FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. Michaela Cichrová

Statistical inference in varying coefficient models

Department of Probability and Mathematical Statistics

Supervisor of the master thesis:	doc. RNDr. Matúš Maciak, Ph.D.
Study programme:	Mathematics
Study branch:	Probability, Mathematical Statistics and Econometrics

Prague 2023

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

I would like to thank doc. RNDr. Matúš Maciak, Ph.D. for his time and valuable advice and my family for their support and patience.

Title: Statistical inference in varying coefficient models

Author: Bc. Michaela Cichrová

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Matúš Maciak, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this master thesis we study varying coefficient models, which is a class of models that allow the coefficients to be smooth functions of some effect-modifying variable. We introduce the models in a broader context and then focus only on longitudinal settings. We consider two spline-based methods to estimate the coefficient functions, the polynomial spline approach and the smoothing spline approach. For the polynomial spline approach, we derive its asymptotic properties, which we use to construct asymptotic confidence intervals and bands. We assess the performance of the confidence bands in a small simulation study, considering two slight modifications of the construction.

Keywords: varying coefficient models, polynomial spline approach, longitudinal data, confidence bands

Contents

Introduction

Regression analysis is a very popular statistical technique used to explore the association between a dependent variable (response) and some independent variables (covariates, predictors) and to determine how the independent variables affect the dependent variable. Depending on the underlying assumptions and specifications, there are three main types of regression models: parametric, nonparametric, and semiparametric.

Parametric regression models assume that the dependent variable has a specific functional form that depends on the covariates and some unknown parameters. An example is the linear regression model, which assumes that the response can be modelled as a linear function of the predictors. Another example is the generalized regression model, which extends the linear regression framework to be able to model various types of dependent variables, such as binary or count data, by using link functions to capture the relationship between the conditional expected value of the response and the linear function of the covariates. Although parametric models are popular as they have the advantage of simplicity and straightforward interpretation, they can be unrealistic in certain applications, and model misspecification may lead to a considerable bias.

In contrast, nonparametric models do not assume any specific functional form for the dependent variable. Nonparametric regression models can fit complex and nonlinear relationships without imposing any restrictive assumptions, however, their usage is limited as the curse of dimensionality makes them practically unusable for high number of dimensions.

To overcome these challenges (the often too restrictive assumptions of the parametric models and the curse of dimensionality of the nonparametric models), semiparametric models have been proposed. They assume that the response has a partially specified functional form that depends on some parametric and some nonparametric components.

One of the most popular semiparametric models is the varying coefficient model introduced by Hastie and Tibshirani (1993). The varying coefficient model can be seen as an extension to the generalized linear model by allowing the coefficients to be smooth functions of some independent variables.

An important issue is the estimation of the coefficient functions. Several estimation methods have been proposed over the years, the most popular are the local polynomial regression and two spline approximation methods: the polynomial spline approach and the smoothing spline approach.

The spline approximation methods use the fact that any smooth function can be approximated using basis splines. Basis splines are composed of a set of basis functions, such as polynomial functions, with each one defined over a unique interval determined by knots. These knots function as points where individual segments come together smoothly to create a continuous curve. The challenge lies in selecting the appropriate number and placement of the knots, as well as in determining the optimal smoothing parameter, which controls the trade-off between the goodness-of-fit and the smoothness of the estimate function in the case of the smoothing spline estimator. The polynomial spline estimator does not take the smoothness of the estimate into account, it is based solely on the goodness-of-fit. The interpretability and nice asymptotic properties of the spline estimation methods make them particularly appealing and they are therefore the focus of this thesis.

The thesis focuses on the varying coefficent models in the longitudinal setting and a construction of simultaneous confidence bands for the coefficient function. In the first chapter, we introduce the varying coefficient models and provide some examples of such models. In the second chapter, we look closely at the coefficient function estimation methods based on spline approximation in longitudinal settings. We introduce different methods for the selection of the smoothing parameters. The third chapter is the main chapter of this thesis and focuses on the confidence intervals and bands for the coefficient functions, using pointwise asymptotic normality as our primary tools. The fourth chapter presents results of a corresponding simulation study.

1. Varying coefficient models

In this chapter, we define the varying coefficient models first in standard setting and in longitudinal setting and provide a range of specific cases of these models. We also discuss the advantages and disadvantages of such models.

