Topics in Functional Data Analysis

Stephen Edward Lane

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

May, 2012

Department of Mathematics and Statistics

The University of Melbourne

Australia

Abstract

As the amount of data captured in experimental and observational situations has grown, so too has the need for more sophisticated tools of analysis. The broad area of functional data analysis (FDA) has received a great deal of attention within the statistics community over the past decade, as a way of dealing with the high dimensional data that is becoming more commonplace in areas such as medical science and biology. High dimensional data has classically been analysed by treating the relevant data sets as multivariate data sets. FDA seeks to analyse these data as if they are in fact functions, observed at either sparse or dense sample points.

Motivated by problems in forest science and capture/recapture experiments, this thesis explores some of the FDA methodology through modelling data sets that contain functional responses or functional covariates.

We show that a nonparametric 'whole of function' approach to predicting conditional probability density functions is a particularly useful alternative to commonly used parametric approaches, especially when density functions have 'non–standard' shapes. The functional prediction approach is further extended by developing functional regression models for functional longitudinal data. We show that these models allow prediction and extrapolation of density functions with arbitrarily (smooth) changing shapes over time, conditional on a functional growth covariate.

A semiparametric population size estimator for mark–recapture data with continuous, time–varying covariates is developed from a functional data perspective. This method allows the full use of data resulting from such experiments, circumventing the usual loss of information in current approaches. We demonstrate that an iterative estimation approach in the form of an EM algorithm outperforms methods that ignore the time variation in the covariates.

Declaration

This is to certify that

i. the thesis comprises only my original work towards the PhD except where indicated in the Preface,

ii. due acknowledgement has been made in the text to all other material used,

iii. the thesis is fewer than 100000 words in length, exclusive of tables, maps, bibliographies and appendices.

Stephen Edward Lane

Preface

This thesis is submitted to the University of Melbourne in total fulfilment of the requirements of the degree of Doctor of Philosophy. No part of this thesis has been submitted in support of any other qualification. Chapter Two has been published as *The functional regression tree method for diameter distribution modelling* in the Canadian Journal of Forest Research (Lane et al., 2010) and was co-authored by Dr Andrew Robinson and Dr Thomas Baker. Chapter Three has been published as *An alternative objective function for fitting regression trees to functional response variables* in Computational Statistics and Data Analysis (Lane and Robinson, 2011) and was co-authored by Dr Andrew Robinson.

Acknowledgements

It is my pleasure to thank the many people who have helped make this thesis possible. In particular, my supervisor Andrew Robinson and co-supervisors Tom Baker and Richard Huggins. Their words of wisdom, effort and the occasional jolt from an electrified cattle prod have had a profound impact on the researcher I am today.

I would also like to thank my friends who have helped by listening to ideas and offering suggestions for various projects along the way. A special thanks to the others in the LB Crew — Jakub and Jason — for helping me to nurture my caffeine dependency.

A final word of thanks to Emma, Hannah, Jude and Angus. Without your constant support, love and the ability to put up with my occasional crankiness, I couldn't have submitted this weighty tome.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and summary of current approaches

Univariate and multivariate data, along with the associated tools to analyse such data sets, have been well understood by statisticians and applied scientists since early last century. With the increase in storage capabilities and processing power of computers (especially personal computers) over the last two to three decades, along with more sophisticated measurement tools, functional data has become a strong research focus of both theoretical and applied statisticians.

In this thesis, we will apply functional data techniques to analyse some problems in the applied biological sciences. The modelling approaches for each of the problems we will present rely on traditional univariate and multivariate parametric methods. We can place each problem within a functional data setting, and it is this setting that is the major focus of this thesis. Further, each of the methods that we describe and develop will essentially be model free, in the sense that we will make use of nonparametric methods in combination with functional data analysis, leading to extremely flexible modelling frameworks.

In Chapters 2 and 3, we will describe methods to model and predict functional responses, in particular, probability density functions. Chapter 2 introduces the functional regression tree method and demonstrates its applicability to estimating size distributions in a forestry context. Chapter 3 will extend the method and detail a result that allows efficient computation of the method. Computational complexity of the resultant algorithm will also be analysed.

In Chapter 4 a method for analysing functions that have a longitudinal aspect will be introduced. Here we model time–varying functions conditional on growth curves, which we will demonstrate by application to a longitudinal forestry data set.

Chapter 5 moves away from functional responses to functional predictors. That is the case in which a measured covariate is in fact itself functional. In this chapter, we investigate the problem in capture–recapture experiments of missing covariate data (specifically where the covariate is functional) through the use of nonparametric smoothing, functional data analysis, and the EM algorithm.

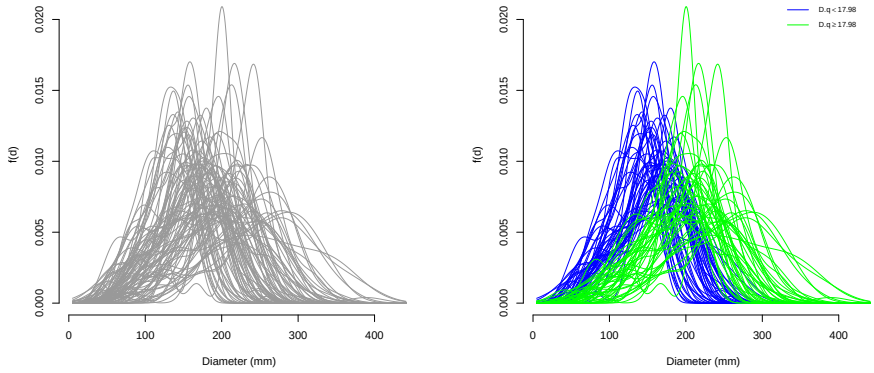
The rest of this introductory chapter provides more detail on the topics to be explored throughout this thesis, starting with a brief explanation of what constitutes functional data, and moving on to explain the current approaches to the problems just described. In the conclusion, we will discuss some future directions that may be taken.

1.1 What is functional data?

Functional data can be quite simply described as data that varies smoothly over some continuum, for example, curves and surfaces. In the current literature on functional data analysis, two main types of functional data receive attention. The first is curve data that has been observed on a fine grid, resulting in a fully observed function. The second is data that is believed to have resulted from some underlying curve that in its entirety is unobservable, yet we have a finite (often quite small) number of observations of the curve over its domain.

Figure 1.1 provides an example of the first type of functional data. Shown in Figure 1.1(a) are nonparametric density estimates of tree diameters within forest stands. There is little information in this figure to lead us to believe that there is any relationship amongst these densities. However, looking at Figure 1.1(b), it is clear that some relationship is apparent. In this figure, those forest stands that have quadratic mean diameter (D.q) less than 18 have their diameter densities coloured blue, whilst those having D.q ≥ 18 are coloured green. How this relationship is estimated will be discussed in Chapters 2 and 3.

As an example of the second type of functional data, Figure 1.2 shows the weights (in grams) of 54 Mountain Pygmy Possums captured over five



(a) All curves (b) Conditional curves

Figure 1.1: Example of fully observed functional data. (a) Tree diameter density curves and (b) Tree diameter density curves conditional on quadratic mean diameter.

nights. When an individual is not captured, there is clearly no way to measure the weight of that individual, so the resulting data are sparse observations from what we assume to be smoothly changing functional forms. In the figure, those individuals that were caught more than once have their successive measured weights connected by lines. Smooth estimates of the weight functions are a result of the methodology presented in Chapter 5, and can be seen in Figure 5.3.

A combination of the two types of functional data is the focus of Chapter 4.

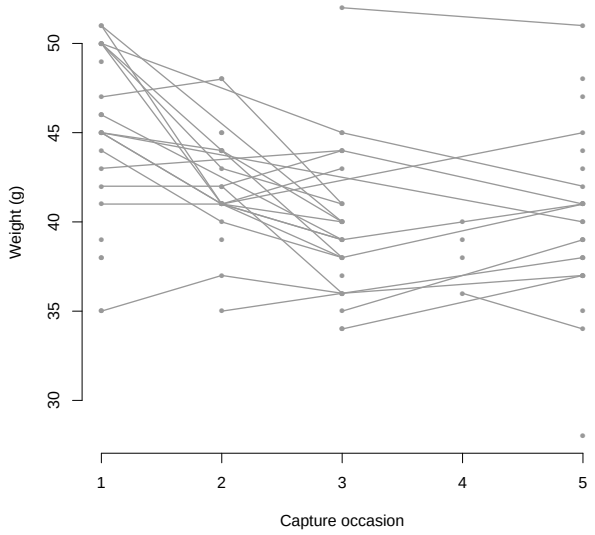


Figure 1.2: Example of sparsely observed functional data.

1.2 Estimating size distributions

This section outlines some current approaches to estimating size distributions that are in common use within the forestry literature. In the context of forest management, a size distribution refers to the probability distribution of a certain tree attribute, often conditional on a set of measured covariates. The common feature of these methods is that in each approach, an underlying parametric family is assumed to generate the distributions. We will further break these methods into two separate sub–methods. The first, which we will term fully–parametric, refers to both the estimation procedure and the form of the size distribution. That is, the estimation is performed, for example, by linear least squares, and that we assume that the distribution of trees within a stand can be described by a known functional form, indexed by a number of parameters. A commonly assumed functional form for size distributions is the Weibull probability density function (PDF)

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} \cdot \exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right] \tag{1.1}$$

where γ, α, β are the location, scale and shape parameters respectively. The Weibull PDF was proposed by Bailey and Dell (1973) as a candidate for mod-

elling size distributions due to its mathematical properties and its flexibility in modelling variously shaped densities.

However, it is common when estimating size distributions, to try a number of different candidate PDFs to fit the data. As an example, Robinson (2004) tested both two– and three–parameter Weibull distributions, Johnson's S_B distribution, and the Chaudry–Ahmad distribution (Chaudry and Ahmad, 1993) before settling on the three–parameter Weibull as the distribution providing the best fit to the data. With this in mind, the discussions in Section 1.2.2 will focus particularly on the modelling framework.

The second stream of parametric estimation I will term partially parametric. Whilst the regression model itself will be parametric, as in the fully parametric case, the size distribution will no longer be assumed to follow a predefined functional form. Section 1.2.3, will describe the percentile method, which is also known as Borders' method, named for the method introduced by Borders et al. (1987).

1.2.1 Some attributes of forestry data

We will now introduce some typical data attributes arising from forest management and its associated studies. The most commonly measured attribute of a tree is its diameter at breast height, where breast height is defined in most countries as 1.3 m. Because it would be prohibitively expensive to measure each tree in a stand, usually only a sample of trees is measured. Characteristics of the whole stand may also be gathered, for example environmental variables such as rainfall, and silvicultural decisions that were made, such as thinning and fertiliser application.

The statistical distribution of the measured size (e.g. tree diameter, basal area, volume) can then be estimated from the sample of trees, and conditional on measured or observed stand covariates, a prediction model can be formed. Size distributions for unmeasured stands can then be predicted by measuring the relevant stand characteristics and plugging them into the predictive model, a simpler and less costly procedure than individual tree measurements.

We now introduce some of the other attributes which we will use. We focus on a particular data set that was made available through the Cooperative Research Centre for Forestry (CRCF, http://www.crcforestry.com.au). The data come from *Eucalyptus globulus* Labill. (Tasmanian Blue Gum) pulpwood plantations consisting of same–design replicated field experiments in

south–western Australia, planted in 1994.

The experiments compare stocking (planting density) treatments ranging from 625 to 2000 trees ha^{-1}. As discussed above, only a sample of trees (within each stand) are measured for diameter, and in this particular experiment, three sample plots were established for each stocking treatment. These sample plots ranged in area from 0.03 to 0.06 ha. All trees in the sample plots were measured for diameter, and a subset of the trees were measured for height (m, height is generally harder and more expensive to measure accurately than diameter). It is from these sample plots we estimate the size distributions; this is what will be called tree–level data.

Using the tree–level data, stand–level (summary) characteristics can be calculated. For this particular data set, these calculations provide us with: density, the number of stems (it is common to measure the number of stems, rather than the number of trees, as some trees may have more than one stem) per hectare; basal area (m^2 ha^{-1}) of the stand (the basal area of a tree is its cross–sectional area at breast height); top height, as the mean height of the 100 largest diameter trees per hectare (m) and total underbark volume (m^3 ha^{-1}).

Trees have been re–measured at approximately two–yearly intervals, with the first measurement at age 1.5 years. This measurement framework at the individual level provides us with a rich longitudinal data set. It is the longitudinal aspect that we ultimately wish to exploit in our predictions, providing us with the capability to extrapolate or interpolate with age. As such, this longitudinal aspect will be explored further in Chapter 4.

1.2.2 Fully–parametric prediction

Parameter prediction

We turn now to describing the approaches that are now in common usage amongst forest biometricians. The parameter prediction method (e.g. Clutter and Bennett, 1965; Hyink and Moser, 1983) of estimating a size distribution proceeds by using the stand–level data as predictors for the parameters of a PDF. This amounts to a set of regression equations

$$\theta_{ij} = f(\beta_{ik} x_{jk}) + \epsilon_{ijk} \tag{1.2}$$

where θ_{ij} is the i^{th} parameter of the j^{th} stand, $i = 1, \ldots, m$ and $j = 1, \ldots, n$, and x_{jk} are the p stand–level covariates for the j^{th} stand, $k = 1, \ldots, p$; β_{ik} are regression parameters to be estimated, and ϵ_{ijk} are random error terms

with mean 0. Note that the specification given by Equation (1.2) allows non–linear models to be estimated (through the choice of f), however this is not common. The method proceeds in the following way:

1. The family of PDFs is chosen, for example, the Weibull PDF as given in Equation (1.1);
2. The parameters of the PDF, θ_{ij} are estimated for each sample plot using the tree–level data. This step is usually performed through the use of maximum likelihood estimates;
3. The regression parameters, β_{ik} are estimated through Equation (1.2), using the values of θ_{ij} obtained in the previous step;
4. The size distribution for the j^{th} stand is then the PDF with parameters $\hat{\theta}_{ij} = f(\hat{\beta}_{ik}x_{jk})$.

The estimation of the regression parameters β_{ijk} in Step 3 will depend not only upon the form of the function f, and the assumptions placed on the regression model, but also on the nature of the data itself. For example, Robinson (2004) suggests that considerable hierarchical structure can often be found in forestry data, and that mixed effects models (Laird and Ware, 1982) provide a suitable modelling framework for such data. In almost all applications of the parameter prediction method for estimating size distributions, the regression parameters β_{ik} are estimated using standard linear models (with or without random effects).

It is likely that the PDF parameters themselves will be correlated, due to the fact that we are estimating each parameter from the same sample. If we estimate the regression parameters β_{ik} independently for each PDF parameter θ_{ij}, then the assumption of independence between these parameters may not be met. A popular method for dealing with this situation when the assumed model is linear, that is, $f(\beta_{ik}x_{jk}) = \beta_{ik}x_{jk}$, is seemingly–unrelated regression (SUR, Zellner, 1962). SUR allows for correlations among the residuals from each of the regression equations to be accounted for by relaxing the independence assumption. Details regarding SUR may be found in Appendix A.3.

Parameter recovery

The parameter recovery method (see, for example, Hyink and Moser, 1983; Burk and Newberry, 1984) of estimating size distributions is quite similar to

the parameter prediction method, in that the size distribution is estimated from a family of PDFs indexed by parameters θ_{ij}. It is the process of how these parameters are estimated that is the difference.

Parameter recovery methods rely on the estimation of percentiles or moments of the size distribution, and are thus applications of the method of moments estimation technique. The procedure begins by estimating the tree–level moments (or percentiles), and relating these to the stand–level characteristics. The parameters of the size distribution are then estimated by matching the predicted (sample) moments (percentiles) to their corresponding population parameters. Because of this matching procedure, the Weibull distribution is the most commonly used size distribution in the forestry literature due to its cumulative distribution function having a closed form expression that makes moment–based methods analytically tractable.

Burk and Newberry (1984) provide an algorithm for recovering the parameters of a Weibull size distribution using sample moments as estimators. Bailey et al. (1989) describe a parameter recovery approach that uses the sample percentiles to recover the population parameters of the Weibull distribution. In their approach, the minimum and median diameters, along with the 25^{th} and 95^{th} diameter percentiles are predicted from the stand–level characteristics, along with quadratic mean diameter (a function of basal area and stems per hectare) through a set of regression equations:

$$d_{qj} = f(\beta_{qk} x_{jk}) + \epsilon_{qjk} \tag{1.3}$$

where d_{qj} is the q^{th} percentile (or quadratic mean diameter) of the the j^{th} stand, and β_{qk} and x_{jk} are as they were in Equation (1.2).

The process of recovering the PDF parameters then follows in a similar vein to that described in the previous section for the parameter prediction method. The first three steps are the same, with prediction of percentiles replacing prediction of parameters. The fourth step becomes:

4. The PDF parameters for the j^{th} stand are recovered from the predicted percentiles. For the Weibull distribution (Equation 1.1), this amounts

to:

$$\hat{\gamma}_j = \frac{n_j^{1/3}\hat{d}_{0j} - \hat{d}_{50j}}{n_j^{1/3} - 1}$$
$$\hat{\beta}_j = \frac{2.343088}{\log(\hat{d}_{95j} - \hat{\gamma}_j) - \log(\hat{d}_{25j} - \hat{\gamma}_j)}$$
$$\hat{\alpha}_j = -\frac{\hat{\gamma}_j \Gamma_{1j}}{\Gamma_{2j}} + \sqrt{\left(\frac{\hat{\gamma}_j}{\Gamma_{2j}}\right)^2 (\Gamma_{1j}^2 - \Gamma_{2j}) + \frac{\hat{d}_{mj}^2}{\Gamma_{2j}}}$$

where $\Gamma_{tj} = \Gamma\left(1 + t/\hat{\beta}_j\right)$, and Γ is the gamma function.

1.2.3 Partially–parametric prediction

Percentile Method

Introduced by Borders et al. (1987), the percentile method of estimating size distributions proceeds by estimating a set of percentiles at the tree–level, and relating these percentiles to the stand–level data. The predicted percentiles form the basis of a nonparametric estimate of the size distribution. This leads to the following set of regression equations:

$$q_{ij} = f(\beta_{ik} x_{jk}) + \epsilon_{ijk} \tag{1.4}$$

where q_{ij} is the i^{th} percentile of the j^{th} stand, $i = 1, \ldots, m$ and $j = 1, \ldots, n$ and x_{jk}, β_{ik} and ϵ_{ijk} are as they were in Equation (1.2). Borders et al. (1987) proposed a set of 12 regression equations for the $\{0, 5, 15, \ldots, 95, 100\}^{\text{th}}$ percentiles. Following their methodology, the 65^{th} percentile is chosen as the 'driver' percentile. Differences between the percentiles are then calculated, and the system of equations becomes

$$d_{65,j} = \beta_{65,k} x_{jk} \qquad \mathbf{d}^*_{q,j} = \mathbf{X}_j \boldsymbol{\beta}_{qk} \tag{1.5}$$

where $d^*_{q,j}$ is the difference between two successive percentiles for the j^{th} stand (for example, $d^*_{75,j} = d_{75,j} - d_{65,j}$), $\mathbf{X}_j$ are the observed stand characteristics, and $\boldsymbol{\beta}_{qk}$ are the model equation parameters. Again the equation parameters are estimated using seemingly unrelated regression (Appendix A.3).

To recreate the diameter distribution for a given plot, we first predict the percentile differences from Equation (1.5), then recover the predicted percentiles (e.g.. $\hat{d}_{75,j} = \hat{d}_{65,j} + \hat{d}^*_{75,j}$). The percentiles are then smoothed (for

example, using a constrained cubic spline) providing a CDF. The PDF is then found by taking the derivative of the spline.

1.2.4 Problems with the current approaches

The methods that we have introduced thus far are certainly fine in theory, and are excellent in practice if the data actually conform to the model that is being proposed. There are, however, problems with a parametric approach to size modelling. At the most basic level, problems will occur when the data do not conform to a single parametric family. For example, diameter distributions are often multi–modal, and in this case, single–family parametric distributions (such as the Weibull) will not provide a suitable fit. It is unclear what steps should be taken when this situation occurs. For instance, Maltamo et al. (2000) visually inspect empirical densities and remove those that are clearly multi–modal before estimating the distribution parameters. This is clearly an inefficient use of data, and ignores what is commonly seen in real–world applications.

The percentile method (Section 1.2.3) in some ways overcomes the problems that occur with prescribing a parametric distribution family, but is not without problems in itself. Firstly, the percentile method requires the estimation of a large number of percentiles to provide an adequate description of the distribution, all of which are estimated from the same set of stand–level data. Estimating the covariance matrix Σ (Equation A.3) when using SUR requires (on average) $2n/m$ observations for each element (where n is the number of observations per equation, and m is the number of equations, see, for example, Beck and Katz (1995)). This can be inefficient for low numbers of observations, and whilst similar problems will occur for the fully–parametric methods, they will not be as severe due to the much lower number of equations being estimated.

A further problem with the percentile method is that for any observation outside the predicted minimum (maximum) for a stand, this observation will have probability 0, with no margin for error. This is in contrast to the fully–parametric approaches which assume the PDF goes to 0 smoothly at the boundaries.

Not only can the parametric family be misspecified, but just as serious a problem is misspecification of the regression relationship (as in Equation 1.2). This is in no way restricted to modelling in the forest science, but is problematic across a wide range of statistical modelling applications.

In Chapters 2-4, we will seek to overcome both misspecification of the parametric family of the response distribution, and misspecification of the regression model, by utilising functional data analysis and nonparametric techniques.

1.3 Population size estimation in capture–recapture experiments

In this section we introduce methods of estimating the population size of a species in a sample area from capture–recapture experiments. A discrete time capture–recapture (CR) study consists of a fixed number of capture occasions where individuals from the population are uniquely marked or tagged upon initial capture and previously captured individuals are noted before being released back into the population. It is assumed that individuals do not lose their marks and behave independently from each other. Individual capture histories are then easily constructed and, under an appropriate model, can be used to estimate population size. We further assume here, and in Chapter 5, that the population is closed. This requires that we assume that the population of interest does not change during sampling. That is, there are no births and deaths among the population of interest, nor any immigration or emigration.

Modelling heterogeneity in the form of individual covariates, such as body weight or gender, is now commonly used in CR models (e.g. Otis et al., 1978; Pollock et al., 1984; Huggins, 1989, 1991) where it is known to reduce bias and increase precision for both model parameters and population estimates (e.g. Pollock, 2002). To account for heterogeneity In these models, the capture probabilities are estimated conditional on individual or environmental covariates such as weight and/or temperature.

The conditional likelihood of Huggins (1989) has been standard methodology for estimating the parameters of CR models and subsequently population sizes since it and an applications paper by the same author (Huggins, 1991) were published. Suppose that we have data from a CR experiment in the form of capture histories for individuals, and associated with an individual, a vector of covariates (e.g. weight). As an example of a capture history for an individual, assume that the CR experiment was conducted over five capture occasions, and that the individual was caught on occasions two and four. This individuals capture history would be $(0, 1, 0, 1, 0)$, where we let

$C_i(j) = 1$ denote the fact that individual i was captured at time j, and 0 otherwise. Assume that for each individual i, the probability of capture at occasion j is p_{ij}. Then for a population of individuals $i = 1, \ldots, N$, and possible capture occasions $j = 1, \ldots, \tau$, the likelihood of the observed data is proportional to

$$L = \prod_{i=1}^{N} \prod_{j=1}^{\tau} p_{ij}^{C_i(j)} (1 - p_{ij})^{(1-C_i(j))} \tag{1.6}$$

Of course, the likelihood given by Equation (1.6) relies on us having captured each individual in the entire population, which is usually not the case, and the population size N becomes the focus of estimation. There are number of further issues with the likelihood in Equation (1.6). The probabilities of capture p_{ij} are most likely unknown and need to be estimated from the data. Maximum likelihood estimates are possible, however in its current form, the likelihood is saturated, that is, the number of parameters is the same as the number of data points. This requires collapsing the number of parameters in some way. As an example, one could set $p_{ij} = p$, so that the probability of capture is the same for all individuals at all capture occasions. However in doing so, we have lost the ability to capture individual heterogeneity.

To allow for heterogeneity between individuals and over time, we can model the capture probabilities p_{ij} conditional on some measured vector of covariates $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})$. Then, letting $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ be the parameters relating the covariates to capture probability, a possible model for the capture probabilities is

$$p_{ij} = H(\mathbf{X}_i \boldsymbol{\beta}) \tag{1.7}$$

where the function H is chosen to ensure that $0 \leq p_{ij} \leq 1$. The logistic function $H(u) = \exp(u)/(1 + \exp(u))$ is common. Thus, we retain the ability to model individual heterogeneity, yet have made the model parsimonious by restricting the number of parameters that need estimating. Estimation of parameters and related inference is now made via the conditional likelihood (Huggins, 1989), which is proportional to

$$L_c = \prod_{i=1}^{D} \pi_i^{-1} \prod_{j=1}^{\tau} p_{ij}^{C_i(j)} (1 - p_{ij})^{(1-C_i(j))} \tag{1.8}$$

where D is the number of distinct individuals captured, and $\pi_i = 1 - \prod_j (1 - p_{ij})$ is the probability of individual i being captured at least once during the CR experiment. Equation (1.8) is known as the conditional likelihood, as it conditions on those individuals that are caught at least once, thus requiring no information on the individuals that were never captured.

When an individual is not captured at a time j, then the parameters in Equation (1.8) cannot be estimated, as the related covariate information X_{ij} will not have been measured. This is the focus of Chapter 5, where we will give more details. We now describe some of the current approaches to handling this situation of missing data.

1.3.1 The problem of missing data in CR experiments

Many researchers have commented on the need for methods in dealing with the inherent missingness of covariates when an individual is not captured (e.g. Pollock, 2002) and several approaches have been proposed. For example, Wang (2005) used a Monte Carlo method in an EM algorithm for continuous–time CR models with missing covariates, Zwane and der Heijden (2008) used multiple imputation to substitute for unobserved values in log–linear CR models and more recently, Xi et al. (2009) used the EM algorithm and the conditional likelihood approach of Huggins (1989). A more straightforward and traditional approach to dealing with missing values is to assume that the covariate is constant across capture occasions. In this naive approach, the measurement taken at the first capture of an individual is then fixed for all capture periods.

Another approach for time–varying covariates is approximation by discrete categories (Nichols et al., 1992). This has the advantage of not requiring how the covariates change over capture occasions, but is disadvantaged by a lack of precision, and ambiguity in the choice of discretisation.

A further possible approach is the use of full likelihood methods. However, these require the specification of the distribution of the covariate and how it changes over time. Complicating the analysis of full likelihood methods is the need to integrate out any missing covariates, a task that is near on impossible with many missing values and many capture occasions. A possible solution then is to use Bayesian methods, such as the approach taken by Bonner and Schwarz (2006) and King et al. (2008). In these analyses, the covariate at time t is assumed to be a realisation of a continuous time Weiner process with time–dependent drift $\mu(t)$. The parameters of the ensu-

ing Markov chain model are estimated using a component–wise Metropolis–Hastings algorithm. Whilst the problem of discretisation (in the Nichols et al. (1992) model) is overcome here, restrictions are still placed on the distribution and functional form of the covariate.

Time–varying individual covariates provide another interesting challenge in CR models. By construction of the model, the probability of capture is conditional on the covariate. However, because we can only measure the covariate when an individual is captured, we see also that the missing data mechanism is missing not at random (MNAR). That is, the probability that we measure the covariate is conditional on the probability of capturing the individual. Having these two processes intertwined in this way means that traditional methods of imputing the missing data may be inefficient.

In Chapter 5 we describe an EM algorithm that accounts for the time–variation of the covariate, and also the probability of missingness. The method that we propose overcomes the problems just discussed by the use of non-parametric smoothing and functional data analysis.

1.4 Chapter synopsis

Chapter 2 will introduce the functional regression tree (FRT) method (Nerini and Ghattas, 2007). The FRT method uses recursive partitioning to estimate a regression model between a functional response and multivariate predictors. We can use such a model to predict functional responses without requiring parametric assumptions about the functional form of both the function and regression relationship. We apply the method to the *E. globulus* data introduced in Section 1.2.1, and compare it to more common approaches in the forest science literature. The results of this comparison show that as both a predictive tool and informal inference tool, the FRT method outperforms the more common parametric approaches. This chapter is based on Lane et al. (2010).

Chapter 3 extends the work of Nerini and Ghattas (2007) by suggesting an alternative criterion for fitting a functional regression tree model. The original criterion of Nerini and Ghattas (2007) targeted node homogeneity, that is, it minimised the deviance within each node of the regression tree. The extension which we detail in Chapter 3 minimises the prediction error, and as such, provides flexible predictions for functional responses conditional on multivariate predictors. The extension of the work of Nerini and Ghattas

(2007) also results in more stable and precise fitting of a functional regression tree model. We will demonstrate this extension using measurements from stands of Douglas–fir (*P. menziesii*) as well as an extensive simulation study. This chapter is based on Lane and Robinson (2011).

Chapter 4 extends the work of Yao et al. (2005b) to the case where the longitudinal response is now a series of functions. The method which we introduce allows the prediction of functions at any time through nonparametric modelling of a sequence of covariance functions. To the best of our knowledge, this is the first time such a model has appeared in the literature to date. We will apply this model to the full longitudinally observed *E. globulus* data. This provides us with the means to predict the evolution of the diameter PDF in stands that have not had diameters directly measured, but have functional covariate information available. Comparing the results of this new approach to a more traditional approach, we find that the prediction error is reduced after allowing for the functional nature of the data. This chapter is currently being edited for submission to the Journal of the American Statistical Association.

Chapter 5 describes a novel use of functional data techniques in capture–recapture (CR) experiments. In this chapter, we combine the work of Yao et al. (2005a) with that of Huggins (1989) allowing the estimation of population size from CR experiments with continuous, time–varying covariates. We will demonstrate the method using data from a CR experiment involving the Mountain Pygmy Possum (*Burramys parvus*) as well as through simulation. The results demonstrate improved performance in terms of precision and variability of both population size estimates and CR model parameters when compared to more common approaches. This new methodology provides an exciting advancement on current techniques, and is currently being edited for submission to Biometrics.

Chapter 2

The functional regression tree method

This chapter introduces the functional regression tree (FRT) method (Nerini and Ghattas, 2007). The FRT method uses recursive partitioning to estimate a regression model between a functional response and multivariate predictors. We can use such a model to predict functional responses without requiring parametric assumptions about the functional form of both the function and regression relationship.

We have already seen in Chapter 1 a number of methods for predicting probability density functions (PDFs) from covariate information. There, we discussed two broad approaches that are currently used within the forest science literature, which we termed fully–parametric and partially–parametric.

Forest stand diameter distributions can adopt a wide variety of shapes, not all of which can be easily matched by specific functional forms. Further, diameter distributions within a single population can be too variable to be matched by a single parametric family. Motivated by this observation, we apply the FRT method to model the diameter distribution of *E. globulus* stands (that were introduced in Section 1.2.1).

We will first outline recursive partitioning and how it may be applied to estimate probability density functions. Secondly, we use the *E. globulus* data to compare the FRT method with parameter prediction and percentile methods. The results of this comparison show that as both a predictive tool and informal inference tool, the FRT method outperforms the more common parametric approaches that are in current use within forest science. This chapter is based on Lane et al. (2010).

2.1 Recursive partitioning

2.1.1 Recursive partitioning for a univariate response

We first review the recursive partitioning procedure for a univariate response (see, for example, Breiman et al., 1984; Hastie et al., 2009). Given observations Y and covariates X, we seek an estimate of $E(Y|X)$ by partitioning $\mathcal{X}$, the space of all possible observations, and estimating the conditional expectation of Y using the mean response within each partition.

The resultant partitioning can be displayed graphically as a tree, an example of which is shown in Figure 2.1. In this example, $\mathcal{X}$ has two partitions, which allocate observations with $x_j < t$ to the left child node, and $x_j \geq t$ to the right child node. The estimate of $E(Y|X)$ is then given by the mean in each of these nodes.

$$\hat{Y} = \begin{cases} \bar{y}_l & x_j < t \\ \bar{y}_r & x_j \geq t \end{cases}$$

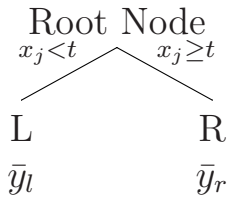


Figure 2.1: A graphical representation of the recursive partitioning procedure. The root node is partitioned into two disjoint child nodes.

This is a simple example of the widely–used CART (Breiman et al., 1984) algorithm for recursive partitioning, which uses binary splits to classify nodes. More generally, starting with all observations (at the root node as in Figure 2.1), we choose a binary split s. The procedure for choosing the best split is as follows. Define the deviance of a node r as

$$D(r) = \sum_{i \in N(r)} (y_i - \hat{y}_r)^2 \tag{2.1}$$

where $\hat{y}_r$ is the fitted value in node r, and $N(r)$ is the set of indices of observations in node r. Then, given a set of possible splits $s \in S$, each of which

splits the node r into child nodes r_1 and r_2, the best split s^* is defined as the split s that satisfies the objective function

$$s^* : s = \underset{s \in S}{\operatorname{argmax}}\ D(r) - D(r_1) - D(r_2) \tag{2.2}$$

This objective translates to minimising the sum of $D(r_1)$ and $D(r_2)$ during the splitting procedure. The value that minimises the deviance (Equation 2.1) is the mean value of the node, and so we set $\hat{y}_r = \bar{y}_r$, the mean of node r. That is, the algorithm guides the splits of the observations into classes *as well as* guiding the prediction for the class.

The procedure then repeats for both the left and right child nodes. That is, a split s^* is chosen to maximise Equation 2.2 for each child node, and so on, until any further splits will result in a child node that contains less than a pre–specified minimum number of observations. The minimum size of nodes may have a large impact on the recursive partitioning algorithm. For example, if the minimum is too small, the partition may be too sparse, however if the minimum is too large, valuable features can be smoothed out. Associated with minimum node size is computational cost, which we discuss in more detail in Chapter 3.

Generally, the trees that result from repeatedly splitting until no possible splits remain are overfit, that is, the space $\mathcal{X}$ is partitioned too finely on the training data. Therefore, predictions that are made for independent data will be less accurate. The established methodology for reducing this overfitting is cost–complexity pruning (Breiman et al., 1984). The pruning procedure introduces a penalty for the size (number of terminal nodes) of a tree, which can be estimated by cross–validation.

