{\mu}} = [J(\boldsymbol{\theta}_i^*)V^{-1}J(\boldsymbol{\theta}_i^*)^T]^{-1}J(\boldsymbol{\theta}_i^*)V^{-1}\bar{\mathbf{y}}^*$$

and

$$\hat{\boldsymbol{\eta}}_i = \Sigma Z^T J(\boldsymbol{\theta}_i^*)V^{-1} [\mathbf{y}_i^* - J(\boldsymbol{\theta}_i^*)^T \hat{\boldsymbol{\mu}}]$$

respectively (Lindstrom & Bates 1990 and Wolfinger 1993, pg. 793). This leads to the following claim, which is also a direct consequence of the fact that every linear combination of multivariate normals is itself normal.

Result 4.4.2 *Let $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ denote a collection of m independent profiles modeled by a nonlinear mixed-effects model (4.3). If σ^2 and Σ are known and $J(\boldsymbol{\theta}_i^*) = \frac{df(\boldsymbol{\theta})}{d\boldsymbol{\theta}}|_{\boldsymbol{\theta}_i=\boldsymbol{\theta}_i^*}$ for some known true value $\boldsymbol{\theta}_i^* = \boldsymbol{\mu}^* + \boldsymbol{\eta}_i^*$, sampling distributions of the structural parameters are approximated by*

- $\hat{\boldsymbol{\mu}} \sim \mathcal{N}_p(\boldsymbol{\mu}, (mJ(\boldsymbol{\theta}_i^*)V^{-1}J(\boldsymbol{\theta}_i^*)^T)^{-1})$, and
- $\hat{\boldsymbol{\eta}}_i \sim \mathcal{N}_q(\mathbf{0}, \frac{m-1}{m}\Sigma Z^T J(\boldsymbol{\theta}_i^*)V^{-1}J(\boldsymbol{\theta}_i^*)^T Z\Sigma)$ independently for $i = 1, \dots, m$.

Since $J(\boldsymbol{\theta}_i^*)$ depends on the true (unknown) value of $\boldsymbol{\theta}_i$, an approximation $\hat{J} = \frac{1}{m}\sum_i J(\hat{\boldsymbol{\theta}}_i)$ can be used. Unknown variance components would also be replaced by their sample ML or REML estimates. While there is no hope of determining the exact distribution of $\hat{\boldsymbol{\eta}}_i$ after so many approximations, in Chapter 5 we present empirical evidence suggesting that (similarly to the linear case) Result 4.4.2 generalizes approximately in the presence of a large amount of data.

Chapter 5

Profile Monitoring

Recall that our ultimate goal is to set forth statistical machinery for detecting curves that are “unusual” during the course of the valve seat insertion process. Process monitoring allows us to do this. The usual course of action in this procedure consists of two phases (Ryan, 1989). In Phase I, a base set of observations is selected as representative of normal operation, and scrutinized for unusual behavior. Detected outliers (if any) are investigated and removed if and only if there are attributable causes of their abnormality. The base set is then re-examined, and control limits indicating bounds for “normal” or “in-control” activity are calculated. Phase II of the analysis involves assessing new curves from ongoing production using the control limits obtained from the base set. The process is flagged as out-of-control whenever new observations fall outside the control limits. Unusual observations are then investigated for potential sources of the problem, and corresponding adjustments to the process are made.

The fact that we are dealing with functional observations complicates the task. Monitoring profiles as high-dimensional multivariate observations is challenging in practice due to such issues as correlation between neighboring values within a single curve, and loss of power as a result of deteriorating covariance estimation in increasing dimensions. A natural extension of the process-monitoring procedure to functional data is to monitor estimated structural parameters instead of the observed data (Williams et. al., 2005). In this chapter, we describe a procedure for monitoring the predicted random effects of a functional observation under the mixed-effects model.

As discussed in Chapter 4, we distinguish between the two different ways in which the fixed and random effects are estimated by calling the former estimates and the latter predictions.

5.1 Chart for Individual Curves

Starting with m curves in Phase I, let $\hat{\boldsymbol{\eta}}_i$ denote predicted random effect vectors for each curve approximately following a $\mathcal{N}_q(\mathbf{0}, W)$ distribution. Then, assuming that $q > 1$, a Hotelling T^2 measure of atypical behavior for each curve is given by:

$$T_i^2 = \hat{\boldsymbol{\eta}}_i^T W^{-1} \hat{\boldsymbol{\eta}}_i.$$

Large values of this test statistic imply that $\hat{\boldsymbol{\eta}}_i$ is far from its expected value of zero, and corresponding curves would be flagged as “unusual”. To define large, we must determine the distribution of T_i^2 .

If $\hat{\boldsymbol{\eta}}_i \sim \mathcal{N}_q(\mathbf{0}, W)$ exactly, this is trivial. Let $W^{-1/2}$ denote a $q \times q$ symmetric standard error matrix of $\hat{\boldsymbol{\eta}}_i$, obtained using eigen-decomposition of W . Then $\boldsymbol{\zeta}_i = W^{-1/2} \hat{\boldsymbol{\eta}}_i \sim \mathcal{N}_q(\mathbf{0}, I_q)$, so that $T_i^2 = \boldsymbol{\zeta}_i^T \boldsymbol{\zeta}_i$ is a sum of q standard normals squared. It follows that $T_i^2 \sim \chi^2(q)$.

Of course W is unknown and the exact distribution of $\hat{\boldsymbol{\eta}}_i$ cannot be determined. A natural solution is to approximate W using a sample covariance matrix of the random predictions:

$$\hat{W} = \frac{1}{m-1} \sum_{i=1}^m (\hat{\boldsymbol{\eta}}_i - \bar{\boldsymbol{\eta}})(\hat{\boldsymbol{\eta}}_i - \bar{\boldsymbol{\eta}})^T,$$

where $\bar{\boldsymbol{\eta}}$ is the average of the predicted random effects. Another possibility is to estimate W by \tilde{W} , where \tilde{W} is a model-based estimate of the covariance matrix for the predicted random effects, found by fitting a mixed-effects model using all of the Phase I data. Based on Result 4.4.1, $\tilde{W} = \frac{m-1}{m} \hat{\Sigma} B_2 \hat{V}^{-1} B_2^T \hat{\Sigma}$ in the linear case, where B_2 is an $n \times q$ matrix of basis functions used to fit the random effects. Similarly for the nonlinear case, using Result 4.4.2, we could set $\tilde{W} = \frac{m-1}{m} \hat{\Sigma} Z^T \hat{J} \hat{V}^{-1} \hat{J}^T Z \hat{\Sigma}$, an approximation generated by linearizing the DE model as stated in §4.4. As before, \hat{J} denotes a $p \times n$ matrix of partial derivatives

of $f(\boldsymbol{\theta})$ at estimated values of the parameters.

Based on simulation results and for the insertion data, both \tilde{W} and \hat{W} produce similar estimates of W . We chose to use \hat{W} , because it is simple and easy to understand. Other possible covariance estimators discussed in Williams et. al. (2005) but not explored in this thesis include the successive differences estimator (Hawkins & Merriam, 1974) and the minimum volume ellipsoid estimator (Rousseeuw, 1984).

Using either \tilde{W} or \hat{W} in place of W , the chi-squared distribution of T_i^2 can be approximated by an F distribution with q and $m - q$ degrees of freedom. This is formalized in the following extension of a well-known result from multivariate statistics (Johnson & Wichern 1992, pg. 237).

Result 5.1.1 *Suppose that $\hat{\boldsymbol{\eta}}_i \sim \mathcal{N}_q(\mathbf{0}, W)$ is an unbiased estimator of a $\boldsymbol{\eta}_i$ ($i = 1, \dots, m$ where m is the total number of profiles) and let $T_i^2 = \hat{\boldsymbol{\eta}}_i^T \hat{W}^{-1} \hat{\boldsymbol{\eta}}_i$ such that $\hat{W} \approx W$ as $m \rightarrow \infty$. Then for curve i ,*

$$F_i = \frac{m - q}{q(m - 1)} T_i^2 \sim F(q, m - q).$$

It follows that curve i is unusual if $F_i \geq F_\alpha(q, m - q)$ for some significance level α . As a general rule throughout the thesis we let $\alpha = 0.0027$. Combining all of the ideas presented thus far, our approach to monitoring individual curves in the force exertion data can be summarized as follows:

PHASE I

1. Select a base set of m curves. Fit a mixed model to these curves. Use it to obtain predictions of the random effects $\hat{\boldsymbol{\eta}}_i$ for each curve and an estimate of the associated covariance W .
2. Calculate Hotelling T^2 statistics for each curve, $T_i^2 = \hat{\boldsymbol{\eta}}_i^T \hat{W}^{-1} \hat{\boldsymbol{\eta}}_i$. Use a QQ plot and/or a Kolmogorov-Smirnov test to verify that $F_i = \frac{m - q}{q(m - 1)} T_i^2 \sim F(q, m - q)$.

3. Confirm that there are no unusual curves in the base set by examining a control chart of the test statistics. Investigate observations corresponding to $F_i \geq F_\alpha(q, m - q)$ for unusual behavior. Remove these if and only if the production process can be adjusted to avoid such observations in the future. $F_\alpha(q, m - q)$ is called an upper control limit (UCL) because it provides an upper bound for plausible values of F_i at the α level of significance.

PHASE II

1. For each new curve, predict its random effect $\hat{\boldsymbol{\eta}}_i$ using variance estimates obtained in Phase I.
2. Calculate Hotelling T^2 statistics for each new profile using \hat{W} from Phase I.
3. Verify that the process is in-control by examining control charts of the test statistics. Flag observations that fall outside of the UCL established in Phase I as “unusual”.

5.2 Chart for Subgrouped Curves

Subgrouping is an important process-monitoring technique for situations when there are small persistent changes within a small subgroup of consecutively observed curves. The basic idea is to monitor the averages of subgrouped data rather than the individual observations themselves (Ryan, 1989), which allows the monitoring procedure to be more sensitive to small departures.

An extension of profile-monitoring to subgrouped data is straightforward. Let m denote the total number of curves in Phase I. Fit a mixed-effects model to these data and obtain predicted random effects for each curve. Group these into k sets of equal size $m_g = m/k$, and calculate corresponding subgroup averages of the predicted random effects. Proceed to monitor the subgrouped averages as if they correspond to individual curves.

The following result formalizes profile-monitoring for subgrouped data (Ryan, 1989).

Result 5.2.1 *Suppose that $\hat{\boldsymbol{\eta}}_i \sim \mathcal{N}_q(\mathbf{0}, W)$ is an unbiased predictor of $\boldsymbol{\eta}_i$ ($i = 1, \dots, m$).*

Group these into k sets S_1, S_2, \dots, S_k of equal size m_g , and let

$$\bar{\hat{\boldsymbol{\eta}}}_g = \frac{1}{m_g} \sum_{i \in S_g} \hat{\boldsymbol{\eta}}_i \quad \text{for } g = 1, \dots, k.$$

denote the means of predicted coefficients for each subgroup. Then a Hotelling T^2 criterion for testing that the functional mean of the curves in subgroup g is “unusual” is given as

$$T_g^2 = \frac{m}{k} \left[\bar{\hat{\boldsymbol{\eta}}}_g^T \hat{W}^{-1} \bar{\hat{\boldsymbol{\eta}}}_g \right]$$

where $\hat{W} \approx W$ as $m \rightarrow \infty$, such that

$$F_g = \frac{m - k - q + 1}{mq - kq - mq/k + q} T_g^2 \sim F(q, m - k - q + 1).$$

Phase I and II steps of the analysis for subgrouped data extend directly from §5.1 by replacing $\hat{\boldsymbol{\eta}}_i$, F_i and T_i^2 with $\hat{\boldsymbol{\eta}}_g$, F_g and T_g^2 as defined in Result 5.2.1, and setting the UCL equal to $F_\alpha(q, m - k - q + 1)$. For the subgrouped data, we let \hat{W} equal the mean of the within-subgroup sample covariance estimates. This is consistent with what is done in multivariate process-monitoring (Ryan, 1989). Another possibility is to let $\hat{W} = \tilde{W}/m_g$, which does not seem to change the results for our data, but might be more appropriate depending on the problem at hand.

An alternative to the subgrouped chart formulated here is to subgroup and average the curves first, then fit a mixed effects model to the subgroup averages, and monitor predicted random effects obtained from the model as you would individual observations. This is not explored here.

5.3 Example

As stated in §2.2, we examine 2,768 valve insertions from January and 1,008 from the last six days in February. February data are used in Phase I and January data in Phase II of the analysis. Note that the process was started on January 7. Although monitoring the

past is not representative of what would happen in reality, this was done for illustrative purposes only, as we expected that the production process stabilized as time progressed. This rationale was discussed in §2.2 and is revisited in §5.4.2.

Our goal is to identify unusual insertions made in January (if any) using the profile-monitoring techniques developed in this thesis. This is accomplished by following the steps outlined in §5.2. Results from the two phases of the subgrouped analysis are summarized below. To save space and avoid repetition only results for the first valve are presented here. R statistical software was used for computation, utilizing the `nlme` package to fit mixed effects models (Pinheiro & Bates, 2001).

5.3.1 Model Specification

Both linear (4.2) and nonlinear (4.8) mixed-effects models are considered. In the linear b-spline case, 20 fixed and four random effects are fit. The number of fixed terms is chosen by examining R^2 goodness of fit measures from fitting a pure fixed effects model (see Figure 3.2). Using twenty b-splines provides an excellent fit of the overall trend in the curves, with the addition of more b-splines offering little improvement. Four random terms are specified for illustrative purposes only, as this seems adequate in capturing curve-specific trends (see Figure 4.2(a)).

For the nonlinear model, we use 12 fixed effects as dictated by the parameterization of the second order DE. Using the motivation in §4.3.2, five of these (α_{i2} , α_{i3} , δ_{i1} , δ_{i2} , and δ_{i3}) are chosen to have random components. The decision is based on the conclusions drawn from the exploratory analysis in §2.3. This is possible due to the interpretability of the parameters in the DE model, which allows random effects to be assigned to the parameters that characterize key sources of variability in the curves as summarized by the first few principal components.

Further discussion of model selection can be found in §5.4.1.

5.3.2 Phase I Analysis

The first part of monitoring involves modeling the February data. B-spline and nonlinear mixed models are fit, producing predicted random coefficients and their estimated covari-

ances. For b-splines, $p = 20$ fixed and $q = 4$ random parameters are fit. For the nonlinear model, there are $p = 12$ fixed parameters, of which $q = 5$ (α_{i2} , α_{i3} , δ_{i1} , δ_{i2} , and δ_{i3}) are chosen to have random components.

For purpose of illustration, for each model, $m = 1,008$ predicted random coefficient vectors are subgrouped into $k = 202$ sets of five parameter vectors per subgroup. An alternative not explored in this thesis would be to group together values across an entire day, or some other time period.

We fit each model to the base set and obtain the predicted random effects for each of the 1,008 curves. For each subgroup ($g = 1, \dots, 202$), corresponding F_g statistics are calculated as defined in Result 5.2.1. Our first goal is to verify that these test statistics follow an appropriate F-distribution. For b-splines, the null hypothesis is that

$$F = \frac{803}{3208} T^2 \sim F(4, 803).$$

The corresponding Kolmogorov-Smirnov test statistic and p-value are 0.0578 and 0.5096, meaning that the null hypothesis should not be rejected. That is, it is not unreasonable to assume that the test statistics F_g follow the $F(4, 803)$ distribution. For the nonlinear model, the test statistic and p-value are 0.0585 and 0.4931 respectively, leading to the same conclusion. Quantile-quantile plots under the linear and nonlinear models in Figures 5.1(a) and 5.1(c) respectively further support our conclusions.

Next, control charts are built for February data using distributional properties of the T^2 statistic based on subgrouped data. Figures 5.1(b) and 5.1(d) display these results for the linear and the nonlinear models. Dashed horizontal lines mark the upper control limits in each case. For b-splines, this is $F_{0.0027}(4, 2218) = 4.08$ at the 0.0027 significance level. For the nonlinear model, $F_{0.0027}(5, 2217) = 3.65$. The fact that all of the observations fall below these values support our assumption that the base set is “in-control”.

5.3.3 Phase II Analysis

The second stage is to examine January data for potential faults in the online production process. Covariance estimates from Phase I are used to calculate predicted values of the random coefficients for both linear b-spline and nonlinear models. In each case, these are

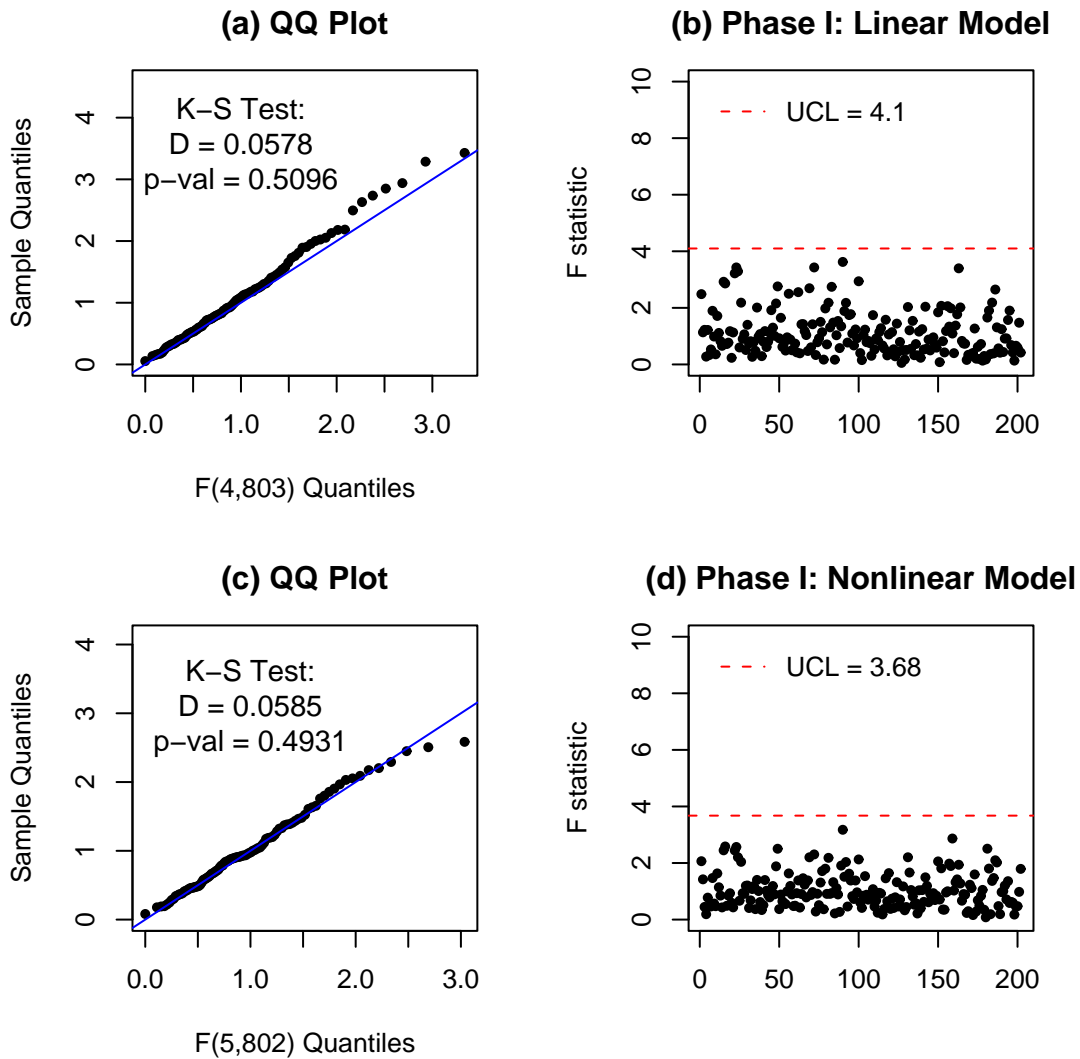


Figure 5.1: Quantile-quantile plots comparing sample distributions of the test statistics under the (a) linear and (c) nonlinear models to the F-distribution. Phase I control charts for the two models are in (b) and (d) respectively.

split into subgroups of five and corresponding subgrouped T^2 statistics obtained as stated in Result 5.2.1.

Subgrouped control charts for the first valve are provided in Figures 5.2(a) and 5.2(c). The first plot is for the b-spline model and the second is for the nonlinear model. As before, observations that exceed upper control bounds for each model (marked as dashed lines) are regarded to be “out-of-control”.

One striking feature of these results is a number of the flagged insertions occur during the first few days of January. This makes perfect sense. Problems are likely to occur in the initial stages of production as the engine-assembly process is set up. Over time these problems tend to be resolved in an effort to improve performance - an intuitive notion that seems to be supported by the control charts. Two other less prominent sets of outlying curves occur in mid-January, and during the last week of that month.

Averages of subgrouped curves that are extreme outliers ($F > 6$) for the linear and nonlinear models are displayed to the right of the control charts in Figures 5.2(b) and 5.2(d) respectively. In the control charts, these are circled in black. Consistency between the fundamentally different b-spline and nonlinear models is reassuring. The seven most outlying subgroups of curves flagged by at least one or both of the linear and nonlinear procedures include two on January 7th, four on the 8th, and one on the 9th.

Looking at the shapes of the averages of the most outlying subgroups of curves, they appear to differ from the rest in the dip of the middle flat part of the profile and have a relatively small amount of force gain at the end. Process experts might be able to attribute these features to specific faults in the manufacturing procedure. Visualization techniques may yield insight into why curves are flagged as outliers. For example, a plot comparing the outlying curve to the mean curve might highlight discrepancies. Also, treating the predicted random effects $\hat{\eta}$ for each curve as “data” and applying exploratory graphical analysis such as parallel coordinate plots and scatterplot matrices might indicate which random effect made the curve appear unusual.

Results for the remaining valves are very similar. Across all valves, the largest outliers tend to correspond to insertions made during the first few days of January.

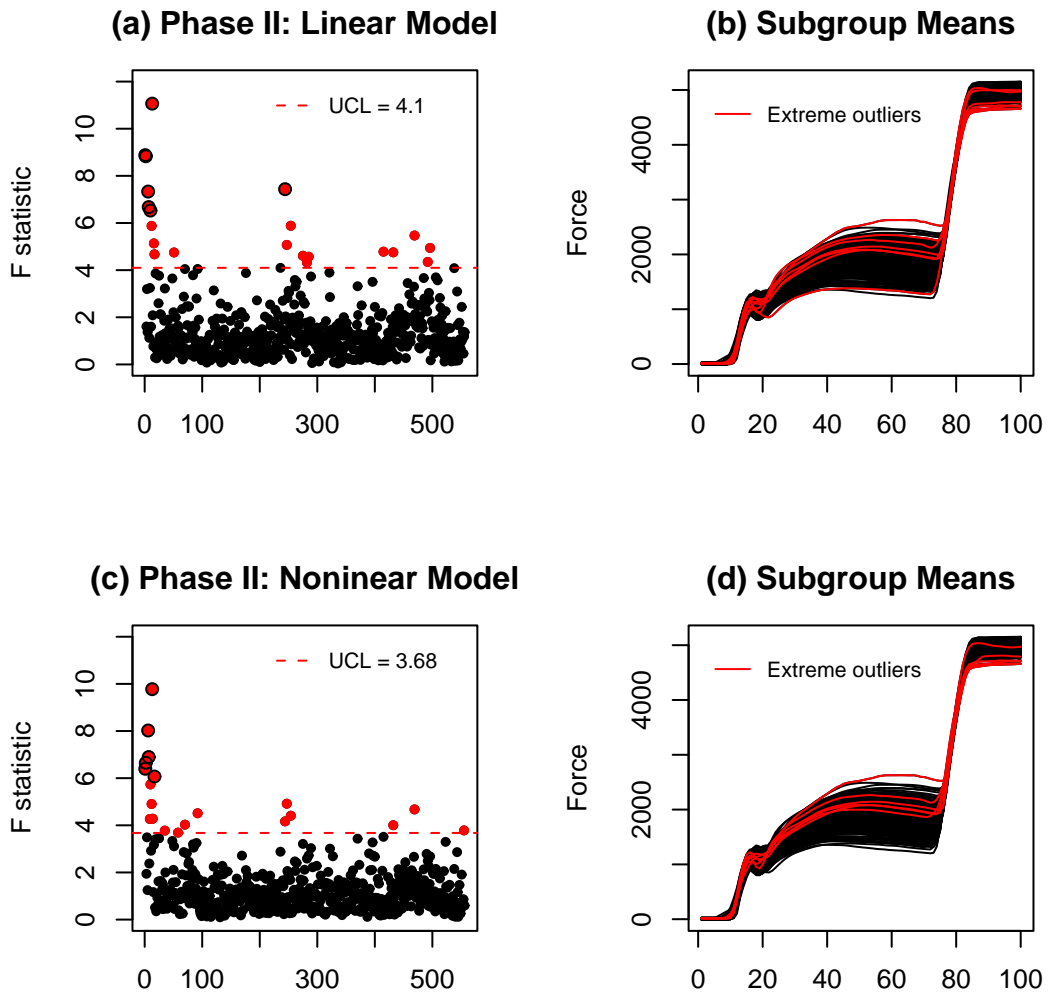


Figure 5.2: Phase II control charts for valve one under the (a) linear b-spline and (c) nonlinear models. Functional means of the outlying subgroups of profiles flagged by each chart are displayed in red alongside all of the subgroup means (black lines) in (b) and (d).

5.4 Discussion

The goal of this chapter was to address key issues involved in monitoring functional process data such as the force exertion example, and to develop tools for accomplishing the task. Using linear and nonlinear mixed-effects models, we utilized Hotelling T^2 statistics to differentiate between “good” and “bad” insertions. Results based on the force exertion data indicate reasonable performance. Some of the underlying issues raised by the proposed methodology and future research are considered next.

5.4.1 Choosing Random Effects

When formulating the mixed effects model, it is not clear how many random effects should be chosen. For both b-spline and nonlinear models, we recommend restricting the number of random effects to be as small as possible. This reduces the computational cost of the procedure, and allows us to monitor compact summaries of the curves in a low dimension.

For nonlinear models, it is also difficult to know which of the parameters should have random components. We suggest using both process knowledge and conclusions drawn from PCA in selecting the fixed parameters that should have associated random components. Another possibility is to fit 12 different models, where each model has a single random effect. We could then use model-selection criteria such as AIC or BIC to identify the best-fitting few, one at a time. Random effects associated with these fits are then selected to be used in the final model. Alternatively, we could adapt a step-wise model-selection approach or use a straightforward fixed-effect regression fit to identify which of the parameters are the most significant, then make these random in a re-fitted mixed-effects model. In both cases, it is important to standardize the parameters as they are not measured on the same scale.

In general, deciding how many random effects should be used and which parameters to make random is a challenging problem. The approach of shrinking a large number of random effects taken by Morris and Carrol (2006) does not seem to be in the spirit of a parsimonious representation of the data. Rice and Wu (2001) suggest the use of information criteria (AIC, BIC, cross-validation) for selection of the number of equally-spaced splines used in a random effects model. Such a technique might be appropriate for the approach

considered here.

5.4.2 Choice of Phase I Data

One problem common to all control chart applications is the choice of a base set in Phase I. For the force exertion data, we monitored the January data and used all 1,008 insertions made in the last six days of February as a baseline set indicative of a well-behaved process. This does not represent what would be done in practice (historical not future data is used as a baseline), but was done for demonstration purposes only due to the suspicion that the insertion-process stabilized as time went on. Even so, with the limited amount of information that we have about our data, it is unclear whether or not February observations are truly representative of “normal operation”. One helpful tactic is to re-fit the combined data after a while and adjust control limits using appropriate sampling and/or subgrouping schemes.

5.4.3 Sequential Charts

The basic idea of monitoring the predicted random effects of a mixed-effects model can easily be extended to include a wide range of other multivariate charts. These include sequential charts and charts to monitor the covariance matrix, with the type of approach used depending on the context of the problem.

We consider one such possible extension - the use of exponentially weighted moving average (EWMA) monitoring procedure for profile data. Following a multivariate version of the EWMA chart developed by Lowry et. al. (1992), our MEWMA procedure monitors weighted averages of the estimated random effects in a way that gives less weight to observations further in the past. More formally, using Lowry’s approach, we monitor

$$MEWMA_i = \mathbf{Z}_i^T \mathbf{W}_Z^{-1} \mathbf{Z}_i \geq UCL.$$

where

$$\mathbf{Z}_i = \lambda \hat{\boldsymbol{\eta}}_i + (1 - \lambda) \mathbf{Z}_{i-1}; \quad \mathbf{Z}_0 = \mathbf{0}; \quad (5.1)$$

and

$$W_Z = Cov(\mathbf{Z}_i) = \frac{\lambda(1 - (1 - \lambda)^{2i})}{2 - \lambda} * W.$$

As before $\hat{\boldsymbol{\eta}}_i$ denotes the i^{th} predicted random effect ($i = 1, \dots, m$) and W is replaced by an appropriate estimate \hat{W} . A corresponding extension to subgrouped data involves using $\bar{\hat{\boldsymbol{\eta}}}_g$ and $W = \hat{W}$, where \hat{W} is the average of the within-subgroup covariance matrices.

A phase II multivariate EWMA chart under the linear model is provided in Figure 5.3. For illustration, the weight parameter is set to $\lambda = 0.2$ and the upper control limit is estimated using MINITAB to provide a false alarm rate of roughly 0.02%. Since the distribution of the \mathbf{Z}_i s is unknown, control limit calculations are complex and beyond the scope of this chapter. Details can be found in Lucas & Saccucci (1990).

Results in Figure 5.3 match our earlier findings in §5.3.3, with the majority of outliers flagged at the beginning of January. A key distinction of the MEWMA chart is that it retains information about previously monitored observations. By weighing past information, the MEWMA chart is more sensitive to small shifts in the process, which explains the relatively smooth shape to the plotted results. The idea is emphasized by connecting adjacent values in Figure 5.3.

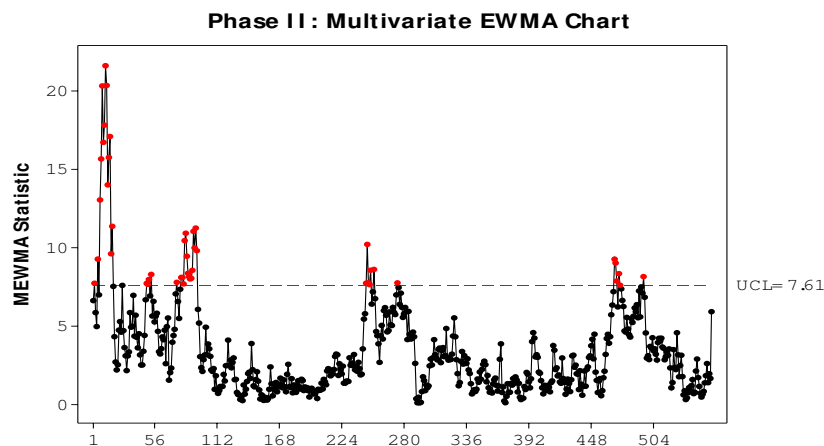


Figure 5.3: Phase II multivariate EWMA control chart under the linear model.