1.1 General framework

We start with a motivating example from the medical field, but varying coefficient models are commonly used in many other areas, such as finance, environmental sciences, and social sciences. One of the key goals in medical studies is to discover how different treatments affect patient outcomes (e.g. blood pressure, or recovery time). It is not sufficient to compare the outcomes between different treatment groups alone, because the effect of the treatment differs greatly depending on a number of patient-specific factors. While interactions can be added to traditional generalized linear models to capture some of the heterogeneity in treatment effects, the choice of interaction terms may not fully account for the complexity of the treatment-outcome relationship. That has the consequence of biased estimates and inaccurate predictions. To address the limitations, varying coefficient models allow the effect of the treatment variable to vary as a smooth function of some patient-specific characteristics, that could be for example age, time since the beginning of the treatment, or white blood cell count.

Formally, let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space, random variable $Y : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be the dependent variable and random variables $R_1, \ldots, R_p, X_1, \ldots, X_p : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, for $p \in \mathbb{N}$, be the covariates. Our goal is to model the conditional expected value of Y given the covariates $X_1, \ldots, X_p, R_1, \ldots, R_p$, denoted as $\mathsf{E}\left(Y \,|\, X_1, \ldots, X_p, R_1, \ldots, R_p\right)$. We assume that the effects of the variables X_j are not fixed, but instead they depend on the corresponding covariates R_j, for $j = 1, \ldots, p$. This dependency is not arbitrary, we model it using a set of real-valued smooth functions $\beta_j : \mathbb{R} \to \mathbb{R}$ such that each predictor's effect can adjust according to its associated covariate. Furthermore, we assume a linear relationship between the predictors and the conditional expectation of the dependent variable.

1.2 Standard model formulation

Let us define the varying coefficient models according to Hastie and Tibshirani (1993).

Definition 1. *Let $p \in \mathbb{N}$ be a constant and $Y, X_1, \ldots, X_p, R_1, \ldots, R_p$ be real-valued random variables. A model is referred to as a varying coefficient model if it takes the form*

$$Y = \sum_{j=1}^{p} \beta_j(R_j) X_j + \varepsilon, \tag{1.1}$$

where $\beta_1(\cdot), \ldots, \beta_p(\cdot)$ are some real-valued smooth functions, ε is a random error term satisfying $\mathsf{E}\left(\varepsilon \,|\, X_1, \ldots, X_p, R_1, \ldots, R_p\right) = 0$ and $var\left(\varepsilon \,|\, X_1, \ldots, X_p, R_1, \ldots, R_p\right) = \sigma^2 > 0$.

According to Definition 1, the conditional expectation of the response Y, i.e $\mathsf{E}\left(Y|, X_1, \ldots, X_p, , R_1, \ldots, R_p\right)$, depends on the covariates $X_1, \ldots, X_p$ through some (unspecified) smooth functions $\beta_1\left(\cdot\right), \ldots, \beta_p\left(\cdot\right)$ of covariates $R_1, \ldots, R_p$. The covariates $R_1, \ldots, R_p$ are known as the effect-modifying (random) variables. The functions $\beta_j\left(\cdot\right)$ for $j = 1, \ldots, p$, $p \in \mathbb{N}$, imply some special kind of interaction between X_j and R_j.

Some well-known regression models can be considered as a special case of Model (1.1).

The first example is the linear regression model (with or without interactions). By assuming that the functions $\beta_j(\cdot)$ in Definition 1 are constant, i.e., $\beta_j(\cdot) \equiv \beta_j$, where $\beta_j \in \mathbb{R}$ for $j = 1, \ldots, p$, Model (1.1) can be rewritten as

$$Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon. \tag{1.2}$$

That is an equation of a linear regression model without interactions. Further, consider the equation of Model (1.1) and set $p = m(m+1)$ for some $m \in \mathbb{N}$. The model can be rewritten by indexing over i and j ranging from 1 to $m+1$ and m, respectively:

$$Y = \sum_{i=1}^{m+1} \sum_{j=1}^{m} \beta_{ij}\left(R_{ij}\right) X_{ij} + \varepsilon. \tag{1.3}$$

Here is how to set the terms

1. Set $X_{ij} = X_i X_j$ for $i = 1, \ldots, m+1$, $j = 1, \ldots, m$ for some random variables $X_1, \ldots, X_{m+1}$. Additionally, $X_{m+1} = 1$ almost surely.