2.1.2 Recursive partitioning for a functional response

Numerous recursive partitioning methods for multivariate data have been proposed. A common theme is some form of dimension reduction prior to the recursive partitioning procedure, and the choosing of an appropriate form for the deviance function D. For example, Yu and Lambert (1999) analysed time–of–day patterns for international phone calls using recursive partitioning in two ways, both of which involved dimension reduction prior to the recursive partitioning scheme. They first fitted natural splines to the curves, then used the coefficients of the basis function as the response variable for recursive partitioning. The second method involved calculating the princi-

pal components of the time–of–day curves and using the first six principal components as the response variable. In both cases, the deviance function was chosen to be the standardised squared–error loss function (Mahalanobis distance). Cariou (2006) analysed electricity load curves in much the same way by applying the partitioning procedure to the first principal component of the electricity load curve, with deviance being measured by the sum of absolute deviations. De'ath (2002) provides an example of using a regression tree (with Euclidean deviance D) to model the distribution of 12 species of hunting spiders in relation to various environmental characteristics.

Nerini and Ghattas (2007) proposed using the whole probability density function (PDF), estimated non–parametrically, as the response variable. Their procedure, which they called a functional regression tree model (FRT), did not involve dimension reduction. They suggested that because interest lies in the PDF as a whole, an appropriate deviance function would be one based on f-divergence (Csiszár, 1967). The authors also provided a simulation study that supported this choice of deviance as opposed to a deviance based on Euclidean distance (such as that in Equation 2.1 for multivariate responses). In general, let the functional response for observations i and j be $Y_i(t)$ and $Y_j(t)$, defined over some range $t \in T$. Then if we can measure the dissimilarity between $Y_i(t)$ and $Y_j(t)$ by $C_{ij} = g(Y_i(t), Y_j(t))$, for some function g, the deviance of a node r can now be written

$$D(r) = \sum_{i \in N(r)} \sum_{j \in N(r)} C_{ij} \tag{2.3}$$

and the FRT procedure operates in exactly the same way as that described previously for a univariate response, but now with deviance function (2.3) used in the objective function (2.2). For PDFs, the natural candidate for the function g is the Kullback–Leibler divergence as proposed by Nerini and Ghattas (2007):

$$g(Y_i(t), Y_j(t)) = \int Y_i(t) \log \left(\frac{Y_i(t)}{Y_j(t)} \right) dt. \tag{2.4}$$

Functions for fitting FRT models with the Euclidean distance deviance criterion are provided in the R library mvpart (De'ath, 2007), which also performs cross–validation to assess the out–of–sample prediction error. The function g in the example above would then become $g = ||Y_i(t) - Y_j(t)||^2$. Our preliminary testing (on the simulated data set detailed in Nerini and Ghattas,

2007) showed that using Euclidean distance as the deviance criterion produced results that were no less accurate than those based on Kullback–Leibler divergence (however we investigate this further in Chapter 3). Therefore the results in this chapter use Euclidean distance as the deviance criterion. Cross–validation can yield highly variable outcomes for tree–based models (see, for example, Merler and Furlanello, 1997), and so instead we estimate the size of the FRT using a bootstrap scheme. In particular we use the $.632+$ version of the bootstrap (Efron and Tibshirani, 1997), which we detail in Appendix B.1.2.

As an example of the way in which the predictions of the functional response are made after the partitioning has occurred, Figure 2.2 displays the probability density functions that were used as the functional responses in the FRT method as applied in the next section. After two levels of partitioning have occurred, we see that the PDFs have been split into four different sections. The PDFs within each partition are more similar to each other, than any other partition. A further example of the partitioning is given in Appendix A.1.

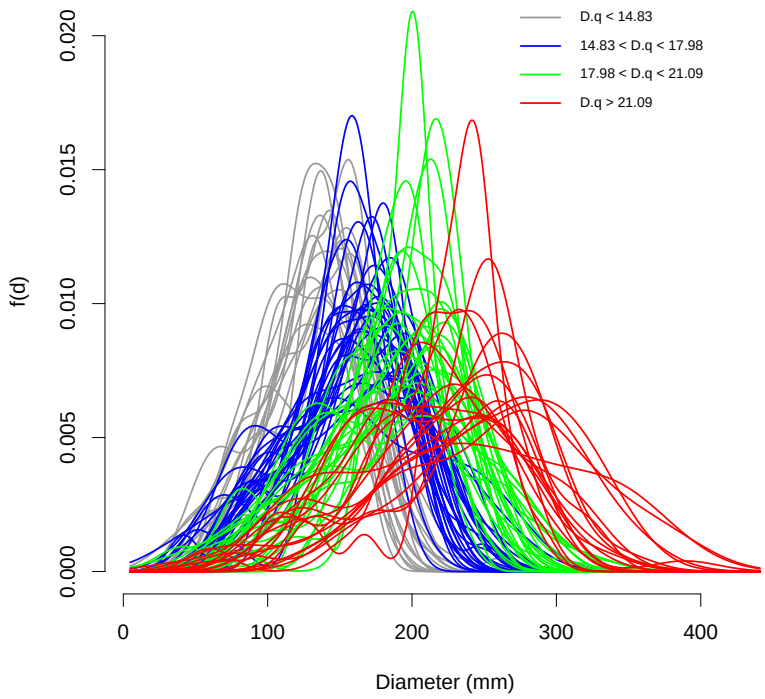


Figure 2.2: Example of PDFs after two levels of the recursive partitioning scheme has been performed.

2.2 Numerical comparisons

In this section we compare the FRT with a parameter prediction (PP) method using the Weibull distribution (see Section 1.2.2 of Chapter 1), and with Borders' (Borders et al., 1987) percentile method (PM, Section 1.2.3, Chapter 1) for predicting diameter distributions. The data used for the comparisons is that which was introduced in the introduction, and we will expand on this in the next section.

All analyses were performed in the open source statistical environment R (R Development Core Team, 2009). The FRT models were fitted using the library mvpart (De'ath, 2007), and the parameter prediction and percentile method models were fit using maximum likelihood and systemfit library (Henningsen and Hamann, 2007).

2.2.1 Data

The *E. globulus* data are from same–design replicated field experiments (a, b, c, d, e) in south–western Australia. The experiments compare stocking (planting density, N_0) treatments of 625, 833, 1000, 1250, 1667, and 2000 trees ha^{-1}. Tree diameter (d, over–bark at 1.3 m height) and total height (h) measurements at age (A) 8 years were used to calculate stand–level characteristics: basal area (G, m^2 ha^{-1}); density, as stems ha^{-1} (N); top height, as the mean height of the 100 largest diameter trees ha^{-1} (H, m) and total underbark volume (V, m^3 ha^{-1}). Quadratic mean diameter (D_q, cm) was calculated from G and N. Table 2.1 gives summary statistics for the stand characteristics over all sites.

Table 2.1: Summary statistics for *E. globulus* stand characteristics. The data are based on 90 plots representing 5 sites $\times$ 6 planting densities.

Characteristic	min – mean – max (sd)
Age (years)	7.81 – 7.84 – 7.90 (0.032)
Basal area (m^2 ha^{-1})	12.63 – 24.91 – 39.53 (6.796)
Density (stems ha^{-1})	438 – 1023 – 1804 (381.2)
Top height (m)	12.98 – 22.00 – 29.23 (4.328)
Volume (m^3 ha^{-1})	68.9 – 201.3 – 399.9 (85.73)
Quadratic mean diameter (cm)	12.31 – 18.24 – 27.27 (3.529)

All characteristics given above for the data were initially included in the model specifications for each of the FRT, PP and PM methods. Specifically, the full model was

$$\begin{aligned}\gamma_i = \alpha_i &+ \beta_0 \cdot \mathrm{N}_{0,i} + \beta_1 \cdot \mathrm{A}_i + \beta_2 \cdot \mathrm{G}_i + \beta_3 \cdot \mathrm{N}_i \\ &+ \beta_4 \cdot \mathrm{H}_i + \beta_5 \cdot \mathrm{V}_i + \beta_6 \cdot \mathrm{D}_{\mathrm{q},i} + \epsilon_i\end{aligned}$$

where α_i is the field experiment coefficient, $\boldsymbol{\beta}$ are the coefficients for stand–level characteristics, the errors ϵ_i are independent and identically distributed, and γ_i denotes the response parameter of interest, that is the shape, scale or location parameter of the Weibull distribution for the parameter prediction method, the p^{th} percentile for the percentile method, or the observed diameter distribution for the FRT method. Details of the fitting methods for the parametric models can be found in Appendix A.

We form the observed diameter distribution $Y_i(d)$ using a kernel density estimate of the diameters for the plot. For each plot $i = 1, \ldots, 90$, let D_{ij} be the j^{th} diameter measurement for the i^{th} sample plot, $j = 1, \ldots, n_i$ (n_i being the number of measurements in plot i). Then the kernel density estimate of the distribution of diameter d for plot i is

$$Y_i(d) = (n_i h_i)^{-1} \sum_{j=1}^{n_i} K\left(\frac{d - D_{ij}}{h_i}\right)$$

where K is the kernel function, for example the standard normal density, and we used the plug–in bandwidth h_i of Silverman (1986).

2.2.2 Goodness–of–fit/lack–of–fit measures

The ability of parametric forms to match the shape of diameter distributions will depend on the regularity of the underlying shape. For example, we expect that the non–parametric methods (FRT and percentile methods) will outperform the parametric method (Weibull parameter prediction method) when sample plots exhibit bimodality. We can estimate the bimodality of the underlying population using the dip statistic (Hartigan and Hartigan, 1985) which provides a measure of unimodality. We would also expect varying results when the distributions exhibit excessive sample skew. To assess skewness, we compare the sample skew with the theoretical values of skew using the predicted parameters of the Weibull distribution. These two values of skew should be approximately equal if little skew is apparent.

To compare the performance of the FRT, PP and PM methods, we use two goodness–of–fit statistics. The first is the root mean squared prediction error between the predicted and observed diameter distribution, given as

$$\text{RMSE}(Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||Y_i - \hat{Y}_i||^2}$$

where n is the number of observations in the sample, Y_i is the observed (kernel density) diameter distribution and $\hat{Y}_i$ is the predicted density from each method. This statistic measures the average discrepancy between each point on the observed and predicted diameter distribution curves. A low value of this statistic indicates that the predicted diameter distribution is close to the observed diameter distribution, across all points.

The second goodness–of–fit statistic is the root mean squared prediction error of volume (V, m^3 ha^{-1}). This statistic measures the discrepancy between observed and predicted volume averaged over all sample stands. For this statistic we use a two–stage process to estimate stand volume. A two–stage process is required as the volume functions that we use are individual–tree volume functions that use individual tree diameter and height as inputs. Heights for some (but not all) trees were measured in the *E. globulus* study, and so we estimated the missing heights using a generalised additive model (mgcv library of R, see Wood, 2006) of height on diameter. We then predicted stem heights for each diameter over the observed range of diameters. Individual tree volume was then estimated using the volume function given in Wong et al. (1999):

$$\hat{v}(h,d) = \frac{2.8737}{100000} \cdot \left(\frac{d}{10}\right)^2 \cdot h + \frac{4.0837}{10000} \cdot \frac{d}{10} \tag{2.5}$$

where d is diameter (mm) and $h = h(d)$ is the height (m) of a tree with diameter d predicted from the non–parametric height model.

Stand volume (m^3 ha^{-1}) for stand i was then calculated as

$$\hat{V}_i = \frac{n_i}{a_i} \int f_i(x)\hat{v}(h,x)dx$$

where $f_i(x)$ is the diameter density function, $\hat{v}(h,x)$ is the individual tree volume function (Equation 2.5), and n_i and a_i are the number of stems measured and the area of stand i respectively. Root mean squared prediction error of

volume per hectare is then given by

$$\text{RMSE}(V) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_i - \hat{V}_i)^2}$$

2.2.3 Results

The FRT method provided the best fit as measured by our goodness–of–fit statistics and graphical comparisons. Based on the goodness–of–fit statistics, the percentile method performed better than the parameter prediction method, however the percentile method resulted in unusual peaks in the diameter distributions in some instances.

Table 2.2 provides a summary of the stand characteristics included in the final model for all modelling procedures; D_q is included in each modelling procedure. Figure 2.3 provides a graphical representation of the functional regression tree fit and predicted diameter distributions for the *E. globulus* data (as the final fit results in 30 terminal nodes, which is unwieldy to display graphically, only a subset of the splits are shown). In this display, splits that reduce deviance the most are shown first, which in this case occurs for D_q at both first and second–level splits.

Table 2.3 gives the goodness–of–fit statistics for the final model specifications of each method. The FRT method is the clear leader based on these statistics, followed by the percentile method and parameter prediction method. A graphical summary of a subset of the model fits from each model is shown in Figure 2.4. Shown in this figure are the observed kernel density estimates used in the model fitting, overlaid on histograms of the true observed diameters. A good overlap is shown between the observed and predicted densities.

Table 2.2: Stand characteristics chosen for the final model specifications for the functional regression tree (FRT), parameter prediction (PP) and percentile methods (PM). For the PP and PM methods, a checkmark indicates the characteristic was present in at least one model component.

Characteristic	Characteristic included in final model?		
	FRT	PP	PM
Field Experiment	√		√
N_0		√	√
A		√	√
G	√		√
N	√	√	√
H	√	√	√
V		√	√
D_q	√	√	√

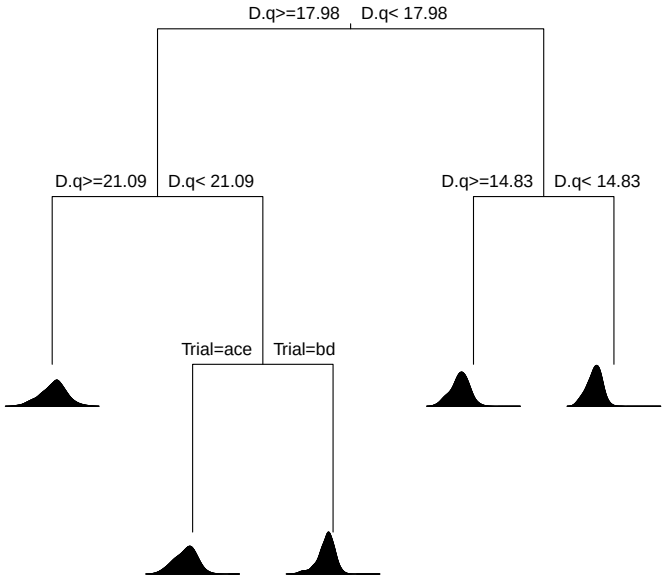


Figure 2.3: Graphical representation of the functional regression tree fit to the *E. globulus* data. Only the first four splits are shown. A graphical representation of the predicted diameter distribution is shown at each node.

Table 2.3: Goodness–of–fit statistics for the functional regression tree (FRT), parameter prediction (PP) and percentile methods (PM). RMSE(y), RMSE$_l(y)$ and RMSE$_u(y)$ represent the average discrepancy between the observed and predicted diameter distributions on a probability scale; RMSE(V) represents the discrepancy between observed and predicted stand volumes (m^3 ha^{-1}).

Statistic	FRT	PP	PM
RMSE(y)	0.0138	0.0324	0.0262
RMSE(V)	24.8	58.0	29.8

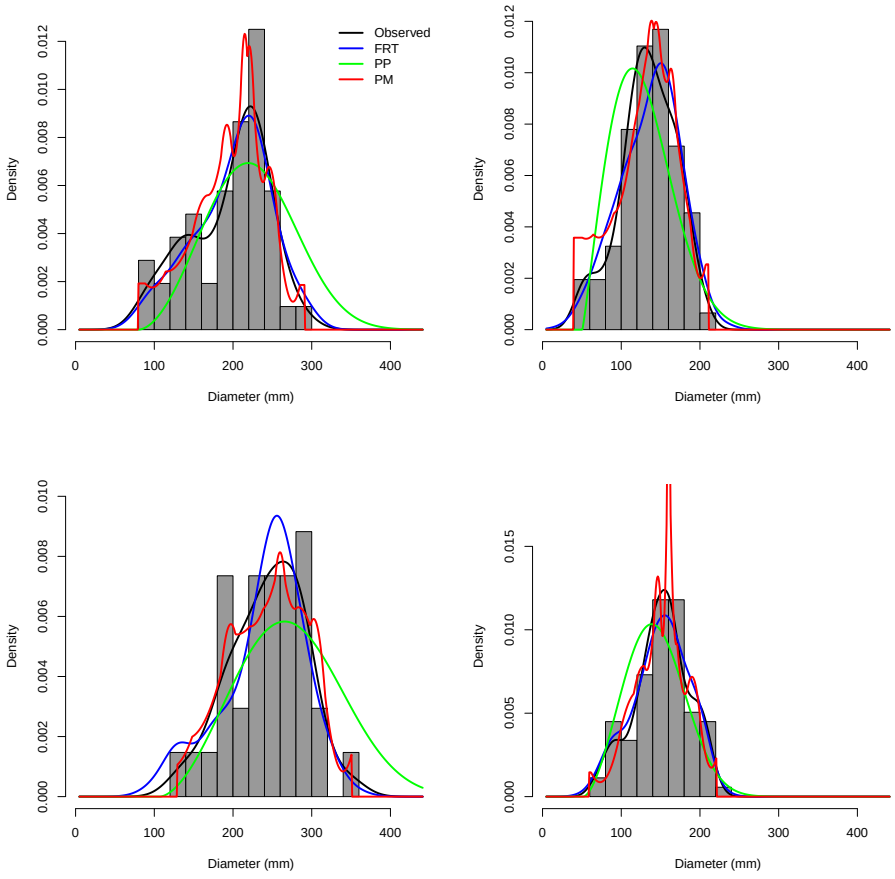
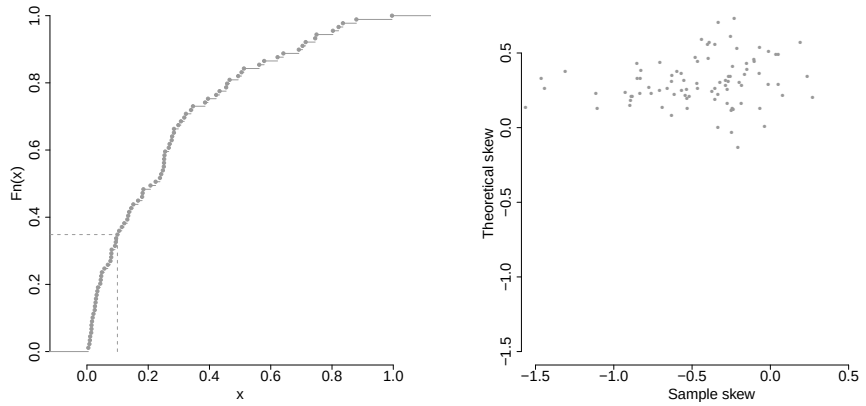


Figure 2.4: Comparison of observed and predicted diameter distributions overlaid on diameter histograms for a range of plots.

Figure 2.5 shows the empirical cumulative distribution function (ECDF) for the p–values of the dip statistic, along with a scatterplot of the sample skew versus the theoretical Weibull distribution skew calculated from the

parameter prediction method estimates. From the ECDF, approximately 35% of the sample plots have a p–value of less than 0.1, indicating that for the other 65% of the plots, there is only modest evidence for unimodality. For the Weibull distribution to provide a suitable fit to the data, we would expect to see that the sample skew and theoretical skew be approximately equal (that is, falling along a 1:1 line) which is not the case.

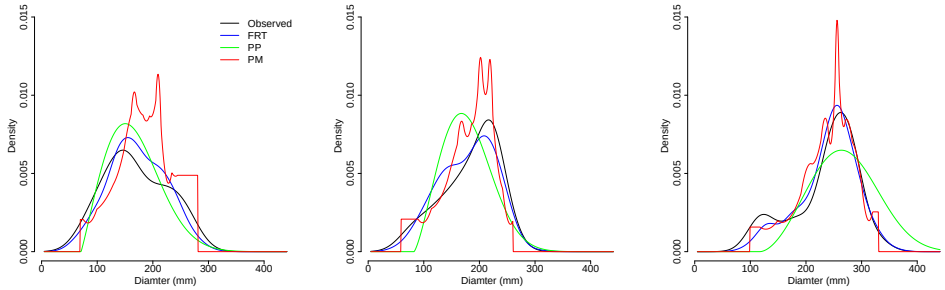


(a) ECDF of dip test p–values

(b) Scatterplot of sample skew *v.* theoretical skew

Figure 2.5: Graphical summaries of selected sample characteristics for the *E. globulus* data: (a) Empirical cumulative distribution function of p–values from the dip test for unimodality and (b) scatterplot of sample skew *v.* theoretical Weibull distribution skew calculated from parameter prediction estimates.

The evidence contained in Figure 2.5 suggests bimodality and excess skewness are features of the data. Figure 2.6 shows the observed and predicted diameter distributions for three sample plots that display these characteristics. The predicted diameter distributions from (4 out of 30) terminal nodes are shown in Figure 2.7, along with the grand mean. The predictions result from the following rules: $23.67 \leq \mathrm{D_q} < 25.46$; $21.92 \leq \mathrm{D_q} < 23.67$ and $\mathrm{G} \geq 27.53$; $21.09 \leq \mathrm{D_q} < 23.67$ and $\mathrm{G} < 27.53$; and $14.83 \leq \mathrm{D_q} < 16.53$, $\mathrm{H} \geq 19.95$ and the plot is in field experiment a, b or c.



(a) (b) (c)

Figure 2.6: Example comparisons of model predictions for three plots that display a range of skewness and bi–modality.

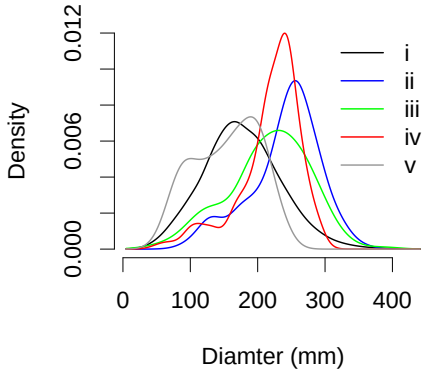


Figure 2.7: Graphical representation of the predictions from (4 out of 30) terminal nodes from the functional regression tree fitted to the *E. globulus* data. The predictions result from the following rules: i) Grand mean; ii) $23.67 \leq \mathrm{D_q} < 25.46$; iii) $21.92 \leq \mathrm{D_q} < 23.67$ and $\mathrm{G} \geq 27.53$; iv) $21.09 \leq \mathrm{D_q} < 23.67$ and $\mathrm{G} < 27.53$; and v) $14.83 \leq \mathrm{D_q} < 16.53$, $H \geq 19.95$ and the plot is in field experiment a, b or c

2.3 Discussion

The results of the goodness–of–fit statistics presented in the previous section indicate that parametric methods which make strong assumptions about the form of the densities may not perform as well as non–parametric methods, such as the FRT and percentile methods considered in this paper. However, basing the choice of method on goodness–of–fit statistics (e.g. Table 2.3) would prove misleading at best. The goodness–of–fit statistics are averages over the whole data set, and thus smooth individual errors. Examination of the predicted diameter distributions (Figure 2.6) shows that (in this case) the percentile method resulted in some predicted diameter distributions with unusually large peaks — something not seen in the observed distributions. The FRT method, however, was not only the clear leader in goodness–of–fit, but examination of the predicted densities shows that the model more closely conforms to the observed densities than the other two methods considered.

An advantage of the FRT method is that the graphical display of the FRT structure (Figure 2.3) can provide an indication of the *importance* of covariates in determining the shape of diameter distributions. Table 2.2 shows that quadratic mean diameter appeared in the final model for all methods, and is likely to play a vital role in determining the shape of the diameter distribution. Traditional regression models provide an indication of how important a covariate is to the response (e.g. p–values), however do not provide an indication of importance *relative to other* covariates. The graphical display of a FRT structure shows the hierarchy of splits, from the root node (all observations) to the terminal nodes (which give the predictions). Figure 2.3 indicates that the quadratic mean diameter is highly important in determining the shape of the diameter distribution. However, this display should be taken as an indication only; there are many interactions at play over the whole FRT structure.

The versatility of the FRT method for fitting a wide variety of diameter distribution shapes is made clear by examining Figure 2.6. This figure compares model predictions for diameter distributions that have skewness and bi–modality present; Figures 2.6(a) and 2.6(c) show diameter distributions where the minor modes are on opposite sides of the major modes in comparison to each other. These quite diverse diameter distribution shapes would be extremely difficult to capture with a single parametric family. A model which fits a mixture of parametric families to each diameter distribution would be likely to capture this diversity, however the analysis of such a model would be extremely complex.

The FRT method, in conjunction with resampling methods, has a number of advantages over traditional regression methods for prediction. As discussed in Section 2.1, the choice of covariate on which to split is made at each stage of the partitioning procedure, and so more complex models, that account for all important factors, can be fitted. The advantage is that these models are chosen 'automatically' by the procedure, where the choice is made by maximising the objective function (2.2) at each split. This means that the FRT method can fit a much broader class of models, as illustrated in Figure 2.3, where the density shown in the 2^{nd} node from the left, results from the prediction rule that the i^{th} plot has $17.98 \leq \text{D}_{\text{q}}(i) < 21.09$ and belongs to field experiment *a*, *c*, or *e*. Traditional regression models cannot provide this level of complexity automatically. As we will investigate in Chapter 3, the choice of dissimilarity (in Equation 2.3) for the FRT method can change the results.

As discussed in Section 2.1.1, minimum node size may impact on the FRT. In the case study presented in Section 2.2, we used the default minimum node size as given by the implementation of the FRT in the R library mvpart (De'ath, 2007); this minimum node size is set to 2. To check the effect of changing the minimum node size, we also fitted the FRT with minimum node sizes 5 and 10, and investigated the error of the FRT (as measured by the RMSE, see Section 2.2). These results indicated that RMSE increased quite dramatically when minimum node size increased, from 0.0138 as in Table 2.3 for the default minimum node size 2, to 0.0185 and 0.0235 for minimum node sizes of 5 and 10 respectively. Given the focus of the method was to predict the PDF, we felt no need to deviate from the default setting.

The choice of how complex a model should be in the FRT method is then made through resampling techniques which minimise a predictive error function (e.g. Hastie et al., 2009, Chapter 7). This step is performed after the initial function fit, and is thus relatively quick to compute. In comparison, in order to select the best subset from a suite of models in the traditional regression approach, we need to repeat the resampling process for each submodel to be tested, which can be computationally demanding if there are a large number of covariates. For the parameter prediction and percentile methods, this process is made even more complex by the need to fit equations for each parameter (parameter prediction) or each percentile (percentile method) simultaneously for each model.

A potential disadvantage with percentile methods is that any diameter

outside the predicted minimum (maximum) diameter (for a stand) has probability 0 whereas the FRT method assigns a (small) probability to these diameters. Whilst we could choose between the methods based on what assumptions we wish to place about the errors in the tails, the FRT method outperforms the percentile method on all goodness–of–fit statistics (Table 2.3).

Further refinement to the FRT may be possible by introducing a weighting scheme that could account for features of interest within the diameter distributions, depending on the context of the analysis, or even to adjust for possible functional outliers. Expert analysis by a forest scientist could identify 'abnormal' PDFs visually, prior to fitting the model. Weights could then be attached to relevant observations, depending on the final goal of the analysis, for example, giving a higher weight to those diameter distributions that may result in optimal volume of harvested timber. A further possibility could be the application of functional depth (e.g. Hyndman and Shang, 2010) to downweight possible outliers prior to fitting the FRT.

The *E. globulus* data used to demonstrate the FRT method comprises single–species stands which adds a layer of uniformity to the expected shape of the distributions. When the data includes mixed species, it is more likely that the size distributions of the samples are non–standard, often displaying both excess skew and multimodality. We would expect that under these circumstances, the FRT approach should provide a superior fit compared with traditional techniques.

Chapter 3

Improving the functional regression tree method

This chapter extends the work of Nerini and Ghattas (2007) by suggesting an alternative criterion for fitting a functional regression tree model. The original criterion of Nerini and Ghattas (2007) targeted node homogeneity, that is, it minimised the deviance within each node of the regression tree. The extension we detail in this chapter minimises the prediction error, and as such, provides flexible predictions for functional responses conditional on multivariate predictors. The extension of the work of Nerini and Ghattas (2007) also results in more stable and precise fitting of a functional regression tree model.

In the previous chapter we found that there was no discernible difference between using Kullback-Leibler divergence and Euclidean distance in the definition of node deviance (Equation 2.3). In this chapter we explore the Kullback-Leibler divergence form of the deviance in more detail and propose an adjustment to the Nerini and Ghattas (2007) procedure that performs considerably better under both simulation conditions and in the analysis of a case–study data set. An efficient parameterisation that leads to a computationally inexpensive bootstrap scheme for model selection is also discussed. Features commonly found in subject–matter data such as: low sample size and low observation numbers; correlation between covariates; nuisance covariates; and non–standard distributions are investigated through a comprehensive simulation study. This chapter is based on Lane and Robinson (2011).

3.1 Review of the FRT method

We will begin this chapter with a brief review of the FRT method that was introduced in Chapter 2. Recall that we have observations $(X_i, Y_i(\cdot)), i = 1, \ldots, N$ independent and identically distributed as $(X, Y(\cdot))$, where $X \in \mathbb{R}^d$ and $Y(\cdot)$ is a functional response variable, $d \in \mathcal{D}$ which we are assuming is a probability density function. We seek to estimate the conditional expectation $E(Y|X)$ by recursively partitioning $\mathcal{X}$, the space of all possible observations of X.

As described in Section 2.1.2, Nerini and Ghattas (2007) proposed using Kullback-Leibler divergence in the calculation of node deviance. As described previously, let a_{ij} be an appropriately defined dissimilarity measure between two functions, $a_{ij} = g(Y_i(\cdot), Y_j(\cdot))$. When we use Kullback-Leibler divergence as the function g, a_{ij} is as in Equation (2.4). The deviance of a given node r is then written as

$$D_1(r) = \sum_{i \in N(r)} \sum_{j \in N(r)} a_{ij} \tag{3.1}$$

and we seek the best split s^* that satisfies the objective function

$$s^* = \underset{s \in S}{\operatorname{argmax}}\ D(r) - D(r_1) - D(r_2) \tag{3.2}$$

Given that one of our main objectives is the prediction of a new PDF conditional on a set of covariates, it would make sense to use an objective function that corresponds to this goal. Accordingly, we suggest a modification of the deviance for this purpose which explicitly includes deviation from the *mean* curve within a node

$$D_2(r) = \sum_{i \in N(r)} \left[\mathrm{KL}(Y_i, \bar{Y}_r) + \mathrm{KL}(\bar{Y}_r, Y_i) \right]. \tag{3.3}$$

Here, $\bar{Y}_r$ is the mean of the densities Y_i in node r, and both KL terms are needed because KL–divergence is asymmetric. Given $\bar{Y}_r$ is also a density, the two KL terms in Equation (3.3) still satisfy the properties of KL–divergence, that is that $\mathrm{KL}(Y_i, Y_j) \geq 0$, with equality only if $i = j$. Note that Nerini and Ghattas (2007) also use the symmetric version of KL in their deviance definition.

Adoption of the deviance in Equation (3.3) seems to go against the usual

restriction on the objective function (Equation 3.2), in that $D_2(r)$ is not necessarily greater than or equal to $D_2(r_1) + D_2(r_2)$ for all partitions r_1 and r_2 of r. This restriction is generally placed on the objective function in order to ensure that splits do not decrease homogeneity. However, we justify the relaxation of this restriction by noting that the deviance $D_2(r)$ measures the predictive capability of the FRT. In contrast, the deviance given by Nerini and Ghattas (2007) measures homogeneity of a node. Thus in essence, the two deviances are seeking a different goal. Given this (and the fact that we seek to maximise the objective function 3.2), any split that results in $D_2(r) < D_2(r_1) + D_2(r_2)$ will just be ignored as providing an inadequate prediction. Conceptually, this results in a modified objective function:

$$s^* = \underset{s \in S}{\operatorname{argmax}}\ D(r) - \min\{D(r_1) + D(r_2), D(r)\}$$

which then satisfies the preceding restriction. The pruning of a tree does use this restriction, however it also relies on a tree having already been grown.

There are generally two steps in the pruning process: the first is recovering the smallest minimising subtree from the maximal tree, the second is pruning the resulting smallest minimising subtree. In either case, a tree has already been grown. Thus if $D(r) < D(r_1) + D(r_2)$ (which, whilst not included in our results, closer inspection of the model fits using $D_2(r)$ in our simulations showed that $D_2(r) > D_2(r_1) + D_2(r_2)$) the split was not considered *during* the growing stage, resulting in all terminal nodes r_1, r_2 satisfying the condition $D(r) \geq D(r_1) + D(r_2)$. Hence, finding the smallest minimising subtree and further, pruning this subtree under the proposed deviance D_2, is accomplished in the usual way.

3.2 Computational details

Recursive partitioning is an exhaustive search method, that is, it looks at *all* possible solutions in order to pick the best one. The computational demands of the procedure, however, can be minimised by noting that the objective function Equation (3.2) with deviance $D_2(r)$ (or indeed, $D_1(r)$), is solely a function of divergences between observations. Setting $\mathcal{A}$ to be a dissimilarity matrix with entries $a_{ij} = \mathrm{KL}(Y_i, Y_j)$, each divergence can be calculated before the partitioning procedure is run. Thus, the value of the objective function at any split s requires only the summation of the appropriate en-

tries from $\mathcal{A}$. This is of course true for any deviance D that can be written as the sum of dissimilarities, as in Equation (3.1). Deviances that have this property include Euclidean distance, $D_1(r)$, and as we show below, $D_2(r)$; in fact, $D_2(r) = D_1(r)/n_r$ (Result 1). Without the relationship between $D_1(r)$ and $D_2(r)$, the value of the objective function (3.2) for any split s using deviance $D_2(r)$ would need to be recalculated at each iteration.

Estimating the KL–divergence between two nonparametric PDFs is a much more computationally demanding task than, for example, Euclidean distance, so without making use of a global dissimilarity matrix $\mathcal{A}$, the cost of calculation within the procedure itself can be prohibitive. For example, we show in Section 3.2.2 that calculating KL–divergence within the recursive partitioning procedure results in $O(pN^3)$ calculations (where p is the number of covariates in the model), compared to $O(N^2)$ when calculated outside the procedure. These computational savings are even more important when resampling is used to estimate the tree–size penalty.