5.4.4 Charts Using Both Fixed and Random Effects

The underlying theme throughout this thesis has been the monitoring of random effects parameters. Another possibility is to monitor a measure of overall variation unexplained by the fitted mixed effects model, such as the residual sum of squares, in addition to the random effects. Monitoring the errors provides a way of detecting more general phenomena that are not represented by the model parameters.

5.4.5 Charts for Multiple Valves

With respect to the force exertion example, we focused our analysis on data from only one of the eight valve seat insertions per head. One possibility would be to extend some of the ideas developed in this thesis to models that consider all eight valves simultaneously.

The easiest way to include information on multiple valves is to include a fixed valve effect in the model. For b-splines this is equivalent to fitting

$$\mathbf{y}_{ij} = B_1\boldsymbol{\mu} + B_2\boldsymbol{\gamma}_j + B_3\boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij},$$

$$\boldsymbol{\eta}_{ij} \sim \mathcal{N}_q(\mathbf{0}, \Sigma) \quad \text{and} \quad \boldsymbol{\epsilon}_{ij} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

where $i = 1, \dots, m$ indexes the insertion, and $j = 1, \dots, 8$ indexes the valve number.

Both the $\boldsymbol{\mu}$ and the valve parameter $\boldsymbol{\gamma}_j$ in this model are fixed, which allows us to control for both the overall trend and any systematic valve effects in the data. The same basic idea can be extended to controlling for the ram effect (we have two rams, each making four valve insertions) by including a fixed effect for the ram rather than the valve.

Chapter 6

Model-Based Clustering

Up until now we have assumed that all production data can be classified in one of two categories: in-control or out-of-control curves. Another approach is to allow for the fact that there may be multiple modes of in-control and out-of-control operation taking place during the production process. Clustering is one such method of grouping data (or in our case - curves) in some meaningful way. In the absence of any other information except for the recorded response values, it provides a way of extracting latent information from the data by organizing it into similar groups.

For functional data, standard clustering techniques can be implemented by treating collections of curves (after suitable dimension reduction) as multivariate observations in n -dimensional space ($n > 1$). The idea is that the clusters may correspond to different modes of operation. For example, data with a large number of similar curves and a remote, small cluster of curves may indicate a single most predominant process and a series of less common operations. Further insight can be gathered by comparing mean curves for each cluster to knowledge of the shape of in-control curves.

In the next two chapters, we review some of the existing techniques for clustering functional data. New methodology is developed in Chapter 8.

6.1 Background

There are several techniques available for clustering data in high dimensions. Two algorithmic ones include hierarchical and partitioning methods. The hierarchical (or agglomerative) approach works by initially treating each observation as a separate cluster and then gradually merging subsets of the data until they form a single set. This sequence of nested clusters can be visualized with a tree-like structure known as a dendrogram. In a second step, branches of the dendrogram are “cut” to result in a K -cluster separation. Partitioning (e.g., K-Means) methods work in a somewhat opposite direction by requiring the user to specify K , the number of clusters to be formed, and subsequently arranging the data into that many groups. This is accomplished by assigning data to clusters that minimize a sum of squared Euclidean distances between the observed values and the cluster centroids to which they belong. The partitioning approach differs from hierarchical methods in that the clusters are not necessarily nested. This is because clusters of size K are not formed by merging two clusters from the $K + 1$ solution as is the case in agglomerative clustering.

A third possibility is model-based clustering (MBC), an approach that assumes a parametric model for the formation of clusters. In Gaussian MBC, for example, each observation is assumed to be sampled from a mixture of K multivariate normal densities. A key appeal of this technique over algorithmic methods is that the presence of a model makes it possible to account for the functional structure of our data. For example, James & Sugar (2003) model functional data using b-splines and assume that the coefficients of this model (rather than the data itself) are distributed as Gaussian mixtures. MBC does present some challenges, including a heavy computational cost and a large number of parameters requiring estimation. This is because the number of unknown parameters in a model is directly linked to curve resolution n . The higher the resolution, the more parameters there are to estimate, making it difficult to identify structure in the data. Estimating $n \times n$ covariance matrices also becomes a challenge, as the number of elements in the matrices increases with n .

In this chapter we illustrate how Gaussian MBC can be used on functional data. We begin with a discussion of techniques for dealing with the high-dimensional aspect of curve data in §6.2, as this poses a particular difficulty in clustering. Fraley & Raftery’s (2002) frequentist implementation of MBC is described next in §6.3. An extension of this model

to functional data is introduced in §6.4, and illustrated using the valve seat insertion data in §6.5. This introduction to MBC serves as a reference and a building block for the alternative clustering models described in Chapters 7 and 8.

6.2 Dimension Reduction

One challenge in applying model-based clustering to the valve seat insertion data is that within-cluster covariance matrices are high dimensional (100×100). This poses a difficulty as the number of parameters that must be estimated increases greatly with the dimensionality of the data. An unrestricted $n \times n$ covariance matrix has $(n^2 + n)/2$ free parameters, and thus $K(n^2 + n)/2$ covariance parameters must be estimated in a K -cluster model. For large n , estimation is particularly difficult if some of the clusters are small, and there are fewer observations than there are unknown variables.

Modeling challenges involving the dimensionality of the observed data are often referred to as the curse of dimensionality. James and Sugar (2003) suggest filtering and regularization as two ways of dealing with the curse of dimensionality in the context of mixture models. The primary difference between the two methods lies in the way in which the covariance structure is estimated during the formation of clusters. In regularization, estimates of the within-cluster covariance matrices Σ_k are constructed directly from the data, and then improved upon by interpolating between a pooled estimate of the overall covariance matrix Σ and the estimates for each cluster. That is,

$$\hat{\Sigma}_k(\lambda) = (1 - \lambda)\hat{\Sigma}_k + \lambda\hat{\Sigma},$$

for some $\lambda \in [0, 1]$ and $\hat{\Sigma} = (\hat{\Sigma}_1 + \dots + \hat{\Sigma}_K)/K$. In filtering, on the other hand, the observed curve \mathbf{y}_i is first projected onto a matrix of p basis functions $B = [\mathbf{b}_1(\mathbf{t})^T, \dots, \mathbf{b}_p(\mathbf{t})^T]$ such that

$$\mathbf{y}_i = B\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i,$$

where $p < n$ and $\boldsymbol{\epsilon}_i$ is assumed to have mean zero and constant variance. The p -dimensional terms $\boldsymbol{\eta}_i$ are estimated using linear least squares regression and used in place of the original data. Thus, in filtering, covariance matrix Σ_k describes variation in the coefficients

$\boldsymbol{\eta}_i$, not the original data. This estimation in a reduced-dimension space may prove simpler than regularization in the original high-dimensional space. Whereas regularization constrains the covariance structure which in turn leads to more stable parameter estimates, the advantage of filtering is that we are reducing the dimensionality of the problem before estimation even begins.

In earlier chapters of the thesis we have used filtering approaches to model the data. We shall continue to use this approach in curve clustering. That is, we model the data using $p = 20$ b-spline basis functions and assume a mixture distribution on the spline coefficients instead of the observed data. This means that each Σ_k is a 20×20 matrix, with $(20^2 + 20)K/2 = 210K$ unknown covariance components instead of $(100^2 + 100)K/2 = 5050K$. Other possible ways to implement filtering include principal component analysis and multidimensional scaling (DeSarbo et. al. 1991, Oh and Raftery 2000).

6.3 A Mixture Model for Multivariate Data

Banfield & Raftery (1993) describe the use of mixtures of multivariate normal densities to model clustered data. A more recent treatment of this technique is given in Fraley & Raftery (2002). In this section, we describe MBC as a tool for grouping curves observed as multivariate vectors of response values occurring at fixed time points.

Consider a set of m independent curves observed on an n -dimensional coordinate system $\mathbf{t} = [t_1, t_2, \dots, t_n]^T$ and denoted by $\mathbf{y}^m = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m\}$. For convenience and the sake of our application, we assume that \mathbf{t} refers to a time component, however the same inferences can be made by substituting a generic \mathbf{x} vector in place of \mathbf{t} .

Suppose that the curves in \mathbf{y}^m can be separated into K disjoint groups. Assuming that K is known for the moment, let \mathbf{z}_i denote a vector of cluster membership indicators for curve i , such that $z_{ik} = 1$ if the curve belongs to the k^{th} cluster and $z_{ik} = 0$ otherwise. Conditional on this label, assume that the observed vector \mathbf{y}_i comes from a multivariate

normal distribution with cluster-specific mean vector and covariance matrix of the forms

$$\boldsymbol{\mu}_k = \begin{bmatrix} \mu_k(t_1) \\ \vdots \\ \mu_k(t_n) \end{bmatrix} \quad \text{and} \quad \Sigma_k = \begin{bmatrix} \Sigma_k(t_1, t_1) & \dots & \Sigma_k(t_1, t_n) \\ \vdots & \ddots & \vdots \\ \Sigma_k(t_n, t_1) & \dots & \Sigma_k(t_n, t_n) \end{bmatrix}$$

respectively. That is,

$$\mathbf{y}_i | \{z_{ik} = 1\} \sim \mathcal{N}_n(\boldsymbol{\mu}_k, \Sigma_k) \quad (6.1)$$

independently for $i = 1, \dots, m$ curves belonging to one of $k = 1, \dots, K$ clusters.

Since we do not know which group each curve belongs to, the \mathbf{z}'_i s are unknown. Letting π_1, \dots, π_K denote probabilities of belonging to each cluster, such that

$$\pi_k = Pr(z_{ik} = 1) \quad \text{and} \quad \sum_{k=1}^K \pi_k = 1,$$

model parameters $\boldsymbol{\theta}_k = (\pi_k, \boldsymbol{\mu}_k, \Sigma_k)$ can be estimated by jointly maximizing the likelihood function for this mixture model:

$$L(\boldsymbol{\theta} | \mathbf{y}^m) = \prod_{i=1}^m \sum_{k=1}^K \pi_k \mathcal{N}_n(\mathbf{y}_i; \boldsymbol{\mu}_k, \Sigma_k) \quad (6.2)$$

with respect to $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$.

Given an observed vector associated with a new curve, Bayes' rule can be used to calculate the probability that the i^{th} curve belongs to cluster k :

$$p(z_{ik} = 1 | \mathbf{y}_i, \hat{\boldsymbol{\theta}}) = \frac{\hat{\pi}_k \mathcal{N}_n(\mathbf{y}_i; \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)}{\sum_{l=1}^K \hat{\pi}_l \mathcal{N}_n(\mathbf{y}_i; \hat{\boldsymbol{\mu}}_l, \hat{\Sigma}_l)}. \quad (6.3)$$

If a predicted label is desired, the cluster label k with the largest posterior probability (6.3) can be returned. Thus, we can set $z_{ik} = 1$ if the value of $p(z_{ik} = 1 | \mathbf{y}_i, \hat{\boldsymbol{\theta}})$ is the largest for all $k \in \{1, \dots, K\}$.

Maximization of (6.2) with respect to the model parameters in $\boldsymbol{\theta}$ is generally attained using the Expectation-Maximization (EM) algorithm. The solution is found by treating

\mathbf{z}_i as missing data, and $\mathbf{x}_i = \{\mathbf{y}_i, \mathbf{z}_i\}$ as independent identically distributed (iid) complete data. Then the likelihood function for complete data is given by

$$L_C(\boldsymbol{\theta}|\mathbf{y}^m, z^m) = \prod_{i=1}^m \prod_{k=1}^K \{\pi_k \mathcal{N}_n(\mathbf{y}_i; \boldsymbol{\mu}_k, \Sigma_k)\}^{z_{ik}}. \quad (6.4)$$

The EM algorithm proceeds by iterating through two steps. In the E-step, conditional expectations of the membership labels are calculated as

$$\hat{z}_{ik} = 1 \times p(z_{ik} = 1|\mathbf{y}_i, \hat{\boldsymbol{\theta}}) + 0 \times p(z_{ik} = 0|\mathbf{y}_i, \hat{\boldsymbol{\theta}}) = p(z_{ik} = 1|\mathbf{y}_i, \hat{\boldsymbol{\theta}}),$$

with the $p(z_{ik} = 1|\mathbf{y}_i, \hat{\boldsymbol{\theta}})$ as specified in (6.3) and components of $\hat{\boldsymbol{\theta}}$ taking on the current estimated values. Then, in the M-step, estimated membership labels are substituted into (6.4), which is maximized to obtain the remaining model parameters $\hat{\boldsymbol{\theta}}$. The process is repeated until parameter estimates from consecutive iterations offer little or no improvement. Provided reasonable starting values, the EM algorithm is likely to converge to the maximum likelihood estimates (MLEs), i.e. the maximum modes of the incomplete likelihood specified in (6.2) (see McLachlan & Peel 2000, §2.14).

6.3.1 Covariance Structure

MBC differs from other clustering methods in that it allows the clusters to vary in magnitude and appearance. Such relationships can be expressed in terms of various constraints on a decomposition of within-cluster covariance matrices. Specifically, an $n \times n$ matrix Σ_k can be written as

$$\Sigma_k = \lambda_k \Gamma_k \Lambda_k \Gamma_k^T \quad k \in \{1, \dots, K\} \quad (6.5)$$

where Γ_k is an orthogonal matrix of eigenvectors, λ_k is the first eigenvalue, and Λ_k is a diagonal matrix with elements $(1, \lambda_{2,k}/\lambda_k, \dots, \lambda_{n,k}/\lambda_k)$, which denote all of the eigenvalues scaled by λ_k . The three elements of the eigen-decomposition - λ_k , Λ_k and Γ_k - describe specific attributes of the k^{th} mixture component, corresponding to its volume, shape and rotation/orientation respectively. Each of these quantities can be constrained in different ways to allow varying levels of flexibility in the specification of the mixture components.

Label	Covariance structure	Description of clusters
EII	$\Sigma_k = \lambda I$	Spherical, equal volume
VII	$\Sigma_k = \lambda_k I$	Spherical, unequal volume
EEI	$\Sigma_k = \lambda \Lambda$	Diagonal, equal volume, equal shape
EVI	$\Sigma_k = \lambda \Lambda_k$	Diagonal, equal volume, varying shape
VEI	$\Sigma_k = \lambda_k \Lambda$	Diagonal, varying volume, equal shape
VVI	$\Sigma_k = \lambda_k \Lambda_k$	Diagonal, varying volume, varying shape
EEE	$\Sigma_k = \Sigma$	Ellipsoidal, equal volume, shape, and orientation
EEV	$\Sigma_k = \lambda \Gamma_k \Lambda \Gamma_k^T$	Ellipsoidal, equal volume and shape, varying in orientation
VEV	$\Sigma_k = \lambda_k \Gamma_k \Lambda \Gamma_k^T$	Ellipsoidal, equal shape, varying in volume and orientation
VVV	$\Sigma_k = \lambda_k \Gamma_k \Lambda_k \Gamma_k^T$	Ellipsoidal, varying in volume, shape and orientation

Table 6.1: Special cases of the covariance structure for the clusters that are implemented in R (Fraley & Raftery, 2002) .

Table 6.1 provides a list of possible restrictions placed on the covariance structures in the `mclust` library in R (Fraley & Raftery, 2006). The labeling convention used in the first column is composed of three letters, with the first describing the desired volume (λ_k), the second - the shape (Λ_k), and the third - the orientation (Γ_k) of the mixture components. The letters can take on three possible values:

- **E**, which implies that corresponding elements of the decomposition are equal across clusters (e.g., $\lambda_k = \lambda$ for all k);
- **V**, which allows the associated part of the decomposition to vary across clusters;
- **I**, represents an identity matrix and is only applicable to defining the shape and orientation (e.g., $\Gamma_k = I$ constrains the off-diagonal elements of the covariance matrix to be zero).

For example, EVI implies that $\lambda_k = \lambda$, $\Lambda_k = \Lambda_k$ and $\Gamma_k = I$, so that $\Sigma_k = \lambda \Lambda_k$, which means that the within-cluster covariance matrices are constrained to be diagonal, with corresponding mixture components of the same size but potentially different shape across clusters. Other special cases include spherically shaped clusters of varying radius ($\Sigma_k = \lambda_k I$), clusters with diagonal covariance matrices of varying magnitude ($\Sigma_k = \lambda_k \Lambda_k$),

and unrestricted elliptically-shaped clusters with a different Σ_k for each mixture component. Covariance structures associated with these three scenarios are increasing in complexity. Spherical clusters, for example, require a maximum of K parameters to estimate the covariance structure, whereas in the unrestricted case this is of order Kn^2 .

6.3.2 Model-Selection

Two prevalent issues in model-based cluster analysis are the choice of the covariance structure and the number of clusters to use. Banfield & Raftery (1993) suggest the use of Bayesian model selection in deciding upon each. Following their notation, suppose that we have a choice between J different models M_1, \dots, M_J . In our case, each of the models uses a different type of covariance structure and a different number of clusters. Further suppose that for each model M_j ($j = 1, \dots, J$) there is an associated prior probability of belonging to that model, $p(M_j)$. Then using Bayes' theorem,

$$p(M_j|\mathbf{y}^m) \propto p(\mathbf{y}^m|M_j)p(M_j)$$

where $p(M_j|\mathbf{y}^m)$ is a posteriori probability of model M_j given the data, \mathbf{y}^m . Banfield & Raftery further note that prior probabilities $p(M_j)$ are usually assumed to be equal and $p(\mathbf{y}^m|M_j)$, referred to as the integrated likelihood, is obtained by integrating over all possible values of the unknown parameters to be estimated in the model, $\boldsymbol{\theta}_j$. In other words,

$$p(\mathbf{y}^m|M_j) = \int p(\mathbf{y}^m|\boldsymbol{\theta}_j, M_j)p(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j. \quad (6.6)$$

The integrated likelihood in (6.6) can be difficult to compute directly, however Schwartz (1978) developed the Bayes' Information Criterion (BIC) as an approximation. The criterion consists of a log-likelihood evaluated at the MLE $\hat{\boldsymbol{\theta}}_j$, and a penalty term as follows:

$$BIC_j = 2 \log p(\mathbf{y}^m|M_j) \approx 2 \log p(\mathbf{y}^m|\hat{\boldsymbol{\theta}}_j, M_j) - v_j \log(m), \quad (6.7)$$

where v_j is the number of independent parameters to be estimated under model M_j . In this framework, the problem of selecting the best model reduces to picking a model that has the highest associated posterior probability given the data, $p(M_j|\mathbf{y}^m)$. When the priors

assigned to each model are equal, this is equivalent to selecting a model associated with the highest integrated likelihood, $p(\mathbf{y}^m|M_j)$. Selecting a model with the highest BIC (6.7) will aid in the selection of the best model.

As outlined in Raftery (1995), the expression in (6.7) is derived using a Taylor series expansion of $g(\boldsymbol{\theta}_j) = \log\{p(\mathbf{y}^m|\boldsymbol{\theta}_j, M_j)p(\boldsymbol{\theta}_j|M_j)\}$ about the MLE $\hat{\boldsymbol{\theta}}_j$, given by

$$\begin{aligned} g(\boldsymbol{\theta}_j) &\approx g(\hat{\boldsymbol{\theta}}_j) + (\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j)^T g'(\hat{\boldsymbol{\theta}}_j) + \frac{1}{2}(\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j)^T g''(\hat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j) \\ &= g(\hat{\boldsymbol{\theta}}_j) + \frac{1}{2}(\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j)^T g''(\hat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j), \end{aligned}$$

where $g'(\hat{\boldsymbol{\theta}}_j)$ and $g''(\hat{\boldsymbol{\theta}}_j)$ are matrices of first and second partial derivatives of $g(\boldsymbol{\theta}_j)$ with respect to elements of $\boldsymbol{\theta}_j$, calculated at their point estimates. The second line in the above expression stems from the fact that $\hat{\boldsymbol{\theta}}_j$ is a solution to $g'(\hat{\boldsymbol{\theta}}_j) = 0$. Also, $g''(\hat{\boldsymbol{\theta}}_j)$ is proportional to $-m\mathcal{I}(\hat{\boldsymbol{\theta}}_j)$, where $\mathcal{I}(\hat{\boldsymbol{\theta}}_j)$ is the expected Fisher information matrix. Thus,

$$\begin{aligned} p(\mathbf{y}^m|M_j) &\approx \int e^{g(\boldsymbol{\theta}_j)} d\boldsymbol{\theta}_j \\ &\approx e^{g(\hat{\boldsymbol{\theta}}_j)} \cdot \int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j)^T m\mathcal{I}(\hat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j)\right\} d\boldsymbol{\theta}_j \\ &= p(\mathbf{y}^m|\hat{\boldsymbol{\theta}}_j, M_j)p(\hat{\boldsymbol{\theta}}_j|M_j) \cdot (2\pi)^{v_j/2} |m\mathcal{I}(\hat{\boldsymbol{\theta}}_j)|^{-1/2}, \end{aligned}$$

because the integrand in the second line of the above equality is a constant multiple of the multivariate normal density.

It follows that

$$\begin{aligned} BIC_j &= 2\log p(\mathbf{y}^m|M_j) \\ &\approx 2\log p(\mathbf{y}^m|\hat{\boldsymbol{\theta}}_j, M_j) + 2\log p(\hat{\boldsymbol{\theta}}_j|M_j) + v_j \log(2\pi) - v_j \log(m) - \log |\mathcal{I}(\hat{\boldsymbol{\theta}}_j)| \\ &\approx 2\log p(\mathbf{y}^m|\hat{\boldsymbol{\theta}}_j, M_j) - v_j \log(m) + O(1). \end{aligned}$$

Fraley & Raftery (2002) claim that while a large number of different techniques have been developed over the years (see Biernacki & Govaert 1999 for a comparison of these techniques), the BIC approach provides good results for model-based clustering. However,

as mentioned earlier, problems arise whenever dealing with high-dimensional data. For the BIC approach, for example, the penalty term $v_j \log(m)$ depends on the number of parameters used to estimate the model. This includes all parameters used to estimate the covariance matrix. Clearly, the simpler the covariance structure, the fewer parameters need to be estimated, which implies that in high-dimensional problems, the criterion will favour the simplest covariance structures (Weakliem 1999, pg. 360). Because simple covariances may be unrealistic for functional data, we use filtering as described in §6.4 to reduce dimensionality and enable the use of more complex covariance structures.

6.4 Extension to Functional Data

One way of incorporating functional information about the shape of the curves into the clustering model is to let

$$\begin{aligned} \mathbf{y}_i &= f(\mathbf{t}; \boldsymbol{\eta}_i) + \boldsymbol{\epsilon}_i; \\ \boldsymbol{\eta}_i | \{z_{ik} = 1\} &\sim \mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n). \end{aligned} \tag{6.8}$$

As with the mixed-effects models in Chapter 4, each n -dimensional observation \mathbf{y}_i is fitted using a random curve component $f(\mathbf{t}; \boldsymbol{\eta}_i)$ plus some *iid* normally distributed vector of white noise $\boldsymbol{\epsilon}_i$ representing roughness in the curve. However, unlike our previous work, random effects in (6.8) depend on the unobserved cluster labels, and thus have cluster-specific means $\boldsymbol{\mu} = \boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k$.

A key distinction between models (6.8) and (6.1) is the fact that the functional model parameters $\boldsymbol{\eta}_i$ rather than the original data \mathbf{y}_i are the quantities being clustered. This is particularly useful when the number of parameters being used to model the shape of each curve is fewer than the number of time points at which each curve is observed ($p < n$), because it simplifies the complexity of the problem and speeds up computation. As such, model (6.8) serves a dual purpose. First, it incorporates functional information into the model. Secondly, it allows us to work in a lower-dimensional subspace of the data.

James & Sugar (2003), Gafney & Smyth (2003), Chudova et. al. (2004), and Zhou & Wakefield (2005) have all considered special cases of (6.8), varying in the functional

specification of $f(\mathbf{t}; \boldsymbol{\eta}_i)$ and the way in which model parameters are estimated. The first paper assumes a b-spline model for each curve, and uses model-based clustering on the b-spline coefficients to group the curves. That is, they let

$$f(\mathbf{t}; \boldsymbol{\eta}_i) = B\boldsymbol{\eta}_i,$$

where B is an $n \times p$ matrix of p known b-spline basis functions recorded at n time points $\mathbf{t} = [t_1, t_2, \dots, t_n]^T$, and $\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \dots, \eta_{ip}]^T$ is a vector of corresponding b-spline coefficients. While James & Sugar (2003) model the data and cluster it simultaneously (see 6.8), a conceptual equivalent of this idea is a two-step procedure in which b-spline coefficients are estimated by least squares regression, and then $\hat{\boldsymbol{\eta}}_i$ are clustered instead of the raw data \mathbf{y}_i in (6.1). An example of the latter for force exertion data is provided in §6.5. Gafney & Smyth (2003) and Chudova et. al. (2004) implement the same b-spline regression model using maximum a posteriori estimation instead of maximum likelihood estimation. Zhou & Wakefield (2005) propose a Bayesian framework for fitting (6.8), implemented for linear curves as a primary example (e.g., instead of a b-spline basis, B has only two columns corresponding to an intercept and a slope term in the model). An adaptation of the Zhou & Wakefield (2005) approach to the b-spline model is developed in §7.1.

6.5 Example

In this section we use model-based clustering as presented in §6.4 to explore latent structures present in the valve seat insertion curves. For computational ease, only the last twenty curves from each of the last six days of production in February are considered. Rather than cluster the seven valves (i.e. seven sets of 120 curves per valve) individually, all of the data are examined as one big group. That is, a total of $20 \times 6 \times 7 = 840$ curves are analyzed. Since we expect that there may be valve effects in the functional data, we hope that some of the between cluster variation will correspond to differences in the valves.

Exploring the Best Model

For each curve in the subset, the filtering technique is used to reduce the observed 100-dimensional data to a 20-dimensional vector of b-spline smoothing coefficients, which are subsequently clustered using the model-based approach. The BIC criterion is used to determine the type of covariance structure and the number of clusters appropriate for these data. We restrict our attention to four non-diagonal covariance structures: EEE, EEV, VEV and VVV, as specified in Table 6.1. Simpler structures were also fit, but found to be highly unlikely according to BIC, likely due to considerable covariance between some of the coefficients.

In Figure 6.1, values of BIC are plotted against the number of clusters for each of the four different models. Large values of BIC indicate the best covariance structure and number of parameters. The plot suggests that a model with a uniform covariance structure and 15 clusters is best-fitting to this data. This model is labeled as EEE, which corresponds to elliptically-shaped clusters of equal volume, shape and orientation (i.e. $\Sigma_k = \Sigma$ in (6.5)). Under the EEV, VEV and VVV schemes, a pronounced jump in the BIC from one to two clusters strongly emphasizes the two-cluster model as best fitting.

As discussed in §6.3.2, BIC's preference for the simpler covariance structure is not surprising, since it explicitly penalizes models with more parameters. This may also explain why BIC specifies a model with only two clusters (and thus very few parameters) as best-fitting when covariance structures are allowed to differ across clusters (EEV, VEV and VVV). While preference for simple models can be useful in general, the penalty poses a challenge for high-dimensional data, since the number of parameters used by a given model increases with the number of clusters and the dimension of the data.

Another problem with this analysis is that it ignores all information contained in the valve labels. This implies that, in the presence of strong differences between insertions made on each valve, clustering can lead to separation by valve rather than any interesting changes in the process. Table 6.2 investigates whether this is the case for the best-fitting model according to BIC by tabulating assigned cluster memberships by valve labels. The results support our suspicion. The eighth cluster is made up entirely of valve two curves, and all insertions assigned to the eleventh cluster are made on the fifth valve. Similar statements can be made for the fifth and tenth clusters, implying pronounced differences

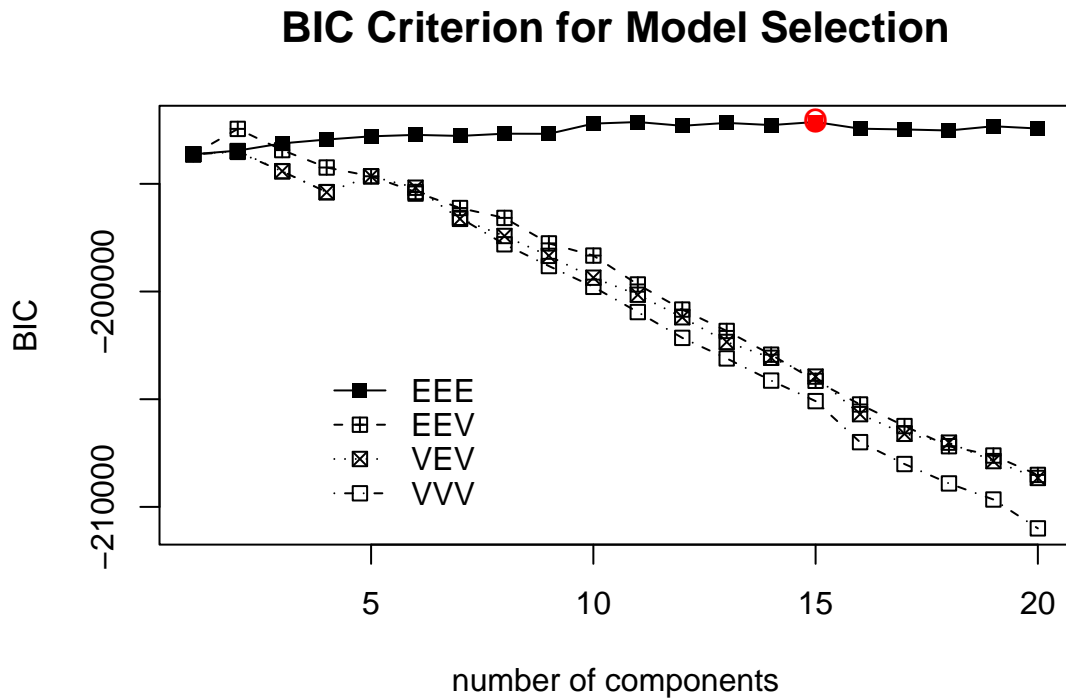


Figure 6.1: The BIC criterion for model selection suggests a 15-cluster model with a common covariance structure across clusters (EEE) as best-fitting to the February data.

between insertions made on each of the seven valves.

Another interesting pattern in Table 6.2 is the fact that none of the clusters have large numbers of observations in both $\{1, 2, 3, 4\}$ (intake valves, with insertions made by ram 1) and in $\{5, 7, 8\}$ (exhaust valves, with insertions made by ram 2). The split seems to correspond to the two types of valve (intake or exhaust). So even in cases where the observations within a cluster come from more than one valve, they typically come from the same type of valve.

	Valves							total
	1	2	3	4	5	7	8	
1	61	5	51	0	1	0	2	120
2	1	4	2	0	0	0	6	13
3	0	0	2	2	0	0	0	4
4	4	9	22	3	0	0	0	38
5	1	0	2	0	101	0	2	106
6	11	3	0	77	0	0	0	91
7	19	2	37	1	2	0	4	65
8	0	63	0	0	0	0	0	63
9	0	0	0	0	5	28	30	63
10	0	34	1	0	0	0	0	35
11	0	0	0	0	10	0	0	10
12	22	0	1	37	0	0	0	60
13	1	0	1	0	1	61	32	96
14	0	0	1	0	0	28	12	41
15	0	0	0	0	0	3	32	35
total	120	120	120	120	120	120	120	840

Table 6.2: Cluster memberships for the 15-cluster EEE model tabulated against the number of curves observed for each valve.