3.2.1 A Kullback–Leibler divergence relation

Consider the deviance definition (3.3), for a node r with mean density $\bar{y}_r$. Then the following result holds

Result 1.

$$\sum_{i \in N(r)} [\mathrm{KL}(y_i, \bar{y}_r) + \mathrm{KL}(\bar{y}_r, y_i)] = \frac{1}{n_r} \sum_{i \in N(r)} \sum_{j \in N(r)} \mathrm{KL}(y_i, y_j)$$

Proof of Result 1. Recognise that

$$\begin{aligned} \mathrm{KL}(y_i, \bar{y}_r) &= \int y_i(t) \log \left(\frac{y_i(t)}{\bar{y}_r(t)} \right) dt \\ &= - \int y_i(t) \log \left(\frac{\bar{y}_r(t)}{y_i(t)} \right) dt \end{aligned}$$

then letting $n_r = |N(r)|$ be the number of observations in node r, and drop-

ping the reference to the integration variable t for simplicity,

$$\begin{aligned}
D_2(r) &= \sum_{i \in N(r)} \mathrm{KL}(y_i, \bar{y}_r) + \mathrm{KL}(\bar{y}_r, y_i) \\
&= \sum_{i \in N(r)} \left\{ -\int y_i \log\left(\frac{\bar{y}_r}{y_i}\right) + \int \bar{y}_r \log\left(\frac{\bar{y}_r}{y_i}\right) \right\} \\
&= \sum_{i \in N(r)} \left\{ -\int y_i \log\left(\frac{\bar{y}_r}{y_i}\right) + \frac{1}{n_r} \int \sum_{j \in N(r)} y_j \log\left(\frac{\bar{y}_r}{y_i}\right) \right\} \\
&= \sum_{i \in N(r)} \left\{ -\int y_i \log\left(\frac{\bar{y}_r}{y_i}\right) + \frac{1}{n_r} \int y_i \log\left(\frac{\bar{y}_r}{y_i}\right) + \frac{1}{n_r} \int \sum_{j \in N(r) \setminus i} y_j \log\left(\frac{\bar{y}_r}{y_i}\right) \right\} \\
&= \frac{1}{n_r} \sum_{i \in N(r)} \left\{ \sum_{j \in N(r) \setminus i} \int y_j \log\left(\frac{\bar{y}_r}{y_i}\right) - (n_r - 1) \int y_i \log\left(\frac{\bar{y}_r}{y_i}\right) \right\} \\
&= \frac{1}{n_r} \sum_{i \in N(r)} \left\{ \sum_{j \in N(r) \setminus i} \int y_i \log\left(\frac{\bar{y}_r}{y_j}\right) - (n_r - 1) \int y_i \log\left(\frac{\bar{y}_r}{y_i}\right) \right\}
\end{aligned}$$

now noting that $\sum_{i \in N(r)} \sum_{j \in N(r) \setminus i} \int y_i \log\left(\frac{\bar{y}_r}{y_j}\right)$ has $(n_r - 1)$ terms equal to $\int y_i \log\left(\frac{\bar{y}_r}{y_j}\right)$ for $j \neq i$,

$$\begin{aligned}
D_2(r) &= \frac{1}{n_r} \sum_{i \in N(r)} \left\{ \sum_{j \in N(r) \setminus i} \int y_i \log\left(\frac{\bar{y}_r}{y_j}\right) - (n_r - 1) \int y_i \log\left(\frac{\bar{y}_r}{y_i}\right) \right\} \\
&= \frac{1}{n_r} \sum_{i \in N(r)} \left\{ \sum_{j \in N(r) \setminus i} \left[\int y_i \log\left(\frac{\bar{y}_r}{y_j}\right) - \int y_i \log\left(\frac{\bar{y}_r}{y_i}\right) \right] \right\} \\
&= \frac{1}{n_r} \sum_{i \in N(r)} \left\{ \sum_{j \in N(r) \setminus i} \int y_i \log\left(\frac{y_i}{y_j}\right) \right\} \\
&= \frac{1}{n_r} \sum_{i \in N(r)} \sum_{j \in N(r)} \mathrm{KL}(y_i, y_j) \\
&= \frac{1}{n_r} D_1(r)
\end{aligned}$$

□

3.2.2 Computational complexity

We now demonstrate the efficiency gains made by calculating KL(y_i, y_j) prior to the recursive partitioning procedure compared with calculation within the procedure. This result applies in general to any deviance which can be written as a sum of dissimilarities between observations. With this in mind, it is easy to see that in terms of dissimilarity calculations, with a total of M observations there are $(M-1)$ operations for each observation, resulting in $M(M-1) = O(M^2)$ operations in total.

Consider now the case where the deviance D in the objective function (3.2) needs to be calculated within the recursive partitioning procedure itself. We discussed in Section 3.1 that while growing the tree, growth continues until all nodes contain at least a minimum number of observations N_{min}, and any further split will result in nodes with less than N_{min} observations. For example, consider a node S with n_s observations and a continuous split variable X_1 as shown in Figure 3.1. Then there are $n_s - 2N_{\text{min}} + 1$ possible splits c_j, such that $x_1 < c_j$ and $x_1 \geq c_j, j = (N_{\text{min}} + 1), \ldots, (n_s - N_{\text{min}} + 1)$ where c_j is the j^{th} order statistic $x_{(j)}$. Then letting n_l be the number of observations sent to the left child node L, $(x_1 < c_j)$, and similarly n_r the number of observations sent to the right child node R (where $n_l + n_r = n_s$), the number of deviance calculations in each node will be $2n_l$ and $2n_r$ respectively for nodes L and R, giving $2(n_l + n_r) = 2n_s$ calculations for the split c_j. Thus for each possible split c_j there will be $2n_s(n_s - 2N_{\text{min}} + 1) = O(n_s^2)$ operations.

To derive the minimal number of operations needed, note that for any node with $N_{\text{min}} \leq n_s < 2N_{\text{min}}$, that node will not be able to be split any further as all possible splits will result in at least one child node with less than N_{min} observations. The minimally split tree is then the tree with splits chosen such that $2N_{\text{min}} - 1$ observations are always sent to either the left (or right) child node, as shown in Figure 3.2 until no further splits can be made.

Given that at each split we send $2N_{\text{min}} - 1$ to either the left (or right) child

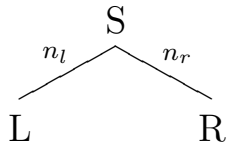


Figure 3.1: Splitting a node S into nodes L and R. n_l observations are sent to the left node L whilst n_r observations are sent to the right node R.

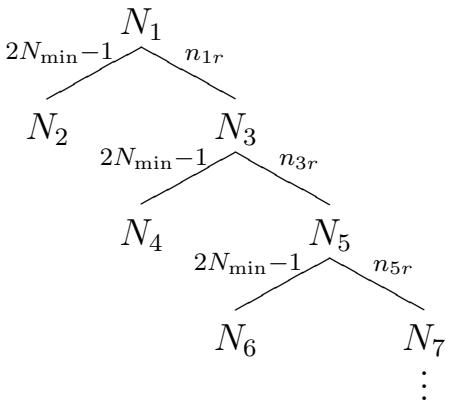


Figure 3.2: An example of a minimally split tree. At each step of the recursive partitioning procedure, $2N_{\min} - 1$ observations are sent to the left node.

node, the minimum number of splits $K_{\min}$ can be found as

$$\begin{aligned} K_{\min} &= \min_K M - K(2N_{\min} - 1) \\ \text{s.t.} \quad & M - K(2N_{\min} - 1) > 0 \\ \Rightarrow \quad & K(2N_{\min} - 1) < M \\ & K < \frac{M}{2N_{\min} - 1} \\ \Rightarrow K_{\min} &= \lfloor K \rfloor \end{aligned}$$

The minimum number of deviance calculations then required is

$$\begin{aligned} &\sum_{k=0}^{K_{\min}-1} 2 \cdot (M - k(2N_{\min} - 1)) \cdot (M - (k+1)(2N_{\min} - 1)) \\ &\approx \frac{2}{3} \frac{M(-4N_{\min}^2 + 4N_{\min} + M^2 - 1)}{2N_{\min} - 1} \end{aligned}$$

with equality when $K_{\min} = \lfloor K \rfloor = K$. To derive the maximum number of splits, $K_{\max}$, note that the maximal tree is the tree with splits chosen such that $N_{\min}$ observations are always sent to either the left (or right) child node, until no further splits can be made. Given this, the maximum number of splits

$K_{\max}$ can be found as

$$\begin{aligned} K_{\max} &= \min_K M - (K+1)N_{\min} \\ \text{s.t.} \quad & M - (K+1)N_{\min} > 0 \\ \Rightarrow \quad & (K+1)N_{\min} < M \\ & K < \frac{M}{N_{\min}} - 1 \\ \Rightarrow K_{\max} &= \lfloor K \rfloor \end{aligned}$$

The maximum number of deviance calculations then required is

$$\begin{aligned} & \sum_{k=0}^{K_{\max}-1} 2 \cdot (M - kN_{\min}) \cdot (M - (k+1)N_{\min}) \\ \approx & \frac{2}{3} \frac{M((N_{\min} \cdot M)^2 - 3M^2 N_{\min} + 3M^2 - N_{\min}^4 + 2N_{\min}^3 - N_{\min}^2)}{(N_{\min} - 1)^3} \end{aligned}$$

with equality when $K_{\max} = \lfloor K \rfloor = K$.

The above calculations are based on a single continuous split variable X_1. For p continuous variables, the number of KL–divergence calculations increases to $O(pM^3)$, which does not represent a substantial penalty for $p \ll M$, however by calculating a 'global' dissimilarity matrix $\mathcal{A}$ (as described earlier), the number of KL–divergence calculations does not change with including extra variables.

3.2.3 Functional standard errors

Nerini and Ghattas (2007) used 20–fold cross–validation (Appendix B.1.1) to estimate the optimally sized tree. Results in Merler and Furlanello (1997) and our own experience indicated that cross–validation proved to be unsuitably variable when used to estimate the tree–size penalty. We opted for the $0.632+$ bootstrap (Efron and Tibshirani, 1997, see Appendix B.1.2 for details) for which the reduction in variation of our fit statistics (Section 3.3) was quite significant when compared with cross–validation. Table B.1 in Appendix B.2 provides a detailed comparison of the standard deviation of the number of terminal nodes (for one of the models to be introduced in Section 3.3) when using cross–validation and bootstrap $0.632+$.

As a by–product of the bootstrap procedure that we used to estimate the tree–size penalty, a point–wise estimate of the standard error curve for a pre-

diction of the PDF can be made. Let this penalty estimate be $\hat{\alpha}$; we apply the estimated penalty to each of the FRT models fit to the $b = 1, \ldots, B$ resamples, resulting in a prediction model for each b: $\hat{P}_{\hat{\alpha},b}$. Letting $\hat{Y}_{i,b}$ be the predicted PDF for the i^{th} plot using model $\hat{P}_{\hat{\alpha},b}$, the standard error can be estimated for $\hat{Y}_i$ by taking the point–wise standard errors of the B (resampled) predicted PDFs, $\hat{Y}_{i,b}$.

3.3 Numerical results

3.3.1 Simulation study

We now describe a simulation study performed to compare the deviances $D_1(r)$, $D_2(r)$ (Equations 3.1 and 3.3) with Euclidean distance, $D_3(r)$:

$$D_3(r) = \sum_{i \in N(r)} ||Y_i - \bar{Y}_r||^2$$
$$\text{where } \bar{Y}_r = n_r^{-1} \sum_{i \in N(r)} Y_i.$$

We expand on the previous work of Nerini and Ghattas (2007) to cover a range of scenarios that may be experienced in practice. In their simulation, Nerini and Ghattas (2007) draw PDF observations from one of four possible distributions. The models in the simulations that we perform assume a specific functional form, however the parameters of these functions are allowed to vary. This is very often the assumption made in practice; for example, the parameter prediction method (e.g. Robinson, 2004; Vanclay, 1994) assumes an underlying population distribution and attempts to estimate sample–specific parameters for each observation.

Recall the setup of the response diameter PDFs in the previous chapter (Section 2.2.1). Let n_i be the number of measurements per observed sample stand. Then for a model $\mathcal{G}$, generate n_i independent and identically distributed samples, denoted $D_{ij}, j = 1, \ldots, n_i$, from the distribution $f_{\mathcal{G}}(\cdot|\mathbf{x}_i)$. A single observation is then given as $(Y_i, \mathbf{X}_i)$, where $Y_i = Y_i(d)$ is the nonparametric kernel density estimate

$$Y_i(d) = (n_i h_i)^{-1} \sum_{j=1}^{n_i} K\left(\frac{d - D_{ij}}{h_i}\right)$$

For a given n_i (the number of samples per observation) and M (the number of observations), the simulation proceeds as described in Table 3.1. For each model, the number of simulations was 100, and each simulation was run for all combinations of observations and samples per observation:

$$M \in \{50, 200\}$$
$$n_i \in \{15, 25, 50, 75, 100, 150, 200, 250, 500, 1000\}$$

Table 3.1: Description of simulation procedure.

1. Generate M observations $(Y_i, \mathbf{X}_i)$ from model G,
2. Fit the FRT model to the observations using each version of deviance: $D_1(r)$, $D_2(r)$ and $D_3(r)$,
3. Generate a further 100 observations (independent of the first M observations) to be used as a test set,
4. Using the FRT fit from step 2, predict $\hat{Y}_t$ for each $\mathbf{X}_t$ in the test set
5. Calculate average Kullback–Leibler divergence (KL_a) between the observed and predicted densities from the training and testing data sets, along with KL_a between the *actual* and predicted densities from the testing set

$$\text{KL}_a = \frac{1}{M} \sum_{t=1}^{M} \Big[\text{KL}(Y_t, \hat{Y}_t) + \text{KL}(\hat{Y}_t, Y_t) \Big]$$

The models $\mathcal{G}$ used in the simulations have been chosen so that they cover a range of conditions that may be seen in practice such as: correlation between the covariates $\mathbf{X}$, nuisance variables, and non–linear relationships in the data generating process. Model 1 is the model that was investigated by Nerini and Ghattas (2007), and Model 2 extends Model 1 with extra distribution possibilities. Models 3 and 4 include four noise variables, with the distribution parameters of Model 4 resulting from a non–linear regression; correlation also exists between the variables. Figure B.1 displays the theoretical distributions possible in Models 1 and 2. The models are described in detail in Table 3.2.

Table 3.2: Description of theoretical models used in simulations.

Model 1, Four terminal nodes This is the model that was investigated in Nerini and Ghattas (2007). Covariates $X_1, \ldots, X_4$ are drawn from the $U(-1, 1)$ distribution. Samples for each observation are drawn from $N(0, 1)$ if $X_1, X_2 \geq 0.5$; $N(0, 1.5)$ if $X_1 < 0.5$ and $X_2 \geq 0.5$; $(2/5)N(1, 0.5) + (3/5)N(-1, 0.5)$ if $X_1 \geq 0.5$ and $X_2 < 0.5$; and $(3/5)N(1, 0.5) + (2/5)N(-1, 0.5)$ if $X_1, X_2 < 0.5$.

Model 2, Nine terminal nodes Covariates $X_1, \ldots, X_4$ are drawn from the $U(-1, 1)$ distribution. Samples for each observation are drawn from Gam$(3, 2)$ if $X_1, X_2 \leq 0.5$ (where Gam(a, b) is the Gamma distribution with shape a and scale b; χ^2_3 if $-0.5 < X_1 \leq 0.5$ and $X_2 \leq -0.5$ (where χ^2_d is the Chi–squared distribution with d degrees of freedom); Gam$(4, 3)$ if $X_1 > 0.5$ and $X_2 \leq -0.5$; χ^2_4 if $X_1 \leq -0.5$ and $-0.5 < X_2 \leq 0.5$; Weib$(4, 8)$ if $-0.5 < X_1, X_2 \leq 0.5$ (where Weib(a, b) is the Weibull distribution with shape a and scale b); Weib$(1.5, 3)$ if $X_1 > 0.5$ and $-0.5 < X_2 \leq 0.5$; $(1/5)$Gam$(4, 3) + (4/5)$Gam$(3, 2)$ if $X_1 \leq -0.5$ and $X_2 > 0.5$; Weib$(1.5, 4.5)$ if $-0.5 < X_1 \leq 0.5$ and $X_2 > 0.5$; and Gam$(4, 4/3)$ if $X_1, X_2 > 0.5$.

Model 3, Weibull mixture regression model Covariates $X_1, \ldots, X_7$ are drawn from the multivariate normal distribution with Cor$(X_i, X_j) = 0.5$ for $i \neq j$ and marginal distributions: $X_1 \sim N(60, 10)$; $X_2 \sim N(1.5, 0.25)$; $X_3, X_4 \sim N(0, 0.5)$; $X_5 \sim N(125, 10)$; and $X_6, X_7 \sim N(0, 1)$. 75% of the samples are drawn from $f_{3,a}(\cdot|X) \sim$ Weib(α_1, β_1) and 25% of the samples are drawn from $f_{3,b}(\cdot|X) \sim 0.25$Weib$(\alpha_1, \beta_2) + 0.75$Weib$(\alpha_2, \beta_3)$. $\alpha_1 = 2 + 0.5X_2 + \epsilon_1$; $\alpha_2 = 4 - 0.5X_2 + \epsilon_1$; $\beta_1 = 3X_1 + 10X_2 + \epsilon_2$; $\beta_2 = 1.5X_1 + 5X_2 + \epsilon_3$; $\beta_3 = 3X_1 + 0.5X_5 + \epsilon_2$. $\epsilon_1 \sim U(0, 1)$; $\epsilon_2 \sim N(0, 10)$; $\epsilon_3 \sim N(0, 5)$.

Model 4, Weibull mixture (nonlinear) regression model Covariates $X_1, \ldots, X_7$ are drawn from the multivariate normal distribution with Cor$(X_i, X_j) = 0.5$ for $i \neq j$ and marginal distributions: $X_1 \sim N(60, 10)$; $X_2 \sim N(1.5, 0.25)$; $X_3, X_4 \sim N(0, 0.5)$; $X_5 \sim N(125, 10)$; and $X_6, X_7 \sim N(0, 1)$. 75% of the samples are drawn from $f_{4,a}(\cdot|X) \sim$ Weib(α_1, β_1) and 25% of the samples are drawn from $f_{4,b}(\cdot|X) \sim 0.25$Weib$(\alpha_1, \beta_2) + 0.75$Weib$(\alpha_2, \beta_3)$. $\alpha_1 = 2 + 0.5X_2 I(X_2 < 1.65) - 0.25X_2 I(X_2 \geq 1.65) + \epsilon_1$; $\alpha_2 = 4 - 0.5X_2 + \epsilon_1$; $\beta_1 = 3X_1 + 10X_2 + \epsilon_2$; $\beta_2 = 1.5X_1 + 5X_2 + \epsilon_3$; $\beta_3 = 3X_1 + 0.5X_5 + \epsilon_2$. $\epsilon_1 \sim U(0, 1)$; $\epsilon_2 \sim N(0, 10)$; $\epsilon_3 \sim N(0, 5)$.

Average Kullback–Leibler divergence (KL_a) was calculated for each run of the simulation. We have chosen to compare KL_a between the observed and predicted densities using the training data in order to gauge if any overfitting was occurring; comparing KL_a between observed and predicted densities in the testing data set provides an indication of how well the method under each deviance extrapolates to independent data; and comparing KL_a between the *actual* and predicted densities in the testing set gives a measure of how close the method under each deviance is to the 'truth'.

Models 1 and 2 differ from the others in that we know how many terminal nodes each fitted FRT should have: four for Model 1, and nine for Model 2. We therefore measure the variability of the effect of each deviance by the root mean squared error (RMSE) of the number of terminal nodes

$$\text{RMSE} = \left[(R - \bar{R}_d)^2 + \text{Var}(R_d)\right]^{1/2}$$

where R is the actual number of terminal nodes, and R_d is a vector containing the observed number of terminal nodes for deviance $d = 1, 2, 3$ in each run of the simulation. Given that we know the number of terminal nodes for Models 1 and 2, the RMSE is an important statistic for these simulations. Whilst KL_a provides us with an idea of the predictive power of the methods, the RMSE provides us with an idea of how well the methods estimate the *known* structure of Models 1 and 2.

3.3.2 Results

The adoption of deviance $D_2(r)$ (Equation 3.3) led to smaller mean (sd) KL_a in almost all simulation models in all KL_a metrics (Table 3.1). Briefly, estimators that were derived from $D_2(r)$ out-performed the estimators that were derived from other measures, in those circumstances in which a clear preference was discernible.

We found no differences between the observed and predicted densities for either training data and testing data for Model 1 (Figure 3.3(a)). However, with an increase in the number of possible distributions from which an observation could be generated (Model 2), a clear difference can be seen when the number of observations is low ($M = 50$, Figure 3.3(b)) and a less discernible difference when $M = 200$. Increasing the complexity of the underlying model (from Model 1 to Model 2) has a large effect on KL_a between the *actual* and predicted densities, with deviance $D_2(r)$ performing the best

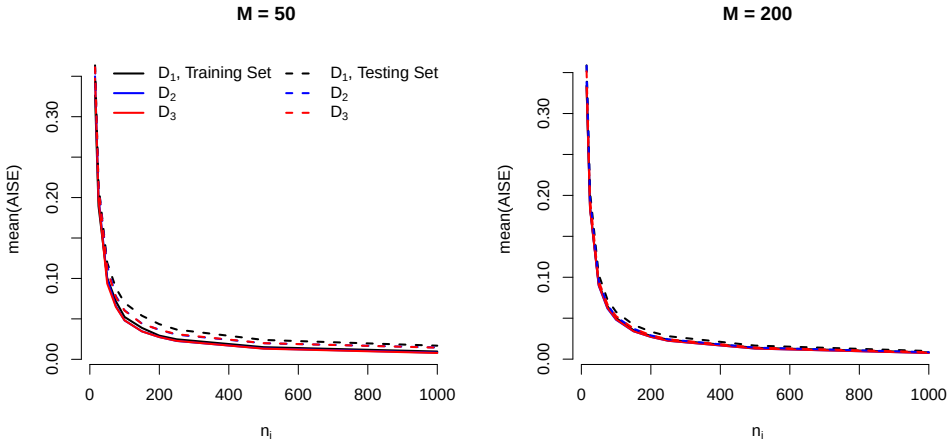
(Figure 3.4).

Table 3.3 provides a comparison of the RMSE for Models 1 and 2, where $M = 200$. As described previously, Model 1 should have four terminal nodes, and Model 2, nine. Deviance $D_2(r)$ generally outperforms the others in this regard.

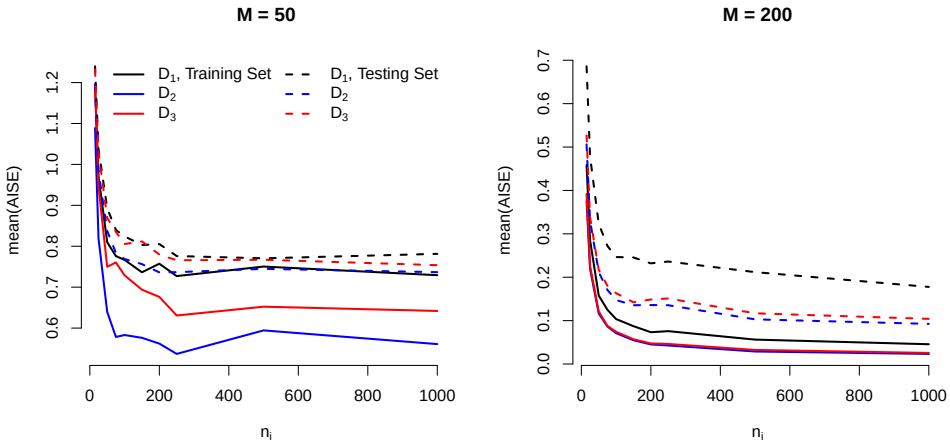
Table 3.3: Comparison of root mean squared error of the number of terminal nodes for Models 1 and 2 using deviances $D_1(r)$, $D_2(r)$ and $D_3(r)$; $M = 200$. n_i is the number of samples generated per observation.

	Model 1			Model 2		
n_i	$D_1(r)$	$D_2(r)$	$D_3(r)$	$D_1(r)$	$D_2(r)$	$D_3(r)$
15	1.40	2.06	1.61	7.35	1.10	1.37
25	2.84	1.05	0.83	8.02	1.11	1.13
50	3.81	0.58	0.63	8.45	1.52	2.08
75	4.09	0.39	0.56	9.06	1.47	2.32
100	4.63	0.14	0.36	8.80	1.82	2.56
150	6.17	0.00	0.10	9.51	1.82	2.72
200	6.99	0.00	0.10	9.29	1.71	2.54
250	6.71	0.00	0.00	11.46	2.24	2.91
500	5.87	0.00	0.00	11.45	2.00	3.56
1000	5.04	0.00	0.00	12.95	2.52	3.33

Allowing the parameters of the underlying distribution to vary via a regression process (Models 3 and 4) has little impact on KL_a between the observed and predicted densities (Figure 3.5). The effect of a non–linear generating process (Model 4) however, has an impact on KL_a between the *actual* and predicted densities, especially for low numbers of observations (Figure 3.6). As was observed in Table 3.3 for Models 1 and 2, deviance $D_2(r)$ appears to result in much more stable fitting of the FRT for Models 3 and 4, as evidenced by the low (and consistent) standard deviation of the number of terminal nodes (Figure 3.7).

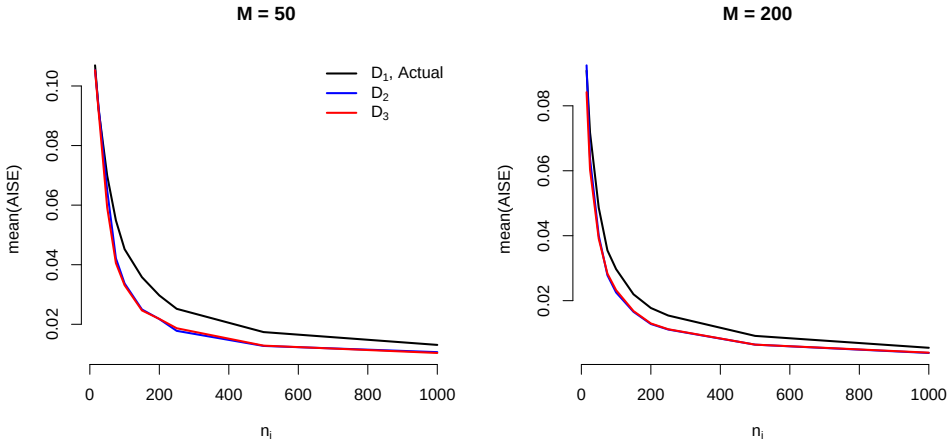


(a) Model 1

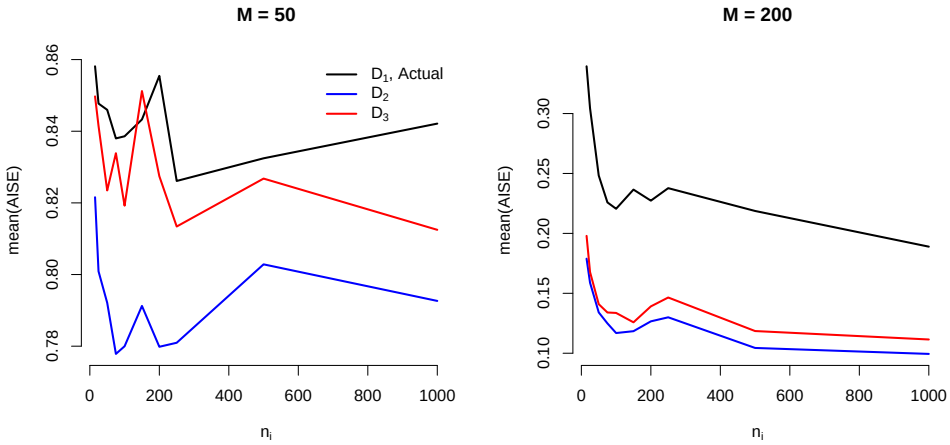


(b) Model 2

Figure 3.3: Mean KL_a between the observed and predicted densities for the training and testing data; Models 1 and 2.

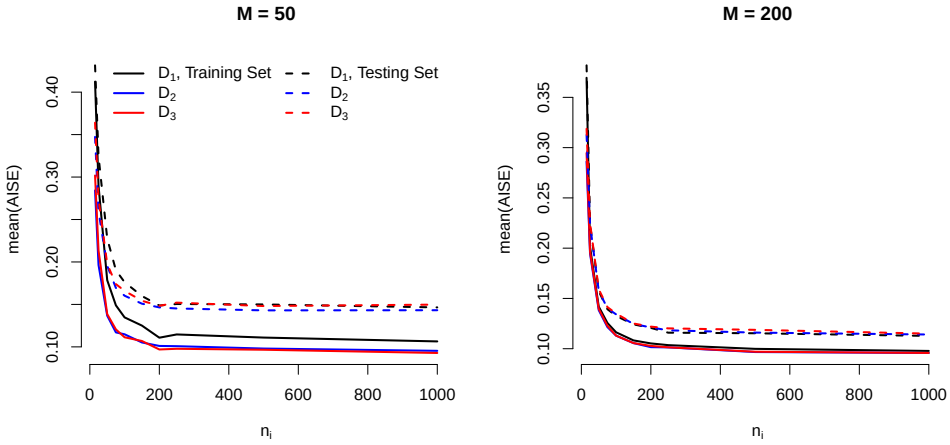


(a) Model 1

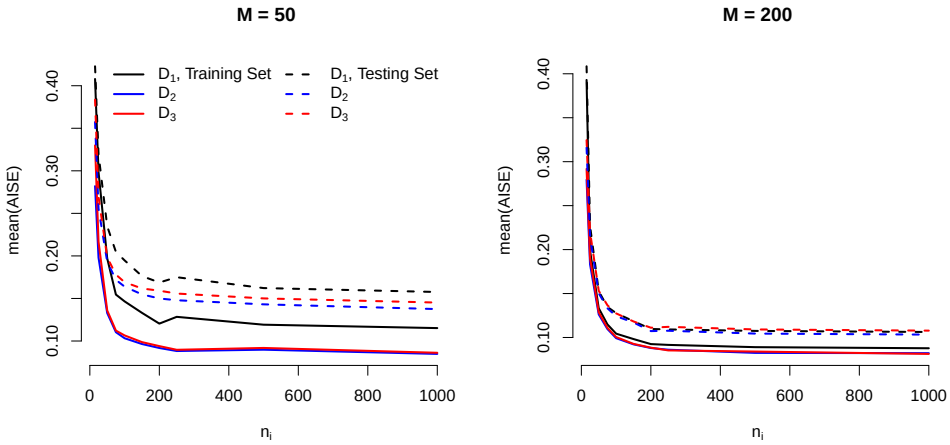


(b) Model 2

Figure 3.4: Mean KL_a between the *actual* and predicted densities for the testing data; Models 1 and 2.

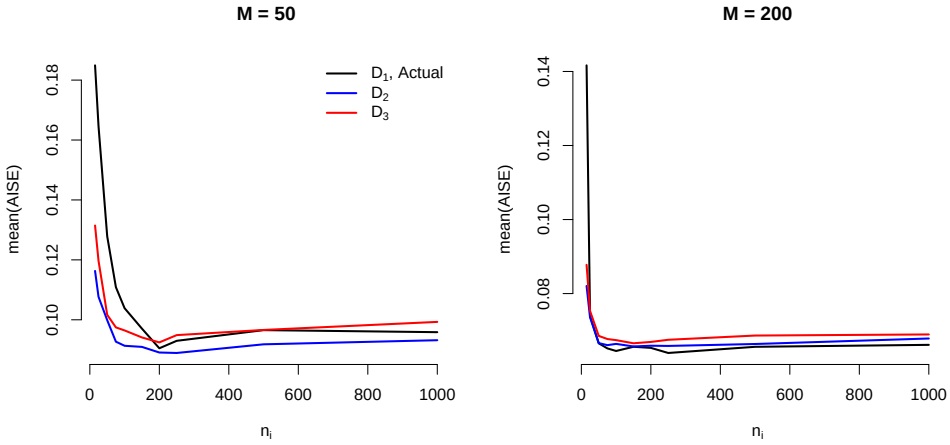


(a) Model 3

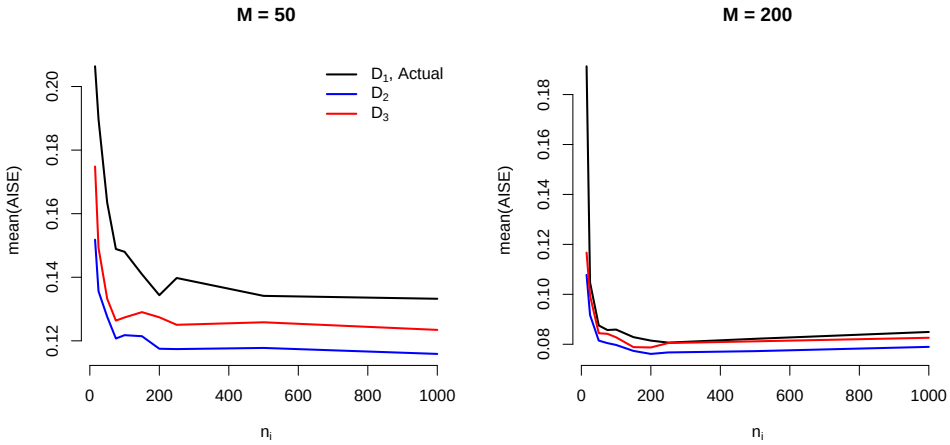


(b) Model 4

Figure 3.5: Mean KL_a between the observed and predicted densities for the training and testing data; Models 3 and 4.

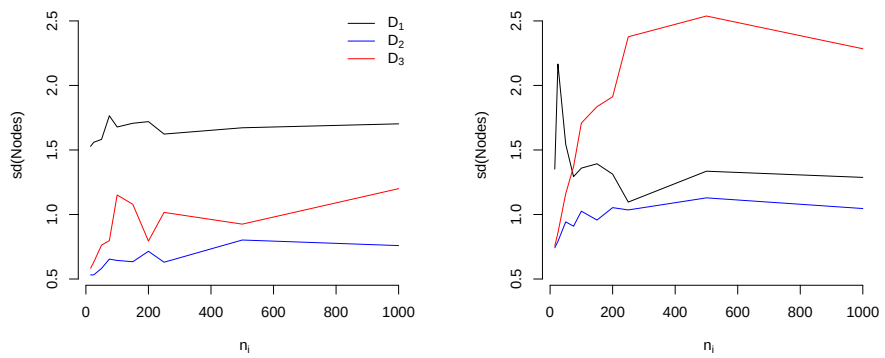


(a) Model 3



(b) Model 4

Figure 3.6: Mean KL_a between the *actual* and predicted densities for the testing data; Models 3 and 4.



(a) Model 3 (b) Model 4

Figure 3.7: Standard deviation of the number of terminal nodes from Models 3 and 4; $M = 200$.

3.3.3 Case study

We applied the FRT method with each deviance to predict tree diameter distributions. The data for this example come from the Intermountain Forest Tree Nutrition Cooperative (IFTNC) and are described in detail in Robinson (2004). Briefly, the data comprise tree diameter measurements (d, over bark at 1.3 m height) on stands of Douglas–fir (*P. menziesii*) in plots in single–aged stands in six regions in north–western USA. Stand basal area (G, m^2 ha^{-1}), density (SPH, stems ha^{-1}), height (H, m) and volume (V, m^3 ha^{-1}) were calculated from the tree–level data. Plot–level characteristics that were measured included soil nutrient levels for carbon, phosphorus, total nitrogen, and mineralised nitrogen; habitat class, soil type, and lithology were identified for each plot. 60 sample plots (10 from each region) were set aside (at random) to provide a test data set. Table 3.4 provides a descriptive summary of some key variables in the training data set, along with a comparison of the same variables in the testing set.

Table 3.4: Summary statistics for *P. menziesii* stand characteristics. The training data are based on 480 plots from 6 regions in north–western USA; the testing data are based on 60 plots from the same regions.

	Training data				Testing data			
Characteristic	min	mean	max	sd	min	mean	max	sd
G (m^2 ha^{-1})	7.26	32.65	76.98	10.45	12.71	32.49	65.83	11.61
SPH (stems ha^{-1})	210	662	2002	310	222	708	1705	329
V (m^3 ha^{-1})	28.18	262.70	753.00	117.75	60.16	254.30	580.80	116.70
H (m)	3.88	7.80	12.29	1.35	4.73	7.59	10.71	1.34

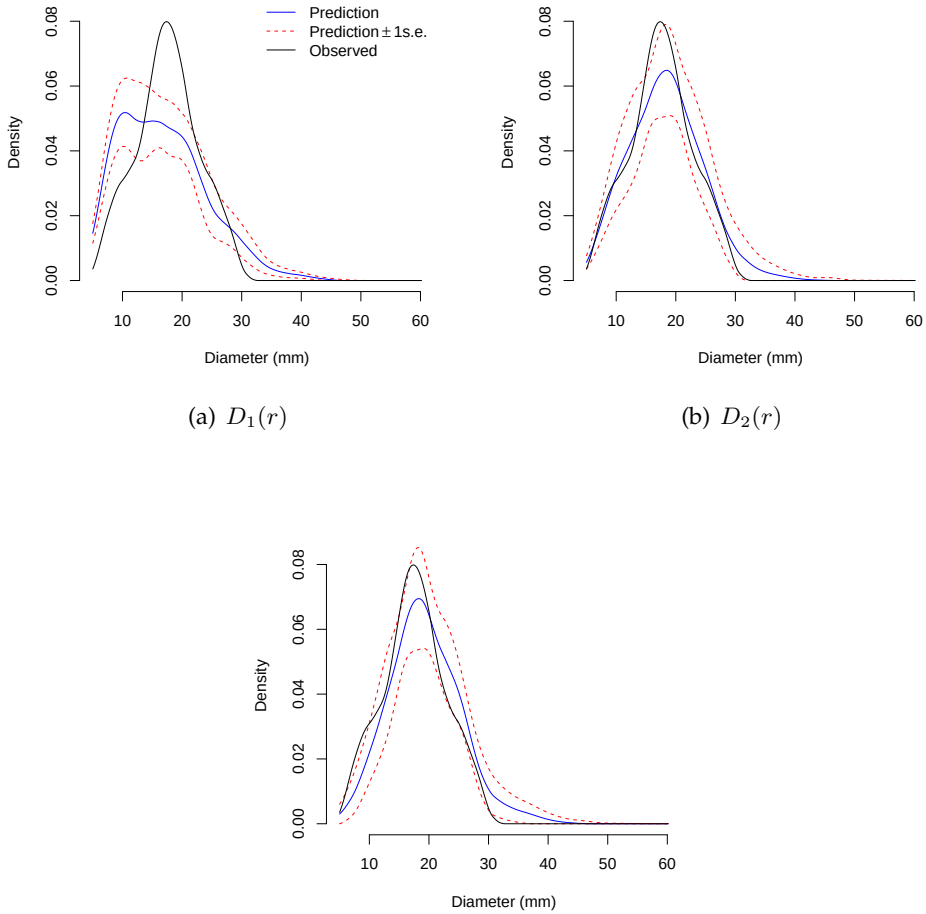
Each deviance was used to fit an FRT model to the IFTNC training data, with the fitted models then used to validate the testing data. Deviance $D_2(r)$ performed well in this example (Table 3.5), yet this was not decisive. Deviance $D_3(r)$ performed the best on the training data set, however when validated on the testing data, did not perform as well as deviance $D_2(r)$, suggesting overfitting had occurred.

We described in Section 3.2.3 a method for estimating standard error curves for predicted distributions. Figure 3.8 shows the predicted and observed diameter distribution for a sample plot from the *testing* data set, along with an estimate of the ± 1 standard error bounds of the PDF under each deviance. This figure confirms the results presented in Table 3.5, that deviance $D_1(r)$ is

Table 3.5: Comparison of KL_a between the training and testing data for *P. menziesii* using deviances $D_1(r)$, $D_2(r)$ and $D_3(r)$.

Training data			Testing data		
$D_1(r)$	$D_2(r)$	$D_3(r)$	$D_1(r)$	$D_2(r)$	$D_3(r)$
0.3316	0.3272	0.2798	0.4680	0.4128	0.4351

outperformed by deviances $D_2(r)$ and $D_3(r)$. The error bounds for $D_1(r)$ are seen to be quite tight, however they fail to include a large proportion of the actual density. To compare the order in which variables are split under each deviance, the fitted FRT structure was produced (Figure 3.9). This showed that under each deviance, height (H) is split first, at close to the same point in each: 7.693, 7.599, and 8.216 m for each deviance respectively. Compare this with the mean height $\bar{H} = 7.80$, and range $H \in (3.88, 12.29)$ m; we see that these split points partition the data into approximately equal cohorts of small and large trees. However, it is to be noted that further splits occur on different variables, so that the interpretation of the fitted FRT structure is vastly different under each deviance.



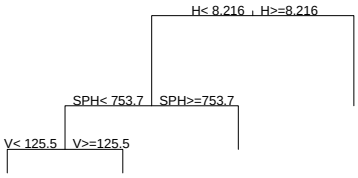
(a) $D_1(r)$

(b) $D_2(r)$

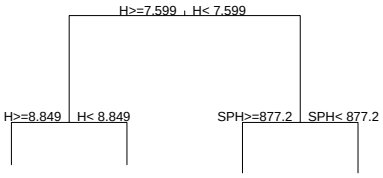
(c) $D_3(r)$

Figure 3.8: Comparison of predicted and observed diameter distributions for a sample plot from the testing data. a) Deviance $D_1(r)$, b) Deviance $D_2(r)$ and c) Deviance $D_3(r)$.

(a) $D_1(r)$



(b) $D_2(r)$



(c) $D_3(r)$

Figure 3.9: Graphical representation of the FRT for each deviance. Only the first three splits are shown.

3.4 Discussion

Restricting the underlying distribution of an FRT model to come from one of only four possible distributions has a large effect on the goodness–of–fit of the model. Model 1 generated data from four possible distributions, and the resulting mean KL_a between the observed and predicted densities (Figure 3.3(a)) and *actual* and predicted densities (Figure 3.4(a)) shows virtually no difference between the three deviances tested. However, increasing the possible distributions to nine (Model 2) has a large effect on the goodness–of–fit of the model. For a low number of observations ($M = 50$), deviance $D_2(r)$ is clearly the best performer, having a much smaller mean KL_a between both observed and predicted densities (Figure 3.3(b)) and *actual* and predicted densities (Figure 3.4(b)). For $M = 200$, the results are much closer for each deviance between observed and predicted densities, however $D_2(r)$ still outperforms the other two deviances when looking at the *actual* and predicted densities.

Letting the underlying distribution vary in its parameters (i.e. Models 3 and 4) did not produce as stark a contrast between the deviances as the fixed distributions (Model 1 and 2) when looking at the mean KL_a between observed and predicted densities: there is no way to pick a clear winner from these statistics. Figure 3.6(a) shows that deviance $D_2(r)$ performs well for $M = 50$ observations, yet there is virtually no difference for $M = 200$ observations. The non–linear parameter generating process within Model 4, however has a considerable effect. The gap between deviances is very clear for low numbers of observations ($M = 50$) with deviance $D_2(r)$ clearly the best, and whilst for a larger number of observations ($M = 200$) the results are closer, $D_2(r)$ still performs better than the other two deviances.

The good performance of $D_2(r)$ in these simulations is reinforced by the results given Table 3.3 and Figure 3.7. Table 3.3 shows that for Models 1 and 2, deviance $D_2(r)$ generally has the lowest RMSE, suggesting that it is not only better able to distinguish between the observed densities, but the size of the resulting FRT model is closest to the expected size (4 terminal nodes for Model 1; 9 terminal nodes for Model 2). The variation in the number of terminal nodes for the other models showed the same patterns.

A bonus of using resampling schemes to estimate the tree–size penalty (B.1), is that we can estimate (point–wise) standard errors for the predicted distributions from a FRT model (Section 3.2.3). Figure 3.8 displays this for one sample plot from the testing data set. Whilst deviance $D_1(r)$ has quite

tight error limits, these fail to include a fair proportion of the actual density. Deviances $D_2(r)$ and $D_3(r)$ do a much better job with most (if not all) the actual distribution lying within the error bounds.

The standard error bands that we have calculated in this paper have been calculated in a point–wise fashion, which does not take fully into account the functional nature of the predicted distributions, however we believe that they are sufficient for providing information about the predictive error of the FRT model. Nerini and Ghattas (2007) use functional principal components analysis (FPCA) on the observed values that make up each node in the FRT to display information on the variation within a node. As mentioned in Section 3.1, the deviance used by Nerini and Ghattas (2007) was focussed on homogeneity, thus the FPCA is only focussed on variation between observed curves. The standard error bands that we have calculated are based on bootstrap resamples, thus they provide an indication of predictive variation for an out–of–sample observation. We are not suggesting that either display of variation is preferred, rather that they complement each other.

An advantage of the FRT model over other more commonly used methods for modelling diameter distributions is that the graphical display of a FRT structure shows the hierarchy of splits, from the root node (all observations) to the terminal nodes (which give the predictions), providing at least an indication of variable importance. We found that the mean height (H) of trees on a plot is highly important in determining the shape of the diameter distribution, with each deviance splitting at almost the same level. Density (SPH), and volume (V) also appear in the FRT models, however differences are obvious at the lower level splits. This sort of variable importance is not readily available in the methods that are mostly used, however it must be taken as an indication only; there are many interactions at play over the whole FRT structure.

We investigated calculation of the curves within the recursive partitioning algorithm itself, i.e. calculating the predicted value for the node based on all individual observations in that node (as opposed to averaging observed curves), however this had no discernible effect on the structure of the FRT and increased computation time immensely, so was not investigated further.

The deviance used in the objective function (Equation (3.2)), could possibly be refined even further to promote qualities of the function that would be desirable in practice. As was discussed in Section 2.3, weighting could be applied prior to the fitting of the FRT, so that 'aberrant' observations would

have less impact on the results of the final FRT model. For the case study presented here, forest researchers could graphically inspect the observations, and those that didn't conform with prior expectations could be penalised when the dissimilarity matrix $\mathcal{A}$ (Section 3.2) is calculated. A procedure such as this could become problematic however, as with a large number of observations all pairwise dissimilarities a_{ij} would need to be adjusted. A possibility similar to the bagging approach discussed above, would be to combine the results from FRT models that were fit using a variety of deviances. Such an ensemble would then effectively 'smooth out' the effect of the choice of deviance. This would remove problems associated with the pairwise adjustments discussed previously, however would come with a computational cost.

In summary, the adjusted deviance that we have proposed ($D_2(r)$, Equation 3.3) provides an improvement over that suggested by Nerini and Ghattas (2007) which we demonstrated using simulated data that covered a wide array of situations that are commonly found in practice. The simulation results were also confirmed when we applied each deviance to an example data set. Nerini and Ghattas (2007) used functional PCA to identify variation around the predicted distribution in each node; we have highlighted estimated standard error curves as an alternative to FPCA, that result from the fitting procedure itself, so that no extra work is needed beyond the model fitting procedure.

Chapter 4

Longitudinal Functional Linear Modelling

We turn now to including a longitudinal aspect into the modelling of the diameter probability density functions (PDFs). This chapter extends the work of Yao et al. (2005b) to the case where the longitudinal response is now a series of functions. The method allows the prediction of functions at any time through nonparametric modelling of a sequence of covariance functions. To the best of our knowledge, this is the first time such a model has appeared in the literature to date. We will apply this model to the full longitudinally observed *E. globulus* data. This provides us with the means to predict the evolution of the diameter PDF in stands that have not had diameters directly measured, but have functional covariate information available. Comparing the results of this new approach to a more traditional approach, we find that the prediction error is reduced after allowing for the functional nature of the data.

In deriving the model introduced in this chapter, we will make use of basis decomposition techniques for sparse functional data (e.g. Yao et al., 2005a,b) along with nonparametric estimation of functional responses (Cardot, 2007). The combination of these methods will allow us to model what are basically short range time series of functions, conditional on longitudinally observed covariates.

Figure 4.1 provides an example of a typical set of tree diameter measurements within two sample stands from the *E. globulus* data introduced previously. In each of these stands, five sets of measurements have been taken over approximately ten years. Displayed in the figure are the individual tree

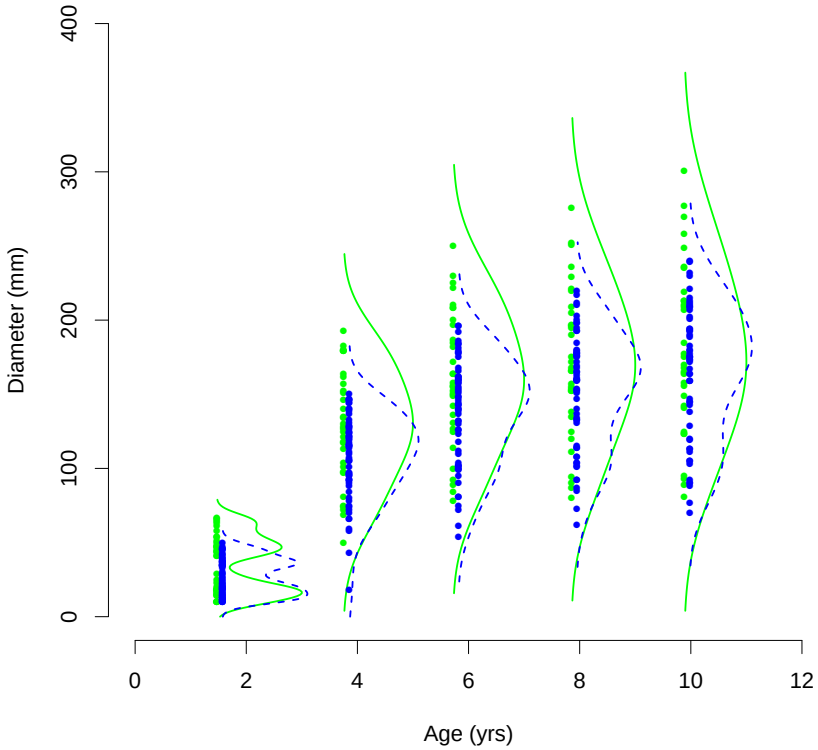


Figure 4.1: Tree diameter measurements in two stands, over time. Kernel density estimates (scaled) are overlaid, showing the change in shape and location over time.

diameters at each measurement age (y–axis) and the age at measurement (x–axis). Overlaid on the figure are kernel density estimates (which have been scaled to fit into the figure, yet retain their individual characteristics) of the diameter probability density functions at each measurement age. We can see that one stand (solid lines/dots) shows changes in spread and location over time, whilst the other (dashed lines/diamonds) also shows clear changes in shape. Predicting this behaviour over time is the focus of this chapter.

The stand attributes that are used in forest growth models are essentially growth curves within each stand, and thus have some unknown correlation structure. Models to predict the diameter PDF from covariate information will need to account for this correlation in the model. A common approach to dealing with this issue in the forestry literature is to first model the stand attribute using algebraic difference equations (ADE) (e.g. Wang and Baker,

2007; Clutter et al., 1983). These model a stand attribute at a fixed time, conditional on its value at a previous time, e.g. $X_2 = f(X_1, T_1, T_2)$ for stand attribute X measured at times T_1 and T_2. The function f is generally a parametric equation whose form is defined by expert opinion. As an example of the functional nature of these attributes, Figure 4.2 shows the evolution of basal area per hectare for each sample stand in the *E. globulus* data. The method we introduce in Section 4.1 will take into account the functional nature of such data in a nonparametric fashion, leading to an extremely flexible modelling approach.

We call such a model as that just described a longitudinal functional linear model (LFLM). This flexible approach will allow us to not only account for longitudinal variation in the diameter densities, but will also allow us to fully utilise the longitudinal information contained in the stand attribute, without requiring a specific parametric form to be assumed.

4.1 Longitudinal functional linear model

4.1.1 Basis representation of the regression model

Let $(X_i, Y_i), i = 1, \ldots, n$ denote the underlying (unobserved) sample pairs of square integrable functions X_i and surfaces Y_i. These pairs are realisations of smooth random functions $X(t)$ and surfaces $Y(d,t)$, with means $E[X(t)] = \eta(t)$ and $E[Y(d,t)] = \mu(d,t)$, where the arguments d and t are in some closed intervals $\mathcal{D}$ and $\mathcal{T}$. With t referencing time on the closed interval $\mathcal{T}$, the longitudinal functional linear model for the conditional expectation of Y is then

$$E[Y(d,t)|X] = \alpha(d,t) + \int_{\mathcal{T}} \beta(d,s,t)X(s)\,\mathrm{d}s \tag{4.1}$$

where the regression function $\beta(d,s,t)$ is a smooth, square integrable function for fixed s,t. After centering X by η, and with $E[Y(d,t)] = \mu(d,t) = \alpha(d,t) + \int_{\mathcal{T}} \beta(d,s,t)\eta(s)\,\mathrm{d}s$ (Yao et al., 2005b), Equation (4.1) becomes

$$E[Y(d,t)|X] = \mu(d,t) + \int_{\mathcal{T}} \beta(d,s,t)\left[X(s) - \eta(s)\right]\,\mathrm{d}s \tag{4.2}$$

Complicating the analysis is the problem that (X_i, Y_i) are not observable in the entirety; that is, we have sparse observations of the processes at times $T_{ij}, j = 1, \ldots, N_i$. From here on, we assume that the observation times T_{ij}

are themselves iid random variables T, on the closed interval $\mathcal{T}$. Thus, for a fixed time T_{ij}, observations are of the form $(X_i(T_{ij}), Y_i(d, T_{ij}))$. Keeping in line with the motivating example of predicting tree diameter density functions, we assume from here on, that for any time T, $Y(d,T)$ is a probability density function. That is, $Y(d,T) \geq 0,\ \forall d \in \mathcal{D}$, and $\int_{\mathcal{D}} Y(d,T)\,\mathrm{d}d = 1$. Further, assume that conditional on $T = t$, $Y(d,t)$ is square integrable and $E(||Y(d,T)||^2|T = t) < \infty$, and that $E\left[Y(d,T)|T = t\right] = \mu(d,t)$ is itself a probability density function. We can define the conditional covariance operator (Cardot, 2007), Γ_t as

$$\Gamma_t Y(d,t) = \int_{\mathcal{D}} \gamma(t,e,d)Y(e,t)\,\mathrm{d}e \tag{4.3}$$

where $(d,e) \in \mathcal{D} \times \mathcal{D}$ and $\gamma(t,e,d) = \text{Cov}\left[Y(d,T), Y(e,T)|T = t\right]$.

Now denote by $(\lambda_k(t), \phi_k(d,t))$ the k^{th} eigenvalue/eigenfunction pair of the conditional covariance operator Γ_t, where $\lambda_1(t) \geq \lambda_2(t) \geq \cdots \geq 0$ and the eigenfunctions are orthonormal (Cardot, 2007), then the conditional covariance function can be written as

$$\gamma(t,e,d) = \sum_{k=1}^{\infty} \lambda_k(t)\phi_k(e,t)\phi_k(d,t). \tag{4.4}$$

Then denoting the *random* principal coefficient functions by

$$b_k(t) = \int_{\mathcal{D}} \left[Y(d,t) - \mu(d,t)\right]\phi_k(d,t)$$

Y can be written as

$$Y(d,t) = \mu(d,t) + \sum_{k=1}^{\infty} b_k(t)\phi_k(d,t). \tag{4.5}$$

The goal of this chapter is to predict an unknown response density surface (i.e. $Y^*(d,t)$, where for any t, Y^* is a density) from *sparse* observations of a new predictor process X^*. To do this, we propose a two–stage functional principal components analysis in which β may be represented by a product combination of basis representations of Y and X, similar to that of the functional linear regression for longitudinal data proposed by Yao et al. (2005b). Estimation of the regression function β will obviously be the key to a useful model. The difficulty lies in the already noted problem that we have sparse (and possibly noisy) observations $(X_i(T_{ij}), Y_i(d, T_{ij}))$.

Consider now, the regression function $\beta(d,s,t)$ in (4.2), where for fixed $(s,t) \in \mathcal{T} \times \mathcal{T}$, $\beta(d,s,t)$ is a square integrable function in $L_2(\mathcal{D})$. Dealing as we are, in probability density functions, in the following we require that $\int_{\mathcal{D}} \int_{\mathcal{T}} \beta(d,s,t) \left[X(s) - \eta(s)\right] \mathrm{d}s \, \mathrm{d}d = 0$, so that $\int_{\mathcal{D}} E\left[Y(d,t)|X\right] = 1$. Under the assumption of square integrability for the responses, Y, the eigenfunction basis $\{\phi_k(d,t)\}_{k=1,\ldots,\infty}$ derived from the covariance operator, Γ_t (4.3), is a complete orthonormal basis for $L_2(D)$. Then for a given $(s,t) \in \mathcal{T} \times \mathcal{T}$

$$\beta(d,s,t) = \sum_{k=1}^{\infty} \beta_k(s,t)\phi_k(d,t) \tag{4.6}$$

for some coefficients $\beta_k(s,t)$. Then, after exchanging integration and summation (Lemma 1), the regression model (4.2) becomes

$$E\left[Y(d,t)|X\right] = \mu(d,t) + \sum_{k=1}^{\infty} \left\{ \int_{\mathcal{T}} \beta_k(s,t) \left[X(s) - \eta(s)\right] \mathrm{d}s \right\} \phi_k(d,t). \tag{4.7}$$

Alternatively, conditioning on X in Equation (4.5), we have the following representation of $E\left[Y(d,t)|X\right]$:

$$E\left[Y(d,t)|X\right] = \mu(d,t) + \sum_{k=1}^{\infty} E\left[b_k(t)|X\right] \phi_k(d,t) \tag{4.8}$$

4.1.2 Functional linear regression for fixed k

We will now fix k and investigate the coefficients of $\phi_k(d,t)$ in detail. Equating the coefficients of $\phi_k(d,t)$ in Equations (4.7) and (4.8), yields (for fixed k) a functional linear regression for longitudinal data (Yao et al., 2005b):

$$E\left[b_k(t)|X\right] = \int_{\mathcal{T}} \beta_k(s,t) \left[X(s) - \eta(s)\right] \mathrm{d}s \tag{4.9}$$

where $E\left[b_k(t)\right] \equiv 0, \forall t \in \mathcal{T}$, and justification for the exchange of integration and summation for Equation (4.9) is given by Lemma 1 and Appendix C.1.

Assume now that the principal coefficient functions $b_k(t)$ and the covariate function $X(t)$ are square integrable functions with covariances $\mathrm{Cov}\left(b_k(s), b_k(t)\right) = \gamma_k(s,t)$ and $\mathrm{Cov}\left(X(s), X(t)\right) = \gamma_X(s,t)$ respectively. Further, assume that both $\gamma_k(s,t)$ and $\gamma_X(s,t)$ have expansions of the type in Equation (4.4), so

that

$$\gamma_k(s,t) = \sum_{l=1}^{\infty} \theta_{kl}\varphi_{kl}(s)\varphi_{kl}(t) \tag{4.10}$$

$$\gamma_X(s,t) = \sum_{m=1}^{\infty} \rho_m\psi_m(s)\psi_m(t) \tag{4.11}$$

where $(\theta_{kl}, \varphi_{kl}(s))$ are the l^{th} eigenvalue/eigenfunction pair of the covariance operator $\Gamma_k b_k(t) = \int \gamma_k(s,t)b_k(s)\,\mathrm{d}s$, and $\theta_{k1} \geq \lambda_{k2} \geq \cdots \geq 0$. Similarly for the expansion of $\gamma_X(s,t)$ (Equation 4.11).

The regression coefficient $\beta_k(s,t)$ is then given by (He et al., 2000):

$$\beta_k(s,t) = \sum_{l=1}^{\infty}\sum_{m=1}^{\infty} \frac{E\left[\zeta_m\xi_{kl}\right]}{E\left[\zeta_m^2\right]}\psi_m(s)\varphi_{kl}(t) \tag{4.12}$$

where ξ_{kl} and ζ_m are the principal coefficients $\xi_{kl} = \int b_k(s)\varphi_{kl}(s)\,\mathrm{d}s$ and $\zeta_m = \int \left[X(s) - \eta(s)\right]\psi_m(s)\,\mathrm{d}s$ respectively.

Given a new functional predictor $X^*(s)$, a prediction for the conditional expectation $E\left[Y^*(d,t)|X^*\right]$ can be made by expanding X^* in terms of its eigenfunctions $\psi_m(s)$, and substituting into Equations (4.9) and (4.8):

$$\begin{aligned} E\left[b_k(t)|X^*\right] &= \int_{\mathcal{T}} \sum_{l=1}^{\infty}\sum_{m=1}^{\infty} \frac{E\left[\zeta_m\xi_{kl}\right]}{E\left[\zeta_m^2\right]}\psi_m(s)\varphi_{kl}(t)\left[X^*(s) - \eta(s)\right]\,\mathrm{d}s \\ &= \sum_{l=1}^{\infty}\sum_{m=1}^{\infty} \frac{E\left[\zeta_m\xi_{kl}\right]}{\rho_m}\zeta_m^*\varphi_{kl}(t) \\ E\left[Y^*(d,t)|X^*\right] &= \mu(d,t) + \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sum_{m=1}^{\infty} \frac{E\left[\zeta_m\xi_{kl}\right]}{\rho_m}\zeta_m^*\varphi_{kl}(t)\phi_k(d,t) \end{aligned} \tag{4.13}$$

where $\zeta_m^* = \int \left[X^*(s) - \eta(s)\right]\psi_m(s)\,\mathrm{d}s$ and $E\left[\zeta_m^2\right] = \rho_m$.

4.1.3 Estimating the components of the LFLM

We turn now to estimation of the relevant components of Equation (4.13). Section 4.1.1 discussed that we do not have complete information on the response density surface $Y(d,t)$ (which is essentially what we are trying to reconstruct), nor the covariate function $X(t)$. We do however, have observations of these functions at a small number of fixed time points for each i. For example, let t_{ij} denote the time of the j^{th} measurement of the i^{th} individual, $j = 1, \ldots, N_i$. The observation at time t_{ij} is then $(Y_i(d,t_{ij}), X_i(t_{ij}))$. We note

that the measurement times t_{ij} do not need to be the same for each i, nor do the number of measurements N_i.

The mean function $\mu(d,t)$ can be estimated by a functional kernel smoother (Cardot, 2007) over the observations $\{Y_i(d,t_{ij}),t_{ij}\}$:

$$\hat{\mu}(d,t) = \sum_{i=1}^{N}\sum_{j=1}^{N_i} W_{ij}(t)Y_i(d,t_{ij}), \text{ where} \tag{4.14}$$
$$W_{ij}(t) = \frac{K_{h_\mu}(t-t_{ij})}{\sum_{i=1}^{N}\sum_{j=1}^{N_i} K_{h_\mu}(t-t_{ij})}$$

and $K_h(u) = (1/h)K(u/h)$ for a positive kernel K with bandwidth $h_\mu = h_\mu(N) \to 0$ as $N \to \infty$. The bandwidth h_μ is selected via leave one group out cross–validation to allow for within–group correlation (Appendix C.2).

We next need to estimate the conditional functional principal components $\phi_k(d,t)$. Following Cardot (2007), let $Y_i \otimes Z_i(d,e) = Y_i(d)Z_i(e)$ for all $(d,e) \in \mathcal{D}\times\mathcal{D}$. The conditional covariance function $\gamma(t,e,d)$ (Equation 4.3) can then be estimated by the bivariate functional kernel smoother

$$\hat{\gamma}(t,e,d) = \sum_{i=1}^{N}\sum_{j=1}^{N_i} W_{ij}(t;h)\left\{Y_i(d,t_{ij}) - \hat{\mu}(d,t)\right\} \otimes \left\{Y_i(e,t_{ij}) - \hat{\mu}(e,t)\right\} \tag{4.15}$$

where the weight function W depends on all the time points and a positive bandwidth h (which may be selected by cross–validation, see Cardot, 2007).

For fixed t, the functional principal components basis can then be found as the spectral decomposition of the estimated covariance operator $\hat{\Gamma}_t$ (Equation 4.3), which results in the estimated eigenvalue/eigenfunction pairs $(\hat{\lambda}_k(t), \hat{\phi}_k(d,t)), k = 1,\ldots,\infty$. That is, the k^{th} eigenvalue/eigenfunction pair satisfy

$$\hat{\Gamma}_t\hat{\phi}_k(d,t) = \hat{\lambda}_k(t)\hat{\phi}_k(d,t) \tag{4.16}$$

The random coefficient functions $b_k(t)$ can now be estimated at times t_{ij} as

$$\hat{b}_k(t_{ij}) = \int_{\mathcal{D}} \left[Y_i(d,t_{ij}) - \hat{\mu}(d,t_{ij})\right]\hat{\phi}_k(d,t_{ij})$$

We now treat the coefficient functions at times t_{ij} above as observations from an unknown underlying function $b_k(t)$, and similarly the covariate func-

tions $X(s)$, and using the expansions in Equations (4.10) and (4.11) write

$$\begin{aligned} b_{k,ij} &= \sum_{l=1}^{\infty} \xi_{ikl}\varphi_{kl}(t_{ij}) \\ X_{ij} &= \eta(t_{ij}) + \sum_{m=1}^{\infty} \zeta_{im}\psi_m(t_{ij}) \end{aligned} \tag{4.17}$$

The estimation of the eigenvalues and eigenfunctions of the principal coefficient functions $b_k(t)$ and the covariate function $X(t)$ are found by estimating the eigenvalues and eigenfunctions of their respective covariance operators $\Gamma_k b_k(t)$ and $\Gamma_X X(t)$ (as in Equation 4.16). Following Yao et al. (2005a) we provide details for the estimation of $E[X(t)] = \eta(t)$ and $(\rho_m, \psi_m(t))$, the m^{th} eigenvalue/eigenvalue pair of $\Gamma_X X(t)$. The components of $b_k(t)$ are found in a similar way.

A locally–linear smoother (e.g. Fan and Gijbels, 1996) is used to estimate $\eta(t)$. Let

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{j=1}^{N_i} K_{h_\eta}(t_{ij} - t)\left[X_i(t_{ij}) - \beta_0 - \beta_1(t - t_{ij})\right]^2$$

for positive kernel K and bandwidth h_η (as in Equation 4.14). Then $\hat{\eta}(t) = \hat{\beta}_0(t)$ is the estimate of $\eta(t)$. The bandwidth h_η can again be chosen by leave one group out cross–validation (Appendix C.2).

For the eigenvalues and eigenfunctions, we need an estimate of the covariance surface $\operatorname{Cov}[X(s), X(t)] = \gamma_X(s,t)$. Let the 'observed' covariance be $C_i(t_{ij}, t_{ij'}) = (X_i(t_{ij}) - \hat{\eta}(t_{ij}))(X_i(t_{ij'}) - \hat{\eta}(t_{ij'}))$ for $j \neq j'$. Note that this means that only those groups i that have at least two observations can be used to estimate $\gamma_X(s,t)$. Let

$$\begin{aligned} \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right) = \underset{\beta_0,\beta_1,\beta_2}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{1\le j\neq j'\le N_i} & K_{h_X}(t_{ij} - s, t_{ij'} - t) \\ & \times \left[C_i(t_{ij}, t_{ij'}) - \beta_0 - \beta_1(t_{ij} - s) - \beta_2(t_{ij'} - t)\right]^2 \end{aligned} \tag{4.18}$$

where $K_{h_X}(s,t)$ is a positive two–dimensional kernel, for example, the product kernel $K_{h_X}(s,t) = (1/h_X^2)K(s/h_X)K(t/h_X)$. Then $\hat{\gamma}_X(s,t) = \hat{\beta}_0(s,t)$ is an estimate of $\gamma_X(s,t)$ and the estimates $(\hat{\rho}_m, \hat{\psi}_m(t))$ of $(\rho_m, \psi_m(t)), m \geq 1$ are found as in Equation (4.16).