Controlling for Valve Differences

Since we already know the valve labels, a clustering method that splits the data by valve is not useful. A way to avoid this is to control for the systematic valve effects within the clustering model, an idea elaborated upon in Chapter 8. An ad-hoc equivalent of this solution for MBC is to subtract off the functional mean curves for each valve from the data prior to the analysis. This is equivalent to fitting the model:

$$\begin{aligned}
 \mathbf{y}_{ij} - \bar{\mathbf{y}}_j &= B\boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij} \\
 \boldsymbol{\eta}_{ij} | \{z_{ij} = k\} &\sim \mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),
 \end{aligned} \tag{6.9}$$

where $j = 1, \dots, 7$ indexes the valve, $i = 1, \dots, 120$ indexes the number of insertions made for each valve, and $\bar{\mathbf{y}}_j = \sum_{i=1}^{120} \mathbf{y}_{ij} / 120$ denote the valve means.

An added advantage of subtracting the valve means from each curve is that the residual curves have a relatively simple shape, allowing fewer b-spline basis functions to be utilized in the model. For illustration purposes, we use only five basis functions to smooth over the residual curves in the force exertion example.

A BIC plot for the fitted models, provided in Figure 6.2, identified the VEV covariance structure with two clusters as best-fitting for the force exertion data. Under EEE, EEV and VVV covariance schemes, a single cluster is identified. While it is not clear if the BIC approach is reliable for multivariate data, intuitively it makes sense that the insertions made during the last six days of the production process are relatively homogenous, as we assume the process to be in-control during that time period. Having said that, our conclusions with respect to the exact number of clusters appear to be inconclusive. If we assume that each cluster corresponds to a different mode of operation in the insertion process, it is unclear from Figure 6.2 if there are only two types of production. An alternative approach to determining the number of groupings of the curve data after accounting for the valve effects is proposed in Chapter 8.

For now, we fit the two-cluster VEV model identified as best-fitting according to BIC. A useful way of visualizing the structure contained in the clusters is to plot the estimated probabilities of belonging to each cluster, $p(z_{ijk} = 1 | \hat{\boldsymbol{\eta}}_{ij}, \hat{\boldsymbol{\theta}})$ as stated in (6.3), for each observation. When displayed in chronological order, the plots help reveal any time dependence that may be present in the formation of clusters. A display of the daily averages of the estimated probabilities plotted by day makes it easier to identify longer term trends. For example, from Figures 6.3(a)-(c), we see that during four of the six days in February, most of the $20 \cdot 7 = 140$ insertions are associated with the first cluster. However, on February 15th and 16th, a change appears to have occurred, with about half of the valve seats inserted in a way that is more characteristic of the second cluster. Functional means of the curves within each cluster, displayed in Figure 6.3(d), suggest that the change occurred because the ram was exerting less force in the middle of the insertion. This is evident from the shape of the cluster mean for the second component, which has the long flat segment between 20 and 80 time units considerably lower than the functional mean for the first cluster.

In summary, model-based clustering of the force exertion data suggests the following:

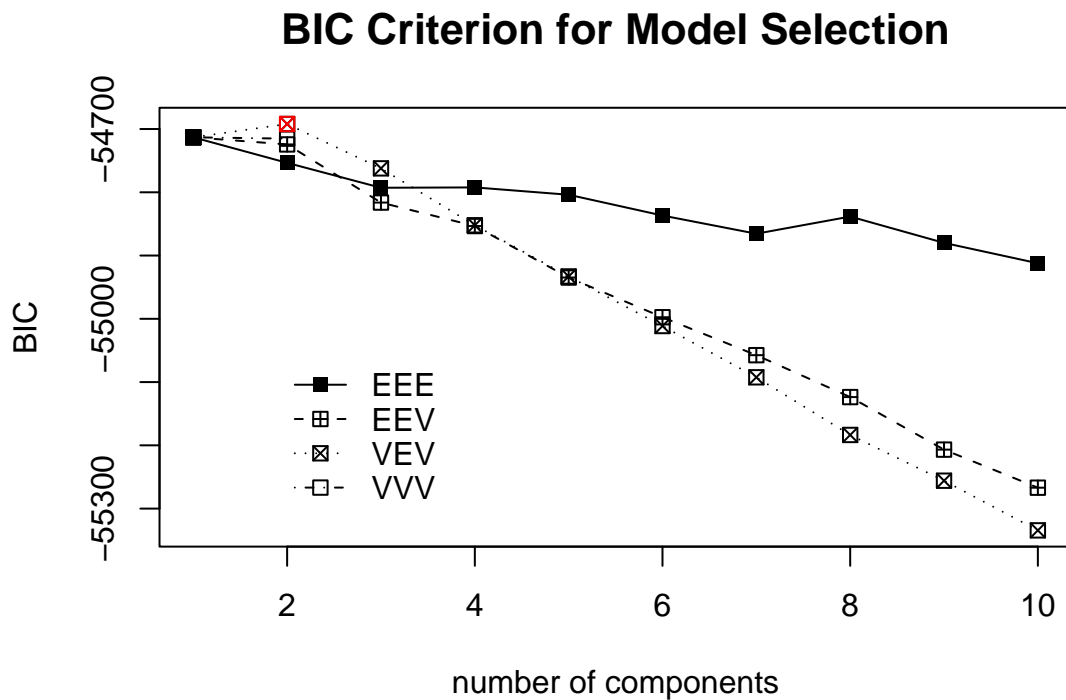


Figure 6.2: BIC plot for model-based clustering of the estimated b-spline coefficients fitted to the residual data after subtracting the valve means.

- a) Insertions made on different valves seem to differ, which must be taken into account in order to produce meaningful clustering results. The fact that the number of clusters dropped dramatically from 15 to 2 when valve effects were incorporated in the model is strong evidence of the impact of systematic effects on the clustering results.
- b) Based on BIC alone, it is not clear exactly how many clusters (i.e. different types of production) govern the insertion process during February 14th through 21st. More effective ways of determining the number of clusters are desirable.
- c) The production process is predominantly homogenous, with about 77% of the curves belonging to the first cluster. The remaining 23% differ in the decreased amount of

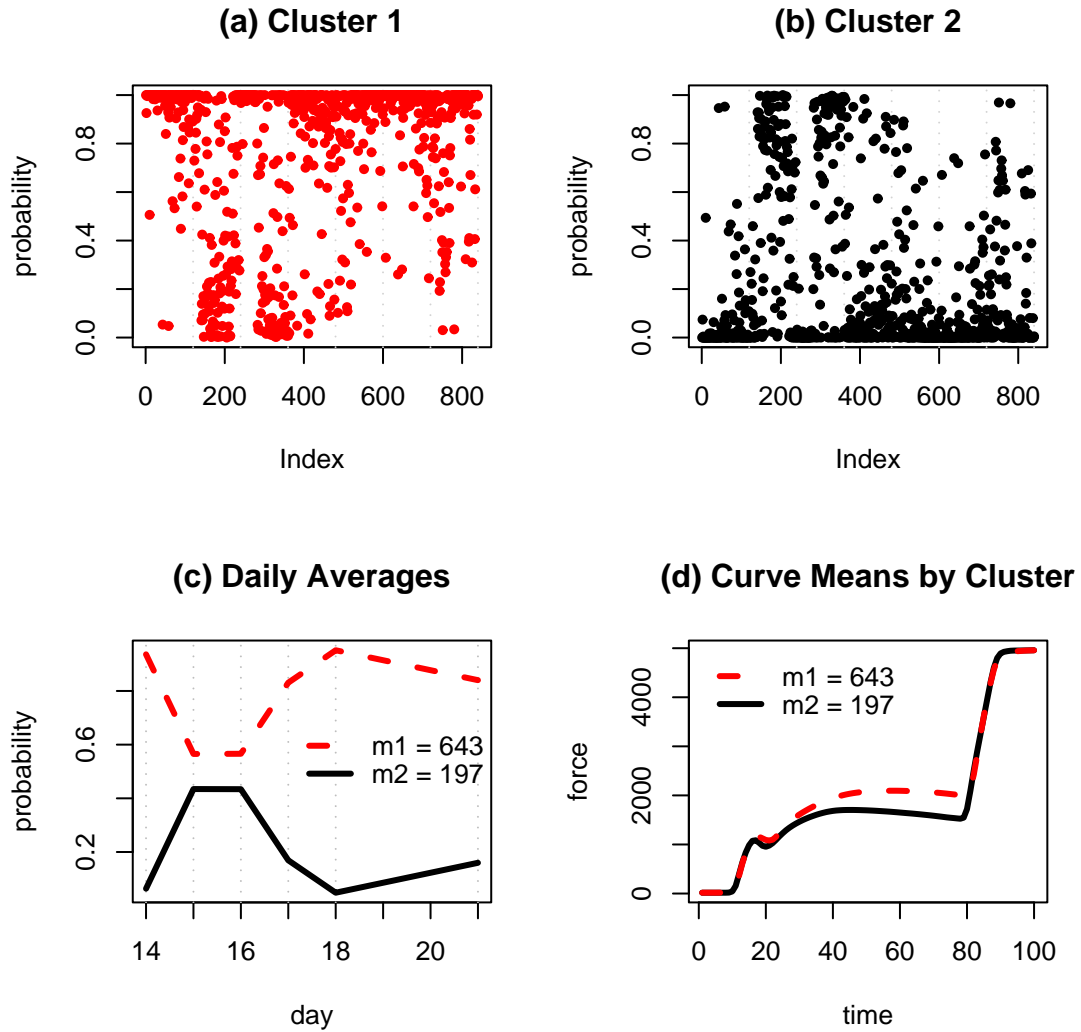


Figure 6.3: MBC results for the best-fitting two cluster fit. Estimated probabilities of belonging to each of the two clusters are plotted by observation index in (a) and (b). Daily averages of these probabilities are shown in (c), and within-cluster covariances are in (d). Within-cluster covariances for this fit have a common shape, but varying volume and orientation.

force applied during the middle of the insertion, a change that occurred primarily on February 15th and 16th.

These results suggest several directions, which will be explored in the next two chapters. These include the need to account for systematic effects (e.g., valve effects), and determining the number of clusters. An important property of curve clustering models for these data is the extent to which they can be used for process analysis and monitoring. Experts in the area might also be able to assign practical meaning to the clusters by looking at their means, which would provide valuable insight into the ranging modes of operation during valve seat insertion. If prediction for a new set of data is the goal, this can also be achieved by clustering a training set representative of the process, then assigning each new curve to a cluster that has the highest associated probability given the data as shown in (6.3). In Chapter 8, we consider one possible application of clustering in the context of profile-monitoring.

Chapter 7

A Bayesian Approach to Clustering

One disadvantage of the model-based clustering approach described in the previous chapter is that it is not always clear how many clusters should be used in the mixture. If we assume that each cluster corresponds to a different process during valve seat insertion, without any expert knowledge, it is impossible to know the correct number of process changes occurring during mainstream production. While BIC is helpful in this regard, it is not always a reliable approximation (e.g., Weakliem 1999 and §6.9.3 in McLachlan and Peel 2000), particularly for high-dimensional data, as it is geared towards choosing simpler models with fewer parameters.

Another drawback is the fact that the clustering model presented in §6.4 does not explicitly take into account valve information. While an ad-hoc solution of subtracting the valve means from the data prior to the analysis is useful, a comprehensive way of accounting for fixed effects within the framework of the model is desirable.

The remainder of this thesis is dedicated to addressing the problems of choosing the number of components and incorporating systematic information into the model. In this chapter, we deal with the first concern by adopting a fully Bayesian approach to clustering via mixtures of multivariate normal densities. All inference about the unknown parameters is based on Markov chain Monte Carlo (MCMC) samples obtained from the posterior distribution. By not relying on approximations such as BIC, more exact inference on the model can be attained. A generalized version of this model that accounts for systematic effects is proposed in Chapter 8.

7.1 Model

In Chapter 6, we considered the following mixture model for clustering curve data:

$$\begin{aligned} \mathbf{y}_i &= B\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i; \\ \boldsymbol{\eta}_i | \{z_i = k\} &\sim \mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n), \end{aligned} \quad (7.1)$$

where $i = 1, \dots, m$ is the curve index, and $k = 1, \dots, K$ indexes the mixture component. In order to emphasize the distinction between known curve indices and unknown cluster indices, cluster membership labels are denoted as $z_i = k$ rather than $z_{ik} = 1$ in the remainder of this thesis. That is, cluster membership for curve i will now be denoted by a label z_i that can take values $1, \dots, K$, rather than K binary indicators.

Up to now, the number of clusters K was assumed to be fixed, and model parameters estimated using maximum likelihood (ML). Bayesian methods provide an alternative approach to estimation and inference, allowing us to treat K as an unknown parameter. A comprehensive overview of Bayesian inference for mixture models can be found in Chapter 4 of McLachlan & Peel (2000), Gelfand et. al. (1990), Green (1995), and Stephens (2000a). Zhou & Wakefield (2005) extended these ideas to clustering curve data using (7.1). Their work is largely restricted to fitting linear curves ($p = 2$), and differs from our problem primarily in the size of the parameter space (in our case, $p = 20$).

Unlike ML, which seeks point estimates of model parameters, the Bayesian paradigm assumes that the unknown parameters are random quantities. Thus, inference about them is based on their posterior distribution, which combines information contained in the data with prior knowledge about the parameters.

For example, let

$$\Theta = (\{z_i\}_{i=1}^m, \{\boldsymbol{\eta}_i\}_{i=1}^m, \{\pi_k\}_{k=1}^K, \{\boldsymbol{\mu}_k\}_{k=1}^K, \{\boldsymbol{\Sigma}_k\}_{k=1}^K, V, R, \sigma^2)$$

denote all of the unknown parameters in (7.1) and $\mathbf{y}^m = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m\}$ the data. Then using Bayes' theorem,

$$p(\Theta | \mathbf{y}^m) = \frac{p(\mathbf{y}^m | \Theta)p(\Theta)}{p(\mathbf{y}^m)} \propto L(\Theta | \mathbf{y}^m)p(\Theta). \quad (7.2)$$

Here, $p(\cdot)$ is used generally to denote a probability density function of its argument. Thus $p(\Theta|\mathbf{y}^m)$ and $p(\Theta)$ are generally not the same distribution. In the remainder of the thesis, the terms “distribution” and “density” are used interchangeably to mean $p(\cdot)$.

According to (7.1), the likelihood function for the curve-clustering model is given by

$$L(\boldsymbol{\eta}, \sigma^2 | \mathbf{y}^m) = \prod_{i=1}^m \mathcal{N}_n(\mathbf{y}_i; B\boldsymbol{\eta}_i, \sigma^2),$$

so that the joint posterior has general form

$$p(\Theta | \mathbf{y}^m) \propto \prod_{i=1}^m \mathcal{N}_n(\mathbf{y}_i; B\boldsymbol{\eta}_i, \sigma^2) \times p(\Theta).$$

In specifying the prior distribution $p(\Theta)$, a hierarchical structure with several independence assumptions will be employed. Before explicitly specifying this prior, we graphically represent the dependence assumptions with a Directed Acyclic Graph (DAG) in Figure 7.1. The figure also summarizes the structure of model (7.1). The structure matches those used by Richardson & Green (1997), Stephens (2000a) and Zhou & Wakefield (2005) for mixture models. Boxed quantities in the plot indicate unknown parameters in the model and directed arrows denote conditional dependence. For example, we see that the amount of variability in the shape of the curve σ^2 is independent of all other parameters, whereas cluster membership labels z_i depend only on the probabilities of belonging to each cluster $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]^T$ and the number of clusters K . Justification of this structure and the exact prior distributions are discussed in detail in §7.2.

Even though the number of clusters K is displayed as an unknown parameter in Figure 7.1, for now let us assume it is fixed. Then the joint posterior distribution for the model

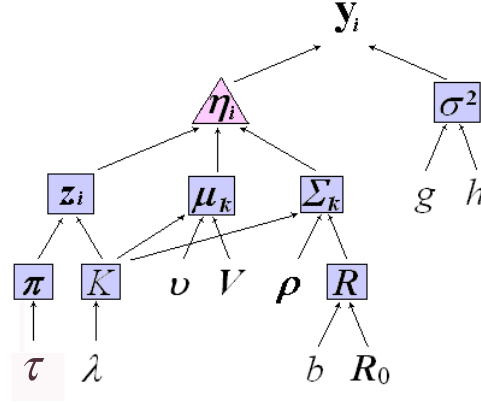


Figure 7.1: DAG of the Bayesian curve-clustering model. The diagram helps visualize the hierarchical dependence amongst the parameters. A triangle frames the quantity being clustered. Quantities not framed by boxes are fixed.

parameters Θ is given by:

$$\begin{aligned}
 p(\Theta|\mathbf{y}^m) &\propto \prod_{i=1}^m \mathcal{N}_n(\mathbf{y}_i; B\boldsymbol{\eta}_i, \sigma^2) \times \\
 &\prod_{k=1}^K \left\{ \prod_{i:z_i=k}^{m_k} \left\{ p(\boldsymbol{\eta}_i|z_i, \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \times p(z_i|\pi_{z_i}) \right\} \times p(\boldsymbol{\mu}_k) \times p(\boldsymbol{\Sigma}_k^{-1}|R) \right\} \\
 &\times p(\boldsymbol{\pi}) \times p(\sigma^{-2}) \times p(R).
 \end{aligned} \tag{7.3}$$

The joint posterior in (7.3) can be used for the purposes of model fitting and drawing inference. For example, Bayesian point estimates of the model parameters can be obtained by calculating conditional expectations, such that

$$E(\Theta|\mathbf{y}^m) = \int \boldsymbol{\theta} \cdot p(\boldsymbol{\theta}|\mathbf{y}^m) \cdot d\boldsymbol{\theta}. \tag{7.4}$$

Marginal posterior distributions of some subset of Θ (say $z_i = k$) can also be estimated by

expressions such as:

$$\begin{aligned}
 p(z_i = k | \mathbf{y}^m) &= \int \dots \int p(\boldsymbol{\theta} | \mathbf{y}_i) \cdot \partial \boldsymbol{\eta} \dots \partial \sigma^2. \\
 &= \int \dots \int p(z_i = k | \boldsymbol{\eta}, \dots, \sigma^2, \mathbf{y}_i) \cdot p(\boldsymbol{\eta}, \dots, \sigma^2 | \mathbf{y}_i) \cdot \partial \boldsymbol{\eta} \dots \partial \sigma^2 \quad (7.5) \\
 &= E(p(z_i = k | \boldsymbol{\eta}, \dots, \sigma^2, \mathbf{y}_i)).
 \end{aligned}$$

A difficulty posed by this approach is the evaluation of the high-dimensional integrals in (7.4) and (7.5), which do not have tractable form and thus cannot be determined analytically. A computational solution to this problem is a Markov chain Monte Carlo (MCMC) procedure presented in §7.3.

The advantage of using Bayesian inference in clustering data is the flexibility that it allows in estimating K . If we treat K as an unknown parameter (Stephens 2000a, Richardson & Green 1997, Cappe et. al. 2003), the posterior could provide inference on the uncertainty in K , via the marginal distribution $p(K | \mathbf{y}^m)$. An estimate of this parameter could then be determined by evaluating its posterior expectation

$$E(K | \mathbf{y}^m) = \sum_{k=1}^{\infty} k \cdot p(K = k | \mathbf{y}^m).$$

The posterior could also be used to infer the amount of variation associated with this approximation.

Detailed discussions pertaining to the strengths and weaknesses of the Bayesian approach can be found in Robert (2001) and Bernardo & Smith (1994), among many others. Key selling points of this methodology are full accountability for uncertainty in the model and the ability to incorporate prior information. Criticisms include computational constraints associated with the implementation of Bayesian approaches using MCMC algorithms and difficulties in specifying the priors. In the context of mixture models, empirical results show superior performance of MCMC over EM algorithms with respect to convergence to good solutions and misclassification error, but not in terms of computation times (Neal 1992 and Dias & Wedel 2004). Specifically, the EM algorithm “often fails to converge to the global maximum of the likelihood surface,” and “is very dependent upon the type

of starting values” (Dias & Wedel, 2004).

7.2 Priors

The first step in fitting a Bayesian model is prior specification. This can often be tricky, particularly for multivariate parameters. Some general guidelines for choosing these distributions are:

1. Priors should represent our knowledge about the problem before observing any data. A relaxed version of this is to allow the use of data in setting the parameters of these distributions (Casella, 1985). In that spirit, for example, we can let $\boldsymbol{\mu}_k \sim \mathcal{N}_p(\boldsymbol{v}, rI_p)$, where \boldsymbol{v} is the midpoint of the estimated b-spline coefficients and r is some measure of their spread (e.g., the mean of the squared range values).
2. In the absence of specific knowledge about the parameters, it is advisable to use *objective* (also called *uninformative*) priors to represent vague or general knowledge. For example, choosing a normal prior that has a large variance can serve as a way of expressing uncertainty about the mean.
3. A good rule of thumb is to avoid restrictive priors, i.e. prior distributions that are too informative about the parameter values. This can be achieved by placing one or more additional levels of prior distribution (called hyperpriors) on some or all of the parameters. For example, we can let

$$\Sigma_k^{-1} | R \sim \mathcal{W}_p(\rho, (\rho R)^{-1}) \text{ where } R \sim \mathcal{W}_p(b, (bR_0)^{-1}).$$

Under the above assumption, the expected value of Σ_k^{-1} is the first moment of the Wishart distribution, $\rho(\rho R)^{-1} = R^{-1}$. Thus, hyperparameter R represents the extent of within-cluster variation in the data. Since this is a $p \times p$ matrix, knowing how to specify it directly without being overly informative is nearly impossible. By assigning a hyperprior to this parameter, we introduce an extra level of variability in our specification of R , and thus make the prior on Σ_k^{-1} less informative.

4. Whenever possible, select *conjugate* priors for computational ease. A prior is said to be conjugate if it results in a posterior distribution that is of the same family (e.g., both the prior and the posterior are normal). If we let Θ_j denote an element of Θ , a *conditionally conjugate* prior is such that both Θ_j and $\Theta_j|\mathbf{y}^m$ are from the same class of distributions.

A summary of prior distributions assigned to the parameters of the curve-clustering model in (7.1) is provided below. These represent the standard choice of priors for mixture models (see Zhou & Wakefield 2005, Stephens 2000a, Richardson & Green 1997, and Gelfand et. al. 1990). We assume that the functional data have been scaled so that their observed range is roughly between 0 and 1. The scaling makes it easier to generalize prior beliefs expressed in this section to a wide variety of functional data, not just the force exertion curves considered in this thesis.

$$\begin{aligned}
 z_i = k|\boldsymbol{\pi} &\sim \mathcal{B}(\boldsymbol{\pi}_k) \\
 \boldsymbol{\pi}|K &\sim \mathcal{D}(\tau_1, \dots, \tau_K) \\
 \boldsymbol{\mu}_k &\sim \mathcal{N}_p(\boldsymbol{\nu}, V) \\
 \Sigma_k^{-1}|R &\sim \mathcal{W}_p(\rho, (\rho R)^{-1}) \\
 R &\sim \mathcal{W}_p(b, (bR_0)^{-1}) \\
 \sigma^{-2} &\sim \Gamma(g, h) \\
 K &\sim \mathcal{P}(\lambda).
 \end{aligned} \tag{7.6}$$

$\mathcal{B}()$, $\mathcal{D}()$, $\mathcal{W}()$, $\mathcal{N}()$, $\Gamma()$ and $\mathcal{P}()$ signify the Bernoulli, Dirichlet, Wishart, multivariate normal, gamma and Poisson distributions respectively. Probability density and mass functions associated with some of these distributions are specified in Appendix A.

The DAG in Figure 7.1 helps visualize hierarchical dependencies imposed on the model by our choice of priors. Quantities not framed by boxes denote the hyperparameters, to which the user must assign fixed values. Letting $r = \|\hat{\boldsymbol{\eta}}_{\max} - \hat{\boldsymbol{\eta}}_{\min}\|^2/p$, where $\hat{\boldsymbol{\eta}}_{\max}$ and $\hat{\boldsymbol{\eta}}_{\min}$ represent upper and lower envelopes of least squares estimates $\boldsymbol{\eta}_i$, recommended values

are:

$$\begin{aligned}
\tau_k &= 1 \text{ for all } k = 1, \dots, K \\
\mathbf{v} &= (\hat{\boldsymbol{\eta}}_{\min} + \hat{\boldsymbol{\eta}}_{\max})/2 \\
V &= rI_p \\
\rho &= b = 2p \\
R_0 &= (100/r)I_p \\
g &= 0.01 \text{ or } 0.001; \quad h = 1/g \\
\lambda &= 3, 5, \text{ or } 10.
\end{aligned}$$

A few key points clarifying the rationale behind prior selection are summarized next. Unless otherwise indicated, these specifications are based on Stephens (2000a).

Proportion Parameters (π_k)

Prior: $\boldsymbol{\pi} \sim \mathcal{D}(\tau_1, \dots, \tau_K)$

The Dirichlet distribution is chosen to describe the probabilities of belonging to each cluster, because it is the conjugate prior of a multinomial distribution. That is, if we let m_k denote the number of observations in each cluster, it is easy to see that

$$[m_1, \dots, m_K] | \boldsymbol{\pi} \sim \mathcal{Mult}(\pi_1, \dots, \pi_K),$$

and by selecting a Dirichlet prior on the $\boldsymbol{\pi}$, we obtain

$$\boldsymbol{\pi} | [m_1, \dots, m_K] \sim \mathcal{D}(\tau_1 + m_1, \dots, \tau_K + m_K).$$

The Dirichlet is also convenient because it is a continuous multivariate distribution satisfying the constraint $\sum \pi_k = 1$. Letting $\tau_0 = \sum \tau_k$, the mean and variance of this distribution are $E(\pi_k) = \frac{\tau_k}{\tau_0}$ and $V(\pi_k) = \frac{\tau_k(\tau_0 - \tau_k)}{\tau_0^2(\tau_0 + 1)}$. Thus, by setting $\tau_k = 1$ for all k , we assume a noninformative prior on $\boldsymbol{\pi}$, with equal expected probabilities of belonging to each cluster. Since $\mathcal{D}(1, \dots, 1) = 1/(K - 1)!$ (see Appendix A), this implies that the prior on

the π_k 's is constant (i.e. uniform) across mixture components.

Within-cluster Mean ($\boldsymbol{\mu}_k$)

Prior: $\boldsymbol{\mu}_k \sim \mathcal{N}_p(\boldsymbol{v}, V)$

If we believe that each cluster of random effect coefficients has its own mean, we want to allow these averages to vary across clusters. Thus, a prior is assumed on the within-cluster means $\boldsymbol{\mu}_k$. The normal distribution is a natural choice for this prior since b-spline coefficients are themselves normal. The hyperparameters of the prior on $\boldsymbol{\mu}_k$ are \boldsymbol{v} and V , which represent the expected value and covariance of $\boldsymbol{\mu}_k$. Thus, by setting $\boldsymbol{v} = (\boldsymbol{\eta}_{\min} + \boldsymbol{\eta}_{\max})/2$, we reflect our belief that within-cluster means will tend to center around the mid-range of the estimated b-spline coefficients (Richardson & Green, 1997).

Specifying V is trickier. The hyperparameter represents the amount of between-cluster variability in the data, and even if we knew how many and how far away the clusters should be, it is difficult to specify each of the $p(p+1)/2$ elements of the matrix. One alternative is to specify V as a diagonal covariance matrix with sufficiently large elements, which will ensure a reasonably flat prior for the $\boldsymbol{\mu}_k$ over the observed range of the data. Based on Stephens (2000a), we set V to be proportional to an identity matrix (i.e. $V = rI_p$), with the constant of proportionality equal to the mean of the squared functional range vector, $r = \|\hat{\boldsymbol{\eta}}_{\max} - \hat{\boldsymbol{\eta}}_{\min}\|^2/p$, which is essentially just an empirical measure of dispersion. This reflects our belief that the amount of variability between the clusters is roughly proportional to the overall variability in the b-spline coefficients, without being too informative about its exact value. Since the squared range of a sample must be larger than its variance, this choice of scaling yields extra dispersion, and gives a reasonably flat prior.

Covariance Parameters (σ^2 , Σ_k and R)

Priors: $\sigma^{-2} \sim \Gamma(g, h)$, $\Sigma_k^{-1} \sim \mathcal{W}_p(\rho, (\rho R)^{-1})$ and $R \sim \mathcal{W}_p(b, (bR_0)^{-1})$

Consider σ^2 - a measure of overall roughness in each curve. While it does not have a simple conditionally conjugate prior, σ^{-2} does (Gelman, 2006). Two commonly used possibilities are the uniform and the gamma priors. We chose to use the latter because it is closely

related to the χ^2 , which is the sampling distribution of the scaled sample variance. The mean and variance of the gamma distribution are $E(\sigma^{-2}) = gh$ and $V(\sigma^{-2}) = gh^2$, so that by setting $g = 1/h$ to be very small (e.g., 0.01 or 0.001), we specify a noninformative (highly variable) prior on σ^2 . Increasing values of g correspond to priors that are less and less objective, with stronger belief put into lower values of σ^2 . Our specification parallels that of Zhou & Wakefield (2005) who suggest setting $g = 0.001$.

Let us now consider within-cluster covariance matrices Σ_k . Since we believe that each mixture component has a different covariance matrix, it is natural to allow these to vary by specifying a prior distribution. A multivariate extension of the gamma distribution is the Wishart, which is conditionally conjugate for precision matrices¹, making it the preferred prior for Σ_k^{-1} . A formal justification of assuming normal priors for mean parameters and Wishart priors for precision matrices is provided in Geiger & Heckerman (2002).

Specifying hyperparameters for high-dimensional Wishart priors is complex. The two parameters of a Wishart distribution correspond to the degrees of freedom (d.f.) and the expected value of the random variable. That is, by assuming that $\Sigma_k^{-1} \sim \mathcal{W}_p(\rho, (\rho R)^{-1})$ as stated in (7.6), we have $E(\Sigma_k^{-1}) = \rho \times (\rho R)^{-1} = R^{-1}$, where ρ is the d.f.. The distribution is only proper for values of $\rho \geq p$ (see Appendix A). As with the χ^2 , smaller d.f. imply more uncertainty. The usual approach is to set these to be as small as possible (e.g., $\rho = p$ or $2p$), in order to ensure that the priors are vague and objective.