Following Yao et al. (2005b), the expectation $E[\zeta_m \xi_{kl}]$ is again estimated as a locally–linear surface smoother. Let $C_i(t_{ij}, t_{ij'}) = (X_{ij} - \hat{\eta}(t_{ij}))(b_{k,ij'})$ be the 'observed' covariance at times $(t_{ij}, t_{ij'})$. Then a smoother similar to that in Equation (4.18) gives an estimate of the cross–covariance surface

$$C(s,t) = \sum_{l=1}^{\infty} \sum_{m=1}^{\infty} E[\zeta_m \xi_{kl}] \psi_m(s) \varphi_{kl}(t)$$

which leads to an estimate of $E[\zeta_m \xi_{kl}]$ as

$$\hat{\sigma}_{klm} = \int_{\mathcal{T}} \int_{\mathcal{T}} \hat{\psi}_m(s) \hat{C}(s,t) \hat{\varphi}_{kl}(t) \, \mathrm{d}s \, \mathrm{d}t$$

We are left now with estimating ζ_m^* in Equation (4.13). Due to the sparse nature of the observations X_{ij}, numerical approximation of the integral estimator $\zeta_m^* = \int [X^*(s) - \eta(s)] \psi_m \, \mathrm{d}s$ will perform poorly in practice. Yao et al. (2005a) use a conditioning argument to come up with a best prediction for ζ_m which they term Principal components Analysis through Conditional Expectation (PACE). The PACE estimate for ζ_m^* is the best linear predictor

$$\hat{\zeta}_m^* = \hat{\rho}_m \hat{\boldsymbol{\psi}}_m^{*T} \hat{\Sigma}_{X^*}^{-1} (\mathbf{X}^* - \hat{\boldsymbol{\eta}}) \tag{4.19}$$

where $\mathbf{X}^* = (X^*(t_1), \ldots, X^*(t_j))^T$ is the vector of observations of the new covariate process $X^*(s)$ at times $t_1, \ldots, t_j$, $\hat{\boldsymbol{\psi}}_m^* = (\hat{\psi}_m(t_1), \ldots, \hat{\psi}_m(t_j))^T$, and $\hat{\boldsymbol{\eta}} = (\hat{\eta}(t_1), \ldots, \hat{\eta}(t_j))^T$. The i, j^{th} entry of the matrix $\hat{\Sigma}_{X^*}$ is the covariance function for X evaluated times (t_i, t_j): $\hat{\gamma}_X(t_i, t_j)$. The best predictor (4.19) assumes normality of the principal component scores ζ_m^*, however Yao et al. (2005a) demonstrate its robustness against non–normality in a simulation study.

4.1.4 Number of included basis functions

In practice, the number of basis functions included in Equation (4.13) will require truncation at fixed values K, L, M. For a fixed k, the number of basis functions included in the functional regression coefficient (Equation 4.12) is chosen by an AIC type criterion (e.g Yao et al., 2005b,a) giving

$$\hat{\beta}_k(s,t) = \sum_{l=1}^{L} \sum_{m=1}^{M} \frac{\hat{\sigma}_{klm}}{\hat{\rho}_m} \hat{\psi}_m(s) \hat{\varphi}_{kl}(t) \tag{4.20}$$

For the number of components included in the regression coefficient (Equation 4.6), K, we suggest a regression sum of integrated squared errors approach. Assume that K is fixed and the respective basis functions and coefficients have been estimated (Section 4.1.3). Then the predicted value of the function $Y(d,t)$ given a new observation $X^*(s)$ is

$$\widehat{Y}_{KLM}(d,t) = \hat{\mu}(d,t) + \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} \frac{\hat{\sigma}_{klm}}{\hat{\rho}_m}\hat{\zeta}^*_m\hat{\varphi}_{kl}(t)\hat{\phi}_k(d,t) \tag{4.21}$$

where $\hat{\zeta}^*_m$ is the conditional FPC score as in Equation (4.19).

For observations $i = 1,\ldots,N$ at times $t_{ij}, j = 1,\ldots,N_i$ we calculate the regression sum of integrated squared errors, ISE_K as

$$\text{ISE}_K = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\int_{\mathcal{D}}\left[Y_i(d,t_{ij}) - \widehat{Y}_i(d,t_{ij})\right]^2 \mathrm{d}d \tag{4.22}$$

We choose the value of K which minimises ISE_K.

This approach decouples the choice of K, from that of L and M. Estimation of the bivariate functional kernel smoother for the conditional covariance $\gamma(t,e,d)$, depends on the bandwidth h in Equation (4.15), which is optimised for a fixed basis dimension, K. Calculating the optimal bandwidth for the bivariate functional kernel smoother (Cardot, 2007) is computationally expensive. Including the optimal choice of K within the whole LFLM would result in a significant increase in computational time.

4.2 Inference

4.2.1 Significance testing

Section 4.1 describes the model and estimation of the relevant components for predicting density functions at a fixed time t, but just as in standard multivariate analysis, we may question whether the functional predictor $X(s)$ actually has an effect on the response $Y(d,t)$. In other words we wish to test

the null hypothesis of no effect of X:

$$H_0 : E\left[Y(d,t)|X\right] = \mu(d,t)$$

vs.

$$H_1 : E\left[Y(d,t)|X\right] = \mu(d,t) + \int_{\mathcal{T}} \beta(d,s,t)\left[X(s) - \eta(s)\right] \mathrm{d}s$$

In order to test the no effect hypothesis, we will follow Cardot et al. (2007) and define an F-statistic in the following way. Let $\hat{\mu}(d,t)$ be the estimator of the mean of $Y(d,t)$ and $\widehat{Y}(d,t)$ be the predicted value of the density Y for a given X, i.e. $\hat{Y}(d,t) = \widehat{E}\left[Y(d,t)|X\right]$ (Equation 4.21). Define

$$\mathrm{RSS}_0 = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\int_{\mathcal{D}} \left[Y_i(d,t_{ij}) - \hat{\mu}(d,t_{ij})\right]^2 \mathrm{d}d$$

$$\mathrm{RSS}_1 = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\int_{\mathcal{D}} \left[Y_i(d,t_{ij}) - \widehat{Y}(d,t_{ij})\right]^2 \mathrm{d}d$$

then

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1} \tag{4.23}$$

Because of complex form of F, the decision on whether to reject H_0 will be based on a permutation p-value. Denote F_{obs} and F_b to be the values of the F-statistic (4.23) for the observed sample, and a sample with the predictor functions randomly permuted, then the permutation p-value is defined to be

$$p = \frac{1}{B}\sum_{b=1}^{B} I\left[F_b > F_{\text{obs}}\right] \tag{4.24}$$

The null hypothesis of no effect is thus rejected for small values of p.

4.2.2 Asymptotic pointwise confidence intervals

Pointwise confidence intervals for fixed t can be calculated for $E\left[Y^*(d,t)|X^*\right]$ by extending the arguments of Yao et al. (2005b). Conditional on the sparse measurements of $X^*(t)$ (X_{ij} in Equation 4.17), $(\tilde{\boldsymbol{\zeta}}^*_M - \boldsymbol{\zeta}^*_M) \sim N(0, \Omega_M)$, where $\tilde{\boldsymbol{\zeta}}^*_M = (\tilde{\zeta}^*_1, \ldots, \tilde{\zeta}^*_M)^T$ (Yao et al., 2005b). Now fix d, t and let $\boldsymbol{\varphi}_{kL} = (\varphi_{k1}(t), \ldots, \varphi_{kL}(t))^T$ for $k \geq 1$. For $l = 1, \ldots, L$ and $m = 1, \ldots, M$, let P_{kLM} be the $L \times M$ ma-

trix with (l, m) entry σ_{klm}/ρ_m and let Λ_{KM} be the $K \times M$ matrix with rows $\boldsymbol{\varphi}_{kL}^T P_{kLM}, k = 1, \ldots, K$. Let $\boldsymbol{\phi}_{dK} = (\phi_1(d,t), \ldots, \phi_K(d,t))^T$. Theorem 3 establishes the asymptotic distribution of $(\widehat{Y}^*_{KLM}(d,t) - E\left[Y^*(d,t)|X^*\right])$ as a mean 0 normal distribution with variance $\hat{\boldsymbol{\phi}}_{dK}^T \hat{\Lambda}_{KM} \hat{\Omega}_M \hat{\Lambda}_{KM}^T \hat{\boldsymbol{\phi}}_{dK}$. Thus, $(1-\alpha)\%$ confidence intervals can be constructed by

$$\widehat{Y}^*_{KLM}(d,t) \pm \Phi^{-1}(1-\alpha/2)\sqrt{\hat{\boldsymbol{\phi}}_{dK}^T \hat{\Lambda}_{KM} \hat{\Omega}_M \hat{\Lambda}_{KM}^T \hat{\boldsymbol{\phi}}_{dK}} \tag{4.25}$$

4.3 Asymptotic properties

In this section we show that the regression function, Equation (4.20) is consistent (Theorem 1), and that the prediction $\widehat{Y}^*_{KLM}(d,t)$ (from sparse measurements of the new functional covariate $X^*(s)$) is consistent for $\widetilde{Y}^*(d,t)$ (Theorem 2). Asymptotic normality of the prediction is established in Theorem 3. Results in this section build on the results of Yao et al. (2005a), Yao et al. (2005b) and Cardot (2007). In particular, we assume that the marginal and joint densities of the measurement times and observations are square integrable, as are the functional response and observations, $Y(d,T)$ and $X(s)$. We refer to for Yao et al. (2005a), Yao et al. (2005b) and Cardot (2007) the more technical assumptions needed in the proofs. Some extensions to the assumptions are also given in Appendix C.4. We begin by demonstrating the convergence of the right hand side of Equation (4.6), required for the exchange of integration and summation in Equation (4.7), and in Theorem 1. We require the following:

A1 $\sum_{k,l,m\geq 1}^{\infty} \sigma_{klm}^2/\rho_m^2 < \infty$, where $\sigma_{klm} = E\left[\zeta_m \xi_{kl}\right]$ and $\rho_m = E\left[\zeta_m^2\right]$.

A2 $\gamma(d,s,t) = \sum_{k,l,m\geq 1}^{\infty} \left|\sigma_{klm}\psi_m(s)\varphi_{kl}(t)\phi_k(d,t)\right|/\rho_m$ is continuous for d, s, t and $\beta_{KLM} = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} \sigma_{klm}\psi_m(s)\varphi_{kl}(t)\phi_k(d,t)/\rho_m$ absolutely converges to $\beta(d,s,t)$ for all $d \in \mathcal{D}$ and $s, t \in \mathcal{T}$ as $K, L, M \to \infty$

Lemma 1. (Yao et al., 2005a) Under (A1) the regression function (Equation 4.6) $\sum_{k,l,m\geq 1} \sigma_{klm}\psi_m(s)\varphi_{kl}(t)\phi_k(d,t)/\rho_m$ converges to $\beta(d,s,t)$. If, in addition, (A2) holds, the convergence is uniform on $\mathcal{D} \times \mathcal{T} \times \mathcal{T}$.

Proof. For the first part, let $\beta_{KLM}(d,s,t) = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} \sigma_{klm}\psi_m(s)\varphi_{kl}(t)\phi_k(d,t)/\rho_m$. Observing the orthonormality of $\{\psi_m(s)\}_{m\geq 1}$, $\{\varphi_{kl}(t)\}_{l\geq 1}$, and $\{\phi_k(d,t)\}_{k\geq 1}$, $\int_{\mathcal{D}}\int_{\mathcal{T}}\int_{\mathcal{T}} \beta_{KLM}(d,s,t)^2 = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} \sigma_{klm}^2/\rho_m^2$. Then letting $\beta = \lim_{K,L,M\to\infty} \beta_{KLM}$, $\int_{\mathcal{D}}\int_{\mathcal{T}}\int_{\mathcal{T}} \left[\beta_{KLM}(d,s,t) - \beta(d,s,t)\right]^2 \,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}d \to 0$.

For the second part, let $\gamma_{KLM}(d,s,t) = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M}|\sigma_{klm}\psi_m(s)\varphi_{kl}(t)\phi_k(d,t)|/\rho_m$. Then from (A2), the sequence $\{\gamma_{KLM}(d,s,t)\}$ is a non–decreasing sequence converging to $\gamma(d,s,t)$. Applying Dini's theorem (e.g. Rudin, 1976) gives uniform convergence on $\mathcal{D}\times\mathcal{T}\times\mathcal{T}$ and thus $\beta_{KLM}(d,s,t)$ converges uniformly to $\beta(d,s,t)$. □

Theorem 1. (Yao et al., 2005b)

$$\lim_{n\to\infty}\int_{\mathcal{D}}\int_{\mathcal{T}}\int_{\mathcal{T}}\left[\hat{\beta}(d,s,t)-\beta(d,s,t)\right]^2\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}d = 0 \text{ in probability.}$$

Proof. Observing the orthonormality of the eigenfunction bases in the following we have,

$$\begin{aligned}
&\int_{\mathcal{D}}\int_{\mathcal{T}}\int_{\mathcal{T}}\left[\hat{\beta}(d,s,t)-\beta(d,s,t)\right]^2\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}d\\
&\quad=\int_{\mathcal{D}}\int_{\mathcal{T}}\int_{\mathcal{T}}\left[\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M}\frac{\hat{\sigma}_{klm}}{\hat{\rho}_m}\hat{\psi}_m(s)\hat{\varphi}_{kl}(t)\hat{\phi}_k(d,t)\right.\\
&\qquad\left.-\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sum_{m=1}^{\infty}\frac{\sigma_{klm}}{\rho_m}\psi_m(s)\varphi_{kl}(t)\phi_k(d,t)\right]^2\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}d\\
&\quad=\int_{\mathcal{D}}\int_{\mathcal{T}}\int_{\mathcal{T}}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M}\left[\frac{\hat{\sigma}_{klm}}{\hat{\rho}_m}\hat{\psi}_m(s)\hat{\varphi}_{kl}(t)\hat{\phi}_k(d,t)-\right.\\
&\qquad\left.\frac{\sigma_{klm}}{\rho_m}\psi_m(s)\varphi_{kl}(t)\phi_k(d,t)\right]^2\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}d\\
&\qquad+\sum_{k=K+1}^{\infty}\sum_{l=L+1}^{\infty}\sum_{m=M+1}^{\infty}\frac{\sigma_{klm}^2}{\rho_m^2}\\
&\qquad+\int_{\mathcal{D}}\int_{\mathcal{T}}\int_{\mathcal{T}}\left[\sum_{k=K+1}^{\infty}\sum_{l=L+1}^{\infty}\sum_{m=M+1}^{\infty}\frac{\sigma_{klm}}{\rho_m}\psi_m(s)\varphi_{kl}(t)\phi_k(d,t)\right]\\
&\qquad\quad\times\left\{\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M}\left[\frac{\hat{\sigma}_{klm}}{\hat{\rho}_m}\hat{\psi}_m(s)\hat{\varphi}_{kl}(t)\hat{\phi}_k(d,t)\right]\right\}\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}d\\
&\quad=A(n)+B(n)+C(n).
\end{aligned}$$

Now by (A1), $B(n)\to 0$ as $n\to\infty$. Slutsky's theorem, the results in Yao et al.

(2005b), (Cardot, 2007) and Appendix C.4, imply that

$$\begin{aligned} A(n) = O_p \Bigg(& \sum_{m=1}^{M} \frac{\delta_m^X A_{\delta_m^X}}{\sqrt{n}h_X^2 - A_{\delta_m^X}} + \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{\delta_{kl}^{b_k} A_{\delta_{kl}^{b_k}}}{\sqrt{n}h_k^2 - A_{\delta_{kl}^{b_k}}} \\ & + \sum_{l=1}^{L} \frac{KM}{\sqrt{n}h_{l1}h_2} + \left[h_{Y_1}^{\beta} + h_{Y_2}^{\alpha} + \left\{ \frac{\log n}{n \min(h_{Y_1}, h_{Y_2})} \right\}^{1/2} \right] \sum_{l=1}^{L} \kappa_l \Bigg) \\ & \xrightarrow{p} 0 \end{aligned}$$

as $n \to \infty$. For $C(n)$, note that by application of the Cauchy–Schwarz inequality, $C(n)^2 \leq A(n) \times B(n)$, and thus $C(n) \xrightarrow{p} 0$ as $n \to \infty$. Combining $A(n)$, $B(n)$ and $C(n)$ completes the proof. □

Recall Equation (4.13) for a new observation of $X^*(t)$, and denote the target prediction (for sparsely observed $X^*(t)$) as $\widetilde{Y}^*(d,t)$. Now let the true conditional FPC score be $\tilde{\zeta}_m^* = \rho_m \boldsymbol{\psi}_m^{*T} \Sigma_{X^*}^{-1} (\mathbf{X}^* - \boldsymbol{\eta})$ where the components are as in Equation (4.19). Then

$$\widetilde{Y}^*(d,t) = \mu(d,t) + \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sum_{m=1}^{\infty} \frac{\sigma_{klm}}{\rho_m} \tilde{\zeta}_m^* \varphi_{kl}(t) \phi_k(d,t).$$

Further, denote by $Y_{KLM}^*(d,t)$ and $\widetilde{Y}_{KLM}^*(d,t)$, finite versions of $E\left[Y^*(d,t)|X^*\right]$ and $\widetilde{Y}^*(d,t)$. We make the following assumptions

A3 $\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sum_{m=1}^{\infty} \sigma_{klm}^2 / \left(\lambda_k(t)\rho_m\right) < \infty$ for fixed t.

A4 The number and locations of the measurements of $X^*(s)$ remain fixed as $n \to \infty$

Then we have the following

Lemma 2. Under assumptions **A3** and **A4**, and as $K, L, M \to \infty$

$$\sup_{d \in \mathcal{D}} E\left[Y_{KLM}^*(d,t) - E\left[Y^*(d,t)|X^*\right]\right]^2 \to 0 \tag{4.26}$$

$$\sup_{d \in \mathcal{D}} E\left[\widetilde{Y}_{KLM}^*(d,t) - \widetilde{Y}^*(d,t)\right]^2 \to 0 \tag{4.27}$$

Proof. In Equation (4.26), note

$$\begin{aligned}
&E\left[Y^*_{KLM}(d,t) - E\left[Y^*(d,t)|X^*\right]\right]^2 \\
&\qquad = E\left[\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M}\frac{\sigma_{klm}}{\rho_m}\zeta^*_m\varphi_{kl}(t)\phi_k(d,t)\right. \\
&\qquad\quad \left. - \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sum_{m=1}^{\infty}\frac{\sigma_{klm}}{\rho_m}\zeta^*_m\varphi_{kl}(t)\phi_k(d,t)\right]^2 \\
&\qquad = \sum_{k=K+1}^{\infty}\sum_{l=L+1}^{\infty}\sum_{m=M+1}^{\infty}\frac{\sigma^2_{klm}}{\rho^2_m}E(\zeta^{*2}_m)\left[\varphi_{kl}(t)\phi_k(d,t)\right]^2 \\
&\qquad = \sum_{k=K+1}^{\infty}\sum_{l=L+1}^{\infty}\sum_{m=M+1}^{\infty}\frac{\sigma^2_{klm}}{\rho_m}\left[\varphi_{kl}(t)\phi_k(d,t)\right]^2
\end{aligned}$$

Now, taking the supremum of the LHS of Equation (4.26) gives

$$\begin{aligned}
&\sup_{d\in\mathcal{D}} E\left[Y^*_{KLM}(d,t) - E\left[Y^*(d,t)|X^*\right]\right]^2 \\
&\qquad = \sum_{k=K+1}^{\infty}\sum_{l=L+1}^{\infty}\sum_{m=M+1}^{\infty}\frac{\sigma^2_{klm}}{\rho_m}\varphi^2_{kl}(t)\sup_d\left[\phi_k(d,t)\right]^2 \\
&\qquad = \sum_{k=K+1}^{\infty}\sum_{l=L+1}^{\infty}\sum_{m=M+1}^{\infty}\frac{\sigma^2_{klm}}{\rho_m\lambda_k(t)}\varphi^2_{kl}(t)\sup_d\left\{\lambda_k(t)\phi^2_k(d,t)\right\}
\end{aligned}$$

Now by the Karhunen-Loève theorem (Loève, 1977), $\sum_{k=1}^{\infty}\lambda_k(t)\phi_k(d,t)\phi_k(e,t)$ converges uniformly in $d, e \in \mathcal{D}$, and thus $\sup_d \lambda_k(t)\phi^2_k(d,t) \to 0$ as $k \to \infty$ and assuming **A4** holds, Equation (4.26) follows. The proof for Equation (4.27) is similar. □

Theorem 2. (Yao et al., 2005b).

Under the conditions of Yao et al. (2005b), Cardot (2007) and **A1–A4**

$$\lim_{n\to\infty}\widehat{Y}_{KLM}(d,t) = \widetilde{Y}^*(d,t)$$

in probability.

Proof. Note that

$$\left|\widehat{Y}^*_{KLM}(d,t) - \widetilde{Y}^*(d,t)\right| \leq \left|\widehat{Y}^*_{KLM}(d,t) - \widetilde{Y}^*_{KLM}(d,t)\right| + \left|\widetilde{Y}^*_{KLM}(d,t) - \widetilde{Y}^*(d,t)\right| \tag{4.28}$$

$$= \text{P1} + \text{P2}$$

Lemma 2 shows that for P2, $\widetilde{Y}^*_{KLM}(d,t) \xrightarrow{p} \widetilde{Y}^*(d,t)$ as $K, L, M, n \to \infty$. From Cardot (2007), we have

$$\sup_d \left|\hat{\mu}(d,t) - \mu(d,t)\right| = O_p(h_{Y_1}^{\beta}) + O_p\left[\left(\frac{\log n}{n h_{Y_1}}\right)^{1/2}\right]$$

as $n \to \infty$, and from Yao et al. (2005b), Lemma (A.1)

$$\left|\hat{\zeta}^*_m - \zeta^*_m\right| = O_p\left(\frac{\delta_m^X A_{\delta_m^X}}{\sqrt{n}h_X^2 - A_{\delta_m^X}}\right)$$

combining these results and by Slutsky's theorem, for P1 we have

$$\left|\widehat{Y}^*_{KLM}(d,t) - \widetilde{Y}^*_{KLM}(d,t)\right| \xrightarrow{p} 0$$

as $n \to \infty$. A further application of Slutsky's theorem for P1 + P2 gives the required result. □

For the next theorem, we make the following assumptions

A5 For all $i = 1, \ldots, n$, $m \geq 1$ and $j = 1, \ldots, N_i$, the FPC scores ζ_{im} are jointly Gaussian

A6 There exists a continuous, positive definite function $\omega(d,e,t)$ such that $\omega_{KLM}(d,e,t) \to \omega(d,e,t)$ as $K, L, M \to \infty$

where $\omega_{KLM}(d,e,t) = \boldsymbol{\phi}_{dK}^T \Lambda_{KM} \Omega_M \Lambda_{KM}^T \boldsymbol{\phi}_{dK}$ (as defined in Section 4.2.2). $\omega_{KLM}(d,e,t)$ is a sequence of continuous positive definite functions, and $\hat{\omega}_{KLM}(d,e,t)$ is an estimate of $\omega_{KLM}(d,e,t)$.

Theorem 3. (Yao et al., 2005b).
Under assumptions **A3–A6**, Lemma (A.1) and assumption B5 of Yao et al. (2005b), for t fixed and all $d \in \mathcal{D}, x \in \mathbb{R}$

$$\lim_{n\to\infty} P\left\{\frac{\widehat{Y}^*_{KLM}(d,t) - E\left[Y^*(d,t)|X^*\right]}{\sqrt{\hat{\omega}_{KLM}(d,d,t)}} \leq x\right\} = \Phi(x)$$

where $\Phi(x)$ is the CDF of the standard normal distribution.

Proof. Note that for fixed K, L, M

$$\widehat{Y}^*_{KLM}(d,t) - Y^*_{KLM}(d,t) = \widehat{Y}^*_{KLM}(d,t) - \widetilde{Y}^*_{KLM}(d,t) + \widetilde{Y}^*_{KLM}(d,t) - Y^*_{KLM}(d,t).$$

We have from the proof of Theorem 2 that $\widehat{Y}^*_{KLM}(d,t) \xrightarrow{p} \widetilde{Y}^*_{KLM}(d,t)$ (P1 in Equation 4.28). Then using the result that $(\tilde{\boldsymbol{\zeta}}^*_M - \boldsymbol{\zeta}^*_M) \sim N(0, \Omega_M)$ (Yao et al., 2005b) we have that

$$\left\{\widehat{Y}^*_{KLM}(d,t) - Y^*_{KLM}(d,t)\right\} \xrightarrow{D} Z_{KLM} \sim N(0, \omega_{KLM}(d,d,t))$$

Under **A6**, we have that $Z_{KLM} \to Z \sim N(0, \omega(d,d,t))$ as $K, L, M \to \infty$. Noting further that

$$\widehat{Y}^*_{KLM}(d,t) - E\left[Y^*(d,t)|X^*\right] = \widehat{Y}^*_{KLM}(d,t) - Y^*_{KLM}(d,t) + Y^*_{KLM}(d,t) - E\left[Y^*(d,t)|X^*\right]$$

and from the Karhunen-Loève theorem, $|Y^*_{KLM}(d,t) - E\left[Y^*(d,t)|X^*\right]| \xrightarrow{p} 0$. Thus

$$\lim_{K,L,M\to\infty} \lim_{n\to\infty} \widehat{Y}^*_{KLM}(d,t) - E\left[Y^*(d,t)|X^*\right] \stackrel{D}{=} Z$$

Now, the convergence of each of the various parts of $\hat{\omega}_{KLM}(d,e,t)$ implies that $\hat{\omega}_{KLM}(d,d,t) \xrightarrow{p} \omega(d,d,t)$. The result follows by application of Slutsky's theorem. □

4.4 Application

4.4.1 Data and comparative methods

For this application section, we use the *E. globulus* data that was introduced in Chapter 1. We briefly review the main aspects of the data as they pertain to the LFLM. The experiment comprised of 90 treatment stands, and in each stand, tree diameter was measured at approximately two–yearly intervals. Not every stand was measured the same number of times: some stands were measured four times, whilst others were measured at seven different times, resulting in a total of 503 measurements. Table 4.1 provides a breakdown of

this measurement schedule. In each stand, tree diameter was measured, and from these, stand–level characteristics were calculated. Using basal area per hectare (m^2 ha^{-1}) as the functional growth covariate, the method described in Section 4.1 is used to predict the probability density function of diameters within stands.

Denote the measured tree diameters in stand $i = 1, \ldots, N$ at time $t_{ij}, j = 1, \ldots, N_i$ by D_{ijp} for $p = 1, \ldots, P_i$ where P_i is the number of trees measured in stand i. As we do not have observations on the true densities, we take as given the response densities $Y_i(d, t_{ij})$ to be the kernel density estimates

$$Y_i(d, t_{ij}) = \frac{1}{P_i h_{ij}} \sum_{p=1}^{P_i} K\left(\frac{d - D_{ijp}}{h_{ij}}\right) \tag{4.29}$$

for a positive kernel K and bandwidth h_{ij}, and assume that P_i is large enough so that the preceding estimate is a good substitute for the true (unknown) density. We denote basal area per hectare in the i^{th} stand at time t_{ij} by $X_{ij} = X_i(t_{ij})$, and with these representations the LFLM follows Equation (4.2).

A comparative approach

We compare the LFLM with the parameter prediction method which was introduced in Section 1.2.2. We briefly recap the method as it will apply to this example. Assume that the underlying parametric distribution family is from the two–parameter Weibull family, $\Pr(D \leq d) = 1 - \exp\left(-\alpha d^{\beta}\right), d > 0$. Given the diameter measurements D_{ijp} for the i^{th} stand at time t_{ij}, the shape and scale parameters of the Weibull density can be estimated by maximum likelihood which we will denote by α_{ij} and β_{ij}. The Weibull parameter estimates α and β may then used as responses in a regression model to relate the parameters to the stand attributes.

In many forestry applications requiring longitudinal analyses such as the example in this paper, regression functions for stand attributes are required as inputs into models for the parameters α and β (e.g. Wang and Baker, 2007). In this comparative section however, we have chosen not to specify the parametric form of stand attributes, but to take a more direct approach. Specifically, we will estimate a nonparametric regression function relating the Weibull parameters α and β using as regressors the basal area of the stand and the time of measurement. We will use generalised additive models (e.g. Wood, 2006)

to fit these regressions which have been formulated as

$$\alpha_{ij} = g_\alpha(X_{ij}, t_{ij}) + \epsilon_{ij}$$
$$\beta_{ij} = g_\beta(X_{ij}, t_{ij}) + \epsilon_{ij}$$

where $g_\alpha(x,t)$ and $g_\beta(x,t)$ are unknown, smooth bivariate functions to be estimated from the data. If $\hat{\alpha}(x,t) = \hat{g}_\alpha(x,t)$ and $\hat{\beta}(x,t) = \hat{g}_\beta(x,t)$ are the predicted values of the Weibull parameters of a stand at time t with basal area per hectare equal to x, then the predicted diameter density is just the Weibull density with parameters $\hat{\alpha}$ and $\hat{\beta}$.

Table 4.1: Number of stands measured in each site at each of a possible seven measurement periods.

	Measurement period						
Site	1	2	3	4	5	6	7
A	18	18	18	18	18	18	18
B	18	18	18	18	18	0	0
C	18	18	18	18	18	0	0
D	18	18	18	18	18	18	0
E	18	18	18	17	0	18	0

Comparative measures

Cross–validation was used to compare the different approaches. The full data set was randomly split into training and testing data sets, with 60 stands in the training data set and 30 stands in the testing data set. Each model was fit on the training data, and then used to predict the diameter densities of the testing data; 20 such randomly selected cross–validation training/testing set combinations were used.

As we are dealing with real data, the true density $Y(d,t)$ is unknown, which leaves the difficulty of comparison between the approaches. We will use two comparative measures here. The first is the prediction integrated squared error (as in Equation 4.22), the second is the empirical (negative) log-likelihood. We compare both of the statistics using the training and testing data sets.

As noted in Section 4.4.1 we have for each stand/time combination, tree diameter measurements D. Using the same notation as previously, for a stand

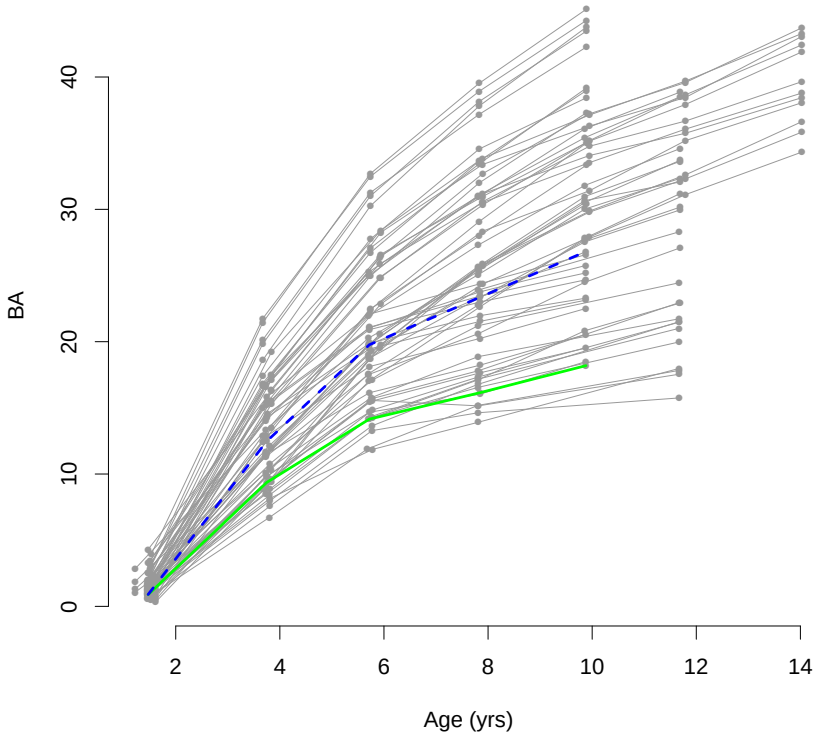


Figure 4.2: Evolution of basal area (m^2 ha^{-1}) over time for each forest stand. Highlighted are the basal area curves for the two stands given in Figure 4.1.

i we have the basal area observations $X_{i1}, \ldots, X_{iN_i}$ at times $t_{i1}, \ldots, t_{iN_i}$ along with diameter measurements $D_{ijp}, p = 1, \ldots, P_i$. The diameter measurements are used to create 'true' observations to compare each method: i) a kernel density estimate to compare the LFLM and ii) a Weibull density (estimated via maximum likelihood) to compare the parametric approach.

For both the LFLM and parametric approaches the predicted sum of squared errors is taken between the 'true' density and the density predicted from the models using the basal area and time information (Equation 4.22). The empirical log-likelihood takes advantage of the fact our responses are densities. For both LFLM and parametric methods, denote the predicted density at time t_{ij}, as $\widehat{Y}_{ij}(d, t_{ij})$. The empirical (negative) log-likelihood $\ell(D)$ is then

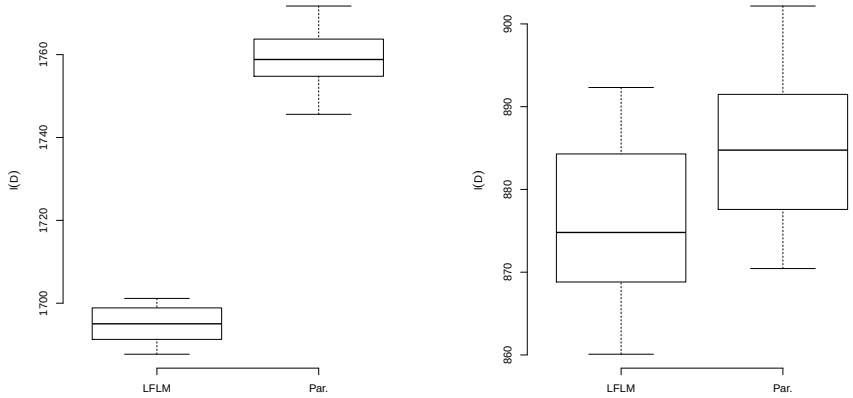
$$\ell(D) = -\sum_{i=1}^{N}\sum_{j=1}^{N_i}\sum_{p=1}^{P_i} \log \widehat{Y}_i(D_{ijp}, t_{ij})/P_i + 2\widehat{K} \tag{4.30}$$

where $\widehat{K}$ is the number of components chosen by Equation (4.22) for the expansion of $Y(d, t)$ in the case of the LFLM, and $\widehat{K} = 2$ for the parametric approach. Note that we are only able to use this statistic in this application as we do have the true diameter measurements D_{ijp}.