The second hyperparameter of a Wishart distribution is a $p \times p$ symmetric matrix characterized by $p(p+1)/2$ components. For large p , it is nearly impossible to know what the elements of this matrix should be. Following Stephens (2000a) and Zhou & Wakefield (2005), one possibility is to center R on a diagonal matrix, but allow it depart from this. This is achieved by assuming a hyperprior on R , such that $R \sim \mathcal{W}_p(b, (bR_0)^{-1})$, where R_0 is a diagonal matrix equal to $100/r$. This implies that the amount of spread in the curves within each cluster is roughly proportional to one tenth the amount of spread in all of the data. The constant multiple of 100 is data-dependent in that it corresponds to a diffuse prior if the observed data has a relatively small range, but it is quite informative if the range of data values is quite large. The problem is avoided by scaling the data by their maximum value, as indicated earlier. This will cause the range of observed values

¹A precision matrix is the inverse of the covariance matrix.

to be roughly between 0 and 1, making our suggested choice of hyperprior sufficiently uninformative. The value of 100 is ten times the one suggested by Stephens, 1997 (pg. 37) to allow for a more diffuse prior.

The importance of specifying objective prior beliefs on Σ_k^{-1} is emphasized by Richardson & Green (1997), who point out the dependence between this parameter and the number of clusters K in the univariate case. These findings extend to the multivariate problem considered in this thesis. When K is treated as a random variable (as is done later in §7.3.3), its posterior distribution will depend on the amount of within-cluster variation. That is, if elements of Σ_k are constrained to be too small, a large number of clusters will be found, whereas if they are forced to be very large, all of the data will be merged into a single cluster. The authors found that by defining an extra level of hierarchy on the hyperprior of the precision parameter (R), we can avoid specifying Σ_k^{-1} too informatively, and estimation of K becomes considerably less sensitive to the choice of prior on Σ_k .

The Number of Clusters (K)

Prior: $K \sim \mathcal{P}(\lambda)$

A Poisson prior on K was chosen following Stephens (2000a). The prior has direct impact on the algorithm used to draw inference from the joint posterior, a topic discussed in more detail in §7.3.

Our choice of hyperparameters for this prior are motivated by the fact that the mean and the variance of a Poisson distribution are $E(K) = V(K) = \lambda$. It is therefore preferable to set λ roughly equal to the expected number of clusters, with larger values resulting in more objective (less informative) priors. In our experience, setting $\lambda = 5$ or 10 in the prior provides a nice range of plausible values for K without being too informative. Following Zhou & Wakefield (2005), these values are larger than those suggested by Stephens (2000a), generating priors that are more diffuse.

7.3 Markov chain Monte Carlo

Having articulated prior beliefs about the parameters, let us return to the problem of making inference. Let $\Theta = (\theta_1, \dots, \theta_d)$ denote some unknown set of parameters with a joint distribution $\pi(\cdot)$. In order to be able to evaluate conditional expectations like (7.4) and (7.5), we must be able to solve complex integrals of general form:

$$E(h(\Theta)) = \int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}. \quad (7.7)$$

Using this expression, point estimates of parameter values can be obtained by setting $h(\Theta) = \Theta$, and marginal distributions of say θ_1 can be estimated by fixing $h(\Theta) = \pi(\theta_1|\theta_2, \dots, \theta_d)$, since $E(\pi(\theta_1|\theta_2, \dots, \theta_d)) = \int \pi(\theta_1|\theta_2, \dots, \theta_d)\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \pi(\theta_1)$.

Due to the multidimensional nature of the integral in (7.7), the expression cannot be evaluated analytically. Markov chain Monte Carlo (MCMC) methods offer a convenient approximation of the solution via simulation.

7.3.1 Review

Two fundamental concepts behind all MCMC methods are the two MCs - Monte Carlo integration and Markov chains. In this section we provide an overview of these ideas, drawing from Tierney 1994 (§2), Roberts 1996 (§3.2) and Stephens 1997 (§2.1). For complete details, we recommend reading Chapters 1, 3 and 4 in Gilks, Richardson & Spiegelhalter (1996). Readers familiar with MCMC theory are encouraged to skip ahead to §7.3.2.

Monte Carlo Integration

Let $\Theta_1, \Theta_2, \dots$ denote an infinite sequence of multidimensional random variables that are independent and identically distributed with density $\pi(\cdot)$ and mean $\boldsymbol{\theta}$. Then according to the strong law of large numbers (SLLN), for $E_\pi(|\Theta_s|) < \infty$,

$$\lim_{N \rightarrow \infty} \bar{\Theta}_N \xrightarrow{a.s.} \boldsymbol{\theta},$$

where $\bar{\Theta}_N = (\Theta_1 + \dots + \Theta_N)/N$ is the sample mean.²

More generally, as $N \rightarrow \infty$,

$$\bar{h}_N(\Theta) = \frac{1}{N} \sum_{i=1}^N h(\Theta_i) \xrightarrow{a.s.} E_{\pi}(h).$$

Based on this result, integrals of the form (7.7) can be approximated using sample averages of draws from $\pi(\cdot)$. The approach is called Monte Carlo integration. It is named after a prominent city in Monaco famous for its casinos, because both gambling and the approximation involve a random component.

Now suppose Θ represents a set of unknown parameters, which we assume to be random. In the context of Bayesian model-fitting, the Monte Carlo method is utilized by drawing repeatedly from the posterior distribution of Θ , then using arithmetic means to estimate marginal posteriors and/or obtain point estimates of the unknown parameters. Sample variances and prediction intervals can also be used to provide measures of uncertainty for these estimates.

Unfortunately, for the curve clustering model, it is not possible to sample directly from the joint posterior distribution presented in (7.3). An alternative is to set up a Markov chain that has a stationary distribution equal to the joint posterior. Although we will no longer be able to evoke the SLLN (successive states of a Markov chain are not independent), under certain conditions, the Ergodic Theorem guarantees that path averages of the chain can be used to approximate expectations. A review of Markov chains and the necessary conditions for the Ergodic Theorem are provided next.

Markov Chains

A Markov chain is defined as a succession of random variables satisfying the property that a future state depends only on the present and not on the past. Formally, a sequence of random variables $\Theta^{(0)} \rightarrow \Theta^{(1)} \rightarrow \dots \rightarrow \Theta^{(N)}$ defined in some state space $\Omega \subseteq \mathbb{R}^p$ is a Markov chain so long as the distribution of $\Theta^{(s+1)}$ given the current value of $\Theta^{(s)}$ is independent of $\Theta^{(0)}, \dots, \Theta^{(s-1)}$.

²a.s. is shorthand for “almost surely”.

Board games provide a simple example (Ash & Bishop, 1972). In Monopoly, at any given time a new move is dictated only by the immediate location on the board. The likelihood of moving to a different location, in this case governed by the roll of the die, is called a *transition probability*. These are dictated by a *transition kernel*. For a fixed measurable set of values $A \subset \Omega$, this is defined as

$$P(\boldsymbol{\theta}, A) = Pr(\Theta^{(s+1)} \in A | \Theta^{(s)} = \boldsymbol{\theta}).$$

The kernel plays a key role in ensuring that $\Theta^{(s)}$ have the same distribution for all values of s . Formally, if $\Theta^{(s)}$ has density π on Ω , then $\Theta^{(s+1)}$ has distribution πP . Thus, a Markov chain will have a *stationary* (or invariant) distribution π if $\pi = \pi P$, or equivalently

$$\Theta^{(s)} \sim \pi \Rightarrow \Theta^{(s+1)} \sim \pi. \quad (7.8)$$

A sufficient criteria for satisfying (7.8) is the detailed balance condition

$$\pi(\boldsymbol{\theta})P(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}')P(\boldsymbol{\theta}', \boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Omega \quad (7.9)$$

which specifies that the probability with which a chain leaves a state must be the same as the one with which it enters it (Cappe et. al., 2003).

Two important properties of a Markov process, which guarantee that a stationary distribution exists, are *irreducibility* and *aperiodicity*. Irreducibility implies that, regardless of the present state, a process can move to any other state in finite time. Aperiodicity requires that the chain does not get caught in cycles. In Monopoly, the two properties guarantee that a) you can always get to any of the 40 locations on the board, and b) you will not continuously return to the same spot, say jail.

Precise definitions of the two concepts are:

- i) A Markov chain is irreducible if for some $s > 0$ and all $A \subset \Omega$,

$$P(\Theta^{(s)} \in A | \Theta^{(0)} = \boldsymbol{\theta}) > 0.$$

ii) An irreducible chain is aperiodic if all of its states have period 1, i.e.

$$\text{gcd}\{s : P(\Theta^{(s)} = \boldsymbol{\theta} | \Theta^{(0)} = \boldsymbol{\theta}) > 0\} = 1,$$

where gcd denotes the greatest common denominator.

An irreducible Markov chain that has aperiodic states is said to be *ergodic*, which leads to the following result (Roberts 1996, pg. 47).

Theorem 7.3.1 (Ergodic Theorem) *Let $\Theta = (\Theta^{(0)}, \dots, \Theta^{(N)})$ denote an ergodic Markov chain defined on state space Ω , with a transition kernel P and a stationary distribution π . Then for any real-valued function $h(\Theta)$ defined on Ω such that $\int |h(\boldsymbol{\theta})| \pi(d\boldsymbol{\theta}) < \infty$,*

$$\bar{h}_N(\Theta) = \frac{1}{N} \sum_{s=1}^N h(\Theta^{(s)}) \xrightarrow{a.s.} E(h(\Theta))$$

as N goes to infinity.

The theorem states that an ergodic average of a stationary Markov chain will converge almost surely to its expected value. That is, $E(h) \approx \bar{h}_N$ for large N .

Combining ideas behind Markov chains and Monte Carlo, MCMC is an approach for estimating expectations involving a set of unknown parameters Θ . This is accomplished by constructing a Markov chain that has the same joint posterior $p(\Theta | \mathbf{y}^m)$ as the stationary distribution. Then, by the Ergodic Theorem, path averages of the chain are used to estimate marginal posteriors and/or any of the unknown parameters.

Convergence

In practice, the first few draws of the MCMC sample (called the *burn-in period*) are discarded in order to ensure that the chain has converged to good solutions and is independent of starting conditions. The Heidelberger & Welch (H-W) test (1983) implemented in the R `coda` library can be used to assess chain convergence of the remaining output and ensure sufficient burn-in.

A null hypothesis for the H-W test is that a chain under consideration has a stationary distribution. This is assessed sequentially. First, the entire chain is tested. If it has

not converged (i.e. the null hypothesis is rejected), 10% of the chain is discarded and the remainder is re-tested. The process is repeated up to four more times until the null hypothesis of convergence is no longer rejected. If the latter does not happen, the conclusion is that the chain does not have a stationary distribution. At each step, the test statistic for evaluating the null hypothesis is a Cramer-von-Mises measure of discrepancy between posterior means of the parameter before and after a fraction of the output was discarded.

The test is univariate, and therefore it is conducted elementwise for parameters that are vectors or matrices. For example, to check the convergence of a $K \times 1$ vector $\boldsymbol{\pi}$, K H-W tests have to be conducted.

Cowles & Bradley (1996) provide a comprehensive review of other MCMC convergence diagnostics that can be used, including those implemented in `coda`. According to their comparative review, most provide similar results. However, the authors caution against using numeric diagnostics alone, and encourage plotting output as well in order to verify convergence visually. Examples of graphical assessment of convergence is provided in §8.5.

7.3.2 Gibbs Sampling Algorithm

Let $\Theta = (\theta_1, \dots, \theta_d)$ denote a multidimensional random variable with a joint distribution $p(\Theta)$. Gibbs sampling is a special type of MCMC method used to draw realizations of Θ from $p(\Theta)$ (Geman & Geman, 1984). It applies to situations where we cannot sample from the joint density function of Θ directly, but are able to draw from the conditional distributions of the θ_i 's given values of the remaining parameters.

For Bayesian problems, the Gibbs algorithm is used to sample from the joint posterior in the following manner. Starting with initial values $\Theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})$, sequentially draw each element of $\Theta^{(s)}$ for $s = 1, 2, \dots, N$ iterations:

$$\begin{aligned} \theta_1^{(s+1)} &\sim p(\theta_1 | \mathbf{y}^m, \theta_2^{(s)}, \dots, \theta_d^{(s)}) \\ \theta_2^{(s+1)} &\sim p(\theta_2 | \mathbf{y}^m, \theta_1^{(s+1)}, \theta_3^{(s)}, \dots, \theta_d^{(s)}) \\ &\vdots \\ \theta_d^{(s+1)} &\sim p(\theta_d | \mathbf{y}^m, \theta_1^{(s+1)}, \dots, \theta_{d-1}^{(s+1)}) \end{aligned}$$

Geman & Geman (1984) show that the resulting sample $(\Theta^{(0)}, \Theta^{(1)}, \dots, \Theta^{(N)})$ forms an irreducible, aperiodic Markov chain with a limiting distribution $p(\Theta|\mathbf{y}^m)$. The latter is easy to see by verifying the stationarity condition in (7.8) (see Stephens 1997, pgs. 19-20).

Gibbs Sampling for the Bayesian Curve Clustering Model

Consider the Bayesian curve clustering model presented in §7.1. Suppose that the number of clusters K is fixed, and let

$$\Theta = (\{z_i\}_{i=1}^m, \{\boldsymbol{\eta}_i\}_{i=1}^m, \{\pi_k\}_{k=1}^K, \{\boldsymbol{\mu}_k\}_{k=1}^K, \{\Sigma_k\}_{k=1}^K, R, \sigma^2)$$

denote all of the unknown parameters in the model. Then using $\Theta^{(0)}$ as starting values, a Gibbs algorithm would proceed as follows for $s = 1, 2, \dots, N$ iterations:

1. Draw $z_i^{(s+1)}$ from $p(z_i|\mathbf{y}^m, \boldsymbol{\eta}_i^{(s)}, \dots, \sigma^{2(s)})$ for $i = 1, \dots, m$.
2. Draw $\boldsymbol{\eta}_i^{(s+1)}$ from $p(\boldsymbol{\eta}_i|\mathbf{y}^m, z_i^{(s+1)}, \boldsymbol{\pi}^{(s)}, \dots, \sigma^{2(s)})$ for $i = 1, \dots, m$.
3. Draw $\boldsymbol{\pi}^{(s+1)}$ from $p(\boldsymbol{\pi}|\mathbf{y}^m, z_i^{(s+1)}, \boldsymbol{\eta}_i^{(s+1)}, \boldsymbol{\mu}_k^{(s)}, \dots, \sigma^{2(s)})$.
4. Draw $\boldsymbol{\mu}_k^{(s+1)}$ from $p(\boldsymbol{\mu}_k|\mathbf{y}^m, z_i^{(s+1)}, \boldsymbol{\eta}_i^{(s+1)}, \boldsymbol{\pi}^{(s+1)}, \Sigma_k^{(s)}, \dots, \sigma^{2(s)})$ for $k = 1, \dots, K$.
- ⋮
8. Draw $\sigma^{2(s+1)}$ from $p(\sigma^2|\mathbf{y}^m, z_i^{(s+1)}, \boldsymbol{\eta}_i^{(s+1)}, \boldsymbol{\pi}^{(s+1)}, \boldsymbol{\mu}_k^{(s+1)}, \Sigma_k^{(s+1)}, R^{(s+1)})$.

The conditional posteriors in 1.-8. are called *full conditionals*. For hierarchical problems, many of these simplify nicely when conditional independencies in the prior and model are utilized. For example, consider the full conditional distribution of $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$. From Figure 7.1, it is clear that $\boldsymbol{\pi}$ is conditionally independent from the data and all parameters except for the z_i 's. This is also evident from the form of the joint posterior (7.3). That is,

$$p(\boldsymbol{\pi}|\mathbf{y}^m, \{z_i\}_{i=1}^m, \{\boldsymbol{\eta}_i\}_{i=1}^m, \{\boldsymbol{\mu}_k\}_{k=1}^K, \{\Sigma_k\}_{k=1}^K, R, \sigma^2) = p(\boldsymbol{\pi}|\{z_i\}_{i=1}^m).$$

In order to obtain the full conditional of $\boldsymbol{\pi}$ up to a proportionality constant, it follows that

$$\begin{aligned} p(\boldsymbol{\pi}|\mathbf{y}^m, \dots) &\propto p(\{z_i\}_{i=1}^m|\boldsymbol{\pi})p(\boldsymbol{\pi}) \\ &= \left\{ \prod_{k=1}^K \prod_{i:z_i=k}^{m_k} p(z_i|\pi_{z_i}) \right\} \cdot p(\boldsymbol{\pi}). \end{aligned}$$

Since $z_i|\boldsymbol{\pi} \sim \mathcal{B}(\pi_{z_i})$ and $\boldsymbol{\pi} \sim \mathcal{D}(\tau_1, \dots, \tau_K)$, we obtain

$$\begin{aligned} p(\boldsymbol{\pi}|\mathbf{y}^m, \dots) &\propto \pi_1^{m_1} \cdot \dots \cdot \pi_{K-1}^{m_{K-1}} \cdot (1 - \pi_1 - \dots - \pi_{K-1})^{m_K} \times \\ &\quad \pi_1^{\tau_1-1} \cdot \dots \cdot \pi_{K-1}^{\tau_{K-1}-1} \cdot (1 - \pi_1 - \dots - \pi_{K-1})^{\tau_K-1} \\ &= \pi_1^{m_1+\tau_1-1} \cdot \dots \cdot \pi_{K-1}^{m_{K-1}+\tau_{K-1}-1} \cdot (1 - \pi_1 - \dots - \pi_{K-1})^{m_K+\tau_K-1} \end{aligned}$$

Thus the full conditional distribution of $\boldsymbol{\pi}$ is $\mathcal{D}(m_1 + \tau_1, \dots, m_K + \tau_K)$, where m_k is the total number of observations in cluster k .

Full conditionals for the remaining parameters of the curve clustering model are derived in Appendix B.1. These are:

$$\begin{aligned} p(z_i = k|\dots) &= \pi_k \cdot \mathcal{N}_n(\mathbf{y}_i; \boldsymbol{\mu}_k, \Sigma_k) \\ \boldsymbol{\eta}_i|\dots &\sim \mathcal{N}_p((\sigma^{-2}B^TB + \Sigma_k^{-1})^{-1}(\sigma^{-2}B^T\mathbf{y}_i + \Sigma_k^{-1}\boldsymbol{\mu}_k), (\sigma^{-2}B^TB + \Sigma_k^{-1})^{-1}) \\ \boldsymbol{\pi}|\dots &\sim \mathcal{D}(m_1 + \tau_1, \dots, m_K + \tau_K) \\ \boldsymbol{\mu}_k|\dots &\sim \mathcal{N}_p((m_k\Sigma_k^{-1} + V^{-1})^{-1}(m_k\Sigma_k^{-1}\bar{\boldsymbol{\eta}}_k + V^{-1}\boldsymbol{\nu}), (m_k\Sigma_k^{-1} + V^{-1})^{-1}) \\ \Sigma_k^{-1}|\dots &\sim \mathcal{W}_p\left(m_k + \rho, \left[\sum_{i:z_i=k}^{m_k} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)(\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T + \rho R \right]^{-1}\right) \\ R|\dots &\sim \mathcal{W}_p\left(b + K\rho, \left[bR_0 + \rho \sum_{k=1}^K \Sigma_k^{-1} \right]^{-1}\right) \\ \sigma^{-2}|\dots &\sim \Gamma\left(\frac{nm}{2} + g, \left[\frac{1}{2} \sum_i (\mathbf{y}_i - B\boldsymbol{\eta}_i)^T (\mathbf{y}_i - B\boldsymbol{\eta}_i) + \frac{1}{h} \right]^{-1}\right) \end{aligned}$$

where $\bar{\boldsymbol{\eta}}_k = \sum_{i:z_i=k}^{m_k} \boldsymbol{\eta}_i / m_k$.

Then a Gibbs sampler for the curve clustering model involves generating an MCMC sample $(\Theta^{(1)}, \dots, \Theta^{(N)})$, with each element of $\Theta^{(s)}$ drawn sequentially from the full conditionals listed above. For large N , the stationary distribution of this chain is roughly equal to the joint posterior. Having discarded the burn-in values, MCMC output is averaged to obtain point estimates of the parameters and can also be used to estimate marginal distributions.

7.3.3 Birth-Death MCMC

One of the most challenging characteristics of our problem is the absence of any information about the type and number of processes that are governing valve seat insertion. This calls for an automated way of selecting the total number of clusters, K . In the context of hierarchical mixture models, this can be accomplished by treating K as one of the unknown parameters, assigning a prior to it and estimating it conditional on the data.

The challenge in allowing K to vary is the fact that this causes the dimensionality of the parameter space to change at each iteration of the MCMC. That is, as K varies, the model contains a larger or smaller number of cluster specific parameters such as π_k , $\boldsymbol{\mu}_k$, and Σ_k . In recent years, several algorithms have been proposed to deal with the trans-dimensional MCMC problem in the context of mixture models. Two dominant ones are reversible jump MCMC (RJMCMC) developed by Green (1995) and birth-death MCMC (BDMCMC) developed by Stephens (2000a). Cappe, Robert & Ryden (2003) show that RJMCMC can be regarded as a special case of BDMCMC, and propose a variation of the latter, called continuous time MCMC.

For its generality and ease in implementation, we only consider Stephens' birth-death approach. Zhou & Wakefield (2005) applied this algorithm to fit the curve clustering model in (7.1). We follow their work closely, making necessary adjustments to improve performance whenever possible. An overview is provided in this section.

The Basics

The assumption behind Stephens' approach to mixture modeling is that new clusters are "born" and old clusters "die" in continuous time. These are regarded as a sequence of

events occurring at some rate $\lambda + \delta$, where λ is the rate of birth and δ is the rate of death per unit of time. Although it can be assigned as any arbitrary quantity, we specify the birth rate λ as the prior mean of K . The progression of birth and death events can then be represented as a Poisson process, with the waiting time until the occurrence of an event following an exponential distribution with mean $1/(\lambda + \delta)$. Stephens proposes simulating this process using the following recipe.

Let K denote a starting value for the number of clusters, and use $\Phi = \{\phi_1, \dots, \phi_K\}$ to denote all cluster-specific parameters, with $\phi_k = \{\pi_k, \mu_k, \Sigma_k\}$ containing the parameters for component k . Proceed to simulate a birth-death process by repeating the following steps for some extended amount of time:

- i) Simulate the time t at which the next event will occur, where $T \sim EXP(1/(\lambda + \delta))$.
- ii) Simulate the type of event x occurring at time t , with

$$X = \begin{cases} 1 & \text{if birth, with } P(X = 1) = \frac{\lambda}{\lambda + \delta}; \\ -1 & \text{if death, with } P(X = -1) = \frac{\delta}{\lambda + \delta}. \end{cases}$$

- iii) Update the number of clusters, $K = K + x$, and make the following adjustments:

- If a birth occurs, propose new cluster parameters ϕ_{new} , and update $\Phi = \Phi \cup \phi_{new}$. Within Φ , adjust the proportion parameters to $\boldsymbol{\pi} = \boldsymbol{\pi}(1 - \pi_{new})$ in order to ensure that elements of $\boldsymbol{\pi}$ sum to one.
- If a death occurs, delete a cluster with the highest death rate δ_k (where $\delta = \sum_k \delta_k$) by setting $\Phi = \Phi \setminus \phi_k$. Scale proportion parameters to $\boldsymbol{\pi} = \frac{\boldsymbol{\pi}}{(1 - \pi_k)}$.

Stephens combines this with the Gibbs sampler (which draws the remaining parameters) by running Gibbs iterations at equally spaced time units of the continuous birth-death simulation. A visualization of this idea is provided in Figure 7.2.

In order to ensure that the stationary distribution of the resulting Markov chain is the joint posterior $p(\Theta, K | \mathbf{y}^m)$, the birth and death rates of the process must satisfy the detailed balance condition dictated by (7.9). That is, it must hold true that for all Θ ,

$$\frac{1}{(K-1)!} p(\Theta \setminus \phi_k, K-1 | \mathbf{y}^m) P(\Theta \setminus \phi_k, \Theta) = \frac{1}{K!} p(\Theta, K | \mathbf{y}^m) P(\Theta, \Theta \setminus \phi_k), \quad (7.10)$$

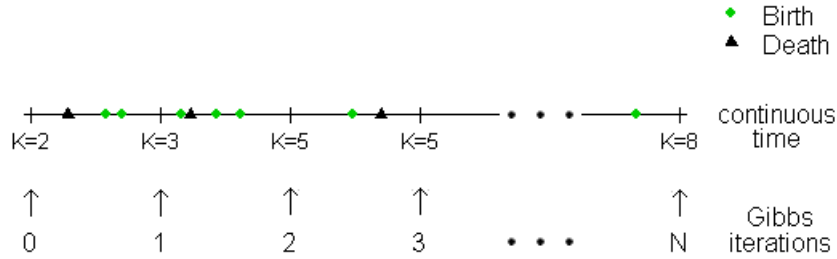


Figure 7.2: A toy example of a birth-death process for $\lambda = 2$ and $\delta = 1$.

where $P(\Theta, \Theta')$ is the rate of transition to a new part of the chain. This implies that the probability of entering a state as a result of a birth must be the same as the probability of leaving it due to a death.

Cappe et. al. (2003) point out that $1/K!$ and $1/(K - 1)!$ in (7.10) denote probabilities of selecting one of $K!$ possible orderings of the elements in Θ , and $(K - 1)!$ possible orderings of $\Theta \setminus \phi_k$. This is necessary because components of Θ lack identifiability, so that $\Theta = (\phi_1, \phi_2, \phi_3, \dots)$ and $\Theta' = (\phi_3, \phi_1, \phi_2, \dots)$, for example, produce the same values of the posterior $p(\Theta, K | \mathbf{y}^m)$.

Letting $p^*(\phi_k) = p^*(\pi_k)p^*(\mu_k)p^*(\Sigma_k^{-1})$ denote the proposal density for new cluster components, the rates of transition to a birth and a death in (7.10) are respectively

$$P(\Theta \setminus \phi_k, \Theta) = \lambda \cdot p^*(\phi_k) \quad \text{and} \quad P(\Theta, \Theta \setminus \phi_k) = \delta_k.$$

It follows that the posterior is a stationary distribution as long as

$$\frac{1}{(K - 1)!} p(\Theta \setminus \phi_k, K - 1 | \mathbf{y}^m) \lambda p^*(\phi_k) = \frac{1}{K!} p(\Theta, K | \mathbf{y}^m) \delta_k,$$

or equivalently

$$\delta_k = \lambda \cdot \frac{p(\Theta \setminus \phi_k, K - 1 | \mathbf{y}^m)}{p(\Theta, K | \mathbf{y}^m)} \cdot \frac{p^*(\phi_k)}{K}$$

for all $k = 1, \dots, K$.

The posterior distribution for the curve clustering model with unknown K is given by:

$$\begin{aligned}
p(\Theta, K | \mathbf{y}^m) &\propto \left\{ \prod_{i=1}^m \mathcal{N}_n(\mathbf{y}_i; B\boldsymbol{\eta}_i, \sigma^2) \right\} \times \left\{ \prod_{i=1}^m \sum_{j=1}^K \mathcal{N}_p(\boldsymbol{\eta}_i; \boldsymbol{\mu}_j, \Sigma_j) P(z_i = j) \right\} \\
&\times \left\{ \prod_{j=1}^K \mathcal{N}_p(\boldsymbol{\mu}_j; \mathbf{v}, V) \times \mathcal{W}_p(\Sigma_j^{-1}; \rho, (\rho R)^{-1}) \right\} \times \mathcal{D}(\boldsymbol{\pi}; 1, \dots, 1) \\
&\times \mathcal{W}_p(V^{-1}; a, (aV_0)^{-1}) \times \mathcal{W}_p(R; b, (bR_0)^{-1}) \times \Gamma(\sigma^{-2}; g, h) \times \mathcal{P}(K; \lambda).
\end{aligned}$$

Let $q(\boldsymbol{\eta} | \Phi) = \prod_i \sum_j \pi_j \mathcal{N}(\boldsymbol{\eta}_i; \boldsymbol{\mu}_j, \Sigma_j)$, and note that $\mathcal{D}(\boldsymbol{\pi}; 1, \dots, 1) = (K-1)!$ and $\mathcal{P}(K; \lambda) \propto \lambda^K / K!$ from the list of distributions in Appendix A. Then the death rates are:

$$\begin{aligned}
\delta_k &= \frac{\lambda}{K} \cdot \frac{q(\boldsymbol{\eta} | \Phi \setminus \phi_k) \left\{ \prod_{j=1 \neq k}^K \mathcal{N}(\boldsymbol{\mu}_j) \mathcal{W}_p(\Sigma_j^{-1}) \right\} \mathcal{D}\left(\frac{\boldsymbol{\pi} \setminus \pi_k}{1 - \pi_k}; 1, \dots, 1\right) \mathcal{P}(K-1; \lambda)}{q(\boldsymbol{\eta} | \Phi) \left\{ \prod_{j=1}^K \mathcal{N}(\boldsymbol{\mu}_j) \mathcal{W}_p(\Sigma_j^{-1}) \right\} \mathcal{D}(\boldsymbol{\pi}; 1, \dots, 1) \mathcal{P}(K; \lambda)} \cdot p^*(\phi_k) \\
&= \frac{\lambda}{K} \cdot \frac{q(\boldsymbol{\eta} | \Phi \setminus \phi_k)}{q(\boldsymbol{\eta} | \Phi)} \cdot \frac{1}{\mathcal{N}(\boldsymbol{\mu}_k) \mathcal{W}_p(\Sigma_k^{-1})} \cdot \frac{1}{K(1 - \pi_k)^{K-1}} \cdot \frac{K}{\lambda} \cdot p^*(\phi_k) \\
&= \frac{q(\boldsymbol{\eta} | \Phi \setminus \phi_k)}{q(\boldsymbol{\eta} | \Phi)} \cdot \frac{p^*(\pi_k) p^*(\boldsymbol{\mu}_k) p^*(\Sigma_k^{-1})}{\text{Beta}(\pi_k; 1, K) \mathcal{N}(\boldsymbol{\mu}_k) \mathcal{W}_p(\Sigma_k^{-1})} \tag{7.11}
\end{aligned}$$

Assuming a $\text{Beta}(1, K)$ proposal for the proportion parameters, and proposals from the priors for the within-cluster mean and covariance components, the death rates for model (7.1) simplify to:

$$\delta_k = \frac{q(\boldsymbol{\eta} | \Phi \setminus \phi_k)}{q(\boldsymbol{\eta} | \Phi)}. \tag{7.12}$$

That is, the death rate for cluster k is simply the probability contribution made by excluding the component from the model, which makes sense intuitively.

Label Switching

A common challenge in fitting Bayesian mixture models is the switching of cluster labels as the MCMC algorithm is run (Richardson & Green 1997, Stephens 2000a). The problem

stems from the fact that the mixture density

$$q(\boldsymbol{\eta}|\Phi) = \prod_i \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{\eta}_i; \boldsymbol{\mu}_k, \Sigma_k) \quad (7.13)$$

in the joint posterior is invariant to the relabeling of the mixture components. That is, the value of $q(\boldsymbol{\eta}|\Phi)$ is the same regardless of whether the cluster labels are $(1, 2, \dots, K)$ or any permutation thereof. This can cause closely related clusters to exchange labels as the MCMC algorithm runs.