4.4.2 Results

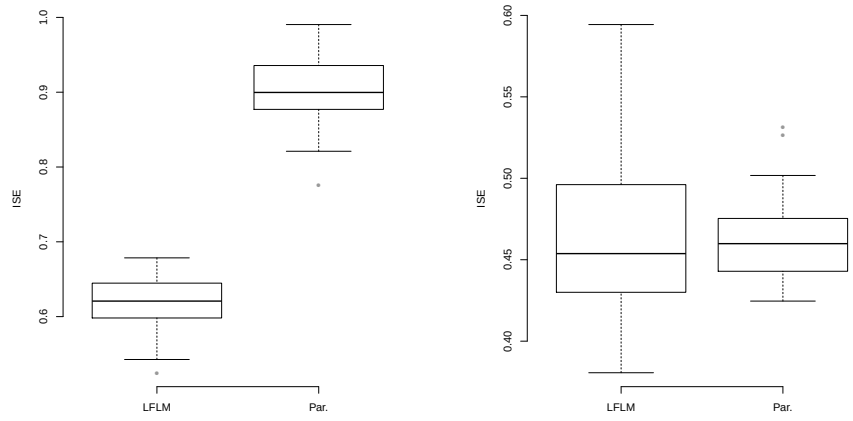
Figure 4.3 shows the distribution of $\ell(D)$ across 50 cross-validation runs for both approaches. The distribution for the training data is shown in Figure 4.3(a), whilst Figure 4.3(b) shows the distribution for the testing data. Similarly, Figures 4.4(a) and 4.4(b) show the distribution of ISE over 50 cross-validation runs for the training and testing data respectively. From Figure 4.3 we see that the LFLM outperforms the parametric approach in terms of negative log-likelihood, both on the training and testing data, however there is increased variability in the testing data results. For the ISE, Figure 4.4 shows that the LFLM outperforms the parametric approach on the training data, but there is no separation between the two approaches for the testing data.

Visualising the regression function $\beta(d, s, t)$ to determine the effect of basal area on the diameter density is obviously impossible, however we can visualise $\beta(d, s, t)$ for fixed values of t. Figure 4.5 displays the estimated regression function $\hat{\beta}(d, s, t = 1.48)$. We can interpret Figure 4.5 as follows: for prediction in early years, the effect of the functional covariate is strong over the whole time domain, yet is restricted to the lower diameters as to be ex-



(a) Training data (b) Testing data

Figure 4.3: Distribution of $\ell(D)$ (Equation (4.30)) across 50 cross-validation runs.



(a) Training data (b) Testing data

Figure 4.4: Distribution of ISE (Equation (4.22)) across 50 cross-validation runs.

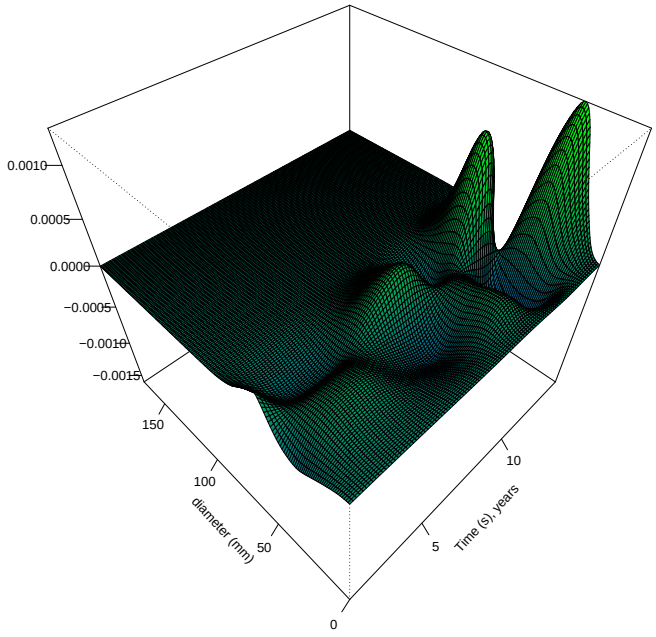
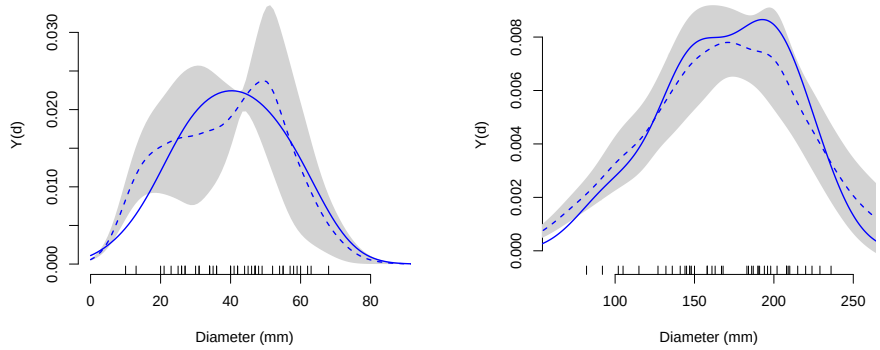


Figure 4.5: Estimated regression function $\hat{\beta}(d, s, t = 1.48)$.

pected.

Due to the heavy computational cost of the permutation test, we calculated the permutation p–value (Equation 4.24) for only one of the cross–validation data sets. Over 10000 permutations, the p–value was found to be 0, indicating that the effect of the functional basal area, $X(t)$ on the diameter density was highly significant. For reference, the 25, 50 and 75% quantiles of the permutation F–statistic were found to be 0.1540, 0.1655 and 0.1790 respectively, compared to the observed $F_{\text{obs}} = 0.2938$.

Predictions (dashed lines) for a stand at age 1.47 and 7.85 years from the testing data set are shown in Figures 4.6(a) and 4.6(b). Also shown in these figures are the 'true' densities (solid lines) and (Bonferroni adjusted) 95% confidence bands using the formula given by Equation (4.25). We use a Bonferroni adjustment so that whole–of–curve confidence can be assessed. Note that the 'true' observation is unobserved as we are dealing with the testing data set.



(a) $t = 1.47$ years (b) $t = 7.85$ years

Figure 4.6: 95% confidence intervals (shaded grey) for $E\left[f(d,t)|X^*\right]$, when a) $t = 1.47$ years, and b) $t = 7.85$ years. The solid line is the observed value, and the dashed line is the predicted value.

4.5 Discussion

The results given in the previous section indicate that the proposed longitudinal functional linear model is a suitable alternative to traditional parametric models such as parameter prediction.

Drawing a parallel between the LFLM and parametric models, we see that both seek to represent a curve by a finite number of parameters. In the functional case, these are the functional principal coefficients, and in the parametric case, the parameters of the assumed distribution. This dimension reduction step can result in more flexible fits under the LFLM model as parameters are not constrained by functional form.

In order to use the method that we have proposed on the forestry data set (Section 4.4) we had to construct our response probability density functions via kernel density estimates (Equation 4.29). We assumed there that the number of measurements per stand/time combination was large enough that the kernel density estimate was indeed a good representation of the true density. Obviously as the number of measurements increases, the estimates get better and better, and this is the case for the parametric models as well, where the MLE will become more precise. There may be room for improvement in models such as the one presented here, and indeed in more traditional forestry models for weighting the regression models by how precise we believe our estimates of the response densities to be. For instance, weighting each ob-

servation (whether the functional response as in the LFLM, or the Weibull parameter in the parametric case) by the number of diameter measurements used in its calculation may provide better model fits in each case.

The permutation test that was performed in Cardot et al. (2007) is less involved than it is in this situation, as interest there was in testing the effect of a covariate on the mean function. They discuss the fact that the bandwidth that is 'optimal' for the prediction of $\hat{\mu}(d,t)$ is not necessarily optimal for significance testing, and go on to recommend that a small grid of bandwidths around the 'optimal' bandwidth be tried. However, due to the complexity of the regression model proposed here, if we were to apply the same principles, we would need to alter not only the mean bandwidth, but also the bandwidths in the covariance function (Equation 4.15), along with those involved in estimation of the functional regression parameter, and also the number of principal components to be used (Equation 4.21) in the prediction model. For this chapter, we conditioned on these values in the permutation test for significance.

A minor issue with the specific application of predicting probability density functions is that due to the truncation of the infinite series representation of the functional coefficient (Equation 4.21) the resulting density predictions may not integrate to one. In this case, all predictions were rescaled so that they were a true probability distribution. For the general functional case, this will not be an issue.

The methods developed in this chapter can also be applied more generally to situations where general functions may be measured on individuals that have a longitudinal nature. With the proliferation of large data sets and more complex measurement schedules, we see methods such as that describe here becoming more and more common.

Figure 4.5 showed the estimated regression function $\hat{\beta}(d,s,t=1.48)$ for fixed t. In an informal sense, what this figure shows aside from the level of impact the functional covariate is having, is that the functional covariate is having an impact across its whole domain. Thus, making use of the whole time course of the functional growth covariate can have a significant impact. This was also confirmed by the results from the permutation test already discussed.

The predictions in Figure 4.6, along with their confidence bands appear to be extremely good estimates of the truth, especially as these are predictions from the testing data set. However poor predictions do occur; Figure C.1

in Appendix C.5 provides examples where the predictions and confidence intervals perform poorly. It is highly likely that the poor performance in this case is due Theorem 3 being an asymptotic result, that is relying on all of $K, L, M \to \infty$ as $n \to \infty$. In the case study, $n = 60$ stands and the number of components K, L, M chosen via cross–validation (Section 4.1.4) was always less than 10, and averaged around 4.

A possible approach to improve the predictions could be to develop a nonparametric approach to the modelling. An approach that has been used for the regular functional linear model is to model the functional principal component scores via an additive model (Müller, 2008). In this way, the predicted function has an additive, rather than linear structure. It is possible that a similar method could be developed to extend the LFLM.

In summary, the longitudinal functional linear model developed in this chapter provides an appealing alternative to current parametric methods of predicting diameter densities due to its flexibility for modelling distributions in situations where a single parametric family is not appropriate.

Chapter 5

Estimating population size from capture–recapture experiments

In this chapter we describe a novel use of functional data techniques in capture–recapture (CR) experiments. The methodology we will describe combines the work of Yao et al. (2005a) with that of Huggins (1989) allowing the estimation of population size from CR experiments with continuous, time varying covariates. As was discussed in Chapter 1 (Section 1.3), there are few existing approaches for dealing with continuous time varying covariates in CR experiments, and as such, this new methodology provides an exciting advancement on current techniques. The results of applying the new method to both simulated and real data demonstrate improved performance in terms of precision and variability of both population size estimates and CR model parameters when compared to more common approaches.

The conditional likelihood of Huggins (1989) has been standard methodology for estimating the parameters of CR models and subsequently population sizes for a number of decades now. When an individual is not caught at a given time point during the experiments duration, it is clear that any individual specific covariates are not measured for that individual. Allowing for heterogeneity in CR models has long been established (e.g. Pollock, 2002) due to the reduction in bias and increased precision of model parameters it affords. It should be clear that if a covariate has an effect on an individuals probability of capture, and that the covariate is also changing over time, then this should be allowed for in the model.

We show in this chapter, that in the case of continuous covariates in CR experiments that a functional data approach within the E–step of an EM al-

gorithm (Dempster et al., 1977) results in improved parameter estimation in the CR model, importantly leading to improved population size estimates.

A difficulty with heterogeneous CR models and missing data methods for parameter estimation is that missingness in covariates is informative by assumption. We have that the probability of capture is conditional on the covariate, however the probability of the covariate being missing is conditional on probability of capture. We demonstrate in Section 5.3 that this can allowed for by weighting estimates of the covariates by their probability of capture.

Usual methods of estimating $E\left[\mathbf{U}_{ij}|\mathbf{X}_{ij}\right]$ in missing data problems require a parametric model for the joint distribution of the missing and observed data. In this chapter, we employ a nonparametric, functional data approach to estimate the missing data. In particular, the Principal component Analysis through Conditional Expectation (PACE) approach for longitudinal data (Yao et al., 2005a) that was introduced in the previous chapter will be utilised. Further, the PACE method is extended to allow for the non–constant weighting introduced by the capture process as discussed above.

5.1 Notation and preliminaries

Consider a closed population consisting of $i = 1, \ldots, N$ individuals where a CR experiment has been conducted over $t \in [1, \tau]$ capture occasions. In this chapter we assume that the timing between each capture occasion is constant, that is $t = 1, \ldots, \tau$. Associated with each individual i, we have (possibly time varying) covariates $\mathbf{X}_{it}$.

Now, denote by $C(t)$ the indicator of capture of an individual at time t. That is, if individual i is caught at occasion t, $C_i(t) = 1$, otherwise $C_i(t) = 0$. Further, let $C = \max_t C(t)$ so that for individual i, $C_i = 1$ denotes that the individual was captured at least once over the τ capture occasions. Then the conditional probability of capture at time t for individual i is $p_{it} = \Pr\left(C_i(t) = 1|\mathbf{X}_{it}\right)$. Letting $\pi_i = \Pr\left(C_i = 1|\mathbf{X}_i\right) = 1 - \prod_{t=1}^{\tau}\left(1 - p_{it}\right)$ an unbiased estimator of the population size is

$$N = \sum_{i=1}^{N} \frac{I\left\{C_i = 1\right\}}{\pi_i} \tag{5.1}$$

N is generally unknown and is often the focus of CR experiments, as it is in this chapter. Further, the p_{it} are unknown and need to be estimated from the data. Thus, given estimates $\hat{p}_{it}$ of the probability that an individual i is

caught at time t, an estimate of the population size is

$$\begin{aligned}\widehat{N} &= \sum_{i=1}^{N} \frac{I\{C_i = 1\}}{\hat{\pi}_i} \\ &= \sum_{i=1}^{D} \frac{1}{\hat{\pi}_i}\end{aligned} \tag{5.2}$$

where $D = \sum_{i \geq 1} C_i$ is the number of distinct individuals captured.

To estimate π_i, a model for the capture process $C(t)$ is needed. In this chapter, we assume that heterogeneity amongst individuals affects their capture probabilities p_{it}, and that this heterogeneity is time varying. In particular, associated with each individual is one or more time varying covariates, which we assume are smooth functions of time. For simplicity of explanation, we give details of the model with one functional covariate $Y(t)$, however extensions to multiple functional covariates are handled in a similar fashion. Assume for the moment that for each individual $i = 1, \ldots, D$, $X_{it} = Y_i(t)$, the values of the covariate function at each capture time t are known, irrespective of whether the individual was actually captured. Further, denote by $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iQ})$ time invariant covariates of individual i, which once individual i is captured, are known for all capture periods. Then a possible model for the capture probabilities is

$$\begin{aligned}p_{it} &= \Pr\left(C_i(t) = 1 | X_{it}, \mathbf{Z}_i\right) \\ &= H\left(\beta_0 + X_{it}\beta_1 + \mathbf{Z}_i\boldsymbol{\alpha}\right)\end{aligned} \tag{5.3}$$

where $H(u) = \exp(u)/(1+\exp(u))$ is the logistic function, and β_0, β_1 and $\boldsymbol{\alpha} \in \mathbb{R}^Q$ are unknown parameters to be estimated. Note that in this formulation the effect of the time varying covariate $Y(t)$ is constant with respect to time of capture. The conditional log–likelihood (Huggins, 1989) of $C_i(t)|X_{it}, \mathbf{Z}_i, C_i = 1$ is then

$$\ell_i(\boldsymbol{\theta}) = -\log(\pi_i) + \sum_{t=1}^{\tau} \left\{ C_i(t) \log\left(\frac{p_{it}}{q_{it}}\right) + \log(q_{it}) \right\} \tag{5.4}$$

where $q_{it} = 1 - p_{it}$, and $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\alpha})'$. Maximising $\ell(\boldsymbol{\theta}) = \sum_i \ell_i(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$

leads to the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, and subsequently

$$\widehat{N} = \widehat{N}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{D} \pi_i(\hat{\boldsymbol{\theta}})^{-1} \tag{5.5}$$

where $\pi_i(\hat{\boldsymbol{\theta}})$ is π_i evaluated at $\hat{\boldsymbol{\theta}}$.

5.2 Estimating the model parameters

It is clear that in Equation (5.3), X_{it} is unknown whenever individual i is not captured that is, when $C_i(t) = 0$. It is thus not possible to estimate the parameters $\boldsymbol{\theta}$ nor the population size N using the *complete data* conditional log–likelihood given in Equation (5.4).

To estimate $\boldsymbol{\theta}$, we propose the use of an EM algorithm. The E–step of the algorithm uses functional principal components analysis (FPCA) to reconstruct an individuals time varying covariate that is incomplete due to non–capture of the individual at various times over the CR experiment. The M–step calculates the maximum likelihood estimates of $\boldsymbol{\theta}$ given expected value of the covariate.

5.2.1 M–step

The M–step requires maximisation of the complete data conditional log–likelihood, conditional on the observed data. Let U_{it} denote the missing data of individual i at time t, whenever $C_i(t) = 0$, and denote by $Q_i(\boldsymbol{\theta}; U_{it})$ the expected complete data conditional log–likelihood of individual i, $Q_i(\boldsymbol{\theta}; U_{it}) = E\left[\ell_i(\boldsymbol{\theta}; X_{it}, U_{it}) | X_{it}, \mathbf{Z}_i, C_i = 1\right]$. Then the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ are

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{D} Q_i(\boldsymbol{\theta}; U_{ij}) \tag{5.6}$$

5.2.2 E–step

The E–step requires the expectation of the complete data conditional log–likelihood, $Q_i(\boldsymbol{\theta}; U_{it})$, used in the M–step (Equation (5.6). Let $\mathbf{X}_{it} = (1, X_{it}, \mathbf{Z}_i)$ so that p_{it} (Equation 5.3) can be written as $p_{it} = H(\mathbf{X}_{it}\boldsymbol{\theta})$. Then in Equation (5.4), $\log(p_{it}/q_{it}) = \mathbf{X}_{it}\boldsymbol{\theta}$. Noting that this term only adds to the likelihood when $C_i(t) = 1$, its expectation is not required for $Q_i(\boldsymbol{\theta}; U_{it})$. The

expected value of the second term of Equation (5.4) cannot be found exactly due to its complex form, so we approximate it by plugging in the expected values of the missing data, which we denote by $E[U_{it}|\mathbf{X}_i, \mathbf{p}_i]$, where $\mathbf{X}_i = (X_{ij})'_{j=1,\ldots,N_i}$ i.e. the observed covariate when captured at times t_{ij}, and similarly $\mathbf{p}_i$. Putting this together, the expected log–likelihood is then

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \sum_{i=1}^{D} Q_i(\boldsymbol{\theta}; U_{it}) \\ &\approx \sum_{i=1}^{D} \left[-\log(\hat{\pi}_i) + \left\{ \sum_{t=1}^{\tau} C_i(t) \mathbf{X}_{it}\boldsymbol{\theta} + \log(\hat{q}_{it}) \right\} \right] \end{aligned} \tag{5.7}$$

where $\hat{q}_{it} = 1 - \hat{p}_{it}$ and $\hat{\pi}_i = (1 - \prod_{t=1}^{\tau} \hat{q}_{it})$ as before, and where now

$$\hat{p}_{it} = H\left(I\left\{C_i(t) = 1\right\} \cdot \mathbf{X}_{it}\hat{\boldsymbol{\theta}} + \left(1 - I\left\{C_i(t)\right\}\right) \cdot \mathbf{U}_{it}\hat{\boldsymbol{\theta}} \right) \tag{5.8}$$

where $\mathbf{U}_{it} = (1, E[U_{it}|\mathbf{X}_i, \mathbf{p}_i], \mathbf{Z}_i)$, that is, $\hat{p}_{it}$ is Equation (5.3) evaluated at the observed data if individual i was captured ($C_i(t) = 1$), otherwise it is evaluated at the expected value $E[U_{it}|\mathbf{X}_i, \mathbf{p}_i]$.

5.2.3 Computational details

To find the MLE $\hat{\boldsymbol{\theta}}$, the EM algorithm iterates over alternate E and M steps until convergence of $\hat{\boldsymbol{\theta}}$. Denote $\hat{\boldsymbol{\theta}}_q$ to be the value of the MLE after the q^{th} iteration of the algorithm, and similarly $E_q[U_{it}|\mathbf{X}_i, \mathbf{p}_i]$. Also let $\hat{\mathbf{p}}_{q,i} = (\hat{p}_{q,i1}, \ldots, \hat{p}_{q,i\tau})$ be the vector of estimated capture probabilities of the i^{th} individual on the q^{th} step.

We start with the E–step: Let $\mathbf{p}_{0,i} = (1, \ldots, 1)$, i.e. $p_{0,it} = 1$ for all i and t and estimate $E_0[U_{it}|\mathbf{X}_i, \mathbf{p}_{0,i}]$ as in Section 5.3. Using $E_0[U_{it}|\mathbf{X}_i, \mathbf{p}_{0,i}]$, in the M–step calculate $\hat{\boldsymbol{\theta}}_1$ (Equation (5.6)) using Equation (5.8). Then for $q = 1, \ldots$

1. E–step
   (a) Calculate $\hat{p}_{q,it} = H(\mathbf{X}_{it}\hat{\boldsymbol{\theta}}_{q-1})$
   (b) Estimate $E_q[U_{it}|\mathbf{X}_i, \hat{\mathbf{p}}_{q,i}]$
2. M–step
   (a) Using $E_q[U_{it}|\mathbf{X}_i, \hat{\mathbf{p}}_{q,i}]$ calculate the MLE $\hat{\boldsymbol{\theta}}_{q+1}$
   (b) Repeat until convergence of the sequence $\left\{\hat{\boldsymbol{\theta}}_q\right\}_{q=1,\ldots}$ is reached

5.3 Estimating $E[U_{it}|\mathbf{X}_i, \mathbf{p}_i]$ via FPCA

We turn now to the estimation of the individual specific time varying covariate $E[U_{it}|\mathbf{X}_i, \mathbf{p}_i]$, required for the EM algorithm described above. Considering a single individual i, let $Y_i(t)$ denote an (unobservable) continuous time varying covariate that is assumed to affect that individuals probability of capture. We observe $X_{ij} = Y_i(t_{ij})$ only when individual i is caught at time $t_{ij}, j = 1, \ldots, N_i$. We further assume that $Y(t)$ is a smooth, random function with mean $E[Y(t)] = \mu(t)$, square integrable over $[1, \tau]$, and covariance function $\gamma(s,t) = \text{Cov}[Y(s), Y(t)]$. As in the previous chapter, we assume that the covariance function can be expanded as an infinite sum of orthonormal basis functions, taken to be the eigenvalue/eigenfunction pairs, $(\lambda_k, \phi_k(t))$: $\gamma(s,t) = \sum_{k\geq 1} \lambda_k \phi_k(s)\phi_k(t)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. In this way, the observation at time t_{ij} of $Y(t)$ in individual i can be written

$$\begin{aligned} X_{ij} &= Y_i(t_{ij}) \\ &= \mu(t_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t_{ij}) \end{aligned} \tag{5.9}$$

for uncorrelated random coefficients ξ_{ik} with $E[\xi_{ik}] = 0$, $\text{Var}[\xi_{ik}] = \lambda_k$.

Following Yao et al. (2005a), assume that the functional principal component scores ξ_{ik} are jointly Gaussian. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iN_i})'$, $\boldsymbol{\mu}_i = (\mu(t_{i1}), \ldots, \mu(t_{iN_i}))'$, $\boldsymbol{\phi}_{ik} = (\phi_k(t_{i1}), \ldots, \phi_k(t_{iN_i}))'$ and $\Sigma_i = \text{Cov}(\mathbf{X}_i, \mathbf{X}_i)$, where the $(j,l)^{\text{th}}$ entry of Σ_i is $\gamma(t_{ij}, t_{il})$. The best predictor of ξ_{ik} is then given by

$$E[\xi_{ik}|\mathbf{X}_i, \mathbf{p}_i] = \lambda_k \boldsymbol{\phi}_{ik}' \Sigma_i^{-1} (\mathbf{X}_i - \boldsymbol{\mu}_i) \tag{5.10}$$

and we set $E[U_{it}|\mathbf{X}_i, \mathbf{p}_i] = \tilde{Y}_i(t)$ where

$$\tilde{Y}_i(t) = \mu(t) + \sum_{k=1}^{\infty} E[\xi_{ik}|\mathbf{X}_i, \mathbf{p}_i]\, \phi_k(t) \tag{5.11}$$

We turn now to the estimation of the components required for $\tilde{Y}_i(t)$. As the missing data mechanism for heterogeneous CR experiments is non–random by assumption, the methods in Yao et al. (2005a) need adjusting to allow for the capture probabilities p_{it}.

5.3.1 Estimating $\mu(t)$

Recall that $Y_i(t)$ is observed *only* when individual i is captured, that is, when $C_i(t) = 1$. However, $E\left[(Y_i(t) - \mu(t)) \cdot C_i(t)\right] = E\left[p_{it} \cdot (Y_i(t) - \mu(t))\right]$, so that

$$E\left[\frac{(Y_i(t) - \mu(t)) \cdot C_i(t)}{p_{it}}\right] = E\left[Y_i(t) - \mu(t)\right] = 0 \tag{5.12}$$

Assume that individuals are arranged, so that for $i = 1, \ldots, D, C_i = 1$, then for known weights w_{it}, the estimating equation

$$\begin{aligned}\sum_{i=1}^{N}\sum_{t=1}^{\tau}\frac{w_{it}}{p_{it}}\left[Y_i(t) - \mu(t)\right]C_i(t) &= \sum_{i=1}^{N} C_i \sum_{t=1}^{\tau}\frac{w_{it}}{p_{it}}\left[Y_i(t) - \mu(t)\right]C_i(t) \\ &= \sum_{i=1}^{D}\sum_{t=1}^{\tau}\frac{w_{it}}{p_{it}}\left[Y_i(t) - \mu(t)\right]C_i(t) \end{aligned} \tag{5.13}$$

has mean 0.

Clearly, knowledge of p_{it} is required in Equation (5.13). Recalling that the estimate of $\mu(t)$ is embedded in the E–step of the EM algorithm (Section 5.2.3), we will substitute the current estimates $\hat{p}_{q,it}$ for p_{it}. We estimate $\mu(t)$ by a local linear smoother, so that $\hat{\mu}_q(s) = \hat{\beta}_0(s)$, where

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0,\beta_1)}{\operatorname{argmin}} \sum_{i=1}^{D}\sum_{j=1}^{N_i}\frac{K_h(s - t_{ij})}{\hat{p}_{q,ij}}\left[X_{ij} - \beta_0 - \beta_1(s - t_{ij})\right]^2 \tag{5.14}$$

is the corresponding weighted least squares function of Equation (5.13), for the q^{th} E–step, where $t_{ij}, j = 1, \ldots, N_i$ is the time of the j^{th} capture of individual i; $K_h(u) = K(u/h)/h$ is positive, symmetric kernel, and $h = h(n)$ a bandwidth, such that $h(n) \to 0$ as $n \to \infty$, where $n = \sum_i N_i$, the total number of captures over the CR experiment. The choice of bandwidth h in Equation (5.14) is made by leave one group out cross–validation (see Appendix C.2).

5.3.2 Estimating $\gamma(s, t)$

We again recall that $Y_i(t)$ is observed *only* when individual i is captured, that is, when $C_i(t) = 1$. However,

$$\begin{aligned}&E\big[\{Y_i(s) - \mu(s)\} \cdot C_i(s) \times \{Y_i(t) - \mu(t)\} \cdot C_i(t)\big] \\ &\qquad = E\left[p_{is} \cdot \{Y_i(s) - \mu(s)\} \times p_{it} \cdot \{Y_i(t) - \mu(t)\}\right]\end{aligned}$$

so that

$$\begin{aligned} E\Big[\frac{C_i(s)}{p_{is}} \cdot \{Y_i(s) - \mu(s)\} \times \frac{C_i(t)}{p_{it}}\{Y_i(t) - \mu(t)\}\Big] \\ = E\big[\{Y_i(s) - \mu(s)\} \times \{Y_i(t) - \mu(t)\}\big] \\ = \gamma(s,t) \end{aligned}$$

Let the 'observed' covariance for individual i be $G_i(s,t) = [Y_i(s) - \mu(s)] \times [Y_i(t) - \mu(t)]$. Then for weights w_{ist}, the estimating equation

$$\sum_{i=1}^{N} \sum_{1 \le s \ne t \le \tau} \frac{w_{ist}}{p_{it}p_{is}} [G_i(s,t) - \gamma(s,t)] C_i(s) C_i(t) \tag{5.15}$$

has mean 0.

Substituting the q^{th} iteration estimates of $\mu(s)$ and p_{is} into Equation (5.15), and we estimate $\gamma(s,t)$ by a local linear smoother, so that $\hat{\gamma}_q(s,t) = \hat{\beta}_0(s,t)$, where

$$\begin{aligned} &(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \\ &\quad \underset{(\beta_0,\beta_1,\beta_2)}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{1 \le j \ne l \le N_i} \frac{K_h(s - t_{ij}, t - t_{il})}{\hat{p}_{q,ij}\hat{p}_{q,il}} [G_{q,i}(t_{ij}, t_{il}) - f_{ijl}(\boldsymbol{\beta}(s,t))]^2 \end{aligned} \tag{5.16}$$

is the corresponding weighted least squares function of Equation (5.15); $f_{ijl}(\boldsymbol{\beta}(s,t)) = \beta_0 - \beta_1(s - t_{ij}) - \beta_2(t - t_{il})$, $G_{q,i}(t_{ij}, t_{il}) = [X_{ij} - \hat{\mu}_q(t_{ij})] \times [X_{il} - \hat{\mu}_q(t_{il})]$, and $K_h(u,v)$ is a two dimensional positive kernel, for example the product kernel.

The eigenfunctions and eigenvalues $(\lambda_k, \phi_k(t))_{k=1,\ldots}$ are found as the spectral decomposition of the covariance operator, $\Gamma Y(t) = \int \gamma(s,t) Y(s)\, \mathrm{d}s$. That is, estimates for the k^{th} eigenvalue/eigenfunction pair are the solutions $\hat{\lambda}_k$ and $\hat{\phi}_k(t)$ of the equations

$$\int_1^{\tau} \hat{\gamma}_q(s,t) \hat{\phi}_k(s)\, \mathrm{d}s = \hat{\lambda}_k \hat{\phi}_k(t) \tag{5.17}$$

where the eigenfunctions are constrained to be orthonormal.

The expectation of the missing data U_{it} for use in the E–step of the EM algorithm (Section 5.2.2) can now be found as the projection onto the function

space spanned by the first K eigenfunctions, $\phi_k(t)$:

$$\widehat{E}_q\left[U_{it}|\mathbf{X}_i,\mathbf{p}_i\right] = \hat{\mu}_q(t) + \sum_{k=1}^{K} \widehat{E}\left[\xi_{ik}|\mathbf{X}_i,\mathbf{p}_i\right]\hat{\phi}_k(t) \tag{5.18}$$

where $\widehat{E}_q\left[\xi_{ik}|\mathbf{X}_i,\mathbf{p}_i\right] = \hat{\lambda}_k\hat{\boldsymbol{\phi}}'_{ik}\hat{\Sigma}_i^{-1}(\mathbf{X}_i - \hat{\mu}_{q,i})$

We choose K in Equation (5.18) so that the fraction of variance explained (FVE(K), Equation 5.19) is at least 99%.

$$\text{FVE}(K) = \frac{\sum_{k=1}^{K}\lambda_k}{\sum_{k=1}^{\infty}\lambda_k} \tag{5.19}$$

5.4 Inference

The population estimate $\widehat{N}$ (Equation 5.5) can be written as

$$\widehat{N} = \sum_{i=1}^{N}\frac{C_i}{\pi_i(\boldsymbol{\theta})} + \sum_{i=1}^{N}\left[\frac{C_i}{\pi_i(\hat{\boldsymbol{\theta}})} - \frac{C_i}{\pi_i(\boldsymbol{\theta})}\right] \tag{5.20}$$

and an unbiased estimate of the variance of the first term in Equation (5.20) is

$$\text{Var}\left[\sum_{i=1}^{N}\frac{C_i}{\pi_i(\boldsymbol{\theta})}\right] = \sum_{i=1}^{D}\frac{1-\pi_i(\boldsymbol{\theta})}{\pi_i(\boldsymbol{\theta})^2} \tag{5.21}$$

Letting $g(u) = \sum_i^D \pi_i(u)^{-1}$, the second term in Equation (5.20) is equal to

$$\begin{aligned}\sum_{i=1}^{N}\left[\frac{C_i}{\pi_i(\hat{\boldsymbol{\theta}})} - \frac{C_i}{\pi_i(\boldsymbol{\theta})}\right] &= g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}) \\ &\approx \{\partial g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}'\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\end{aligned} \tag{5.22}$$

so that

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^{N}\left[\frac{C_i}{\pi_i(\hat{\boldsymbol{\theta}})} - \frac{C_i}{\pi_i(\boldsymbol{\theta})}\right]\right) &\approx \text{Var}\left(\{\partial g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}'\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right) \\ &= \{\partial g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}'\,\text{Var}(\hat{\boldsymbol{\theta}})\,\{\partial g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}\end{aligned} \tag{5.23}$$

The covariance of the two terms in Equation (5.20) is 0 (Huggins, 1989), so that the variance of the population estimator $\widehat{N}$ is the sum of the two terms

in Equations (5.21) and (5.23).

We require now an estimate of Var($\hat{\boldsymbol{\theta}}$). If there were no missing data, i.e. X_{it} was known irrespective of individual i being captured at time t, then standard methods for the observed information matrix $I(\boldsymbol{\theta}) = -\partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ can be used, and the variance of $\hat{\theta}_i$ could be estimated as $I^{-1}(\hat{\boldsymbol{\theta}}; \mathbf{X})_{ii}$.

However, due to the missing data, the observed information matrix is analytically difficult to calculate, and in the case of the conditional likelihood for CR experiments, its expectation even more so. We instead suggest that a nonparametric bootstrap be used to estimate both the covariance matrix of $\hat{\boldsymbol{\theta}}$ and Var($\widehat{N}$), and that the respective confidence intervals (CIs) be taken as the bootstrap percentile CIs. In this way, inference about both $\boldsymbol{\theta}$ and N is achieved.