Since the modes of (7.13) tend to be more uniquely identified in high-dimensional space, the problem becomes less prominent as the number of basis functions used to fit the random effects increases. Nonetheless, we recommend applying a relabeling algorithm proposed by Stephens (2000b) to ensure that the cluster labels remain consistent.

Stephens' (2000b) algorithm works by considering all $K!$ possible permutations of the labels and picking the best one at every step of the MCMC. The permuted labels are chosen by minimizing the Kullback-Leibler (K-L) difference between the present estimate of (7.13) and its target value. The target value of the density is calculated as an average over the first few iterations. Complete details can be found on pg. 8 of Stephens (2000b).

BDMCMC Sampling for the Bayesian Curve Clustering Model

A complete BDMCMC algorithm for the Bayesian curve clustering model is provided below. It follows the recipe prescribed by Stephens (2000a) and adopted by Zhou & Wakefield (2005).

Begin by setting the initial values $K^{(0)}$, $\Phi^{(0)}$, $\boldsymbol{\eta}^{(0)}$, $\sigma^{2(0)}$ and $R^{(0)}$. We suggest using least squares to approximate $\boldsymbol{\eta}^{(0)}$ and partitioning clustering methods, such as K-means, to set initial values for cluster-dependent parameters. The latter involves running K-means on the estimated b-spline coefficients $\hat{\boldsymbol{\eta}}^{(0)}$, getting cluster labels for each curve, and then calculating sample means and covariances of the least squares estimates $\{\hat{\boldsymbol{\eta}}_i^{(0)}, i \in \text{cluster } k\}$. Initial values for the remaining parameters can be obtained as draws from their respective priors.

Proceed to repeat the following steps for $s = 1, 2, \dots, N$ iterations.

1. Sample $K^{(s+1)}$ and $\Phi^{(s+1)}$ by running the birth-death process for a fixed time t_0 . That is, start with $K^{(s+1)} = K^{(s)}$ and $\Phi^{(s+1)} = \Phi^{(s)}$, then repeat steps i)-iii) for $t \leq t_0$.

i) Simulate the time to the next jump, $t \sim EXP(1/(\lambda + \delta))$, where λ is a fixed birth rate and $\delta = \sum_{k=1}^K \delta_k$ is the overall death rate calculated according to (7.12).

ii) Simulate the type of jump, $X = I(\text{birth}) \sim \mathcal{B}\left(\frac{\lambda}{\lambda + \delta}\right)$.

iii) If a death occurs, identify cluster k corresponding to the largest death rate δ_k . Adjust proportion parameters $\boldsymbol{\pi}^{(s+1)} = \frac{\boldsymbol{\pi}^{(s+1)}}{1 - \pi_k^{(s+1)}}$ to sum to one. Update $K^{(s+1)} = K^{(s+1)} - 1$ and $\Phi^{(s+1)} = \Phi^{(s+1)} \setminus \phi_k^{(s+1)}$.

If a birth occurs, propose a new cluster $\phi_{new} = (\pi_{new}, \boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$ from the proposal distribution $p^*(\phi)$, e.g.

$$\begin{aligned} \pi_{new} &\sim \text{Beta}(1, K^{(s)} + 1), \\ \boldsymbol{\mu}_{new} &\sim \mathcal{N}(\boldsymbol{v}^{(s)}, V^{(s)}), \\ \boldsymbol{\Sigma}_{new} &\sim \mathcal{W}(\rho, (\rho R^{(s)})^{-1}). \end{aligned} \tag{7.14}$$

Update $K^{(s+1)} = K^{(s+1)} + 1$ and $\Phi^{(s+1)} = \Phi^{(s+1)} \cup \phi_{new}$.

2. Simulate $z_i^{(s+1)}$ and $\boldsymbol{\eta}_i^{(s+1)}$ for $i = 1, \dots, m$ from their full conditionals.

3. Simulate $R^{(s+1)}$ and $\sigma^{2(s+1)}$ from their respective full conditionals.

4. Simulate cluster-specific parameters $\phi_k^{(s+1)} = \{\pi_k^{(s+1)}, \boldsymbol{\mu}_k^{(s+1)}, \boldsymbol{\Sigma}_k^{(s+1)}\}$ from their full conditionals.

An improvement incorporated into this algorithm is the use of K-means to set initial values on the cluster-specific parameters. While these results still depend on the arbitrarily-chosen initial choice of K , we found the approach to be helpful in practice. Reasonable choice of initial values will lead to chains that burn in faster.

Chapter 8

Bayesian Random Effects Clustering

8.1 Model

In this chapter, we extend the Bayesian curve clustering model to account for systematic changes in the process, such as valve or ram effects. We do this by expanding the basic model in (7.1) as follows:

$$\begin{aligned} \mathbf{y}_{ij} &= B_1 \boldsymbol{\gamma}_j + B_2 \boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij}; \\ \substack{n \times 1} & \quad \substack{n \times p} \quad \substack{p \times 1} & \quad \substack{n \times q} \quad \substack{q \times 1} & \quad \substack{n \times 1} \end{aligned} \quad (8.1)$$
$$\boldsymbol{\eta}_{ij} | z_{ij} = k \sim \mathcal{N}_q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{and} \quad \boldsymbol{\epsilon}_{ij} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Here $i = 1, \dots, c$ could index the engine head (which we will call part) and $j = 1, \dots, v$ the valve number, so that the total number of observed curves is $c \times v = m$.

The new parameters in this model are the $\boldsymbol{\gamma}_j$'s, which are unknown quantities that do not depend on the cluster label. We can think of (8.1) as a mixed effects mixture model similar to the linear mixed effects model (4.2) in Chapter 4. In fact, for $K = 1$, the two models are equivalent. As before, the role of $\boldsymbol{\gamma}_j$ is to account for fixed systematic changes in the process, while the clustering is performed on the random $\boldsymbol{\eta}_{ij}$'s. In each case, a different set of basis functions (B_1 and B_2) is used to estimate the two terms in the model. We shall refer to our new approach as Bayesian random effects clustering (BREC).

In the context of force exertion data, model (8.1) allows for the possibility that, consistently throughout the process, each of the $v = 7$ valve seats may have been differently

inserted. An equivalent ad-hoc solution would be to subtract off the smoothed mean curves for each valve from all observations associated with that valve, and then cluster b-spline coefficients obtained from smoothing the residual curves as was done in §6.5. If several types of systematic effects are present (such as both valve and ram effects), subtracting off mean curves can get cumbersome, whereas the comprehensive structure of (8.1) is naturally appealing. Furthermore, by formalizing the existence of fixed effects in the model, we are able to accurately estimate the curve roughness component σ^2 .

Another important advantage of BREC is its generality. Previous work in clustering curves such as James & Sugar (2003), Zhou & Wakefield (2005), and Gaffney & Smyth (2003) are all special cases of (8.1) with $B_1\boldsymbol{\gamma}_j = \mathbf{0}$. For the valve seat example, we could also set $\boldsymbol{\eta}_{ij} = \boldsymbol{\eta}_i$, which would be equivalent to simultaneously clustering all of the seven insertions on the same part after the valve effects have been removed. Another possibility mentioned earlier would be to account for any differences associated with the two types of rams that simultaneously insert the intake and the exhaust valve seats. This can be done by including another fixed component in the model corresponding to the ram effect, e.g., letting $l = 1, 2$ index the two types of rams,

$$\begin{aligned} \mathbf{y}_{ijl} &= B_0\boldsymbol{\alpha}_l + B_1\boldsymbol{\gamma}_j + B_2\boldsymbol{\eta}_{ijl} + \epsilon_{ij}; \\ \boldsymbol{\eta}_{ijl}|z_{ijl} = k &\sim \mathcal{N}_q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{and} \quad \epsilon_{ijl} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n). \end{aligned} \tag{8.2}$$

An added bonus of this representation over say (7.1) is a type of dimension reduction that occurs at the clustering level. That is, while a large number of basis functions are necessary to smooth over the valve means in the fixed effects, only a few need be used to smooth over the residual curves in the random effects (i.e. $q \ll p$). As such the quantities being clustered are in a low-dimensional subspace, which leads to improved performance in model-fitting (James & Sugar 2003 and Zhou & Wakefield 2005).

As in Chapter 7, a Bayesian approach to inference is adopted. This will enable assessment of uncertainty in all model parameters, particularly the number of clusters K . Prior specification and posterior estimation using BDMCMC are extended from Chapter 7 in §8.2 and §8.3. Possible applications of the BREC model in practice are discussed in §8.4, and synthetic and real data are analyzed in §8.5 to show accuracy and efficacy of our methodology. A discussion completes this chapter in §8.6.

8.2 Priors

Model (8.1) is characterized by the following set of unknown parameters:

$$\Theta = (\{z_{ij}, \boldsymbol{\eta}_{ij}\}_{i=1}^c, \boldsymbol{\gamma}_j\}_{j=1}^v, \boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K, R, \sigma^2, K).$$

A hierarchy of prior distributions assigned to these quantities includes:

$$\begin{aligned} z_{ij} = k | \boldsymbol{\pi} &\sim \mathcal{B}(\pi_k) \\ \boldsymbol{\pi} | K &\sim \mathcal{D}(\tau_1, \dots, \tau_K) \\ \boldsymbol{\gamma}_j &\sim \mathcal{N}_p(\mathbf{w}, W) \\ \boldsymbol{\mu}_k &\sim \mathcal{N}_q(\mathbf{0}, V) \\ \Sigma_k^{-1} | R &\sim \mathcal{W}_q(\rho, (\rho R)^{-1}) \\ R &\sim \mathcal{W}_q(b, (bR_0)^{-1}) \\ \sigma^{-2} &\sim \Gamma(g, h) \\ K &\sim \mathcal{P}(\lambda). \end{aligned} \tag{8.3}$$

The list of priors specified in (8.3) is similar to (7.6), with one additional prior placed on the systematic components $\boldsymbol{\gamma}_j$ and a change to a distribution centered on zero for the cluster means $\boldsymbol{\mu}_k$. A normal prior was chosen for the fixed effects in order to ensure conditional conjugacy of the posterior distribution. Cluster means of the random effects are assumed to have an expected value of zero, because these are based on the random effects, which now model the residual curves after subtracting off the valve means. A graph summarizing prior dependencies among the parameters is displayed in Figure 8.1.

Suggested values for the hyperparameters are summarized below. As before, if we let \mathbf{x}_{\max} and \mathbf{x}_{\min} denote vectors of length d containing maximum and minimum values for each row of some $d \times m$ matrix $\mathbf{x} = [\mathbf{x}_1^T, \dots, \mathbf{x}_m^T]$, then the scalar $r_{\mathbf{x}} = \|\mathbf{x}_{\max} - \mathbf{x}_{\min}\|^2/d$

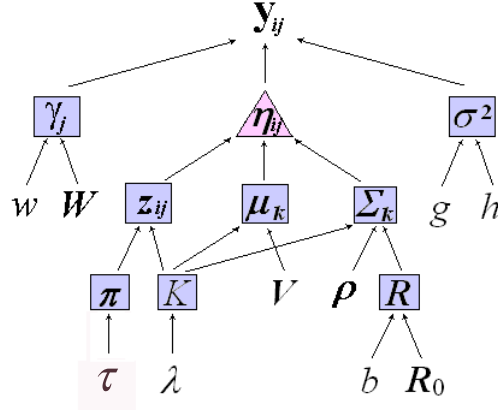


Figure 8.1: DAG of the Bayesian random effects clustering model. The diagram helps visualize the hierarchical dependence amongst the parameters and the data. A triangle frames the quantity being clustered. Quantities not framed by boxes are fixed, or in the case of \mathbf{y}_{ij} , are observed.

represents a measure of spread in the data similar to that of Stephens, 1997 (pg. 37).

$$\begin{aligned}
 \mathbf{w} &= (\hat{\gamma}_{\min} + \hat{\gamma}_{\max})/2; \\
 W &= r_{\hat{\gamma}} I_p; \\
 \tau_k &= 1 \text{ for all } k = 1, \dots, K; \\
 V &= r_{\hat{\eta}} I_q; \\
 R_0 &= (100/r_{\hat{\eta}}) I_q; \quad \rho = b = 2q; \\
 g &= 0.01 \text{ or } 0.001; \quad h = 1/g; \\
 \lambda &= 3, 5, \text{ or } 10.
 \end{aligned}$$

Values of $\hat{\gamma}_{\min}$, $\hat{\gamma}_{\max}$, $r_{\hat{\gamma}}$, and $r_{\hat{\eta}}$ are calculated based on the estimated fixed effect and predicted random effect coefficients. These prior parameters can be specified empirically using least squares to fit (8.1), such that $\hat{\gamma}_j = (B_1^T B_1)^{-1} B_1^T \bar{\mathbf{y}}_j$ and $\hat{\boldsymbol{\eta}}_{ij} = (B_2^T B_2)^{-1} B_2^T (\mathbf{y}_{ij} - B_1 \hat{\boldsymbol{\gamma}}_j)$.

Our rationale in specifying the prior distribution on the fixed effects is the same as that of assigning the prior on the random effects. That is, we want to centre the fixed effects

prior on the midpoint of the sampling distribution ($\mathbf{w} = (\hat{\gamma}_{\min} + \hat{\gamma}_{\max})/2$), and choose a covariance W that is sufficiently large so that the prior is diffuse over the range of the estimated coefficients.

8.3 Estimation

Letting c denote the number of parts and v the number of insertions made on each part as indicated earlier, a total of $m = c \cdot v$ observed curves are considered. Then the likelihood function for model (8.1) is:

$$L(\boldsymbol{\gamma}^m, \boldsymbol{\eta}, \sigma^2 | \mathbf{y}^m) = \prod_{i=1}^r \prod_{j=1}^v \mathcal{N}_n(\mathbf{y}_{ij}; B_1 \boldsymbol{\gamma}_j + B_2 \boldsymbol{\eta}_{ij}, \sigma^2).$$

Combining this information with our prior beliefs (for fixed K), we get the following joint posterior for the unknown parameters:

$$\begin{aligned} p(\Theta | \mathbf{y}^m) &\propto \prod_{i=1}^r \left\{ \prod_{j=1}^v \mathcal{N}_n(\mathbf{y}_{ij}; B_1 \boldsymbol{\gamma}_j + B_2 \boldsymbol{\eta}_{ij}, \sigma^2) \times \mathcal{N}_p(\boldsymbol{\gamma}_j; \mathbf{w}, W) \right\} \times \\ &\quad \prod_{k=1}^K \left\{ \prod_{ij: z_{ij}=k}^{m_k} \left\{ \mathcal{N}_q(\boldsymbol{\eta}_{ij}; z_{ij}, \boldsymbol{\mu}_{z_{ij}}, \Sigma_{z_{ij}}) \times \mathcal{B}(z_{ij}; \pi_{z_{ij}}) \right\} \mathcal{N}_q(\boldsymbol{\mu}_k; \mathbf{0}, V) \times \mathcal{W}_q(\Sigma_k^{-1}; \rho, (\rho R)^{-1}) \right\} \\ &\quad \times \mathcal{D}(\boldsymbol{\pi}; \tau_1, \dots, \tau_K) \times \mathcal{W}_q(R; b, (bR_0)^{-1}) \times \Gamma(\sigma^{-2}; g, h). \end{aligned} \quad (8.4)$$

This can then be used to obtain full conditional distributions of the parameters, which are summarized below. Complete derivations are provided in Appendix B.2.

$$\begin{aligned} p(z_{ij} = k | \dots) &\propto \mathcal{N}(\boldsymbol{\eta}_{ij}; \boldsymbol{\mu}_k, \Sigma_k) \cdot \pi_k \\ \boldsymbol{\pi} | \dots &\sim \mathcal{D}(m_1 + \tau_1, \dots, m_K + \tau_K) \\ \boldsymbol{\eta}_{ij} | \dots &\sim \mathcal{N}_p \left((\sigma^{-2} B_2^T B_2 + \Sigma_k^{-1})^{-1} (\sigma^{-2} B_2^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j) + \Sigma_k^{-1} \boldsymbol{\mu}_k), \left[\sigma^{-2} B_2^T B_2 + \Sigma_k^{-1} \right]^{-1} \right) \\ \boldsymbol{\gamma}_j | \dots &\sim \mathcal{N}_q \left((c\sigma^{-2} B_1^T B_1 + W^{-1})^{-1} (c\sigma^{-2} B_1^T (\bar{\mathbf{y}}_j - B_2 \bar{\boldsymbol{\eta}}_j) + W^{-1} \mathbf{w}), \left[c\sigma^{-2} B_1^T B_1 + W^{-1} \right]^{-1} \right) \\ \boldsymbol{\mu}_k | \dots &\sim \mathcal{N}_p \left(m_k (m_k \Sigma_k^{-1} + V^{-1})^{-1} \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k, \left[m_k \Sigma_k^{-1} + V^{-1} \right]^{-1} \right) \end{aligned}$$

$$\begin{aligned}\Sigma_k^{-1}|\dots &\sim \mathcal{W}_p\left(m_k + \rho, \left[\sum_{i:z_{ij}=k}^{m_k} (\boldsymbol{\eta}_{ij} - \mu_k)(\boldsymbol{\eta}_{ij} - \mu_k)^T + \rho R\right]^{-1}\right) \\ R|\dots &\sim \mathcal{W}_p\left(b + \rho K, \left[bR_0 + \rho \sum_{k=1}^K \Sigma_k^{-1}\right]^{-1}\right) \\ \sigma^{-2}|\dots &\sim \Gamma\left(\frac{nm}{2} + g, \left[\frac{1}{2} \sum_{j=1}^v \sum_{i=1}^c (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij})^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij}) + \frac{1}{h}\right]^{-1}\right).\end{aligned}$$

The list is nearly identical to the full conditionals summarized in §7.3.2. Obvious changes include the addition of new parameters $\boldsymbol{\gamma}_j$, replacing all B with B_2 and \mathbf{y}_i with $\mathbf{y}_{ij} - B_1^T \boldsymbol{\gamma}_j$. Gibbs steps are adjusted accordingly. Then, allowing the number of clusters to vary, BDMCMC is used to generate a large sample of unknown parameters from the joint posterior $p(\Theta, K|\mathbf{y}^m)$. The algorithm is presented below.

BDMCMC Algorithm

Begin by calculating initial values of the estimated fixed effects and predicted random effects:

$$\boldsymbol{\gamma}_j^{(0)} = (B_1^T B_1)^{-1} B_1^T \bar{\mathbf{y}}_j \quad \text{and} \quad \boldsymbol{\eta}_{ij}^{(0)} = (B_2^T B_2)^{-1} B_2^T (\mathbf{y}_{ij} - B_1 \hat{\boldsymbol{\gamma}}_j)$$

Cluster $\boldsymbol{\eta}_{ij}^{(0)}$ using K-means, and use the results to set initial values for cluster-dependent parameters ($z_{ij}^{(0)}$, $\boldsymbol{\pi}^{(0)}$, $\boldsymbol{\mu}_k^{(0)}$, and $\Sigma_k^{(0)}$). Obtain the remaining starting values ($R^{(0)}$, and $\sigma^{2(0)}$) by drawing from the priors. Repeat the following steps for $s = 1, 2, \dots, N$ iterations.

1. Obtain $K^{(s+1)}$ by simulating a birth-death process for $t_0 = 1$ units of time, as outlined in steps i)-iii) in §7.3.3.
2. Draw $z_{ij}^{(s+1)}$, $\boldsymbol{\eta}_{ij}^{(s+1)}$ and $\boldsymbol{\gamma}_j^{(s+1)}$ from their full conditionals ($i = 1, \dots, c; j = 1, \dots, v$).
3. Draw $R^{(s+1)}$ and $\sigma^{2(s+1)}$ from their respective full conditionals.
4. Draw $\pi_k^{(s+1)}$, $\boldsymbol{\mu}_k^{(s+1)}$, $\Sigma_k^{(s+1)}$ from their full conditionals ($k = 1, \dots, K^{(s+1)}$).

Having obtained sample values of the parameters $\Theta^{(1)}, \dots, \Theta^{(N)}$, the first N_b of these are discarded as burn-in, and the remaining posterior samples are used for inference. Convergence of the parameters (and an appropriate value for N_b) can be assessed visually

using trace plots of the simulated values against iteration number or numerically using convergence diagnostics.

8.4 Applications

Before we consider some examples, it is worth re-emphasizing the practical relevance of our approach. In this section, we describe three general applications of curve clustering, and specify how our model can be used in each context.

8.4.1 Mode Detection

Any process observed over the span of several days, months or years begs the question: how does it change (if at all) over time? For functional response data with nothing more recorded than the curves themselves, an answer can be obtained from clustering. By grouping curves with similar features, we are able to identify important changes in the process. To investigate whether or not these changes are time-dependent, probabilities of belonging to each cluster over time can be examined.

In our experience, clustering is especially effective at detecting shifts between reasonably well-defined modes. We tend to think of these modes as corresponding to distinct states of the system. An example would be if we observe that most curves belong to one cluster one day and another cluster on another day, indicating a sudden change in the production process. If instead the change is a gradual shift over time, clustering might not be the best tool to detect this, and sequential profile monitoring tools such as EWMA charts described in §5.4.3 can be used instead.

In Chapter 6, two graphical tools used to identify and assess potential changes in the production process are:

- Predicted probabilities of belonging to each cluster and daily averages thereof plotted in chronological order (see Figures 6.3(a)-(c)). Such plots make it possible to identify which clusters of curves correspond to one or more changes in the process.
- Plots of functional means for each cluster (see Figure 6.3(d)). By looking at the shapes of the within-cluster means, particularly with the help of subject specialists,

it may be possible to determine possible causes for a change in the process.

The two types of visual display can also be used to detect outliers. A more comprehensive solution to the problem is proposed in §8.4.3.

An underlying assumption behind the above-mentioned approach to mode detection is that the data are clustered according to some unobservable structure in the data, rather than any known systematic changes, such as the valve or ram effects in the valve seat insertion example. An important advantage of the BREC model over other clustering techniques is that it is able to control for these systematic changes.

8.4.2 Bayesian Prediction

Another interesting problem is that of prediction for future observations. In manufacturing, for example, it is often important to make instant decisions about the data at the time of production. The usual approach (see Chapter 5) is to compare new data to some well-established base set. Clustering can be applied in the same manner. Having clustered a subset of curves, a new observation \mathbf{y}_{ij} can be classified to the cluster that has the highest associated probability of belonging to that group - $Pr(z_{ij} = k | \mathbf{y}_{ij})$. In the context of valve insertion, if we assume that the clusters correspond to different types of assembly (e.g., “good”, “bad” and varying degrees thereof), assigning new observations to one of these groups provides important information about the current state of the process.

Prediction calculations for the BREC model are presented here. From (8.1),

$$\mathbf{y}_{ij} | z_{ij} = k \sim \mathcal{N}_n(B_1\boldsymbol{\gamma}_j + B_2\boldsymbol{\mu}_k, B_2\Sigma_k B_2^T + \sigma^2 I_n).$$

Then using Bayes rule, for any new insertion on valve j with observed force values \mathbf{y}_{new} , the probability of belonging to cluster k having observed the data is given by

$$p_{new}(k) = Pr(z_{new} = k | \mathbf{y}_{new}) = \frac{\pi_k \cdot \mathcal{N}_n(\mathbf{y}_{new} | \boldsymbol{\mu}_k^*, \Sigma_k^*)}{\sum_{l=1}^K \pi_l \cdot \mathcal{N}_n(\mathbf{y}_{new} | \boldsymbol{\mu}_l^*, \Sigma_l^*)}, \quad (8.5)$$

where

$$\boldsymbol{\mu}_k^* = B_1\boldsymbol{\gamma}_j + B_2\boldsymbol{\mu}_k \quad \text{and} \quad \Sigma_k^* = B_2\Sigma_k B_2^T + \sigma^2 I_n$$

for $k = 1, \dots, \hat{K}$.

Using (8.5), a new curve is assigned to the cluster that has the highest estimated value of $p_{new}(k)$. Since the parameters required to evaluate (8.5) are unknown, the posterior must be used to estimate them. A simple approach is to calculate posterior means of all parameters in (8.5), and “plug in” these values to obtain $\hat{p}_{new}(k)$. The term $\hat{p}_{new}(k)$ could also be estimated as a posterior expectation, which can be evaluated based on MCMC output as follows:

$$\hat{p}_{new}(k) = \frac{1}{N - N_b} \sum_{s=N_b}^N \frac{\pi_k^{(s)} \cdot \mathcal{N}_n(\mathbf{y}_{new} | \boldsymbol{\mu}_k^{*(s)}, \Sigma_k^{*(s)})}{\sum_{l=1}^K \pi_l^{(s)} \cdot \mathcal{N}_n(\mathbf{y}_{new} | \boldsymbol{\mu}_l^{*(s)}, \Sigma_l^{*(s)})}, \quad (8.6)$$

where $s = N_b, \dots, N$ denotes the iteration numbers after the burn-in period,

$$\boldsymbol{\mu}_k^{*(s)} = B_1 \boldsymbol{\gamma}_j^{(s)} + B_2 \boldsymbol{\mu}_k^{(s)} \quad \text{and} \quad \Sigma_k^{*(s)} = B_2 \Sigma_k^{(s)} B_2^T + \sigma^{2(s)} I_n.$$

Expression (8.6) is used to predict cluster labels in this thesis.

An obvious limitation of this approach is that a new curve may not follow any known process, and thus would not belong to any of the clusters. Observations that do not belong to any of the known clusters should be considered outliers. A profile monitoring technique for detecting outliers using the clustering model is proposed next.

8.4.3 Profile Monitoring

One way to determine if a new observation is an outlier with respect to a base set of clustered curves is to verify that it is far away from all of the existing clusters. In order to formalize this, we must define a measure of distance. A natural candidate for multivariate quantities is the Mahalanobis distance, defined as

$$T_k^2 = (\hat{\boldsymbol{\eta}}_{ij} - \hat{\boldsymbol{\mu}}_k)^T \hat{\Sigma}_k^{-1} (\hat{\boldsymbol{\eta}}_{ij} - \hat{\boldsymbol{\mu}}_k).$$

The metric represents the distance of the predicted random effects from the k^{th} cluster mean scaled by the k^{th} within-cluster covariance. As such, a new observation is likely to

belong to the cluster corresponding to the lowest value of T_1^2, \dots, T_K^2 . Let

$$D_{ij} = \min(T_1^2, \dots, T_K^2) \quad (8.7)$$

denote this distance to the closest cluster. The higher this value the less likely it is that a curve belongs to any of the clusters, since it is far away from all of them. Thus, one possible way to detect outliers is to monitor changes in the D_{ij} .

Using statistical process control methodology outlined in Chapter 5, profile-monitoring using a clustered base set proceeds in two phases. The first phase involves the analysis and assessment of in-control data, with the results used in the second phase to classify new curves. The steps of the proposed procedure are outlined below.

PHASE I

1. Cluster data that is assumed to be “in-control”. Obtain $\hat{\boldsymbol{\eta}}_{ij}$, $\hat{\boldsymbol{\mu}}_k$ and $\hat{\Sigma}_k$ as posterior means estimated from the MCMC output.
2. For each curve in the base set, calculate D_{ij} metrics as stated in (8.7).
3. Let UCL be some value so that $D_{ij} < UCL$ for all or most ij . This represents a threshold value for unusual behaviour.

PHASE II

1. For each new curve, obtain point estimates of the random effects

$$\hat{\boldsymbol{\eta}}_{ij} = (B_2^T B_2)^{-1} B_2^T (\mathbf{y}_{ij} - B_1 \hat{\boldsymbol{\gamma}}_j),$$

where $\hat{\boldsymbol{\gamma}}_j = (B_1^T B_1)^{-1} B_1^T \bar{\mathbf{y}}_j$ and $\bar{\mathbf{y}}_j = \sum_{i=1}^c \mathbf{y}_{ij} / c$.

2. Calculate D_{new} using estimates of $\hat{\boldsymbol{\mu}}_k$ and $\hat{\Sigma}_k$ from Phase I.
3. Flag curves that have $D_{new} \geq UCL$ as outliers, where UCL is the upper control limit obtained in Phase I.

Since the above procedure involves monitoring Mahalanobis distances, we shall refer to this profile monitoring procedure as the distance approach.

8.5 Examples

In this section, we use empirical evidence to answer two important questions regarding our BREC approach: does it work, and is it useful? Synthetic data are used to address the first concern (since it requires knowledge of the truth), and real data to examine the latter.

R statistical software (R Development Core Team, 2006) was used to implement the BDMCMC algorithm for fitting the BREC model. The Heidelberger & Welch test (1983) as implemented in the `coda` library was utilized in assessing chain convergence and sufficient burn-in. A description of this convergence criteria can be found in §7.3. Having obtained parameter estimates, Stephens (2000b) relabeling algorithm described in §7.3.3 was run to ensure that the cluster labels remain consistent. We implemented the method in C and ran in R using a wrapper function. The WNLIB ANSI C subroutine library written by Naylor & Chapman (2006) was used to minimize the K-L divergence criteria with respect to the labels.

8.5.1 Synthetic Data

Our key objective in simulating force exertion data was to ensure that it closely resembles the real valve seat insertion problem. To ensure this, we used model-based clustering (MBC) results from Chapter 6 to separate the February data (which we suspect to be well-behaved) into four clusters and obtain plausible estimates for various parameters of the BREC model. More precisely, we followed these steps in simulating new data:

1. Using a subset of curves from the last six days of production in February, calculate two average curves: one for intake and one for exhaust type of valve insertion. Smooth over these average curves to obtain the fixed-effect b-spline coefficients. These will correspond to the true systematic effects in our model (γ_j).
2. Let B_1 denote the matrix of $p = 20$ b-spline basis functions used for fixed effects

smoothing in step 1. Subtract $B_1\gamma_j$'s from the force curves, and use least-squares to estimate b-spline coefficients for this residual data.

3. Cluster the second set of b-spline coefficients using MBC. The four-cluster EEE model appears to fit best according to BIC. Calculate proportions of curves belonging to each cluster (π_k), within-cluster means ($\boldsymbol{\mu}_k$) and within-cluster covariances ($\boldsymbol{\Sigma}_k$).
4. Generate new data according to the model:

$$\begin{aligned}\boldsymbol{\epsilon}_{ij} &\sim MVN(\mathbf{0}, 10^2 I_{100}) \\ \boldsymbol{\eta}_{ij} | z_{ij} = k &\sim MVN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \mathbf{y}_{ij} &= B_1\gamma_j + B_2\boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij}\end{aligned}$$

for $i = 1, \dots, 198$, $j = 1, 2$, and $k = 1, 2, 3, 4$. The columns of B_2 are set to contain $q = 10$ basis functions.

Following these steps, a total of 396 curves split into four clusters with $m_1 = 320$, $m_2 = 36$, $m_3 = 20$ and $m_4 = 20$ observations in each one were generated. The data (\mathbf{y}_{ij}) and the random effect coefficients ($\boldsymbol{\eta}_{ij}$) obtained in step 4 are plotted according to the cluster labels in Figure 8.2. Looking at the plot of the random effects, clustering this data is clearly not a trivial problem since the curves within each cluster are highly variable and the three smaller clusters overlap with the bigger one. In fact, when we use MBC to cluster the 396 true random effects, none of the top three BIC models get the correct number of clusters or the correct covariance structure. The best-fitting model according to BIC is a two-cluster model with an EEE covariance structure (ellipsoids of equal shape, volume and direction), which merges the first three clusters into one. As indicated previously, BIC gives preference to simpler models (with fewer clusters and a simple covariance structure) not necessarily on merit, but because such models have fewer parameters.

BREC appears to do considerably better in estimating K . Figure 8.3(b) shows the predicted marginal distribution of this parameter obtained by running 10,000 iterations of the BDMCMC algorithm, and discarding the first half as burn-in. The plot points clearly to $K = 4$, the true number of clusters. Trace plot in Figure 8.3(a) indicates a rapid

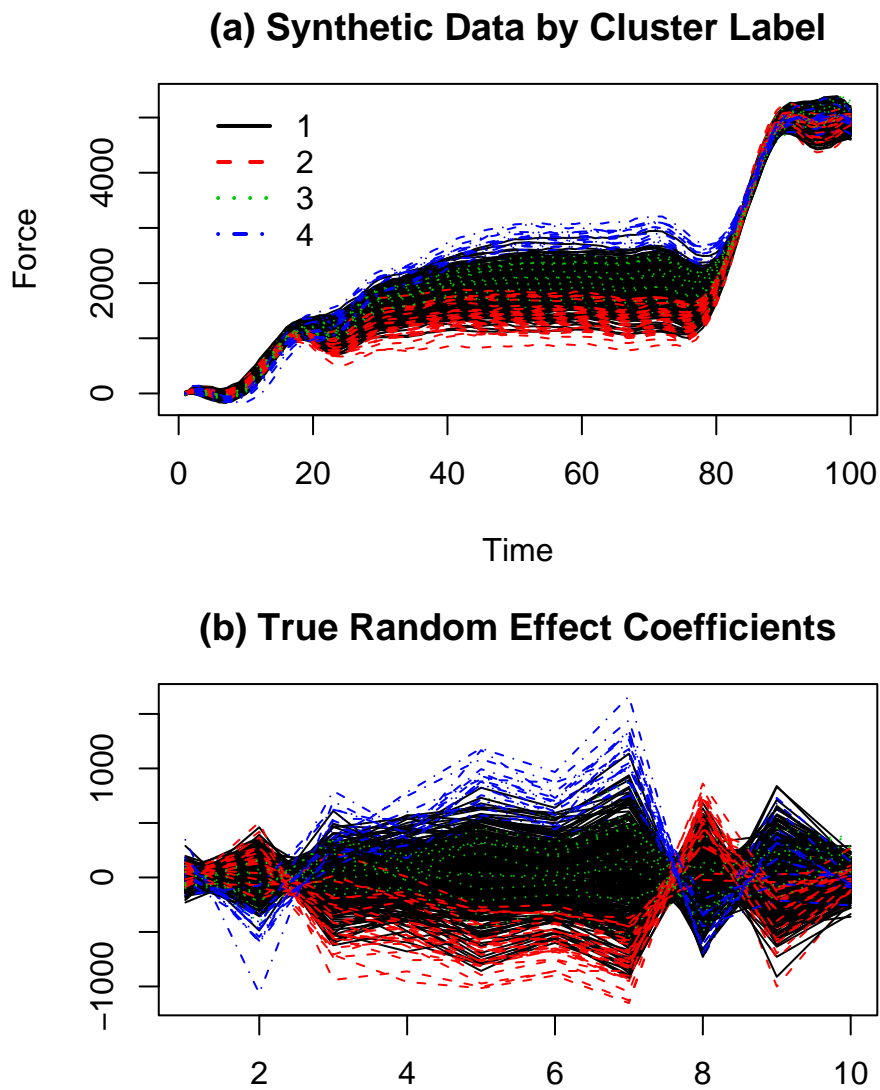


Figure 8.2: (a) Simulated data and (b) b-spline coefficients of the random effects used to generate the data.

convergence to this value from a starting point that was purposely set to be much higher than the truth.

Estimation of the remaining parameters is also good. Table 8.1 summarizes the true versus predicted number of curves assigned to each of the four clusters. It shows that only 19 of the 396 curves have been mislabeled, producing a misclassification rate of 4.8%.

		Predicted labels				m_k
		1	4	2	3	
True labels	1	314	2	4	0	320
	2	4	32	0	0	36
	3	5	0	15	0	20
	4	4	0	0	16	20

Table 8.1: True vs. predicted cluster labels obtained from BREC.

Estimated proportion and roughness parameters are presented below the true values in Table 8.2. Based on the results in Table 8.1, cluster labels (1, 2, 3, 4) for the estimated values of $\boldsymbol{\pi}$ were changed to (1, 4, 2, 3) in order to allow for a comparison with the true labels. The correspondence between the two rows is reassuring.

	π_1	π_2	π_3	π_4	σ^2
True	0.80	0.09	0.05	0.05	100.0
Estimated	0.80	0.10	0.06	0.04	100.4

Table 8.2: Selected true and estimated parameter values obtained from BREC.

However, the true challenge of this model lies in estimating the eight multidimensional within-cluster mean and precision matrix parameters. To investigate whether or not this task was achieved, we plotted the estimated versus true elements of each parameter in Figure 8.4. For guidance, a solid line with slope one and intercept zero is included in each plot. We see that, for the most part, the points in each plot fall close to the line, indicating a close correspondence between the point estimates and the truth.

As a final check, convergence of all of the above-mentioned parameters was investigated using trace plots and H-W convergence diagnostics implemented in the R `coda` library and

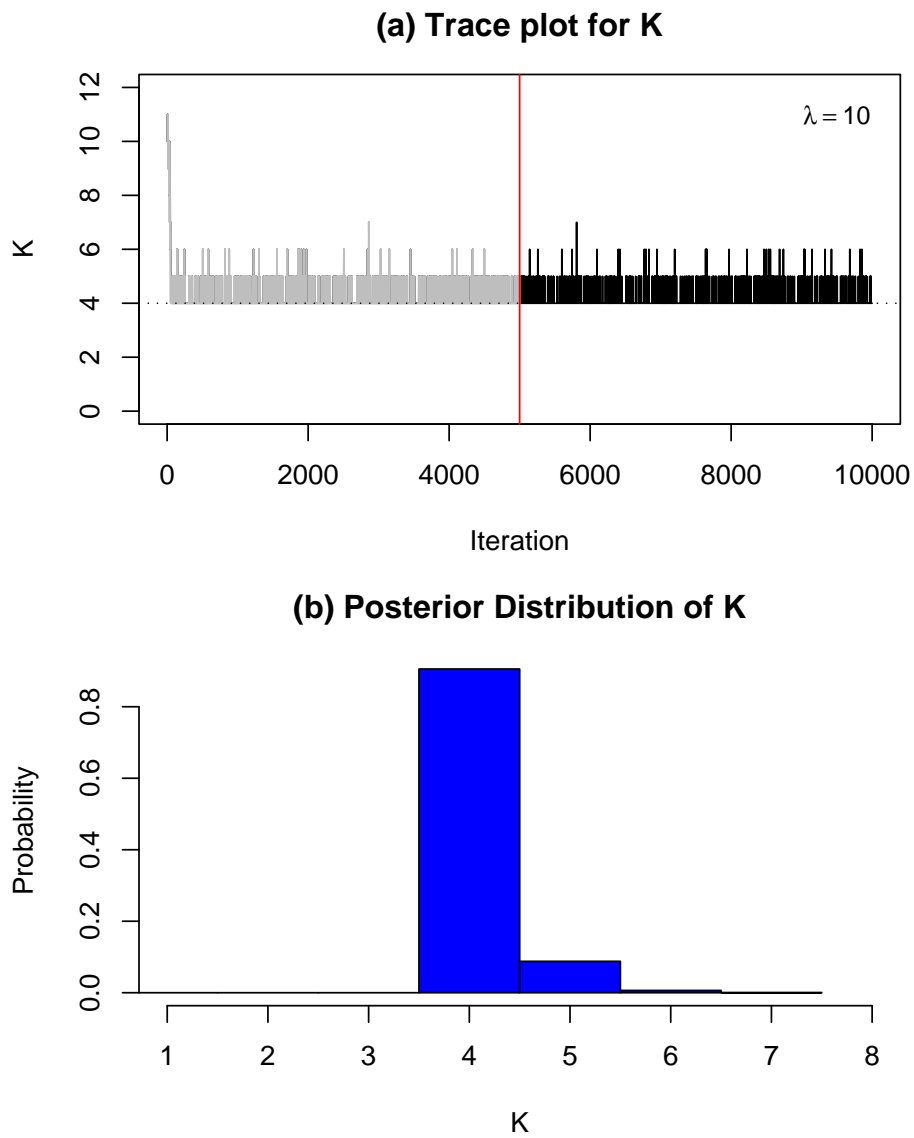


Figure 8.3: Results for synthetic data. (a) Trace plot and (b) estimated probability mass function for the number of clusters K . The vertical line in the top plot marks the end of the burn-in period.

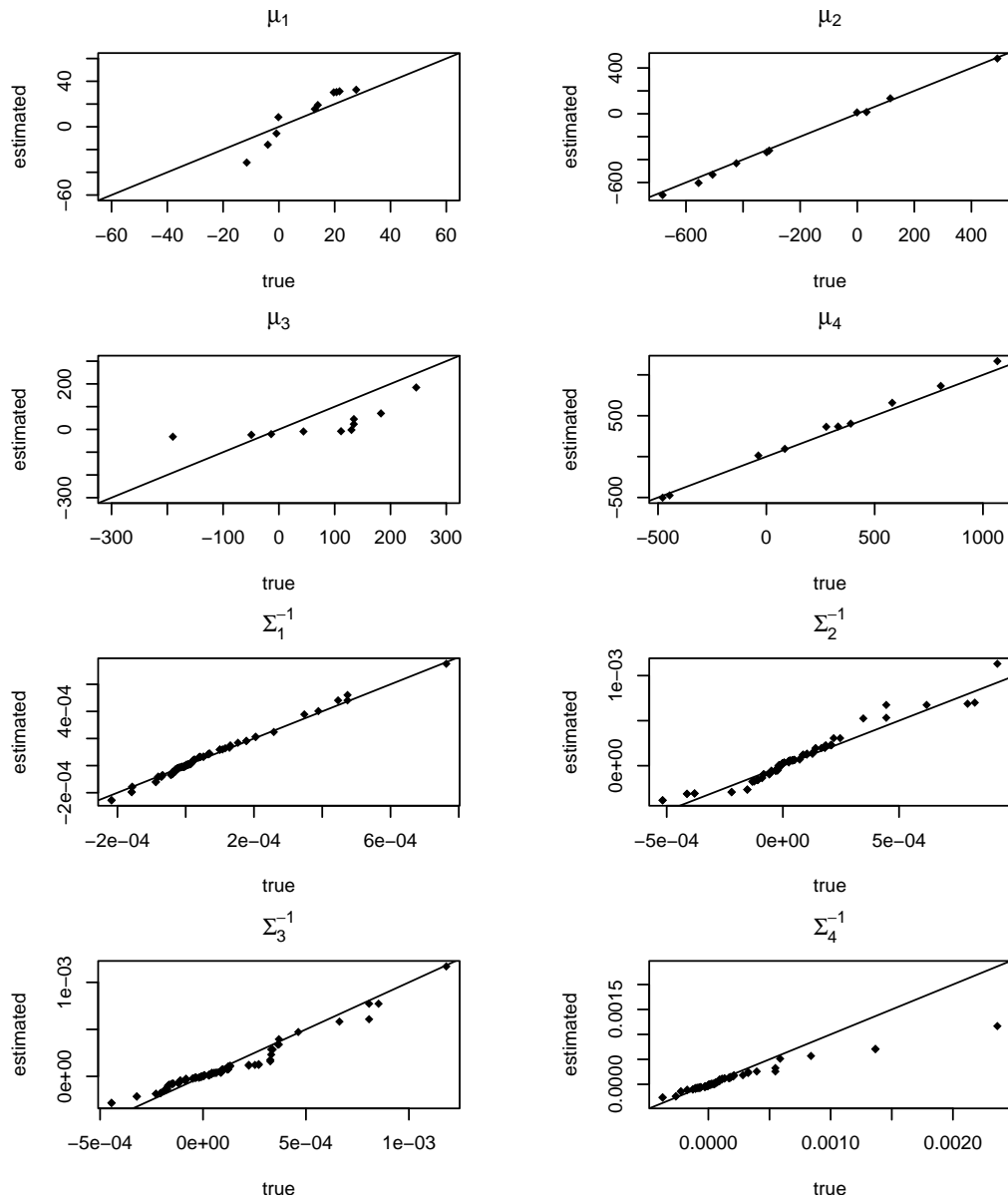


Figure 8.4: Results for synthetic data: point estimates of the within-cluster mean vectors and precision matrices strung out into vectors and plotted against their true values.

summarized in §7.3. Both pointed to sufficient burn-in and convergence of the parameters to stationary values. A sample of the trace and estimated posterior density plots for σ , π_3 , and two randomly selected elements of the $\boldsymbol{\mu}_2$ vector and the Σ_1^{-1} matrix are included in Figure 8.5. Dashed vertical lines mark true values of the parameters in the density plots. The fact that these do not fall in the tails of the posterior distribution provides reassurance that the chain converged close to the true solutions.

In conclusion, empirical results based on synthetic data suggest that our model is both competitive and reliable in estimating the number of clusters and other model parameters. This performance is especially impressive in comparison to model-based clustering since MBC was solving a simpler problem, that of clustering the random effects as if they were data. BREC was inferring the model from the observed data only, a more challenging task.

8.5.2 Force Exertion Data

Let us now consider the real data. As before, we proceed under the assumption that profiles from the last six days in February are “in-control,” and attempt to detect unusual behaviour in January. We carry out process monitoring in two steps, as outlined in §8.4.

Phase I Analysis

Our goal at the preliminary stage of the analysis is to cluster February data, and examine it for unusual behaviour. We used $p = 20$ and $q = 5$ basis functions to model the fixed and random effects respectively. That is, the clustering took place in a five-dimensional space, which is a significant reduction from the 100-dimensional space in which the original data was observed.

The BDMCMC algorithm was run for 20,000 iterations, with the first 10,000 discarded as burn-in. Draws for K converged very quickly, with high posterior probability (0.953) that $K = 3$. According to trace plots, convergence of the remaining parameters was also achieved (see Figure 8.6 for sample trace plots). The Heidelberger & Welch tests were also passed for all of the parameters.

The plots in Figures 8.7 help gain a better understanding of the results, and check the data for the presence of any possible changes in the process. Predicted probabilities

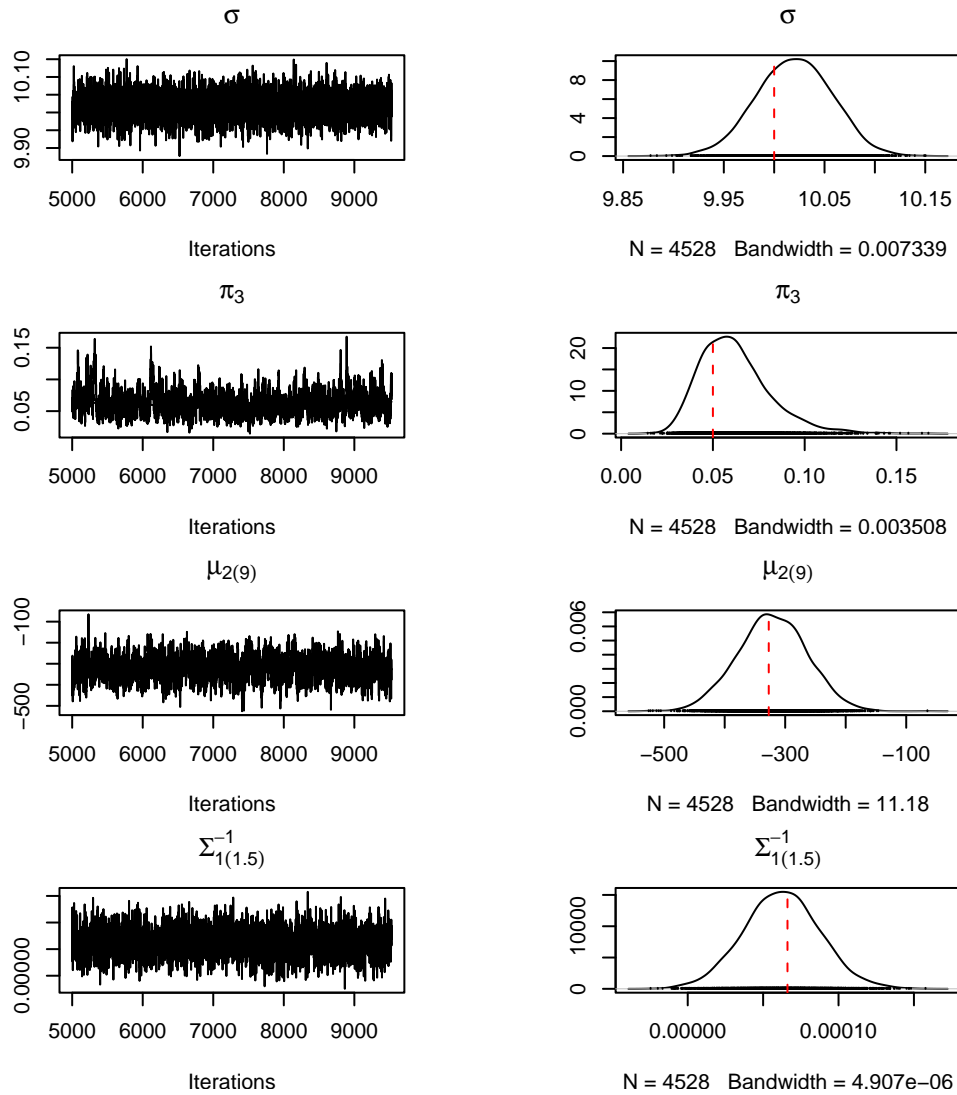


Figure 8.5: Results for synthetic data: trace and estimated density plots for selected parameters of the BREC model. True values of the parameters are marked by vertical dashed lines in the density plots.

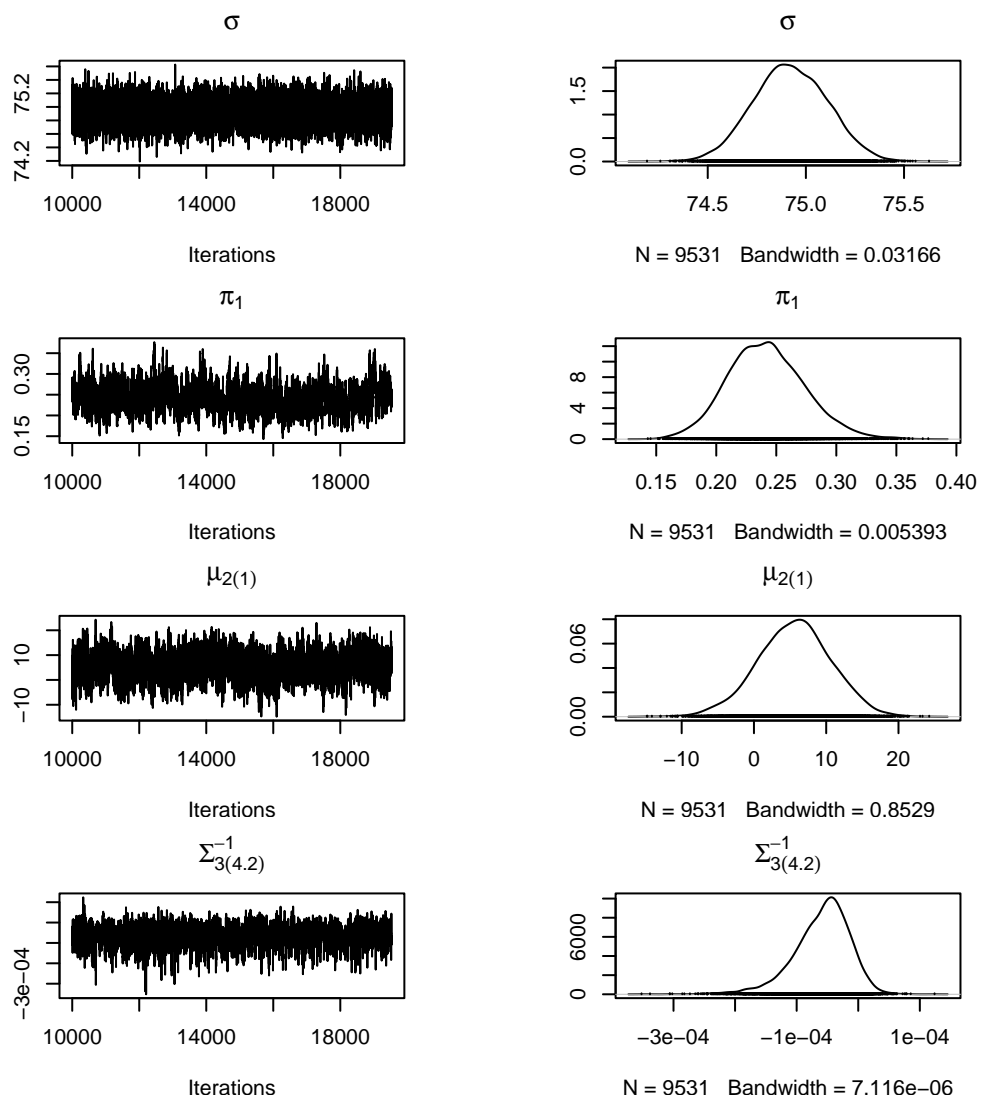


Figure 8.6: Results for February data: trace and estimated density plots for σ^2 , π_1 , and two randomly selected elements of the $\boldsymbol{\mu}_2$ vector and Σ_3^{-1} matrix.

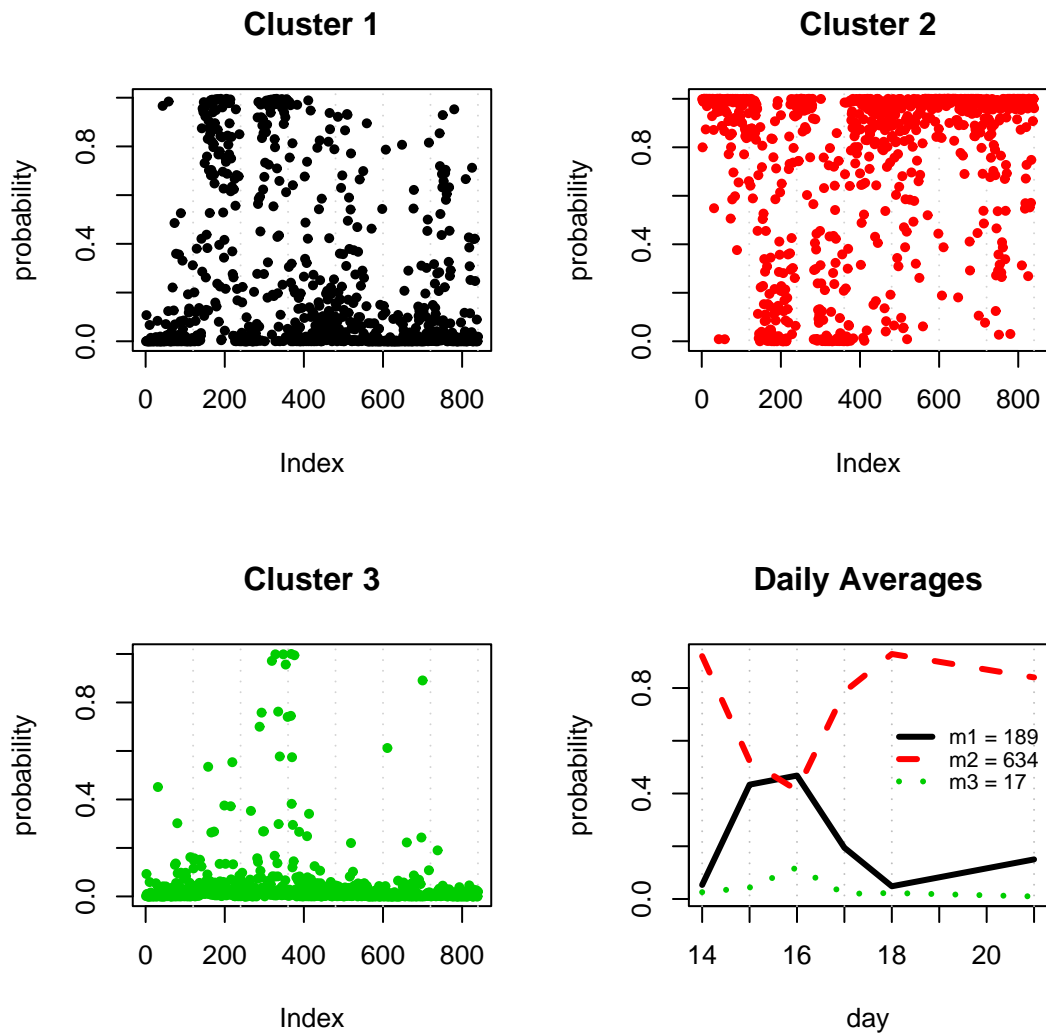


Figure 8.7: Predicted probabilities of belonging to each cluster plotted in chronological order by observation number for February data. Daily averages of these probabilities are displayed in the lower right panel.

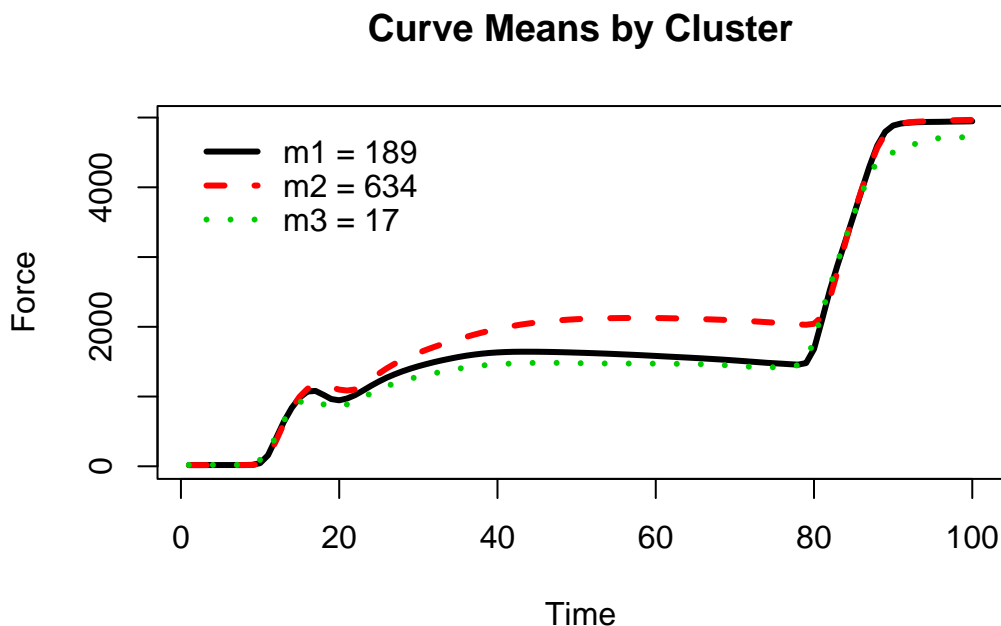


Figure 8.8: Functional means of the curves in each cluster, with cluster labels obtained by fitting the BREC model to February data.

of belonging to each cluster, $\hat{p}_{ij}(k)$, plotted by observation number are displayed. Daily averages of the predicted probabilities are summarized in the lower bottom panel. The plots show that the majority of the data during the six days in February belongs to the second cluster. An exception is the sudden spike in the number of curves assigned to the first cluster on February 15th and 16th.

Plots of curve averages for each cluster in Figure 8.8 indicate that the clusters differ mostly in the amount of force exerted at the mid-point of the insertion (between 20-80 time units). This is evident from the ranging levels of the means in the long flat segment of the curves.

These results seem to agree with our MBC findings in §6.5. A comparison of Figures 6.3, 8.7 and 8.8 reveals that the first two clusters found by the BREC approach are roughly the same as the two found by the MBC of the estimated b-spline coefficients after

subtracting the fixed valve effects. Unlike the previous analysis, an extra cluster was found by the present approach. BIC's preference for simpler models may explain why only two instead of three clusters were found to be best-fitting in §6.5.

As a final step in the Phase I analysis of February data, a control chart of the D_{ij} metrics based on (8.7) is provided in Figure 8.9(a), with the upper control limit set to $UCL = 42$ and marked by a dashed horizontal line. This number was chosen as twice the maximum value of the D_{ij} , driven by the assumption that most of the February data is in-control. This is supported by the fact that these values are relatively homogenous, with none of the distance measures drastically higher than the rest.

Our overall conclusion for the February data is that there are no significant drifts or outliers. The process appears to be governed by three different types of insertions, which differ primarily in the amount of force exerted at the mid-point and the end of the insertion. Parameter estimates from Phase I and the specified UCL are further retained for use in the analysis of January data.

Phase II Analysis

At the final stage, we scan the January data for outliers. Predicted probabilities of belonging to each cluster, $\hat{p}_{new}(k)$, calculated as stated in (8.6) for the new data are plotted in Figure 8.10. Daily averages of these probabilities are also shown in the lower right panel of Figure 8.10. The plots suggest nothing unusual at the start of January, however this could be because of the restriction that the new observations have to belong to just one of the clusters.

A better approach to outlier detection for new data based on the BREC results is the distance control chart obtained by following the three steps of Phase II analysis as outlined in §8.4.3. A plot of the resulting control chart is provided in Figure 8.9(b). As expected, most of the outliers identified by the chart occur at the beginning of the month, before production had a chance to stabilize.