Due to the correlation within an individual's functional covariate, resampling for the bootstrap is performed at the individual level. Let $\mathbf{W}_i = (\mathbf{C}_i, \mathbf{X}_i)$ be the observed capture and covariate data for individual $i = 1, \ldots, D$, and let $\hat{F}$ be the empirical distribution of $\mathbf{W}$. Then the bootstrap proceeds as follows

Step 1 Generate a bootstrap sample $\mathbf{W}^*$ from $\hat{F}$ as

$$\mathbf{W}_1^*, \ldots, \mathbf{W}_D^* \overset{\text{iid}}{\sim} \hat{F}$$

Step 2 Using the EM algorithm (Section 5.2), estimate the MLE, giving $\hat{\boldsymbol{\theta}}^*$, and subsequently $\widehat{N}^*$

Step 3 The bootstrap covariance matrix of $\hat{\boldsymbol{\theta}}^*$ is then given by

$$\text{Cov}^*(\hat{\boldsymbol{\theta}}^*) = E^*\left[(\hat{\boldsymbol{\theta}} - E^*(\hat{\boldsymbol{\theta}}))(\hat{\boldsymbol{\theta}}^* - E^*(\hat{\boldsymbol{\theta}}^*))'\right]$$

where E^* denotes expectation of the distribution $\hat{F}$. The variance of $\widehat{N}^*$ is found similarly.

Steps 1 and 2 are repeated B times, giving B independent realisations of $\hat{\boldsymbol{\theta}}^*$ and $\widehat{N}^*$, say $\hat{\boldsymbol{\theta}}_1^*, \ldots, \hat{\boldsymbol{\theta}}_B^*$ and $\widehat{N}_1^*, \ldots, \widehat{N}_B^*$ respectively. Then the covariance of

$\hat{\boldsymbol{\theta}}$ and $\widehat{N}$ can be estimated by

$$\text{Cov}(\hat{\boldsymbol{\theta}}) \approx \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}_b^* - \bar{\boldsymbol{\theta}}^*)(\hat{\boldsymbol{\theta}}_b^* - \bar{\boldsymbol{\theta}}^*)' \tag{5.24}$$

$$\text{Var}(\widehat{N}) \approx \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{N}_b^* - \bar{N}^*) \tag{5.25}$$

where $\bar{\boldsymbol{\theta}}^* = (1/B)\sum_b \hat{\boldsymbol{\theta}}_b^*$ and $\bar{N}^* = (1/B)\sum_b \widehat{N}_b^*$.

$(1-\alpha)\%$ confidence intervals can be estimated by the bootstrap percentile intervals $(t^*_{\alpha/2}, t^*_{1-\alpha/2})$, where t^*_α is the α percentile of the bootstrap statistic t^*, estimated by the $\alpha \times B$ order statistic $t^*_{(\alpha\times B)}$, where $t^*_{(b)}$ is the b^{th} order statistic of the bootstrap estimates $t^*_1, \ldots, t^*_B$.

5.5 Numerical results

5.5.1 Simulation study

In this section, we explore the EM approach developed in the previous section through simulation. The simulation explores various relationships between the functional covariate and capture probability, along with various functional forms of the covariate. For each scenario in the simulation, the basic procedure is as follows:

Step 1 Generate $i = 1, \ldots, N$ individual functional covariates, $X_i(t)$

Step 2 For each i, calculate the probability of capture at capture occasions $j = 1, \ldots, \tau$ as $p_{ij} = H(f_{\mathcal{G}}[X_i(j)])$, where $f_{\mathcal{G}}$ is a linear function, dependent on the scenario (see Table 5.1) and $H(u)$ is the inverse logistic function

Step 3 For each i, j, simulate capture of an individual by comparing p_{ij} to a uniform random variable: if $p_{ij} \leq 0.5$ then the individual is not captured, that is, we set $C_i(j) = 0$; otherwise if $p_{ij} > 0.5$, $C_i(j) = 1$

Step 4 Those individuals that are not captured at any capture occasion j, i.e. $C_i = \sum_j C_i(j) = 0$ are removed from the data, leaving D unique individuals. Individuals that are captured at least once, have their capture history recorded, along with the functional covariate when captured. For example, if there are 5 capture occasions, and individual i was caught at capture occasions 2 and 4, then their capture history is $\{0, 1, 0, 1, 0\}$ and their corresponding functional covariate is $\{-, X_i(2), -, X_i(4), -\}$

Step 5 The relevant models are fit to the D individuals that are caught at least once.

The various scenarios for the simulation study are given in Table 5.1. The number of simulations in each scenario was 100. For each scenario, we have fit the functional model developed in this chapter, as well as two other models for comparison. The first, which we call the naive model, we assume that the covariate X is not time varying, and take the value of X_i to be the value of the covariate at the time of first capture. For example if individual i was caught first at capture occasion 3, then we would set $X_{ij} = X_{i3}$ for all $j = 1, \ldots, \tau$. The conditional log–likelihood given by Equation (5.4) is used to estimate the relevant parameters. The second model, which we call the gold standard model, assumes the covariate X is time varying and that at any capture occasion j, its value is known, irrespective of whether individual i was captured at that occasion. Again, parameters are estimated using the conditional log–likelihood, Equation (5.4).

In each of the models, we take the relationship between covariate and probability of capture to be of the same form as the data generation mechanism (Table 5.1). That is, we use the log–likelihood to find the MLEs $(\hat{\alpha}, \hat{\beta})$, where

$$p_{ij} = H(\alpha + \beta X_{ij})$$

Figure 5.1(a) displays the functional covariates $X_i(t)$ from one run of Scenario 2, whilst Figure 5.1(b) displays the corresponding probability of capture for each individual at each capture occasion (note this is unknown, and what we are in fact estimating). The blue points indicate observations when an individual was captured, whilst the grey points represent uncaptured individuals. In this scenario, there is an inverse relationship between the functional covariate and probability of capture.

Table 5.2 summarises the results from Scenarios 1 and 2. Shown in the table are the mean (sd) over 100 simulations, and the results for the EM/FPCA method are given at the converged parameter values. The results show that parameter estimates using the EM/FPCA method are extremely close to those of the gold standard. The naive approach results in parameter estimates that are quite different from the true parameters, especially in Scenario 2. This bias in the estimates resulted in population size estimates far from the true population sizes for the naive method, whilst the gold standard and EM/FPCA

Table 5.1: CR simulation scenarios. In each of the following scenarios, the functional covariate X_{ij} is generated by Equation (5.9). Parameter values have been chosen so that the percentage of captured individuals D/N is between approximately 60% and 80%. All simulations are carried out over $\tau = 5$ capture occasions.

Scenario 1 Observations are generated with mean process $\mu(t) = 3 + 4\exp(-t/2)$, with covariance function derived from two eigenfunctions $\phi_1(t) = \sin(\pi t/4)/\sqrt{2}$ and $\phi_2(t) = -\cos(\pi t/4)/\sqrt{2}$. The eigenvalues are set at $\lambda_1 = 1, \lambda_2 = 0.5^2$, and $\lambda_k = 0, k \geq 3$, and the functional principal component scores, $\xi_{ik} \sim N(0, \lambda_k), k = 1, 2$. The linear relationship relating X to capture probabilities (Step 3 of the simulation procedure) is $p_{ij} = H(-3.5 + 0.5X_{ij})$.

Scenario 2 Observations are generated with mean process $\mu(t) = 45.0 + 2t + 2\exp(-(t-3)^2)$, and the covariance function is as in Scenario 1. The eigenvalues are set at $\lambda_1 = 2, \lambda_2 = 1$, and $\lambda_k = 0, k \geq 3$, and the functional principal component scores, $\xi_{ik} \sim N(0, \sigma = \lambda_k), k = 1, 2$. The linear relationship relating X to capture probabilities (Step 3 of the simulation procedure) is $p_{ij} = H(1.5 - 0.05X_{ij})$.

Scenario 3 Constant model. Here the true covariate is not functional, and is generated as $X_{ij} = X_i \sim N(3, 1)$. The linear relationship relating X to capture probabilities (Step 3 of the simulation procedure) is $p_{ij} = H(-2.5 + 0.5X_i)$ for each capture occasion $j = 1, \ldots \tau$.

Scenario 4a Misspecified model. In this model, the functional covariate is generated as in Scenario 1, but is not related to the probability of capture. A constant probability of capture is set for each individual at $p_i = 0.2$.

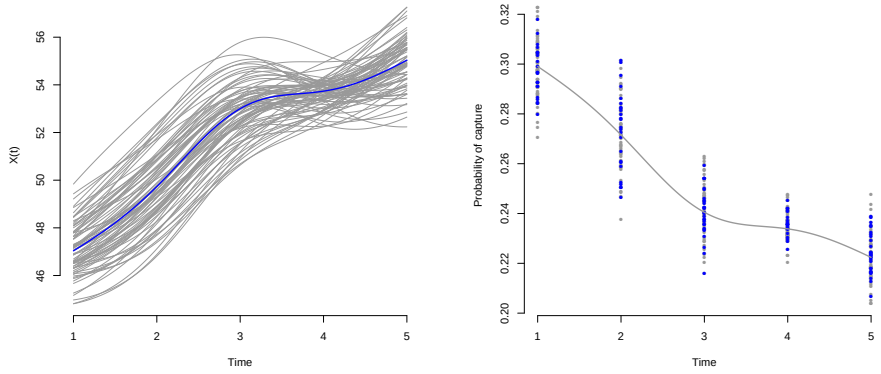
Scenario 4b Misspecified model as in Scenario 4a, but with covariate generated as in Scenario 2.

---

methods were extremely close to the truth.

Coverage of the parameter β_1 is close to the nominal coverage of 95% (Table 5.2) under Scenario 1 in all methods. However, due to the heavy bias of the naive method parameter estimates as already discussed, coverage is poor for the naive method under Scenario 2.

Figure 5.2 displays a comparison between the parameter estimates after one iteration of the EM algorithm, and at convergence of the EM algorithm. Figure 5.2(a) displays a boxplot of $\hat{\beta}_1$ from Scenario 1, whilst Figure 5.2(b)



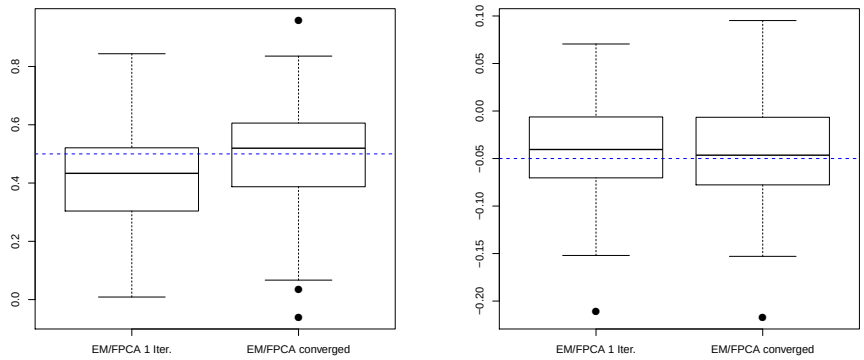
(a) True functional covariates, $X_i(t)$

(b) Observed X_{ij} vs. true capture probabilities p_{ij}

Figure 5.1: Example data from simulation Scenario 2.

Table 5.2: Results of simulations under Scenario 1 and 2. Shown in the table are the mean (sd) over 100 simulations. The nominal coverage of parameter β_1 is 95%.

N	Method	$\hat{N}$	$\hat{\beta}_0$	$\hat{\beta}_1$	Coverage (β_1)
	Scenario 1				
50	EM/FPCA	50.42 (9.2)	-3.487 (0.88)	0.496 (0.20)	0.97
	Naive	58.64 (16.1)	-3.591 (1.29)	0.442 (0.25)	0.98
	Gold	50.08 (9.0)	-3.602 (0.74)	0.526 (0.17)	0.94
100	EM/FPCA	102.40 (13.0)	-3.469 (0.64)	0.485 (0.14)	0.97
	Naive	118.30 (21.9)	-3.704 (1.07)	0.462 (0.20)	0.92
	Gold	102.29 (13.0)	-3.474 (0.60)	0.489 (0.13)	0.91
	Scenario 2				
50	EM/FPCA	50.92 (5.8)	1.199 (2.72)	-0.045 (0.05)	0.95
	Naive	60.20 (13.2)	7.412 (3.58)	-0.173 (0.07)	0.56
	Gold	50.91 (5.8)	1.395 (2.54)	-0.049 (0.05)	0.90
100	EM/FPCA	101.48 (8.0)	1.388 (1.85)	-0.048 (0.04)	0.94
	Naive	117.26 (16.0)	7.324 (2.34)	-0.171 (0.05)	0.19
	Gold	101.46 (8.0)	1.565 (1.80)	-0.052 (0.03)	0.90



(a) Scenario 1 (b) Scenario 2

Figure 5.2: Comparison of $\hat{\beta}_1$ for the EM/FPCA method after 1 iteration and convergence. (a) Scenario 1 and (b) Scenario 2.

Table 5.3: Results of simulations under Scenario 3. Shown in the table are the mean (sd) over 100 simulations. The nominal coverage of parameter β_1 is 95%.

N	Method	$\hat{N}$	$\hat{\beta}_0$	$\hat{\beta}_1$	Coverage (β_1)
50	EM/FPCA	51.78 (8.3)	-2.447 (0.82)	0.464 (0.24)	0.96
	Naive/Gold	53.38 (11.3)	-2.549 (0.89)	0.492 (0.25)	0.91
100	EM/FPCA	100.40 (8.4)	-2.433 (0.52)	0.478 (0.15)	0.98
	Naive/Gold	101.29 (9.4)	-2.481 (0.52)	0.491 (0.15)	0.94

displays a boxplot of $\hat{\beta}_1$ from Scenario 2. Both figures show that the weighted estimators for $\mu(t)$ (Equation 5.14) and $\gamma(t, s)$ (Equation 5.16) used in the EM algorithm, result in convergence of the parameters to their true values.

Table 5.3 summarises the results from Scenario 3. Recall that under Scenario 3, the covariate $X_{ij} = X_i$ for all capture periods, i.e. constant. The EM/FPCA method performs as well as the naive/gold standard methods in this scenario. Table 5.4 summarises the results from Scenarios 4a and 4b, which are models that include a covariate that is truly time varying, however is not related to the probability of capture. In both of these scenarios, the true $\beta_0 = -1.3863$ and $\beta_1 = 0$. The EM/FPCA performs as well as the gold standard in these scenarios, both in terms of parameter estimation and population estimation, however the naive method is seen to perform poorly.

Table 5.4: Results of simulations under Scenario 4a and 4b. Shown in the table are the mean (sd) over 100 simulations. The nominal coverage of parameter β_1 is 95%.

N	Method	$\hat{N}$	$\hat{\beta}_0$	$\hat{\beta}_1$	Coverage (β_1)
	Scenario 4a				
50	EM/FPCA	51.50 (12.2)	-1.286 (0.96)	-0.026 (0.22)	0.94
	Naive	67.20 (48.8)	-3.866 (1.62)	0.535 (0.32)	0.47
	Gold	51.52 (12.1)	-1.415 (0.80)	0.004 (0.18)	0.93
100	EM/FPCA	102.16 (10.2)	-1.359 (0.67)	-0.010 (0.16)	0.93
	Naive	116.04 (18.3)	-3.471 (0.90)	0.451 (0.18)	0.26
	Gold	102.22 (10.3)	-1.433 (0.49)	0.007 (0.11)	0.95
	Scenario 5a				
50	EM/FPCA	52.27 (9.4)	-2.007 (2.69)	0.011 (0.05)	0.94
	Naive	70.04 (25.0)	8.638 (4.57)	-0.202 (0.09)	0.31
	Gold	52.31 (9.5)	-1.535 (2.40)	0.002 (0.05)	0.94
100	EM/FPCA	101.29 (12.3)	-1.671 (2.36)	0.005 (0.05)	0.86
	Naive	125.02 (26.4)	7.728 (3.03)	-0.183 (0.06)	0.08
	Gold	101.34 (12.3)	-1.356 (2.01)	-0.001 (0.04)	0.90

5.5.2 Case study: Mountain Pygmy Possum

We now demonstrate the use of the time varying EM/FPCA method to a data set concerning the Mountain Pygmy Possum. This data set (Huggins and Hwang, 2007) is from a CR experiment conducted at Mount Hotham, Australia, over five nights in November, 2000. The animals of interest in this experiment were the Mountain Pygmy Possum (*Burramys parvus*), which is listed as an endangered species in Australia, and was actually thought to be extinct, until rediscovered in 1966. At each capture, the weight of the individual possum was recorded, and in total, 54 unique possums were captured. Figure 5.3 provides a graphical summary of the weight of individual possums as recorded when captured over the five night period. Also shown is the overall mean weight, $\hat{\mu}(t)$, as fit by the EM algorithm, and the individual fitted weights $\widehat{Y}_i(t)$, also from the EM algorithm, for the fitted model.

Examining Figure 5.3, it appears that the weight of individual possums may be time varying. There is a decrease in weight between capture occasions one to four, and an increase between capture occasions four and five. Given this, we expect the model based on first recorded weight (naive model) to not perform as well as the EM/FPCA method. We fit both a linear model and a quadratic model under each method, and also calculate the AIC for each

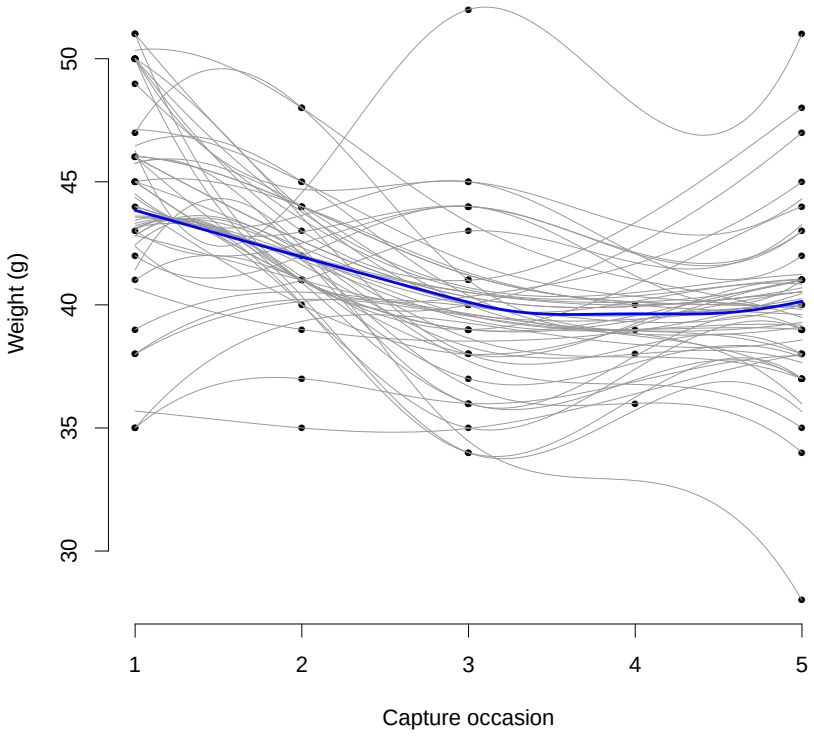


Figure 5.3: Weight of Mountain Pygmy Possums over a five night period. The blue curve is the estimated mean weight.

method. The results are given in Table 5.5.

Using AIC (Table 5.5), the best model is chosen to be the quadratic EM/FPCA model. Figure 5.4 plots the relationship between the weight of the Pygmy Possums, and their probability of capture, as fitted by the quadratic EM/FPCA model (shown in blue) and the linear naive model (shown in green). Overlaid in the figure is a kernel density estimate of the weights of the possums (shown in black). Also shown in the figure are 95% pointwise confidence intervals for the probability of capture. Under the EM/FPCA method, the predicted population (95% CI) was 57.92 (55.8, 62.5) and for the naive method, predicted population (95% CI) was 60.29 (56.6, 73.1).

Table 5.5: Results of fitting linear and quadratic models to the Pygmy Possum data, using the naive and EM/FPCA methods. Bootstrap 95% confidence intervals are given in parentheses.

Model	Method	$\hat{\alpha}$	$\hat{\beta}_0 X_{ij}$	$\hat{\beta}_1 X_{ij}^2$	AIC
Constant	—	-0.29 (-0.66, 0.11)			368.17
Linear	Naive	-4.34 (-8.60, -1.81)	0.09 (0.03, 0.20)		357.48
	EM/FPCA	-2.13 (-5.37, 1.20)	0.04 (-0.04, 0.12)		368.53
Quadratic	Naive	-14.22 (-47.56, 8.61)	0.56 (-0.51, 2.05)	-0.01 (-0.02, 0.01)	358.36
	EM/FPCA	58.41 (5.99, 193.35)	-2.87 (-9.28, -0.31)	0.03 (0.01, 0.11)	349.31

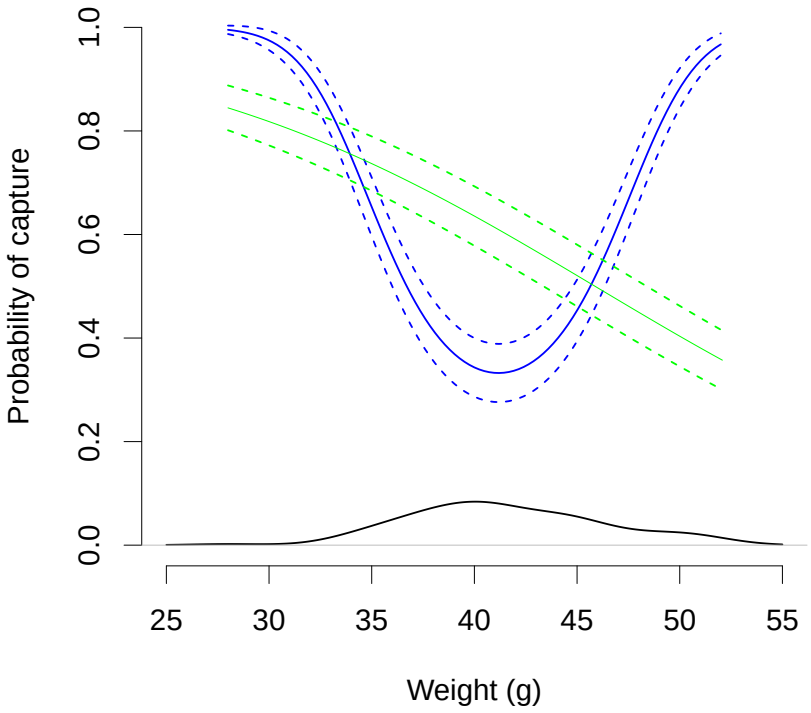


Figure 5.4: Relationship between probability of capture and weight of Pygmy Possums using the naive (green) and EM/FPCA (blue) methods.

5.6 Discussion

We have demonstrated in this chapter that the use of a functional principal components analysis within an EM algorithm provides a more than satisfactory solution to the problem of time varying covariates within capture–recapture experiments. Compared with Bayesian approaches such as that presented by Bonner and Schwarz (2006), our EM/FPCA method does not require parametric specification of the time varying function. As such, arbitrary dependence on time in the functional covariate is possible. Further, being essentially a random effects model, the EM/FPCA method adapts to individual heterogeneity in a straightforward manner.

The simulation results presented in Section 5.5.1 clearly demonstrate the robustness of the EM/FPCA method. Where the covariate was truly time varying, the EM/FPCA outperformed the naive method considerably, with excellent coverage of the fitted parameters providing confidence for subsequent inference in the case study. Population estimates were also extremely good, and overall, the EM/FPCA was almost at a par with the gold standard method. Where the true covariate was not time varying, but actually constant over capture occasions (Scenario 3, Table 5.1) the EM/FPCA method again performed just as well as the naive/gold standard method. Whilst this result only comes from one simulation study, it gives confidence that the EM/FPCA will still provide solid inference when the true covariate is constant.

It is perhaps not surprising that the EM/FPCA method performs well in the constant case (Scenario 3). A constant can still be viewed as a function over time, albeit non–varying. The advantage of the EM/FPCA is (as discussed earlier) its ability to adapt to individual heterogeneity, and so constant models under the EM/FPCA method lead to the 'filling in' of missing data, resulting in effectively the same inferential properties as that of the standard conditional likelihood inference.

A somewhat surprising result was that in Scenarios 4 and 5. When a time varying nuisance parameter existed, the naive method was completely thrown off, and as a result, extremely poor coverage of the parameter estimates would suggest that the power of the naive model was considerably low in these situations. The EM/FPCA model performed well, with good mean parameter estimates and coverage. This suggests that when a time varying covariate is present, but does not influence the probability of capture, the EM/FPCA method would correctly fail to reject the null hypothesis of no effect.

The simulation results give us confidence in the models that are fit to the Pygmy Possum data. The confidence intervals given in Table 5.5 lead us to reject the null hypothesis of no effect, and further accept that there is a quadratic relationship between (the logit of) capture probability and the weight of the possum, which is taken to be time varying.

In summary, the EM/FPCA provides a worthy addition to the capture–recapture statistician's toolkit. Not only is it appropriate to use the information inherent in a time varying covariate to its fullest through a method such as that described, the EM/FPCA is also robust to model misspecification and can handle non time varying covariates with similar ease.

Chapter 6

Possible future directions

Functional data analysis is an exciting area of statistics that is under continuous development. In this thesis we have described a number of modelling approaches and demonstrated how they may be applied to some common problems in forest biometry and capture–recapture experiments. Of course, the methodologies presented are not restricted to the applications described in the preceding chapters. As an example, many longitudinal data sets in the biological sciences in general are characterised by repeated measurements on individuals within distinct groups. In situations where repeated observations on an individual over time are not possible (e.g. destructive sampling methods may be required for measurements, or individuals within a sample can alter rapidly), the method presented in Chapter 4 can still be applicable, as there we are interested in the distribution of individuals over time, not specific individuals.

The response variable of interest is likewise not restricted to be of a certain form as were the examples in this thesis. The improved functional regression tree method presented in Chapter 3 could be applied to more general functions, with the appropriate choice of dissimilarity matrix, $\mathcal{A}$. For example, the spectra resulting from the analysis of a sample data by mass spectrometry may be considered as functional data, and analysed as such.

The methods presented in Chapters 2 to 4 could possibly be extended to better account for experimental conditions. The *E. globulus* data introduced in Section 1.2.1 were collected under a replicated block design. It may be possible to extend those methods by allowing for correlation within blocks in some manner, thus having a shrinkage effect on node deviance. One possible approach for doing so would be to use an EM algorithm at each stage of

the splitting procedure. However, adding to the computational complexity within each node (e.g. Section 3.2.2) would quickly become unmanageable computationally.

Further enhancements of the methodology of Chapters 2 and 3 were commented on in the discussions of those chapters. We discussed the use of expert opinion in adjusting the dissimilarity matrix, $\mathcal{A}$, prior to fitting the models. Whilst this could allow for a vast range of adjustments (we gave the example of adjusting PDFs that are more favourable to harvesting for timber volume), the burden on the analyst would increase rapidly with sample size, as all pairwise dissimilarities a_{ij} would need to be adjusted. On top of the burden to the analyst, the subjectivity of the process of adjusting the dissimilarities may also be questioned.

Objective methods of adjusting the dissimilarity matrix $\mathcal{A}$ in Chapters 2 and 3 provide possible future research directions. For example, dissimilarities could be augmented by a measure of modality, such as the dip statistic used in Section 2.2.3. Such a statistic could be combined with a standard dissimilarity measure to provide a further discriminatory effect between functional observations. A possibility in the cases described in Chapters 2 and 3 where the response functions are calculated as kernel density estimates, could be weighting based on the number of observations used to calculate the estimate. Such an adjustment would place more weight on functional responses based on larger sample sizes, acting effectively as a certainty adjustment.

We discussed in Chapter 3 that choice of deviance itself in the calculation of the dissimilarity matrix could have an impact on the results of the fitted model. An ensemble method was proposed as one way of dealing with this choice. In this approach, we would fit multiple models using different dissimilarities, and combine the results. This would essentially have the effect of smoothing out the choice of deviance, but would also require significant increases in computational time. Computational time could however be minimised by implementing the approach in parallel. This would be interesting area for future research.

The method described in Chapter 4 could possibly be further enhanced by including time–invariant covariates into the model. Presenting a model statement is quite simple in this scenario, for example by including a linear term, e.g. $\mathbf{Zb}$ in the model statement, Equation (4.1), that accounts for covariates that do not change over time. However, estimating the required parameters

in such a model does not appear to have a natural solution at this stage.

Experimental design, and specifically the timing of measurements may provide some interesting outcomes in terms of not only precision, but also cost of operations when used in conjunction with the Longitudinal Functional Linear Model (Chapter 4). The method in that chapter made use of data pooling in order to overcome issues with low numbers of observations over time, and sparseness. In the application presented, however, the measurement schedule meant that although observations were sparse, pooling was not used in its most optimal way, as the times of measurement were not completely dense over the time domain of interest (e.g. Yao et al., 2005a, for more details on pooling to overcome sparseness). Dense observations over the time period of interest could be achieved if instead of fixed time observations, measurements of the trees within stands were designed to be taken at staggered intervals. Such a measurement schedule could result in fewer measurements being needed, yet also result in greater precision due to the dense nature of the measurements.

The application of functional data analysis techniques to capture–recapture data in Chapter 5 is the first of its type to the best of our knowledge. As such, many extensions seem possible. For example, as the E–step of the algorithm is essentially divorced from the capture process model, more sophisticated capture process models could be used. An example is to use a nonparametric method to model the effect of the functional covariate, however care would need to be taken to ensure that model identifiability issues did not occur.

The EM/FPCA method of Chapter 5 may further be extended to analyse open populations. Here, one generally has many more capture occasions spread over a larger amount of time, and so time–varying covariates would seem to occur quite naturally. However open populations models require the modelling of not only the capture process, but also methods for dealing with births, deaths and immigration/emigration and as such will likely be much more complex than closed populations to analyse.

Bibliography

Bailey, R. and Dell, T. (1973). Quantifying diameter distributions with the Weibull function. *Forest Science*, 19(2):97–104.

Bailey, R. L., Burgan, T. M., and Jokela, E. J. (1989). Fertilized midrotation–aged slash pine plantations: stand structure and yield prediction models. *Southern Journal of Applied Forestry*, 13(2):76–80.

Beck, N. and Katz, J. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, pages 634–647.

Bonner, S. J. and Schwarz, C. J. (2006). An extension of the cormack-jolly-seber model for continuous covariates with application to microtus pennsylvanicus. *Biometrics*, 62(1):142–149.

Borders, B., Souter, R., Bailey, R., and Ware, K. (1987). Percentile-based distributions characterize forest stand tables. *Forest Science*, 33(2):570–576.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, California.

Burk, T. and Newberry, J. (1984). A simple algorithm for moment-based recovery of Weibull distribution parameters. *Forest Science*, 30(2):329–332.

Cardot, H. (2007). Conditional functional principal components analysis. *Scandinavian Journal of Statistics*, 34(2):317–335.

Cardot, H., Prchal, L., and Sarda, P. (2007). No effect and lack-of-fit permutation tests for functional regression. *Computational Statistics*, 22(3):371–390.

Cariou, V. (2006). Extension of multivariate regression trees to interval data. application to electricity load profiling. *Computational Statistics*, 21(2):325–341.

Chaudry, M. A. and Ahmad, M. (1993). On a probability function useful in size modelling. *Canadian Journal of Forest Research*, 23:1679–1683.

Clutter, J., Fortson, J., Pienaar, L., Brister, G., and Bailey, R. (1983). *Timber management: a quantitative approach*, volume 6. Wiley New York.

Clutter, J. L. and Bennett, F. A. (1965). Diameter distributions in old–field slash pine plantations. *Georgia Forest Research Council Rep.*, 13:1 – 9.

Csiszár, I. (1967). Information type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318.

De'ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, 83(4):1105–1117.

De'ath, G. (2007). *mvpart: Multivariate partitioning*. rpart by Terry M Therneau and Beth Atkinson. R port of rpart by Brian Ripley. Some routines from vegan – Jari Oksanen. Extensions and adaptations of rpart to mvpart by Glenn De'ath.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. R. Statist. Soc. B*, 39(1):1–38.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the. 632+ bootstrap method. *Journal of the American Statistical Association*, pages 548–560. David.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Hartigan, J. and Hartigan, P. (1985). The dip test of unimodality. *Ann. Statist.*, 13(1):70–84.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, second edition.

He, G., Muller, H., and Wang, J. (2000). Extending correlation and regression from multivariate to functional data. *Asymptotics in statistics and probability: papers in honor of George Gregory Roussas*, page 197.

Henningsen, A. and Hamann, J. D. (2007). systemfit: A package for estimating systems of simultaneous equations in r. *Journal of Statistical Software*, 23(4):1–40.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133.

Huggins, R. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, pages 725–732.

Huggins, R. and Hwang, W.-H. (2007). Non-parametric estimation of population size from capture-recapture data when the capture probability depends on a covariate. *J. Roy. Statist. Soc. Ser. C*, 56(4):429–443.

Hyink, D. and Moser, J. (1983). A generalized framework for projecting forest yield and stand structure using diameter distributions. *Forest Science*, 29(1):85–95.

Hyndman, R. and Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45.

King, R., Brooks, S. P., and Coulson, T. (2008). Analyzing complex capture-recapture data in the presence of individual and temporal covariates and model uncertainty. *Biometrics*, 64(4):1187–1195.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lane, S., Robinson, A., and Baker, T. (2010). The functional regression tree method for diameter distribution modelling. *Can. J. For. Res.*, 40(9):1870–1877.

Lane, S. E. and Robinson, A. P. (2011). An alternative objective function for fitting regression trees to functional response variables. *Computational Statistics & Data Analysis*, 55(9):2557 – 2567.

Loève, M. (1977). *Probability theory*. Springer-Verlag, New York, fourth edition.

Maltamo, M., Kangas, A., Uuttera, J., and Torniainen, T. (2000). Comparison of percentile based prediction methods and the Weibull distribution in describing the diameter distribution of heterogeneous scots pine stands. *Forest Ecology and Management*, 133:263–274.

MATLAB (2011). *version 7.13.0.564 (R2011b)*. The MathWorks Inc., Natick, Massachusetts.

Merler, S. and Furlanello, C. (1997). Selection of tree-based classifiers with the bootstrap 632+ rule. *Biometrical Journal*, 39(3):369–382.

Müller, H.-G. (2008). Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544.

Nerini, D. and Ghattas, B. (2007). Classifying densities using functional regression trees: Applications in oceanology. *Computational Statistics & Data Analysis*, 51(10):4984–4993.

Nichols, J., Sauer, J., Pollock, K., and Hestbeck, J. (1992). Estimating transition probabilities for stage-based population projection matrices using capture-recapture data. *Ecology*, 73(1):306–312. doi: 10.2307/1938741.

Otis, D., Burnham, K., White, G., and Anderson, D. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monogr*, (62):3–135.

Pollock, K. (2002). The use of auxiliary variables in capture-recapture modelling: An overview. *J. of Appl. Stats.*, 29(1):85–102.

Pollock, K., Hines, J., and Nichols, J. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, pages 329–340.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robinson, A. (2004). Preserving correlation while modelling diameter distributions. *Can. J. For. Res.*, 34(1):221–232.

Rudin, W. (1976). *Principles of mathematical analysis*. McGraw-Hill Book Co., New York, third edition. International Series in Pure and Applied Mathematics.

Silverman, B. W. (1986). *Density Estimation*. Chapman & Hall, London, UK.

Vanclay, J. K. (1994). *Modelling Forest Growth and Yield: Applications to Mixed tropical Forests*. CAB International.

Wang, Y. (2005). A semiparametric regression model with missing covariates in continuous-time capture-recapture studies. *Australian & New Zealand Journal of Statistics*, 47(3):287–297.

Wang, Y. and Baker, T. G. (2007). A regionalised growth model for *Eucalyptus globulus* plantations in south–eastern Australia. *Australian Forestry*, 70(2):93–107.

Wong, J., Baker, T., Duncan, M., Stackpole, D., and Stokes, R. (1999). Individual tree volume functions for plantation eucalypts. Department of Natural Resources and Environment, Victoria.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Xi, L., Watson, R., Wang, J., and Yip, P. (2009). Estimation in capture–recapture models when covariates are subject to measurement errors and missing data. *Canadian Journal of Statistics*, 37(4):645–658.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.*, 33(6):2873–2903.

Yu, Y. and Lambert, D. (1999). Fitting trees to functional data, with an application to time-of-day patterns. *Journal of Computational and Graphical Statistics*, 8(4):749–762.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368.

Zwane, E. and der Heijden, P. V. (2008). Capture-recapture studies with incomplete mixed categorical and continous covariates. *Journal of data science*, 6:557–572.

Appendix A

FRT Appendix 1

A.1 A further FRT example

We provide here another demonstration of the way in which the FRT method is fit using recursive partitioning. Recall that we have covariates $\mathbf{X}$, which we wish to use to estimate $E(Y(d)|\mathbf{X})$ by partitioning the space $\mathcal{X}$ of all possible observations. Assume that we have two covariates, X_1 and X_2. Then the recursive partitioning of $\mathcal{X}$ defines a hypercube in $\mathbb{R}^2$. Figure A.1 displays a possible partitioning of $\mathbb{R}^2$, from the FRT method. The first split is labelled **a**, which splits the variable X_1 into two regions, $X_1 < a$ and $X_1 \geq a$ (as in Figure 2.1). This split would have been the best split resulting from the objective function, Equation (2.2). The next split is at **b**, which splits X_2 and defines two regions R_1 and R_2. The next split results in four defined regions.

The functional observations that are associated with each region, can then be used to predict a new observation. For example, suppose we have a new observation, $\mathbf{X}^*$. We find from our partitioning that $\mathbf{X}^* \in \mathrm{R}_2$. Then the predicted function $\widehat{Y}^*(d)$ will be that associated with region R_2 from the partitioning, as is displayed in Figure A.2.

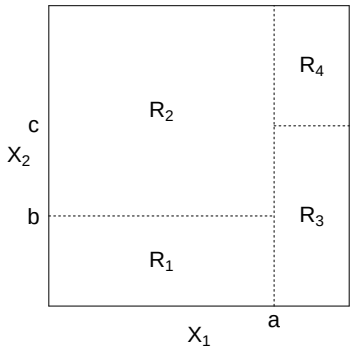


Figure A.1: Possible partitioning of $\mathbb{R}^2$ from the FRT method. The three splits **a**, **b**, **c** define four regions $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, $\mathbf{R}_4$.

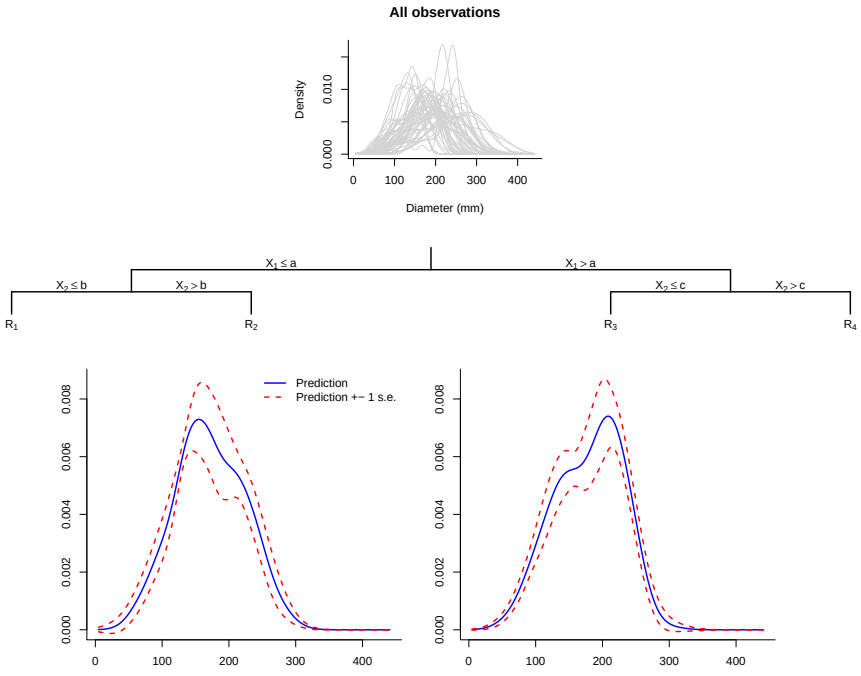


Figure A.2: Example prediction resulting from a new observation $\mathbf{X}^* \in \mathrm{R}_2$ from Figure A.1.

A.2 Prediction methods

We give here details of the process used to fit the parameter prediction and percentile methods using the *E. globulus* data as an example. As given in Section 2.2.1, the full model is written as

$$\begin{aligned} y_i = \alpha_i + \gamma_i + \beta_1 \cdot \mathrm{A}_i + \beta_2 \cdot \mathrm{G}_i + \beta_3 \cdot \mathrm{N} \\ + \beta_4 \cdot \mathrm{H}_i + \beta_5 \cdot \mathrm{V}_i + \beta_6 \cdot \mathrm{D}_{\mathrm{q},i} \end{aligned} \tag{A.1}$$

where α_i and γ_i are coeffiecients for site and stocking treatment respectively and y_i denotes the response parameter of interest, that is the shape, scale or location parameter of the Weibull distribution for the parameter prediction method or the p^{th} percentile for the percentile method.

For the parameter prediction method, the scale and shape parameters of the Weibull density function (Equation 1.1, Section 1.2) are estimated for each sample plot using maximum likelihood, with the location parameter being defined as the minimum diameter in each plot.

The Weibull parameters are then related to the stand characteristics by Equation (A.1), and the parameters are then estimated using seemingly unrelated regression (Zellner, 1962, Appendix A.3). The process is completed by removing terms from the model equation for each parameter until all remaining terms are significant.

For the percentile method, the response parameter y_i in the model equation (A.1) are the $\{0, 0.05, 0.15, \ldots, 0.95, 1\}^{\text{th}}$ percentile differences for each sample plot. Following Borders et al. (1987) methodology, the 65^{th} percentile is chosen as the 'driver' percentile. Differences between the percentiles are then calculated, and the system of equations becomes

$$d_{65} = X\beta \qquad\qquad d_q^* = X\beta$$

where d_q^* is the difference between the q^{th} and $(q-1)^{\text{th}}$ percentile for the sample plot (for example, $d_{75}^* = d_{75} - d_{65}$), X are the observed sample characteristics, and β are the model equation parameters. Again the equation parameters are estimated using seemingly unrelated regression with terms being removed (in turn from each equation) until all remaining terms are significant.

To recreate the diameter distribution for a given plot using the parameter prediction, we first predict the parameters using the final model, use these

as the parameters for the Weibull PDF and calculate the densities over a fine grid. For the percentile method, we first predict the percentile differences, then recover the actual percentiles (eg. $d_{75} = d_{65} + d^*_{75}$). A constrained cubic spline is then used to interpolate the percentiles. The derivative of the spline can then be found to calculate the densities.

A.3 Seemingly–unrelated regression

Seemingly–unrelated regression (SUR) is a method popular in econometrics that allows the (efficient) estimation of multiple dependent variables from the same data. This is the situation that occurs in Section 2.2.1 and detailed above, where the parameter prediction, parameter recovery and percentile methods relied on the one data set to estimate all parameters and percentiles from a family of PDFs.

Consider the regression equation given in Equation (1.2) and in particular, its linear formulation. Zellner (1962) describes a procedure that results in estimators of β_{ijk} that are more efficient than the case in which we were to treat each of the i equations turn–by–turn. Following Zellner (1962), I will outline the SUR procedure in terms of the parameter prediction method detailed above. First, write Equation (1.2) in matrix form as

$$\begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \vdots \\ \boldsymbol{\theta}_m \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_m \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{bmatrix} \tag{A.2}$$

where $\boldsymbol{\theta}_i$ is an $n \times 1$ vector containing the 'observed' values of the i^{th} PDF parameter, X_i is an $n \times p$ matrix of observed stand–level variables, $\boldsymbol{\beta}_i$ is an $p \times 1$ vector of regression coefficients and $\boldsymbol{\epsilon}_i$ is an $n \times 1$ vector of random error terms with mean 0. It is further assumed that the $m \times n$ error vector on the RHS of Equation (A.2) has variance–covariance matrix Σ equal to

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix} \otimes \mathbf{I}_n \tag{A.3}$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix $\sigma_{ii'} = E(\epsilon_{ij}\epsilon_{i'j})$, $j = 1, \ldots, n$, $i = 1, \ldots, m$.

Zellner (1962) show that by pre–multiplying both sides of Equation (A.2) by a matrix H, such that $E(H\boldsymbol{\epsilon}\boldsymbol{\epsilon}'H') = H\Sigma H' = I$, the system can be transformed into single equation regression model that is solvable by generalised least squares (GLS).

It is generally the case that Σ is unknown and needs to be estimated. Zellner (1962) suggest a two–stage estimator whereby $\sigma_{ii'}$ in Equation (A.3) is estimated by the residuals from the single equation, ordinary least squares estimator. That is, $\hat{\sigma}_{ii'} = s_{ii'} = (\boldsymbol{\theta}_i - X_i\hat{\boldsymbol{\beta}}_i)'(\boldsymbol{\theta}_i - X_i\hat{\boldsymbol{\beta}}_i)$. With $\hat{\Sigma}$ now in place, the SUR estimator of $\boldsymbol{\beta}$ is the GLS estimator

$$\hat{\mathbf{b}} = (X'\hat{\Sigma}X)^{-1}X'\hat{\Sigma}^{-1}\boldsymbol{\theta}$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_1'\boldsymbol{\theta}_2' \cdots \boldsymbol{\theta}_m']'$, and X is the block–diagonal matrix $\{X_i\}$ in Equation (A.2). Details regarding the asymptotic distributions of $\hat{\mathbf{b}}$ and its corresponding variance–covariance matrix resulting from this two–stage procedure can be found in Zellner (1962).

Appendix B

FRT Appendix 2

B.1 Cost–complexity pruning

As mentioned in Section 3.1, the optimal sized tree is chosen by cost–complexity pruning. Without reproducing the minutiae of the procedure (which may be found in Breiman et al., 1984), a finite (nested) sequence of subtrees is found by the procedure from which one can select the optimal sized tree through resampling techniques. This sequence of subtrees is the set of trees $\mathcal{T}_\alpha$ which minimise the cost-complexity measure

$$D_\alpha = D(T) + \alpha|T|$$

and can be indexed by a finite sequence $\alpha = \alpha_0, \ldots, \alpha_S$, where $\alpha_0 = 0$ gives the maximal tree (for which there are no more possible splits), and $\alpha_S = \infty$ gives the root node. Here $D(T)$ is the deviance of the tree T (the sum of the terminal node deviances), and $|T|$ is the size (number of terminal nodes). We then choose the tree which minimises some measure of the error, either through cross–validation or bootstrap as explained next. Note, some authors (e.g. Hastie et al., 2009) advocate using an adjustment for the standard error of the loss function to choose the optimal tree, known as the 1-s.e. rule. We have found this definition to over–prune the trees when the response is a PDF and so instead we use the minimum error to prune the tree.

B.1.1 Cross–validation

For cross–validation, we first split the data into K subsets and proceed as follows:

- Remove the first subset and fit the full model to the remaining $K-1$ subsets

- For each α_s above, prune the full model, compute the predicted value $(\hat{f}^k_{\alpha_s}(x_i))$ for the observations in the removed subset and calculate

$$E_{\alpha_s}(k) = \sum_{i \in k^{\text{th}} \text{ subset}} L(y_i, \hat{f}^k_{\alpha_s}(x_i))$$

  the cross–validation error for the k^{th} subset, and some loss function L

- Repeat for each of the remaining $K-1$ subsets

The choice of the optimal sized tree is then the tree T_α corresponding to α_s which minimises the cross–validation error

$$\widehat{\text{Err}}^{\text{cv}}_{\alpha_s} = \sum_{k=1}^{K} E_{\alpha_s}(k)$$

Nerini and Ghattas (2007) use Euclidean distance as the loss function L. Our simulations have shown that this can still be an appropriate choice for L, however we advocate using the Kullback–Leibler divergence as used in the splitting criterion. That is, we recommend

$$L = \text{KL}(y_i, \hat{f}^k_{\alpha_s}(x_i)) + \text{KL}(\hat{f}^k_{\alpha_s}(x_i), y_i) \tag{B.1}$$

$$\text{where KL}(y_i, y_j) = \int y_i(t) \log\left(\frac{y_i(t)}{y_j(t)}\right) dt$$

B.1.2 Bootstrap $0.632+$

Let $o_i = (y_i, \mathbf{x}_i) \sim F, i = 1, \ldots, n$ be the observed data, where y_i is the response variable of interest (this paper assumes that y_i is a probability density function) and $\mathbf{x}_i$ are the corresponding covariates. Then assuming $\hat{F}$ to be the epmirical distribution which places mass of $1/n$ on each o_i, B bootstrap samples o^b_i are drawn from $\hat{F}$, and the full model is fit to each of the B boostrap replicates. Similar to cross–validation described above, for each α_s we prune the full model and define the leave–one–out bootstrap estimate of error, and

apparent error rate respectively as

$$\widehat{\text{Err}}_{\alpha_s}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}_{\alpha_s}^{*b}(x_i))$$
$$\overline{\text{err}}_{\alpha_s} = \sum_{r} L(y_i, \hat{f}_{\alpha_s}(x_i))$$

where $\hat{f}_{\alpha_s}^{*b}(x_i)$ is the predicted value for $x_i \in N(r)$ for the tree grown on the b^{th} bootstrap sample, C^{-i} is the set of bootstrap replicates that do not contain the i^{th} observation, $|C^{-i}|$ is the number of bootstrap replicates that do not contain the i^{th} observation and $\hat{f}_{\alpha_s}(x_i)$ is the predicted value for tree grown using the whole data set. Define as well, the no–information error rate as

$$\hat{\gamma}_{\alpha_s} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} L(y_i, \hat{f}_{\alpha_s}(x_i))$$

and the relative overfitting rate as

$$\hat{R}_{\alpha_s} = \begin{cases} (\widehat{\text{Err}}_{\alpha_s}^{(1)} - \overline{\text{err}}_{\alpha_s})/(\hat{\gamma}_{\alpha_s} - \overline{\text{err}}_{\alpha_s}) & \text{if } \widehat{\text{Err}}_{\alpha_s}^{(1)}, \overline{\text{err}}_{\alpha_s} > \hat{\gamma}_{\alpha_s} \\ 0 & \text{otherwise.} \end{cases}$$

Letting $\widehat{\text{Err}}_{\alpha_s}^{(1)'} = \min(\widehat{\text{Err}}_{\alpha_s}^{(1)}, \hat{\gamma}_{\alpha_s})$ the 0.632+ bootstrap estimate of error (Efron and Tibshirani, 1997) for tree $\mathcal{T}_{\alpha_s}$ is given by

$$\widehat{\text{Err}}_{\alpha_s}^{(0.632+)} = \widehat{\text{Err}}_{\alpha_s}^{(0.632)} + (\widehat{\text{Err}}_{\alpha_s}^{(1)'} - \overline{\text{err}}_{\alpha_s})\frac{0.368 \cdot 0.632 \cdot \hat{R}_{\alpha_s}}{1 - 0.368 \cdot \hat{R}_{\alpha_s}} \qquad \text{(B.2)}$$
$$\text{where } \widehat{\text{Err}}_{\alpha_s}^{(.632)} = 0.368\overline{\text{err}}_{\alpha_s} + 0.632\widehat{\text{Err}}_{\alpha_s}^{(1)}$$

We choose the optimal sized tree to be the tree T_α corresponding to α_s which minimises (B.2), again choosing loss function L as that in Equation (B.1).

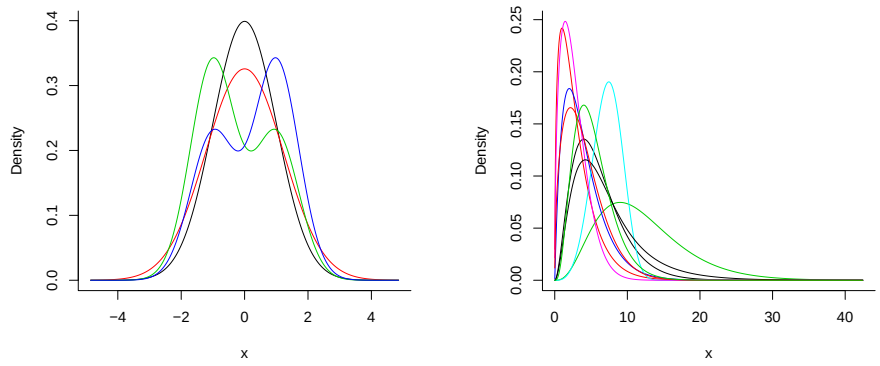
B.2 Further simulation results for FRT

In this appendix we include extra results from the simulation study (Section 3.3.1). We noted in Section 3.2.3 that cross–validation proved to be too variable to estimate the tree–size penalty and so we used the bootstrap $0.632+$. Table B.1 compares the standard deviation of the number of terminal nodes in Model 3 for each deviance, when using both the bootstrap $0.632+$ and cross–

validation to estimate the tree–size penalty. Figure B.1 displays the theoretical distributions used for Models 1 and 2 in the simulation study.

Table B.1: Comparison of root mean squared error of the number of terminal nodes for Model 3 when using bootstrap $0.632+$ (Appendix B.1.2) vs. cross–validation (Appendix B.1.1); $M = 200$.

	Bootstrap $0.632+$			Cross–validation		
m_i	$D_1(r)$	$D_2(r)$	$D_3(r)$	$D_1(r)$	$D_2(r)$	$D_3(r)$
15	0.53	0.58	1.53	1.38	0.99	3.44
25	0.53	0.63	1.56	1.69	1.43	1.91
50	0.58	0.76	1.58	7.52	1.77	2.58
75	0.65	0.80	1.77	2.20	1.95	3.26
100	0.64	1.15	1.68	1.90	1.58	1.73
150	0.63	1.08	1.71	2.86	2.07	1.77
200	0.71	0.79	1.72	1.97	2.10	1.48
250	0.63	1.02	1.62	1.77	1.78	3.39
500	0.80	0.93	1.67	2.50	2.14	1.49
1000	0.76	1.20	1.70	2.22	1.87	1.71



(a) Model 1 (b) Model 2

Figure B.1: Theoretical distributions used in the simulation study. Models 1 and 2 (see Table 3.2 for details).

Appendix C

LFLM Appendix

C.1 Conditional expectation of the FPC scores

Noting that $\phi_l(d,t)$, $X(s)$ and $\eta(s)$ are a smooth bounded functions in $L_2(\mathcal{D})$, swapping integration and summation is not an issue. The integral form of the conditional expectation of the FPC scores $b_k(t)$ in Equation (4.9) is given by

$$\begin{aligned}
E\left[b_l(t)|X\right] &= \int_{\mathcal{D}} E\left\{f(d,t) - \mu(d,t)|X\right\}\phi_k(d,t)\,\mathrm{d}d \\
&= \int_{\mathcal{D}} \left[\sum_{k=1}^{\infty}\left\{\int_{\mathcal{T}} \beta_k(t,s)\left(X(s)-\eta(s)\right)\,\mathrm{d}s\right\}\phi_k(d,t)\right]\phi_l(d,t)\,\mathrm{d}d \\
&= \int_{\mathcal{T}} \sum_{k=1}^{\infty}\beta_k(t,s)\left(X(s)-\eta(s)\right)\left\{\int_{\mathcal{D}}\phi_k(d,t)\phi_l(d,t)\,\mathrm{d}d\right\}\,\mathrm{d}s \\
&= \int_{\mathcal{T}} \beta_l(t,s)\left(X(s)-\eta(s)\right)\,\mathrm{d}s
\end{aligned}$$

C.2 Leave one group out cross–validation

To allow for any within–group correlation, we choose bandwidths for the LFLM by leave one group out cross–validation. For the mean smoother, $\hat{\mu}(d,t)$ (Equation 4.14), let $\hat{Y}_{-i}(d,t_{ij})$ be the predicted curve of the i^{th} observation at time t_{ij}, made by removing all observations in group i from the

smoother:

$$\hat{Y}_{-i}(d,t_{ij}) = \sum_{k\neq i}^{N}\sum_{j=1}^{N_k} W_{kj}(t)Y_k(d,t_{kj}), \text{ where}$$
$$W_{kj}(t) = \frac{K_{h_\mu}(t-t_{kj})}{\sum_{k\neq}^{N}\sum_{j=1}^{N_k} K_{h_\mu}(t-t_{kj})}$$

The cross-validation choice of the bandwidth h_μ is then

$$\hat{h} = \underset{h_\mu}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{j=1}^{N_i} \left|\left|Y_i(d,t_{ij}) - \hat{Y}_{-i}(d,t_{ij})\right|\right|^2$$

C.3 Permutation testing

As described in Section 4.2.1, the significance of the functional predictor can be tested via a permutation p–value. We describe the implementation of this test briefly in this appendix. The 'observed' value of the F–statistic is found directly by fitting the functional regression model (4.2) and calculating F_{obs} using Equation (4.23).

Under the null hypothesis that the functional predictor X has no effect, the permutation principle says that the pairing of any particular $(X, Y(d,t))$ is random. To hold with this principle, we define our permutations over the group–level (stands in our application) so that a permutation of the functional predictors is performed. Specifically, denote as in Section 4.1, the functional response at time t_{ij} as $Y_i(d,t_{ij})$ and the functional predictor as $X_i(t_{ij})$, and further denote the whole–of–stand observation as

$$\left(\{Y_i(d,t_{i1}),\ldots,Y_i(d,t_{iN_i})\},\{X_i(t_{i1}),\ldots,X_i(t_{iN_i})\}\right) = (\mathbf{Y}_i,\mathbf{X}_i)$$

for $i = 1,\ldots,N$. Further, let $\boldsymbol{\chi} = \{\mathbf{X}_1,\ldots,\mathbf{X}_N\}$ and let $\boldsymbol{\chi}_b$ be a random permutation of the set $\boldsymbol{\chi}$ with elements $\{\mathbf{X}_{b,k(1)},\ldots,\mathbf{X}_{b,k(N)}\}$, where $k(i)$ denotes the index after permutation. The typical data pair after the b^{th} random permutation becomes $\left(\mathbf{Y}_i,\mathbf{X}_{b,k(i)}\right)$, which are used to estimate $\beta(d,s,t)$ in the functional linear model (4.2). We make note here that it is not necessary that $Y(d,t)$ and $X(t)$ are measured the same number of times (and indeed at the same time), see Yao et al. (2005b) for details. This is important for the permutation test, as it is likely that under permutation, the lengths of $\mathbf{Y}$ and $\mathbf{X}$ will be different.

Following Cardot et al. (2007), we will also condition on the bandwidths used in the mean and covariance smoothers (Equations 4.14 and 4.15) as well as the dimensions of the functional coefficient (Equation (4.21)). Using the estimated values of these from fitting the 'observed' model, we then fit $b = 1, \ldots, B$ permutation models of the form

$$E\left[Y(d)|t, X\right] = \mu(d,t) + \int_{\mathcal{T}} \beta_b(d,s,t)\left[X_b(s) - \eta(s)\right] \mathrm{d}s \tag{C.1}$$

where X_b and β_b denote the fact that we have used permuted data in fitting the model. Letting $\hat{Y}_b(d,t) = \widehat{E}\left[Y(d,t)|X_b\right]$ be the predicted value of the density $Y(d,t)$ under the permutation model (C.1), RSS_1 in Equation (4.23) becomes

$$\mathrm{RSS}_{1b} = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\int_{\mathcal{D}}\left[Y_i(d,t_{ij}) - \hat{Y}_b(d,t_{ij})\right]^2 \mathrm{d}d$$

and

$$F_b = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_{1b}}{\mathrm{RSS}_{1b}}.$$

The p–value (4.24) follows directly.

C.4 Further results

This appendix provides an extension to the results of Yao et al. (2005b) and Cardot (2007) that are required for Theorem 1. The term $\sum_{k=1}^{K}\sum_{l=1}^{L}\delta_{kl}^{b_k}A_{\delta_{kl}^{b_k}}\left(\sqrt{n}h_k^2 - A_{\delta_{kl}^{b_k}}\right)^{-1}$ in $A(n)$ of Theorem 1 represents the convergence of the eigenfunctions of the functional principal coefficients $b_k(t)$. Specifically, following Yao et al. (2005b), $\hat{\varphi}_{kl}(t)$ can be chosen such that

$$\sup_{t\in\mathcal{T}}|\hat{\varphi}_{kl}(t) - \varphi_{kl}(t)| = O_p\left(\frac{\delta_{kl}^{b_k}A_{\delta_{kl}^{b_k}}}{\sqrt{n}h_k^2 - A_{\delta_{kl}^{b_k}}}\right)$$

and it follows from this and Equations (38) and (39) of Yao et al. (2005b) that

$$|\hat{\sigma}_{klm} - \sigma_{klm}| = O_p\left(\max\left\{\frac{\delta_m^X A_{\delta_m^X}}{\sqrt{n}h_X^2 - A_{\delta_m^X}}, \frac{\delta_{kl}^{b_k}A_{\delta_{kl}^{b_k}}}{\sqrt{n}h_k^2 - A_{\delta_{kl}^{b_k}}}, \frac{1}{\sqrt{n}h_{l1}h_2}\right\}\right)$$

From Cardot (2007) and Yao et al. (2005b), we have the following:

Lemma 3.

$$\sup_d \left|\hat{\phi}_k(d,t) - \phi_k(d,t)\right| = O_p\left[\left(h_{Y_1}^{\beta} + h_{Y_2}^{\alpha} + \left\{\frac{\log n}{n\min(h_{Y_1}, h_{Y_2})}\right\}^{1/2}\right)\kappa_l\right]$$

Proof. Note that from Cardot (2007), we have

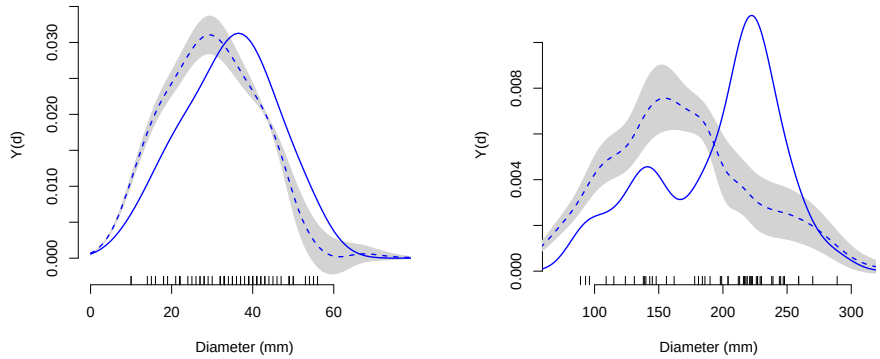
$$\begin{aligned}
&\left|\hat{\lambda}_k(t)\hat{\phi}_k(d,t) - \lambda_k(t)\phi_k(d,t)\right| \\
&\quad= \left|\int_{\mathcal{D}} \hat{\gamma}(t,e,d)\hat{\phi}_k(e,t)\,\mathrm{d}e - \int_{\mathcal{D}} \gamma(t,e,d)\phi_k(e,t)\,\mathrm{d}e\right| \\
&\quad\leq \int_{\mathcal{D}}\left|\hat{\gamma}(t,e,d) - \gamma(t,e,d)\right| \cdot \left|\hat{\phi}_k(e,t)\right|\mathrm{d}e \\
&\qquad+ \int_{\mathcal{D}}\left|\gamma_k(t,e,d)\right| \cdot \left|\hat{\phi}_k(e,t) - \phi_k(e,t)\right|\mathrm{d}e \\
&\quad\leq \left\{\int_{\mathcal{D}}\left[\hat{\gamma}(t,e,d) - \gamma(t,e,d)\right]^2\,\mathrm{d}e\right\}^{1/2} \\
&\qquad+ \left\{\int_{\mathcal{D}}\gamma_k^2(t,e,d)\,\mathrm{d}e\right\}^{1/2} \cdot \left|\left|\hat{\phi}_k(e,t) - \phi_k(e,t)\right|\right|_{\mathcal{H}}
\end{aligned}$$

Now by Theorem 1 and Corollary 1 of Cardot (2007), and assuming, without loss of generality, $\lambda_k(t) > 0$, and that $\kappa_j = \sup_t \delta_j$ is a bounded sequence where δ_j are as defined in Cardot (2007)

$$\left|\frac{\hat{\lambda}_k(t)\hat{\phi}_k(d,t)}{\lambda_k(t)} - \phi_k(d,t)\right| = O_p\left[\left(h_{Y_1}^{\beta} + h_{Y_2}^{\alpha} + \left\{\frac{\log n}{n\min(h_{Y_1}, h_{Y_2})}\right\}^{1/2}\right)\kappa_l\right]$$

uniformly in d, where h_{Y_1} and h_{Y_2} are the bandwidths used in the functional mean smoother (Equation 4.14) and functional covariance smoother (Equation 4.15) respectively. Combined with Corollary 1 of Cardot (2007), the result follows. □

C.5 Extra figures



(a) $t = 1.61$ years (b) $t = 7.90$ years

Figure C.1: 95% confidence intervals (shaded grey) for $E\left[f(d,t)|X^*\right]$, when a) $t = 1.61$ years, and b) $t = 7.90$ years. The solid line is the observed value, and the dashed line is the predicted value.

Appendix D

Code Appendix

This appendix provides instructions on how to replicate the simulation results in Chapters 3 and 5. The code to perform the various calculations is in the attached CD. The code relating to Chapter 5 makes use of the *PACE* package (http://anson.ucdavis.edu/~ntyang/PACE/) and has been included in the attached CD for ease of use. References to scripts etc. assume that the user has saved the folder `Code` in the attached CD to their home directory on a Mac/Unix-alike system.

D.1 Chapter 3 code

The code to run the simulations for this chapter is included in the `Chapter_3` folder of the attached CD, and requires R (R Development Core Team, 2009) to run. The following packages need to be installed on the machine used to run the code

```
require(MASS)
require(mvpart)
require(nor1mix)
require(flexmix)
require(caTools)
```

Make sure that an R session is started within the same folder that the scripts are located, or the working directory is changed to that folder, e.g.

```
setwd("~/Code/Chapter_3")
```

then source the driver file (which includes all the necessary scripts and functions)

```
source("Main.r")
```

To run the simulation for Model 1 (Table 3.2), with $M = 50$ observations and $n_i = 25$ samples per observation the following may be used (note that for a large number of repetitions and/or bootstraps, this may take some time)

```
M <- 50
n.i <- 25
boots <- 50                    # number of
                               # bootstraps
numreps <- 100                 # number of
                               # simulation reps
sims <- run.sim(data.type = 1, n.i, M, B = boots,
                numreps = numreps)
results <- compile.fun()
show(results)
```

D.2 Chapter 5 code

The code to run the simulations for this chapter is included in the `Chapter_5` folder of the attached CD, and requires MATLAB (MATLAB, 2011) to run. In the following code examples, we assume that the code has been saved in the user's home directory of a Mac/Unix-alike system. The scripts to run each simulation are included in the `Chapter_5` folder, with filenames corresponding to the scenarios simulated (see Table 5.1). We describe the code for Scenario 1a here. The first thing to do is to tell MATLAB where all relevant code exists

```
addpath(genpath('~/Code/Chapter_5/PACE/'));
addpath(genpath('~/Code/Chapter_5/Func_Pred'));
```

We can then use the following to run the Scenario 1a simulation with 100 repetitions, true population $N = 50$ and 100 bootstrap replications (note that this may take some time)

```
% Scenario 1a
model = 'Mu1';
% Number of simulation reps
nreps = 100;
% Number actually in the population
ncohort = 50;
% Number of bootstraps
B = 100;
% Run the simulation
Sim_Data;
```

The next lot of code summarises the simulation results

```
% Viewing the results.
% Mean of N from EM/FPCA, gold standard and naive
% methods.
mean([N_2 N_3 N_4])
% Mean of beta1 from EM/FPCA, gold standard and
% naive methods.
mean([par2(:, 2) par3(:, 2) par4(:, 2)])
% 95% Bootstrap CI of beta1 from EM/FPCA, gold
% standard and naive methods.
mean(CI_par_fpca(:, 3:4))
mean(CI_par_gold(:, 3:4))
mean(CI_par_first(:, 3:4))
```