Profiles corresponding to unusual behaviour are highlighted in black and red in Figure 8.9(c). The two colours represent varying degrees of severity in the nature of the outliers, with black marking curves that correspond to $D_{ij} \geq UCL$ and red marking extreme outliers with $D_{ij} \geq 3 \cdot UCL$. By looking at these curves, we see that one of the most prominent

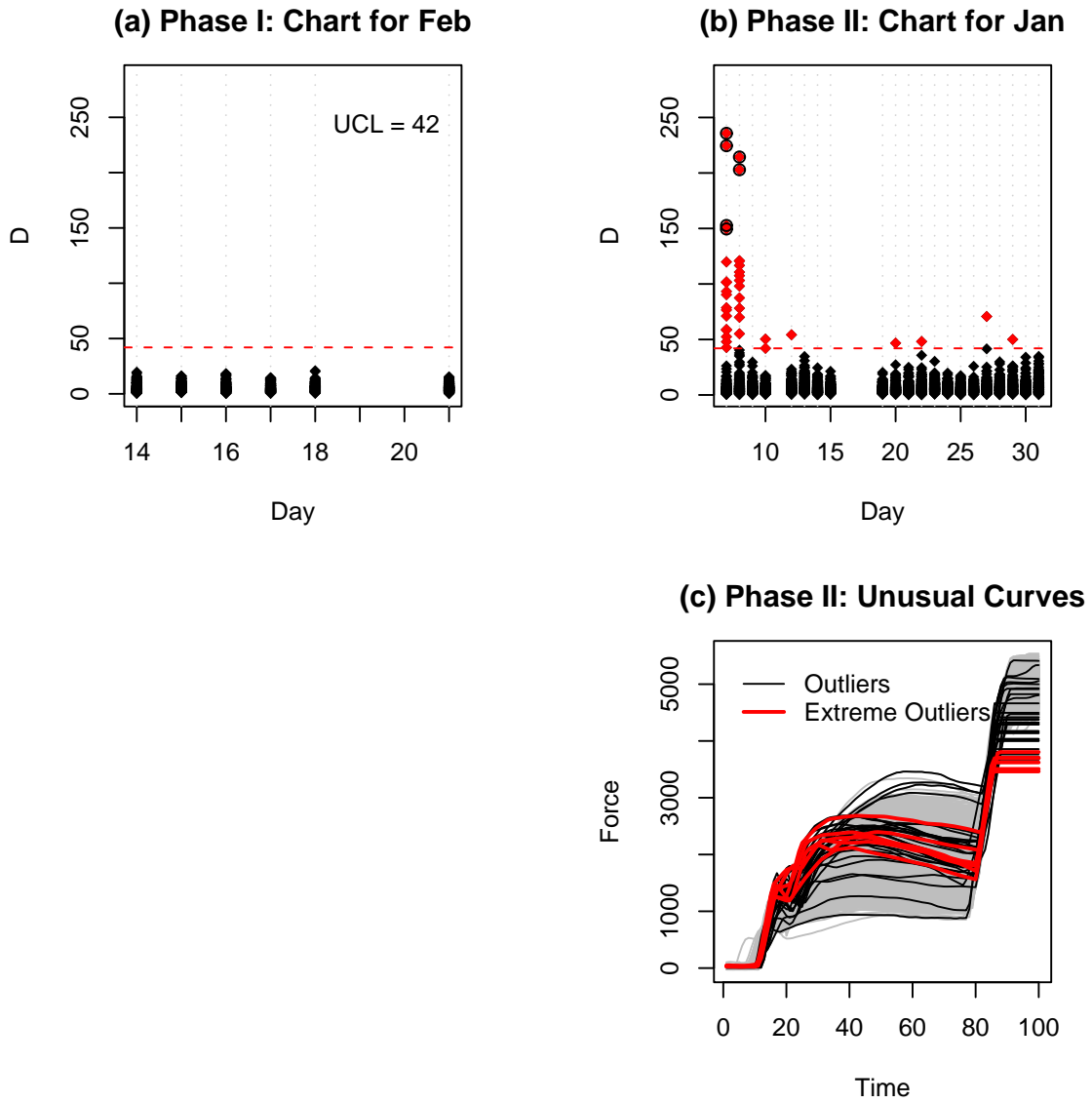


Figure 8.9: Distance control charts for (a) February and (b) January data. The dashed horizontal line corresponds to the UCL. Outliers exceeding this value are displayed in black, and severe outliers in red, alongside the remaining profiles highlighted in gray in (c).

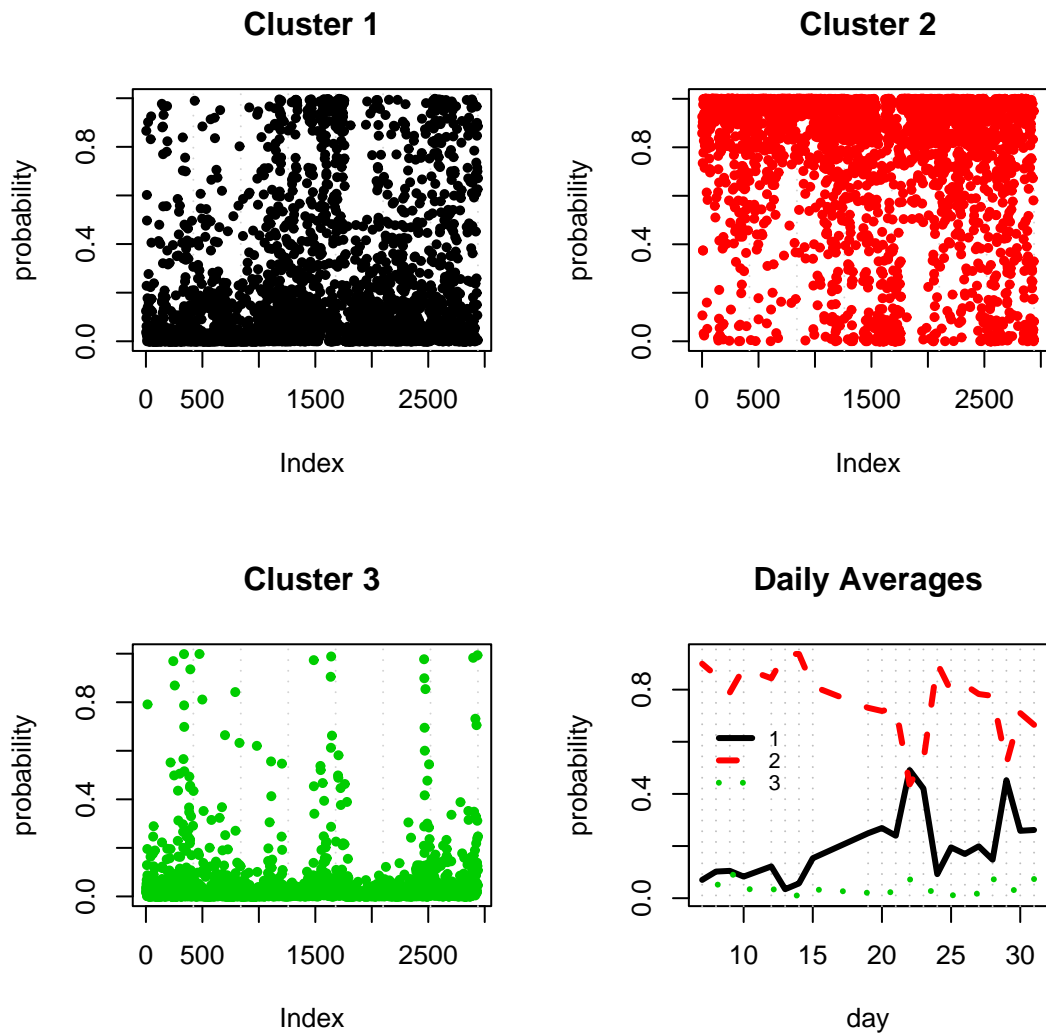


Figure 8.10: Predicted probabilities of belonging to each cluster plotted in chronological order by observation number for January data. Daily averages of these probabilities are displayed in the lower right panel.

differences between them and the remainder of the data is the amount of force applied at the end of the insertion, with the largest outlier showing the smallest force as demonstrated by the level of the short flat end part of the curve. While experts would be able to shed more light into the nature of these differences, it seems plausible that the outliers are “bad” insertions because an insufficient amount force was applied at the crucial end stage of the process when the ram makes contact with the engine head and pushes the valve seat into the cylinder head.

These results match our findings from §5.3.3. Correspondence between Figures 8.9(b)-(c) and 5.2 is highly reassuring. An important advantage of the distance chart and the BREC model is that it can be used to monitor all valves simultaneously, rather than separately monitoring each valve.

8.6 Discussion

Using two different examples, we have now demonstrated that the clustering model proposed in this chapter is both useful and effective. Miscellaneous topics related to this model and possible extensions of the basic idea are considered next. Although we do not implement all of the discussed approaches, a brief summary of how this can be achieved is presented in each case.

Multiple Fixed Effects

Controlling for more than one known effect is easy in the context of our proposed model. This is accomplished by adding corresponding fixed effects terms. An example is the adjustment for ram effects in the valve seat insertion example, as shown in (8.2).

Choice of Basis

Throughout the thesis, we have used b-spline basis functions for linear modeling of functional data. However, our formulation does not preclude other possibilities. For example,

we can model the curves using the linear relationship

$$\mathbf{y}_i = X\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i,$$

where X is an $n \times p$ matrix containing the first p principal components, which can be obtained from PCA on the data itself. An obvious difficulty with this approach is that the data is used twice; first to estimate X and then to estimate the coefficients $\boldsymbol{\eta}_i$. An alternative is to use a subset of the observed curves or some other base set (e.g., historic data) in calculating the basis.

Choice of Covariance Structure

A simplifying assumption behind much of our work is that the residual curves from each model follow a multivariate normal distribution with vector mean zero and a diagonal covariance matrix $\sigma^2 I_n$. This implies that all recordings being made are independent within a curve and between curves. This is somewhat restrictive as observed curve values that are close to each other will likely be more similar than measurements made on opposite ends of the profile. Furthermore, consecutive curves may belong to the same cluster. By tweaking the assumptions made on the error covariance structure of the model, we can take some of these issues into account.

One possibility to consider is the Gaussian process model of the form:

$$\mathbf{y}_{ij}(\mathbf{t}) = B_1\boldsymbol{\gamma}_j + B_2\boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij}(\mathbf{t});$$

$$\boldsymbol{\eta}_{ij}|z_{ij} = k \sim \mathcal{N}_q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{and} \quad \boldsymbol{\epsilon}_{ij}(\mathbf{t}) \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 K(\mathbf{t}, \mathbf{t}')).$$

The term $K(\mathbf{t}, \mathbf{t}')$ represents a correlation matrix, the form of which can be specified in such a way so as to induce correlation between observed components of the force vector that are close. Similar correlation constraints can also be imposed on the specified priors to achieve the same goal.

Nonlinear Clustering

Our present clustering model is linear with respect to both fixed and random effects. While this was sufficient for illustration purposes, it is possible to extend this to fit nonlinear mixed effects models. This is particularly interesting because of the importance of differential equations in modeling dynamic systems that often arise in the context of functional data (Ramsay & Silverman, 2005). Such models are entirely nonlinear, and require careful consideration if model parameters are allowed to be random. MCMC implementation details for nonlinear hierarchical models can be found in Bennett et. al. (1996).

Birth Proposals

Recall from our introduction to BDMCMC that we followed Stephens (2000a) and Zhou & Wakefield (2005) by proposing new cluster parameters $\boldsymbol{\mu}_{new}$ and Σ_{new} from their respective priors as summarized in (7.14). Although this provided favourable results, it may be possible to improve the algorithm by developing smarter birth proposals. For example, we may wish to run a partitioning clustering method (such as K-means) prior to the analysis, and then propose births from the set of pre-determined mixture components corresponding to these clusters.

Another possibility is to change the proposal distribution for the cluster mean to $\boldsymbol{\mu}_{new} \sim \mathcal{N}(\bar{\boldsymbol{\eta}}^{(0)}, S_{\boldsymbol{\eta}}^{(0)})$, where $\bar{\boldsymbol{\eta}}^{(0)}$ and $S_{\boldsymbol{\eta}}^{(0)}$ are the mean and covariance of some initial estimate of the b-spline coefficients. This is different than sampling from the prior in that the proposed cluster centres are likely to be within the true range of $\boldsymbol{\eta}_{ij}$ values. This should lead to plausible clusters being born more often and thus faster convergence.

From (7.11), adjusted death rates for a BDMCMC algorithm in which new cluster-specific parameters $\boldsymbol{\phi}_{new} = \{\pi_{new}, \boldsymbol{\mu}_{new}, \Sigma_{new}\}$ are proposed from some function $p^*(\boldsymbol{\phi}_{new})$ are given by

$$\delta_k = \frac{q(\boldsymbol{\eta}|\Phi \setminus \phi_k)}{q(\boldsymbol{\eta}|\Phi)} \cdot \frac{p^*(\boldsymbol{\phi}_k)}{\mathcal{Beta}(\pi_k; 1, K) \mathcal{N}(\boldsymbol{\mu}_k) \mathcal{W}_p(\Sigma_k^{-1})}.$$

Cape et. al. (2003) extend the idea even further by including two more types of proposals: a split of a cluster and a merge of two clusters. Their method is called continuous time MCMC, of which BDMCMC is a special case. While we did not require such a level

of complexity in fitting our model, exploring the idea of better proposals would be an interesting topic to consider as part of future research.

Chapter 9

Summary and Conclusions

A connecting theme behind the ideas presented in this thesis is the analysis of functional process data such as the valve seat insertion example. Our contributions to this area are summarized in four key points listed below.

- A curve-registration technique for aligning valve seat insertion data is presented in Chapter 2. This pre-processing tool allows for better comparison of the observed data, and plays an integral role in ensuring validity of all subsequent analysis.
- In Chapter 3, two approaches for modeling individual force exertion curves are presented. The first is a b-spline model, which is linear with respect to its parameters, and the second is a novel differential equations model, which is nonlinear and based on the dynamics of the process. The advantage of the latter is that it is more parsimonious and the parameters can be interpreted in the context of the problem. A mixed effects model extending these ideas to fit multiple curves is developed in Chapter 4. The inclusion of both fixed and random effects allows for a flexible model capable of fitting each of the curves individually, at the same time borrowing strength from all of the data. It is also valuable as a dimension-reduction tool, since generally very few random effects are needed to model individual differences from the fixed overall trend.
- A new curve-clustering technique is presented in Chapter 8 as a practical tool for identifying changes in the process, such as drifts over time or outliers. The model

utilizes fixed effects to control for known systematic changes in the process and random effects to identify unobservable latent structures in the data. The novelty of this approach over similar methods, such as model-based clustering (Chapter 6) and its Bayesian equivalent (Chapter 7), is the ability to control for known differences in the data and the dimension reduction that using mixed-effects entails. The model is fit using a Bayesian approach, which allows us to incorporate prior knowledge about the parameters and assess uncertainty associated with estimated quantities. This is particularly useful in that it provides a way of estimating the number of clusters following Stephens' birth-death MCMC paradigm.

- Three new profile-monitoring tools for detecting outlying curves are developed in Chapters 5 and 8. The first two are extensions of a multivariate Hotelling T^2 chart to functional data fitted using linear and nonlinear mixed effects. The third is a Mahalanobis distance chart based on the curve-clustering model. The latter differs from the previous two in that multiple types of “in-control” processes are assumed to exist in Phase I of the analysis.

Efficacy and practicality of our methodology is demonstrated using synthetic and real data throughout the thesis. Conclusions for the valve seat insertion data, in particular, are consistent and point to outliers in early January, at which time the insertion process was initiated. In February, the process was verified to be “in-control” in Chapter 5, using statistical methodology developed in §4.4. These well-behaved data were classified into at least three different processes in Chapter 8.

The questions answered in this thesis suggest more questions and future research possibilities. Many of these have been considered at various parts of the thesis and are summarized next.

With respect to profile monitoring, one unresolved issue is that of choosing the random effects in the model. Depending on which and how many of the parameters are allowed to be random, the chart will differ. One way to address this issue is to fit several models with different sets of random effects, then use model-selection criteria such as AIC or BIC to identify the best fitting one. Other possible extensions of the profile monitoring idea are to develop sequential charts such as CUSUM or EWMA and/or monitor registration

points from curve-registration. One possible EWMA chart for detecting process drifts in functional data that occur gradually over time is proposed and briefly discussed in §5.4.3.

Whenever mixed-effects models are used for curve data, our underlying assumption has been that the residual curves from each model follow a $\mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$ distribution. A better choice may be to assume a Gaussian spatial process model, which accounts for time-dependent correlation in the observed data, or a time-series model, which allows for serial correlation. In Bayesian curve-clustering, similar covariances structures may be appropriate for our choice of priors. Other curve-clustering possibilities include the inclusion of multiple fixed effects, different choices of birth proposals for the BDMCMC algorithm, and clustering using differential equations. Possible implementations of some of these ideas are proposed in §8.6.

Appendix A

Commonly Used Distributions

- **Beta**

If π follows a $\text{Beta}(\alpha, \beta)$ distribution, then for $\alpha, \beta > 0$,

$$p(\pi; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1 - \pi)^{(\beta-1)}.$$

Aside: $p(\pi; 1, K) = K(1 - \pi)^{K-1}$.

- **Dirichlet**

Let $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]^T \sim \mathcal{D}(\delta_1, \dots, \delta_K)$ such that $\sum \pi_k = 1$. Then

$$p(\boldsymbol{\pi}; \boldsymbol{\delta}) = \frac{\Gamma(\delta_1 + \dots + \delta_k)}{\Gamma(\delta_1 \cdot \dots \cdot \delta_k)} \cdot \pi_1^{\delta_1-1} \cdot \dots \cdot \pi_{k-1}^{\delta_{k-1}-1} \cdot (1 - \pi_1 - \dots - \pi_{k-1})^{\delta_k-1}.$$

Aside: Since $\Gamma(x) = (x - 1)!$, $p(\boldsymbol{\pi} | [1, \dots, 1]) = (K - 1)!$.

- **Gamma**

If w follows a $\Gamma(\alpha, \beta)$ distribution, then for $\alpha, \beta > 0$,

$$p(w; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot w^{\alpha-1} e^{-w/\beta}.$$

- **Multivariate Normal**

If a n -dimensional vector \mathbf{y} follows a $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ distribution, then

$$p(\mathbf{y}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-n/2} \cdot |\Sigma|^{-\frac{1}{2}} \cdot \exp \left\{ -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})^T \right\}.$$

- **Wishart**

If a p -dimensional square matrix V is $\mathcal{W}_p(\mathbf{a}, A)$, then for $a \geq p$,

$$p(V; \mathbf{a}, A) = C \cdot |A|^{-a/2} \cdot |V|^{(a-p-1)/2} \cdot \exp \left\{ -\frac{1}{2} \cdot \text{tr}(A^{-1}V) \right\}$$

is the density of V , where $C^{-1} = 2^{ap/2} \pi^{p(p-1)/4} \prod_{s=1}^p \Gamma((a+1-s)/2)$.

Appendix B

Full Conditionals

In this section we derive full conditional densities for the Bayesian heirarchical curve-clustering model described in §7.1 and the BREC model developed in §8.1.

B.1 Bayesian Clustering Model

1. $z_i = k$ ($i = 1, \dots, m$)

$$p(z_i = k | \boldsymbol{\eta}_i, \boldsymbol{\mu}_k, \Sigma_k, \pi_k) \propto p(\boldsymbol{\eta}_i | z_i = k, \boldsymbol{\mu}_k, \Sigma_k) \cdot p(z_i = k | \pi_k) = \mathcal{N}(\boldsymbol{\eta}_i; \boldsymbol{\mu}_k, \Sigma_k) \cdot \pi_k.$$

$$\text{That is, } p(z_i = k | \boldsymbol{\eta}_i, \boldsymbol{\mu}_k, \Sigma_k, \pi_k) = \frac{\pi_k \mathcal{N}(\boldsymbol{\eta}_i; \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(\boldsymbol{\eta}_i; \boldsymbol{\mu}_l, \Sigma_l)}.$$

2. $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]^T$

$$\begin{aligned} p(\boldsymbol{\pi} | \mathbf{y}^m, \mathbf{z}'_i s) &\propto \left\{ \prod_{k=1}^K \prod_{i: z_i=k}^{m_k} p(z_i | \pi_{z_i}) \right\} \cdot p(\boldsymbol{\pi}) \\ &\propto \pi_1^{m_1} \cdot \dots \cdot \pi_{K-1}^{m_{K-1}} \cdot (1 - \pi_1 - \dots - \pi_{K-1})^{m_K} \times \\ &\quad \pi_1^{\delta-1} \cdot \dots \cdot \pi_{K-1}^{\delta-1} \cdot (1 - \pi_1 - \dots - \pi_{K-1})^{\delta-1} \\ &= \pi_1^{m_1+\delta-1} \cdot \dots \cdot \pi_{K-1}^{m_{K-1}+\delta-1} \cdot (1 - \pi_1 - \dots - \pi_{K-1})^{m_K+\delta-1} \end{aligned}$$

That is, the full conditional density of π is

$$\boldsymbol{\pi} \sim \mathcal{D}(m_1 + \delta, \dots, m_K + \delta).$$

3. $\boldsymbol{\eta}_i$ ($i = 1, \dots, m$)

$$\begin{aligned} p(\boldsymbol{\eta}_i | \mathbf{y}^m, z_i = k, \boldsymbol{\mu}_k, \Sigma_k, \sigma^2) &\propto p(\mathbf{y}_i | \boldsymbol{\eta}_i, \sigma^2) \cdot p(\boldsymbol{\eta}_i | z_i = k, \boldsymbol{\mu}_k, \Sigma_k) \\ &\propto \exp \left\{ -\frac{1}{2} \left\{ \sigma^{-2} (\mathbf{y}_i - B\boldsymbol{\eta}_i)^T (\mathbf{y}_i - B\boldsymbol{\eta}_i) + (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k) \right\} \right\} \end{aligned}$$

Inside the exponent,

$$\sigma^{-2} (\mathbf{y}_i - B\boldsymbol{\eta}_i)^T (\mathbf{y}_i - B\boldsymbol{\eta}_i) = \sigma^{-2} \mathbf{y}_i^T \mathbf{y}_i - \sigma^{-2} \mathbf{y}_i^T B\boldsymbol{\eta}_i - \sigma^{-2} \boldsymbol{\eta}_i^T B^T \mathbf{y}_i + \sigma^{-2} \boldsymbol{\eta}_i^T B^T B \boldsymbol{\eta}_i$$

$$\text{and } (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k) = \boldsymbol{\eta}_i^T \Sigma_k^{-1} \boldsymbol{\eta}_i - \boldsymbol{\eta}_i^T \Sigma_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\eta}_i + \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\mu}_k.$$

Let $D_k = (\sigma^{-2} B^T B + \Sigma_k^{-1})^{-1}$ and use “...” to denote constant terms not involving $\boldsymbol{\eta}_k$. Then the sum of the above two terms reduces to

$$\begin{aligned} &\boldsymbol{\eta}_i^T (\sigma^{-2} B^T B + \Sigma_k^{-1}) \boldsymbol{\eta}_i - \boldsymbol{\eta}_i^T (\sigma^{-2} B^T \mathbf{y}_i + \Sigma_k^{-1} \boldsymbol{\mu}_k) - (\sigma^{-2} \mathbf{y}_i^T B + \boldsymbol{\mu}_k^T \Sigma_k^{-1}) \boldsymbol{\eta}_i + \dots \\ &= \boldsymbol{\eta}_i^T D_k^{-1} \boldsymbol{\eta}_i - \boldsymbol{\eta}_i^T (\sigma^{-2} B^T \mathbf{y}_i + \Sigma_k^{-1} \boldsymbol{\mu}_k) - (\sigma^{-2} B^T \mathbf{y}_i + \Sigma_k^{-1} \boldsymbol{\mu}_k)^T \boldsymbol{\eta}_i + \dots \\ &= \boldsymbol{\eta}_i^T D_k^{-1} \boldsymbol{\eta}_i - \boldsymbol{\eta}_i^T D_k^{-1} D_k (\sigma^{-2} B^T \mathbf{y}_i + \Sigma_k^{-1} \boldsymbol{\mu}_k) - (\sigma^{-2} B^T \mathbf{y}_i + \Sigma_k^{-1} \boldsymbol{\mu}_k)^T D_k D_k^{-1} \boldsymbol{\eta}_i + \dots \\ &= (\boldsymbol{\eta}_i - D_k (\sigma^{-2} B^T \mathbf{y}_i + \Sigma_k^{-1} \boldsymbol{\mu}_k))^T \cdot D_k^{-1} \cdot (\boldsymbol{\eta}_i - D_k (\sigma^{-2} B^T \mathbf{y}_i + \Sigma_k^{-1} \boldsymbol{\mu}_k)) + \dots \end{aligned}$$

It follows that the full conditional densities of $\boldsymbol{\eta}_i$ are

$$\boldsymbol{\eta}_i | \{\mathbf{y}_i, z_i = k, \boldsymbol{\mu}_k, \Sigma_k, \sigma^2\} \sim \mathcal{N}(D_k (\sigma^{-2} B^T \mathbf{y}_i + \Sigma_k^{-1} \boldsymbol{\mu}_k), D_k).$$

4. $\boldsymbol{\mu}_k$ ($k = 1, \dots, K$)

$$\begin{aligned} p(\boldsymbol{\mu}_k | \boldsymbol{\eta}, z_i = k, \Sigma_k, \sigma^2, V) &\propto \left\{ \prod_{i:z_i=k}^{m_k} p(\boldsymbol{\eta}_i | z_i, \boldsymbol{\mu}_k, \Sigma_k) \right\} \cdot p(\boldsymbol{\mu}_k) \\ &\propto \exp \left\{ -\frac{1}{2} \sum_{i:z_i=k}^{m_k} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k) - \frac{1}{2} (\boldsymbol{\mu}_k - \boldsymbol{v})^T V^{-1} (\boldsymbol{\mu}_k - \boldsymbol{v}) \right\} \end{aligned}$$

As before inside the $\exp\{\dots\}$ we get

$$\sum_{i:z_i=k}^{m_k} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k) = \sum_{i:z_i=k}^{m_k} \{ \boldsymbol{\eta}_i^T \Sigma_k^{-1} \boldsymbol{\eta}_i - \boldsymbol{\eta}_i^T \Sigma_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\eta}_i + \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\mu}_k \}$$

$$\text{and } (\boldsymbol{\mu}_k - \boldsymbol{v})^T V^{-1} (\boldsymbol{\mu}_k - \boldsymbol{v}) = \boldsymbol{\mu}_k^T V^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T V^{-1} \boldsymbol{v} - \boldsymbol{v}^T V^{-1} \boldsymbol{\mu}_k + \boldsymbol{v}^T V^{-1} \boldsymbol{v}.$$

Let $\bar{\boldsymbol{\eta}}_k = \frac{1}{m_k} \sum_{i:z_i=k}^{m_k} \boldsymbol{\eta}_i$, $C_k = (m_k \Sigma_k^{-1} + V^{-1})^{-1}$ and use “...” to denote all leftover terms that do not contain $\boldsymbol{\mu}_k$. Then the two term up above sum to

$$\begin{aligned} &\boldsymbol{\mu}_k^T (m_k \Sigma_k^{-1} + V^{-1}) \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k + V^{-1} \boldsymbol{v}) - (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k + V^{-1} \boldsymbol{v})^T \boldsymbol{\mu}_k + \dots \\ &= \boldsymbol{\mu}_k^T C_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T C_k^{-1} C_k (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k + V^{-1} \boldsymbol{v}) - (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k + V^{-1} \boldsymbol{v})^T C_k C_k^{-1} \boldsymbol{\mu}_k + \dots \\ &= (\boldsymbol{\mu}_k - C_k (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k + V^{-1} \boldsymbol{v}))^T \cdot C_k^{-1} \cdot (\boldsymbol{\mu}_k - C_k (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k + V^{-1} \boldsymbol{v})) + \dots \end{aligned}$$

It follows that the full conditional densities of $\boldsymbol{\mu}_k$ are

$$\boldsymbol{\mu}_k | \{ \boldsymbol{\eta}, z_i = k, \Sigma_k, \sigma^2, V \} \sim \mathcal{N}(C_k (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k + V^{-1} \boldsymbol{v}), C_k).$$

5. Σ_k^{-1} ($k = 1, \dots, K$)

$$\begin{aligned}
p(\Sigma_k^{-1} | \boldsymbol{\eta}, z_i = k, \boldsymbol{\mu}_k, R) &\propto \left\{ \prod_{i:z_i=k}^{m_k} p(\boldsymbol{\eta}_i | z_i, \boldsymbol{\mu}_k, \Sigma_k) \right\} \cdot p(\Sigma_k^{-1} | R) \\
&\propto |\Sigma_k^{-1}|^{m_k} \exp \left\{ -\frac{1}{2} \sum_{i:z_i=k}^{m_k} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k) \right\} |\Sigma_k^{-1}|^{\frac{(\rho-n-1)}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\rho R \Sigma_k^{-1}) \right\} \\
&= |\Sigma_k^{-1}|^{(m_k + \rho - n - 1)/2} \exp \left\{ -\frac{1}{2} \text{tr} \left(\left(\sum_i (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)(\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T + \rho R \right) \Sigma_k^{-1} \right) \right\}.
\end{aligned}$$

It follows that the full conditional densities of Σ_k are

$$\Sigma_k^{-1} | \{\boldsymbol{\eta}, z_i = k, \boldsymbol{\mu}_k, R\} \sim \mathcal{W} \left(m_k + \rho, \left(\sum_{i:z_i=k}^{m_k} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)(\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T + \rho R \right)^{-1} \right).$$

6. R

$$\begin{aligned}
p(R | \Sigma_1, \dots, \Sigma_K) &\propto p(R) \cdot \prod_{k=1}^K p(\Sigma_k^{-1} | R) \\
&\propto |bR_0|^{b/2} |R|^{(b-n-1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(bR_0 R) \right\} \\
&\quad \prod_{k=1}^K |\rho R|^{\rho/2} |\Sigma_k^{-1}|^{(\rho-n-1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\rho R \Sigma_k^{-1}) \right\} \\
&\propto |R|^{(b-n-1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(bR_0 R) \right\} |\rho R|^{K\rho/2} \exp \left\{ -\frac{1}{2} \text{tr} \left(\left(\rho \sum_{k=1}^K \Sigma_k^{-1} \right) R \right) \right\} \\
&= |R|^{(b+K\rho-n-1)/2} \exp \left\{ -\frac{1}{2} \text{tr} \left(\left(bR_0 + \rho \sum_k \Sigma_k^{-1} \right) R \right) \right\}.
\end{aligned}$$

This is a Wishart pdf, and thus the full conditional density of R is

$$R|\{\Sigma_1, \dots, \Sigma_K\} \sim \mathcal{W} \left(b + K\rho, \left(bR_0 + \rho \sum_{k=1}^K \Sigma_k^{-1} \right)^{-1} \right).$$

7. σ^{-2}

$$\begin{aligned} p(\sigma^2 | \mathbf{y}^m, \boldsymbol{\eta}) &\propto \prod_{i=1}^m p(\mathbf{y}_i | \boldsymbol{\eta}_i, \sigma^2) \cdot p(\sigma^{-2}) \\ &\propto (\sigma^{-2n})^{m/2} \exp \left\{ -\frac{1}{2} \sigma^{-2} \sum_{i=1}^m (\mathbf{y}_i - B\boldsymbol{\eta}_i)^T (\mathbf{y}_i - B\boldsymbol{\eta}_i) \right\} \cdot \sigma^{-2(g-1)} \exp \left\{ -\sigma^{-2} \frac{1}{h} \right\} \\ &= \sigma^{-2(nm/2+g-1)} \exp \left\{ -\sigma^{-2} \left(\frac{1}{2} \sum_i (\mathbf{y}_i - B\boldsymbol{\eta}_i)^T (\mathbf{y}_i - B\boldsymbol{\eta}_i) + \frac{1}{h} \right) \right\}. \end{aligned}$$

It follows that the full conditional density of σ^{-2} is

$$\sigma^{-2} | \{\mathbf{y}^m, \boldsymbol{\eta}\} \sim \Gamma \left(\frac{nm}{2} + g, \left(\frac{1}{2} \sum_i (\mathbf{y}_i - B\boldsymbol{\eta}_i)^T (\mathbf{y}_i - B\boldsymbol{\eta}_i) + \frac{1}{h} \right)^{-1} \right).$$

B.2 Bayesian Random Effects Clustering

1. $z_{ij} = k$ ($i = 1, \dots, c$ and $j = 1, \dots, v$)

$$p(z_{ij} = k | \boldsymbol{\eta}_{ij}, \boldsymbol{\mu}_k, \Sigma_k, \pi_k) \propto p(\boldsymbol{\eta}_{ij} | z_{ij} = k, \boldsymbol{\mu}_k, \Sigma_k) \cdot p(z_{ij} = k | \pi_k) = \mathcal{N}(\boldsymbol{\eta}_{ij}; \boldsymbol{\mu}_k, \Sigma_k) \cdot \pi_k.$$

$$\text{That is, } p(z_{ij} = k | \boldsymbol{\eta}_{ij}, \boldsymbol{\mu}_k, \Sigma_k, \pi_k) = \frac{\pi_k \mathcal{N}(\boldsymbol{\eta}_{ij}; \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(\boldsymbol{\eta}_{ij}; \boldsymbol{\mu}_l, \Sigma_l)}.$$

$$2. \boldsymbol{\pi} = [\pi_1, \dots, \pi_K]^T$$

$$\begin{aligned} p(\boldsymbol{\pi} | \mathbf{y}^m, z'_{ij}, s) &\propto \left\{ \prod_{k=1}^K \prod_{ij: z_{ij}=k}^{m_k} p(z_{ij} | \pi_{z_{ij}}) \right\} \cdot p(\boldsymbol{\pi}) \\ &= \pi_1^{m_1 + \delta - 1} \cdot \dots \cdot \pi_{K-1}^{m_{K-1} + \delta - 1} \cdot (1 - \pi_1 - \dots - \pi_{K-1})^{m_K + \delta - 1} \end{aligned}$$

Thus, $\boldsymbol{\pi} \sim \mathcal{D}(m_1 + \delta, \dots, m_K + \delta)$.

$$3. \boldsymbol{\eta}_{ij} \quad (i = 1, \dots, c \text{ and } j = 1, \dots, v)$$

$$\begin{aligned} p(\boldsymbol{\eta}_{ij} | \mathbf{y}_{ij}, z_{ij} = k, \boldsymbol{\gamma}_j, \boldsymbol{\mu}_k, \Sigma_k, \sigma^2) &\propto p(\mathbf{y}_{ij} | \boldsymbol{\gamma}_j, \boldsymbol{\eta}_{ij}, \sigma^2) \cdot p(\boldsymbol{\eta}_{ij} | z_{ij} = k, \boldsymbol{\mu}_k, \Sigma_k) \\ &\propto \exp \left\{ -\frac{1}{2} (\sigma^{-2} (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij})^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij}) + (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k)) \right\} \end{aligned}$$

where

$$\begin{aligned} &\sigma^{-2} (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij})^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij}) + (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k) \\ &= \sigma^{-2} (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j)^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j) - \sigma^{-2} (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j)^T B_2 \boldsymbol{\eta}_{ij} - \sigma^{-2} \boldsymbol{\eta}_{ij}^T B_2^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j) \\ &\quad + \sigma^{-2} \boldsymbol{\eta}_{ij}^T B_2^T B_2 \boldsymbol{\eta}_{ij} + \boldsymbol{\eta}_{ij}^T \Sigma_k^{-1} \boldsymbol{\eta}_{ij} - \boldsymbol{\eta}_{ij}^T \Sigma_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\eta}_{ij} + \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\mu}_k \\ &= \boldsymbol{\eta}_{ij}^T (\sigma^{-2} B_2^T B_2 + \Sigma_k^{-1}) \boldsymbol{\eta}_{ij} - \boldsymbol{\eta}_{ij}^T (\sigma^{-2} B_2^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j) + \Sigma_k^{-1} \boldsymbol{\mu}_k) \\ &\quad - (\sigma^{-2} (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j)^T B_2 + \boldsymbol{\mu}_k^T \Sigma_k^{-1}) \boldsymbol{\eta}_{ij} + \dots \end{aligned}$$

Letting $\mathbf{y}_{ij}^* = \mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j$ and $D_k = (\sigma^{-2} B_2^T B_2 + \Sigma_k^{-1})^{-1}$,

$$\begin{aligned} &\boldsymbol{\eta}_{ij}^T D_k^{-1} \boldsymbol{\eta}_{ij} - \boldsymbol{\eta}_{ij}^T (\sigma^{-2} B_2^T \mathbf{y}_{ij}^* + \Sigma_k^{-1} \boldsymbol{\mu}_k) - (\sigma^{-2} B_2^T \mathbf{y}_{ij}^* + \Sigma_k^{-1} \boldsymbol{\mu}_k)^T \boldsymbol{\eta}_{ij} + \dots \\ &= \boldsymbol{\eta}_{ij}^T D_k^{-1} \boldsymbol{\eta}_{ij} - \boldsymbol{\eta}_{ij}^T D_k^{-1} D_k (\sigma^{-2} B_2^T \mathbf{y}_{ij}^* + \Sigma_k^{-1} \boldsymbol{\mu}_k) - (\sigma^{-2} B_2^T \mathbf{y}_{ij}^* + \Sigma_k^{-1} \boldsymbol{\mu}_k)^T D_k D_k^{-1} \boldsymbol{\eta}_{ij} + \dots \\ &= (\boldsymbol{\eta}_{ij} - D_k (\sigma^{-2} B_2^T \mathbf{y}_{ij}^* + \Sigma_k^{-1} \boldsymbol{\mu}_k))^T \cdot D_k^{-1} \cdot (\boldsymbol{\eta}_{ij} - D_k (\sigma^{-2} B_2^T \mathbf{y}_{ij}^* + \Sigma_k^{-1} \boldsymbol{\mu}_k)) + \dots \end{aligned}$$

Thus, $\boldsymbol{\eta}_{ij} | \{\mathbf{y}_{ij}, z_{ij} = k, \boldsymbol{\gamma}_j, \boldsymbol{\mu}_k, \Sigma_k, \sigma^2\} \sim \mathcal{N}(D_k (\sigma^{-2} B_2^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j) + \Sigma_k^{-1} \boldsymbol{\mu}_k), D_k)$.

4. $\boldsymbol{\gamma}_j$ ($j = 1, \dots, v$)

$$\begin{aligned} p(\boldsymbol{\gamma}_j | \mathbf{y}_{ij}, z_{ij} = k, \boldsymbol{\eta}_{ij}, \mathbf{w}, W, \sigma^2) &\propto \prod_{i=1}^c p(\mathbf{y}_{ij} | \boldsymbol{\gamma}_j, \boldsymbol{\eta}_{ij}, \sigma^2) \cdot p(\boldsymbol{\gamma}_j | \mathbf{w}, W) \\ &\propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^c (\sigma^{-2} (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij} - B_1 \boldsymbol{\gamma}_j)^T (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij} - B_1 \boldsymbol{\gamma}_j) + (\boldsymbol{\gamma}_j - \mathbf{w})^T W^{-1} (\boldsymbol{\gamma}_j - \mathbf{w})) \right\}, \end{aligned}$$

where

$$\begin{aligned} &\sigma^{-2} (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij} - B_1 \boldsymbol{\gamma}_j)^T (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij} - B_1 \boldsymbol{\gamma}_j) + (\boldsymbol{\gamma}_j - \mathbf{w})^T W^{-1} (\boldsymbol{\gamma}_j - \mathbf{w}) \\ &= \sigma^{-2} (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij})^T (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij}) - \sigma^{-2} (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij})^T B_1 \boldsymbol{\gamma}_j - \sigma^{-2} \boldsymbol{\gamma}_j^T B_1^T (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij}) \\ &\quad + \sigma^{-2} \boldsymbol{\gamma}_j^T B_1^T B_1 \boldsymbol{\gamma}_j + \boldsymbol{\gamma}_j^T W^{-1} \boldsymbol{\gamma}_j - \boldsymbol{\gamma}_j^T W^{-1} \mathbf{w} - \mathbf{w}^T W^{-1} \boldsymbol{\gamma}_j + \mathbf{w}^T W^{-1} \mathbf{w} \\ &= \boldsymbol{\gamma}_j^T (\sigma^{-2} B_1^T B_1 + W^{-1}) \boldsymbol{\gamma}_j - \boldsymbol{\gamma}_j^T (\sigma^{-2} B_1^T (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij}) + W^{-1} \mathbf{w}) \\ &\quad - (\sigma^{-2} (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij})^T B_1 + \mathbf{w}^T W^{-1}) \boldsymbol{\gamma}_j + \dots \end{aligned}$$

Letting $\mathbf{y}_j^* = \sum_{i=1}^c (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij}) = c(\bar{\mathbf{y}}_j - B_2 \bar{\boldsymbol{\eta}}_j)$ and $D = (c\sigma^{-2} B_1^T B_1 + W^{-1})^{-1}$,

$$\begin{aligned} &p(\boldsymbol{\gamma}_j | \mathbf{y}_{ij}, z_{ij} = k, \boldsymbol{\eta}_{ij}, \mathbf{w}, W, \sigma^2) \\ &\propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\gamma}_j^T D^{-1} \boldsymbol{\gamma}_j - \boldsymbol{\gamma}_j^T (\sigma^{-2} B_1^T \mathbf{y}_j^* + W^{-1} \mathbf{w}) - (\sigma^{-2} B_1^T \mathbf{y}_j^* + W^{-1} \mathbf{w})^T \boldsymbol{\gamma}_j + \dots) \right\} \\ &= \exp \left\{ -\frac{1}{2} (\boldsymbol{\gamma}_j^T D^{-1} \boldsymbol{\gamma}_j - \boldsymbol{\gamma}_j^T D^{-1} D (\sigma^{-2} B_1^T \mathbf{y}_j^* + W^{-1} \mathbf{w}) - (\sigma^{-2} B_1^T \mathbf{y}_j^* + W^{-1} \mathbf{w})^T D D^{-1} \boldsymbol{\gamma}_j + \dots) \right\} \\ &= \exp \left\{ -\frac{1}{2} ((\boldsymbol{\gamma}_j - D (\sigma^{-2} B_1^T \mathbf{y}_j^* + W^{-1} \mathbf{w}))^T \cdot D^{-1} \cdot (\boldsymbol{\gamma}_j - D (\sigma^{-2} B_1^T \mathbf{y}_j^* + W^{-1} \mathbf{w})) + \dots) \right\}. \end{aligned}$$

Thus, $\boldsymbol{\gamma}_j | \{\mathbf{y}_{ij}, z_{ij} = k, \boldsymbol{\eta}_{ij}, \mathbf{w}, W, \sigma^2\} \sim \mathcal{N}(D(c\sigma^{-2} B_1^T (\mathbf{y}_{ij} - B_2 \boldsymbol{\eta}_{ij}) + W^{-1} \mathbf{w}), D)$.

5. $\boldsymbol{\mu}_k$ ($k = 1, \dots, K$)

$$\begin{aligned}
p(\boldsymbol{\mu}_k | \boldsymbol{\eta}, z_{ij} = k, \Sigma_k, \sigma^2) &\propto \left\{ \prod_{ij: z_{ij}=k}^{m_k} p(\boldsymbol{\eta}_{ij} | z_{ij} = k, \boldsymbol{\mu}_k, \Sigma_k) \right\} \cdot p(\boldsymbol{\mu}_k) \\
&\propto \exp \left\{ -\frac{1}{2} \sum_{ij: z_{ij}=k}^{m_k} (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k) - \frac{1}{2} \boldsymbol{\mu}_k^T V^{-1} \boldsymbol{\mu}_k \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \left(\sum_{ij: z_{ij}=k}^{m_k} \boldsymbol{\eta}_{ij}^T \Sigma_k^{-1} \boldsymbol{\eta}_{ij} - \boldsymbol{\eta}_{ij}^T \Sigma_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\eta}_{ij} + \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T V^{-1} \boldsymbol{\mu}_k \right) \right\}.
\end{aligned}$$

Letting $\bar{\boldsymbol{\eta}}_k = \frac{1}{m_k} \sum_{ij: z_{ij}=k}^{m_k} \boldsymbol{\eta}_{ij}$, $C_k = (m_k \Sigma_k^{-1} + V^{-1})^{-1}$ and denoting terms that do not contain $\boldsymbol{\mu}_k$ by "...",

$$\begin{aligned}
p(\boldsymbol{\mu}_k | \boldsymbol{\eta}, z_{ij} = k, \Sigma_k, \sigma^2) &\propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\mu}_k^T (m_k \Sigma_k^{-1} + V^{-1}) \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k) - (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k)^T \boldsymbol{\mu}_k + \dots) \right\} \\
&= \exp \left\{ -\frac{1}{2} (\boldsymbol{\mu}_k^T C_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T C_k^{-1} C_k (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k) - (m_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k)^T C_k C_k^{-1} \boldsymbol{\mu}_k + \dots) \right\} \\
&= \exp \left\{ -\frac{1}{2} ((\boldsymbol{\mu}_k - m_k C_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k)^T \cdot C_k^{-1} \cdot (\boldsymbol{\mu}_k - m_k C_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k) + \dots) \right\}.
\end{aligned}$$

It follows that $\boldsymbol{\mu}_k | \{\boldsymbol{\eta}, z_{ij} = k, \Sigma_k, \sigma^2\} \sim \mathcal{N}(m_k C_k \Sigma_k^{-1} \bar{\boldsymbol{\eta}}_k, C_k)$.

6. Σ_k^{-1} ($k = 1, \dots, K$)

$$\begin{aligned} p(\Sigma_k^{-1} | \boldsymbol{\eta}, z_{ij} = k, \boldsymbol{\mu}_k, R) &\propto \left\{ \prod_{ij:z_{ij}=k}^{m_k} p(\boldsymbol{\eta}_{ij} | z_{ij} = k, \boldsymbol{\mu}_k, \Sigma_k) \right\} \cdot p(\Sigma_k^{-1} | R) \\ &\propto |\Sigma_k^{-1/2}|^{m_k} \exp \left\{ -\frac{1}{2} \sum_{ij:z_{ij}=k}^{m_k} (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k) \right\} |\Sigma_k^{-1}|^{\frac{(\rho-n-1)}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\rho R \Sigma_k^{-1}) \right\} \\ &= |\Sigma_k^{-1}|^{(m_k + \rho - n - 1)/2} \exp \left\{ -\frac{1}{2} \text{tr} \left(\left(\sum_{ij} (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k)(\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k)^T + \rho R \right) \Sigma_k^{-1} \right) \right\}. \end{aligned}$$

Thus $\Sigma_k^{-1} | \{\boldsymbol{\eta}, z_{ij} = k, \boldsymbol{\mu}_k, R\} \sim \mathcal{W} \left(m_k + \rho, \left(\sum_{ij:z_{ij}=k}^{m_k} (\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k)(\boldsymbol{\eta}_{ij} - \boldsymbol{\mu}_k)^T + \rho R \right)^{-1} \right)$.

7. R

Full conditional of R and derivations thereof are as shown in point 6 of §B.1.

8. σ^2

$$\begin{aligned} p(\sigma^2 | \mathbf{y}^m, \boldsymbol{\gamma}, \boldsymbol{\eta}) &\propto \prod_{i=1}^c \prod_{j=1}^v p(\mathbf{y}_{ij} | \boldsymbol{\gamma}_j, \boldsymbol{\eta}_{ij}, \sigma^2) \cdot p(\sigma^{-2}) \\ &\propto (\sigma^{-2n})^{m/2} \exp \left\{ -\frac{1}{2} \sigma^{-2} \sum_{i=1}^c \sum_{j=1}^v (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij})^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij}) \right\} \cdot \\ &\quad \sigma^{-2(g-1)} \exp \left\{ -\sigma^{-2} \frac{1}{h} \right\} \\ &= \sigma^{-2(nm/2 + g - 1)} \exp \left\{ -\sigma^{-2} \left(\frac{1}{2} \sum_{ij} (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij})^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij}) + \frac{1}{h} \right) \right\}. \end{aligned}$$

Thus $\sigma^{-2} | \dots \sim \Gamma \left(\frac{nm}{2} + g, \left(\frac{1}{2} \sum_{ij} (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij})^T (\mathbf{y}_{ij} - B_1 \boldsymbol{\gamma}_j - B_2 \boldsymbol{\eta}_{ij}) + \frac{1}{h} \right)^{-1} \right)$.

Bibliography

- [1] Abramovich F. and Angelini C. (2006). “Testing in Mixed-effects FANOVA Models,” *Journal of Statistical Planning and Inference*, 136, 4326-4348.
- [2] Alt, F.B. (1982). “Multivariate Quality Control: State of the Art,” *Transactions of the 1982 ASQC Quality Congress*, Detroit, MI, 886-893.
- [3] Ash, R. B. and Bishop, R. L. (1972). “Monopoly as a Markov Process,” *Mathematics Magazine*, 45(1), 26-29.
- [4] Abraham, C., Cornillon, P., Matzner-Lober, E. and Molinari, N. (2000). “Unsupervised Curve Clustering Using B-splines,” Technical Report No. 00-04. Department of Mathematics, University of Montpellier.
- [5] Andrieu, C., De Freitas, N., Doucet, A., Jordan, M. I. (2003). ”An Introduction to MCMC for Machine Learning,” *Machine Learning*, 50, 5-43.
- [6] Antoniadis, A. and Sapatinas, T. (2004). “Estimation and Inference in Functional Mixed-effects Models,” Technical Report, Department of Mathematics and Statistics, University of Cyprus, Nicosia.
- [7] Banfield, J.D. and Raftery, A.E. (1993). “Model-Based Gaussian and Non-Gaussian Clustering,” *Biometrics*, 49(3), 803-821.
- [8] Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.
- [9] Bennett, J.E., Racine-Poon, A. and Wakefield, J.C. (1996). “MCMC for Nonlinear Hierarchical Models,” In *Markov chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter (editors), 339-358, Chapman and Hall: New York.
- [10] Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997). “Inference in Model-based Cluster Analysis,” *Statistics and Computing*, 7, 1-10.

- [11] Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.
- [12] Biernacki C. and Govaert G. (1999). "Choosing Models in Model-based Clustering and Discriminant Analysis," *Journal of Statistical Computation and Simulation*, 64, 49-71.
- [13] Cappe, O., Robert, C. P. and Ryden, T. (2003). "Reversible Jump, Birth-and-Death, and More General Continuous Time MCMC Samplers," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 65(3), 679-700.
- [14] Casella, G. (1985). "An Introduction to Empirical Bayes Data Analysis," *American Statistician*, 39(2), 83-87.
- [15] Casella, G. and George, E. I. (1992). "Explaining the Gibbs Sampler," *American Statistician*, 46(3), 167-174.
- [16] Clifford, P. and Nicholls, G. (1994). "Comparison of Birth-and-Death and Metropolis-Hastings Markov Chain Monte Carlo for the Strauss Process." Ph.D. Thesis, Department of Statistics, Oxford University.
- [17] Chudova, D., Hart, C., Mjolsness, E. and Smyth, P. (2004). "Gene Expression Clustering with Functional Mixture Models," *Advances in Neural Information Processing*, 16.
- [18] Chudova, D., Gaffney, S. and Smyth, P. (2003). "Probabilistic Models for Joint Clustering and Alignment of Multidimensional Curves," *Proceedings of the Nineteenth Conference on Uncertainty and Artificial Intelligence*.
- [19] Cowles, M. K., Bradley P. C. (1996). "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91(434), 883-904.
- [20] De Sarbo, W.S., Howard, D.J. and Jedidi, K. (1991). "MULTICLUS: A New Method for Simultaneously Performing Multidimensional Scaling and Cluster Analysis," *Psychometrika*, 56, 121-136.
- [21] Dias, J. G. and Wedel, M. (2004). "An Empirical Comparison of EM, SEM and MCMC Performance for Problematic Gaussian Mixture Likelihoods," *Statistics and Computing*, 14(4), 323-332.
- [22] de Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer.
- [23] Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: Wiley.

- [24] Fox, J., (2002). "Linear Mixed Models," *Appendix to An R and S-PLUS Companion to Applied Regression*.
- [25] Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Discriminant Analysis and Density Estimation," *Journal of the American Statistical Association*, 97, 611-631.
- [26] Fraley, C. and Raftery, A. E. (2006). "MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-Based Clustering," Technical Report, Department of Statistics, University of Washington.
- [27] Gaffney, S. and Smyth, P. (2004). "Joint Probabilistic Curve Clustering and Alignment," *Advances in Neural Information Processing*, 17.
- [28] Gaffney, S. and Smyth, P. (2003). "Curve Clustering with Random Effects Regression Mixtures," *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.
- [29] Gallant, A. R. (1975). "Nonlinear Regression," *American Statistician*, 29(2), 73-81.
- [30] Ganesan, R. and Das T. K. (2002). "Wavelet Based Multiscale Statistical Process Monitoring-Literature Review and Research Extensions," *IIE Transactions on Quality and Reliability*.
- [31] Gardner, M. M., Lu, J., Gyurcsik, R. S., Wortman, J. J., Hornung, B. E., Heinisch, H. H., Rying, E. A., Rao, S., Davis, J. C. and Mozumder, P. K. (1997). "Equipment Fault Detection Using Spatial Signatures," *IEEE Transactions on Components, Packaging, and Manufacturing Technology - Part C*, 20, 295-304.
- [32] Geiger, D. and Heckerman, D. (2002). "Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions," *The Annals of Statistics*, 30(5), 1412-1440.
- [33] Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, 85(412), 972-985.
- [34] Gelman, A. (2006). "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515-533.
- [35] Geman, S. and Geman, D. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

- [36] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in Practice*, Chapman and Hall: New York.
- [37] Green, P. J. (1995). "Reversible Jump Markov chain Monte Carlo Simulation and Bayesian Model Determination," *Biometrika*, 82(4), 711-732.
- [38] Greenwood, M. C. (2004). "Functional Data Analysis for Glaciated Valley Profile Analysis," Ph.D. Thesis, Department of Statistics, University of Wyoming.
- [39] Guo, W. (2002), "Functional Mixed-effects Models," *Biometrics*, 58, 121-128.
- [40] Harville, D. A. (1990). "BLUP (Best Linear Unbiased Prediction) and Beyond," *Advances in Statistical Methods for Genetic Improvement of Livestock*, 239-276, New York: Springer.
- [41] Hastie, T., Tibshirani, R. and Friedman, J. (2002). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.
- [42] Hawkins, D. M. and Merriam, D. F. (1974). "Zonation of Multivariate Sequences of Digitized Geologic Data," *Mathematical Geology*, 6(3), 263-269.
- [43] Hotelling, H.H. (1947). "Multivariate Quality Control Illustrated by the Air Testing of Sample Bombsights," *Techniques of Statistical Analysis*, 111-184.
- [44] Heidelberger P. and Welch P.D. (1983). "Simulation Run Length Control in the Presence of an Initial Transient," *Operations Research*, 31, 1109-1145.
- [45] Jasra, A., Holmes, C.C. and Stephens, D. (2005). "MCMC and the Label-switching Problem in Bayesian Mixture Models," *Statistical Science*, 20(1), 50-67.
- [46] James, G. and Sugar, C. (2003). "Clustering for Sparsely Sampled Functional Data," *Journal of the American Statistical Association*, 98, 397-408.
- [47] Jeong M. K. and Lu J. C. (2004). "Statistical Process Control Charts for Complicated Functional Data," To appear in the *International Journal of Production Research*.
- [48] Jin, J. and Shi, J. (2001). "Automatic feature extraction of waveform signals for in-process diagnostic performance improvement," *Journal of Intelligent Manufacturing*, 12, 257-268.
- [49] Johnson R. A. and Wichern D. W. (1992). *Applied Multivariate Statistical Analysis*. New York: Prentice Hall.

- [50] Jones, M. C. and Rice J. A. (1992). "Displaying the Important Features of Large Collections of Similar Curves," *The American Statistician*, 46(2), 140-145.
- [51] Kang, L. and Albin, S. L. (2000). "On-line Monitoring When the Process Yields a Linear Profile," *Journal of Quality Technology*, 32(4), 418-426.
- [52] Kim, K., Mohamoud, M. A. and Woodall, W. H. (2003). "On the Monitoring of Linear Profiles," *Journal of Quality Technology*, 35, 317-328.
- [53] Koulis, T., Ramsay, J. O. and Levitin, D. (2006). "Input-Output Systems in Psychoacoustics," Submitted to *Psychometrika*.
- [54] Lada, E. K., Lu, J. and Wilson, J. R. (2002). "A Wavelet-Based Procedure for Process Fault Detection," *IEEE Transactions on Semiconductor Manufacturing*, 15, 79-90.
- [55] Laird, N. M. and Ware, J. H. (1982). "Random Effects Model for Longitudinal Data," *Biometrics*, 38(4), 963-974.
- [56] Lange, N. and Ryan, L. (1989). "Assessing Normality in Random Effects Models," *The Annals of Statistics*, 17(2), 624-642.
- [57] Lau, J. W. and Green, P. J. (2007). "Bayesian Model Based Clustering Procedures," To appear in the *Journal of Computational and Graphical Statistics*.
- [58] Lindstrom, M. J. and Bates, D. M. (1988). "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83 (404), 1014-1022.
- [59] Lindstrom, M. J. and Bates, D. M. (1990). "Nonlinear Mixed Effects Models for Repeated Measures Data," *Biometrics*, 46(3), 673-687.
- [60] Lowry, C. A., Woodall, W. H., Champ, C. W. and Rigdon, S. E. (1992). "A Multivariate Exponentially Weighted Moving Average Control Chart," *Technometrics*, 34(1), 46-53.
- [61] Lucas J.M. and Saccucci M.S. (1990). "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements," *Technometrics*, 32, 1-12.
- [62] Mahmoud, M. A. and Woodall, W. H. (2004). "Phase I Monitoring of Linear Profiles with Calibration Applications," Submitted to *Technometrics*.

- [63] Medvedovic, M. and Sivaganesan, S. (2002). "Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles," *Bioinformatics*, 18, 1194-1206.
- [64] Medvedovic, M., Yeung, K. Y. and Bumgarner, R. E. (2004). "Bayesian Mixture Model Based Clustering of Replicated Microarray Data," *Bioinformatics*, 20, 1222-1232.
- [65] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons.
- [66] Morris, J. S., Arroyo, C., Coull, B. A., Ryan, L. M. and Gortmaker, S. L. (2006). "Using Wavelet-Based Functional Mixed Models to characterize Population Heterogeneity in Accelerometer Profiles: A Case Study," *Journal of the American Statistical Association*, 101(476), 1352-1364.
- [67] Morris, J. S. and Carroll, R. (2006). "Wavelet-based Functional Mixed Models," *Journal of the Royal Statistical Society: Series B*, 68(2), 179-199.
- [68] Naylor, W. and Chapman, B. (2006). "WNLIB Library," *World Wide Web*, <http://www.willnaylor.com/index.html#research>.
- [69] Neal, R. M. (1992). "Bayesian mixture modeling." In *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, C. R. Smith, G. J. Erickson and P. O. Neudorfer (editors), Kluwer Academic Publishers: Dordrecht, 197-211.
- [70] Oh, M. S. and Raftery, A. E. (2007). "Model-based Clustering with Dissimilarities: A Bayesian Approach," To appear in the *Journal of Computational and Graphical Statistics*.
- [71] Ogunnaike, B. A. and Ray, W. H. (1994). *Process Dynamics, Modeling, and Control*. New York: Oxford University Press.
- [72] Pinheiro, J. and Bates, D.M. (2001). *Mixed Effects Models in S and S-PLUS*. New York: Springer-Verlag.
- [73] R Development Core Team (2006). "R: A language and environment for statistical computing," *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [74] Raftery A. E. (1995). "Bayesian Model Selection in Social Research," *Sociological Methodology*, 25, 111-163.

- [75] Ramsay, J. O. (2000). "Differential equation models for statistical functions," *Canadian Journal of Statistics*, 28, 225-240.
- [76] Ramsay, J. O. and Silverman, B. W. (1997, 2005). *Functional Data Analysis*. New York: Springer.
- [77] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*. New York: Springer-Verlag.
- [78] Ramsay, J. O. and Cao, J. (2006). "Smoothing, Nuisance Parameters and Profiling in Functional Data Analysis," Submitted to *Computational Statistics and Data Analysis*.
- [79] Ramsay, J. O. , Hooker, G, Cao, J. and Campbell D. (2006). "Estimating Differential Equations," Submitted to *Journal of the Royal Statistical Society, Series B*.
- [80] Rousseeuw, P.J. (1984). "Least median squares of regression," *Journal of the American Statistical Association*, 79, 871-880.
- [81] Rice J. A. and Wu C. O. (2001). "Nonparametric mixed effects models for unequally sampled noisy curves," *Biometrics*, 57, 253-259.
- [82] Richardson, S. and Green, P. (1997). "On Bayesian analysis of mixtures with unknown number of components (with discussions)," *Journal of the Royal Statistical Society, Series B*, 59, 731-792.
- [83] Robert, C. (2001). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer-Verlag.
- [84] Roberts, G. O. (1996). "Markov chain concepts related to sampling algorithms". In *Markov chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter (editors), pp 45-58, Chapman and Hall: New York.
- [85] Ryan, T. P. (1989). *Statistical Methods For Quality Improvement*. New York: Wiley.
- [86] Smith, A. F. M. and Roberts, G. O. (1993). "Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods" *Journal of the Royal Statistical Society, Series B*, 55(1), 3-23.
- [87] Stephens, M. (1997). "Bayesian Methods for Mixtures of Normal Distributions," Ph.D. Thesis, Department of Statistics, Oxford University.

- [88] Stephens, M. (2000a). "Bayesian Analysis of Mixture Models with Unknown Number of Components - An Alternative to Reversible Jump," *Annals of Statistics*, 28-1, 40-74.
- [89] Stephens, M. (2000b). "Dealing with Label Switching in Mixture Models," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 62(4), 795-809.
- [90] Tierney, L. (1994) . "Markov chains for exploring posterior distributions (with discussion)," *Annals of Statistics*, 22, 1701-1762.
- [91] Walker, E. and Wright, W. P. (2002). "Comparing Curves with Additive Models," *Journal of Quality Technology*, 34(1), 118-129.
- [92] Weakliem, L.D. (1999). "A Critique of the Bayesian Information Criterion for Model Selection," *Sociological Methods Research*, 27, 359-397.
- [93] Williams, J.D., Woodall, W. H. and Birch, J.B. (2005). "Phase I Analysis of Nonlinear Product and Process Profiles," Submitted to *Technometrics*.
- [94] Wolfinger, R., (1993). "Laplace's Approximations to Nonlinear Mixed Models," *Biometrika*, 80(4), 791-795.
- [95] Woodall, W. H., Spitzner, D. J., Montgomery, D. C. and Gupta S. (2004). "Using Control Charts to Monitor Process and Product Quality Profiles," *Journal of Quality Technology*, 36(3), 309-320.
- [96] Zhou, C. and Wakefield, J. (2005). "A Bayesian hierarchical mixture model for curve partitioning," *Biometrics*, 62(2), 515-526.