2. Set $\beta_{ij}(\cdot) \equiv \beta_{ij} \in \mathbb{R}$, i.e., as a constant function. Also, set

   (a) $\beta_{ij} = 0$ when $i \leq j$.

   (b) $\beta_j = \beta_{(m+1)j}$.

Substituting these terms into Equation (1.3) yields

$$Y = \sum_{j=1}^{m} \beta_j X_j + \sum_{i=1}^{m} \sum_{j=1}^{i-1} \beta_{ij} X_i X_j + \varepsilon, \tag{1.4}$$

which is an equation of a linear regression model with interactions.

The generalized additive model is another example of varying coefficient models. This model assumes that the response variable Y depends nonlinearly on p covariates $R_1, \ldots, R_p$, and their effects are additive and smooth. The model equation can be expressed as

$$Y = \beta_1(R_1) + \cdots + \beta_p(R_p) + \varepsilon. \tag{1.5}$$

This model can be considered as a special case of Model (1.1) by setting $X_j = 1$ almost surely for $j = 1, \ldots, p$. For more details, see Hastie and Tibshirani (1986).

Definition 1 is the first definition of the varying coefficient model provided in the literature and is sufficient for the purpose of this thesis. However, it is important to note that the definition could be extended to allow for multivariate effect-modifying random vectors.

Another extension is to allow for some correlation structure in the model.

1.3 Longitudinal model and other additional extensions

Let us return to the motivating example. Medical studies are often conducted by collecting data from the same subjects repeatedly over time. This results in within-subject correlated observations. Several methods are available for analysing longitudinal data, including generalized estimating equations (GEE) (see Hardin and Hilbe, 2003) and mixed-effect regression models (see Hardin and Hilbe, 2003). The varying coefficient models framework can be extended to provide an alternative approach to longitudinal data modelling.

We define the varying coefficient model according to Huang et al. (2002). Extension of the definition by using some link function is also possible.

Definition 2. *Let $\{Y(t), t \in \mathbb{R}\}$ be a real-valued response stochastic process and let $\{\mathbf{X}(t) = (X_1(t), \ldots, X_p(t)), t \in \mathbb{R}\}$ be an $\mathbb{R}^p$-valued covariate stochastic process. We define the longitudinal varying coefficient model as follows:*

$$Y(t) = \sum_{j=1}^{p} \beta_j(t) X_j(t) + \varepsilon(t), \tag{1.6}$$

where $\{\varepsilon(t), t \in \mathbb{R}\}$ is a zero-mean stochastic process and $\{\varepsilon(t), t \in \mathbb{R}\}$ and $\{\mathbf{X}(t) = (X_1(t), \ldots, X_p(t)), t \in \mathbb{R}\}$ are independent.

The functions $\beta_j(\cdot)$ capture the time-varying effects of the covariates on the response. We primarily concentrate on this longitudinal extension in this thesis.

Since their introduction 30 years ago, varying coefficient models have gained significant popularity due to their flexibility and ability to capture complex relationships. There are numerous other examples of those models, let us list just some of them.

1. **Survival Analysis**: In survival analysis, the varying coefficient model can be used to analyze the relationship between the hazard function and some time-varying covariates. In this case, the model is often referred to as a time-varying Cox proportional hazards model. (see Fisher and Lin, 1999)

2. **Spatial Analysis**: In spatial analysis, the varying coefficient model can be used to model the relationship between the response variable and spatially-varying covariates. An example is the geographically weighted regression model that allows the relationships between the response variable and the predictors to vary across space. (see Brunsdon et al., 1996)

3. **Dynamic Regression Models**: These models capture complex time-varying relationships in time series data. They are able to handle non-linear patterns. The threshold autoregressive (TAR) model is one example, allowing for different autoregressive processes based on a threshold variable. Such models are widely used in finance. (see Cai et al., 2000b)

1.4 Pros and Cons

Choosing an appropriate model is a critical step in the statistical analysis. The choice depends on the structure of the data, and the question at hand. When considering the use of the varying coefficient model, it is important to weigh the pros and cons of the approach to determine its suitability for the specific problem.

Pros:

1. **Flexibility**: Varying coefficient models can capture complex, non-linear relationships between the response variable and covariates. By allowing the coefficients to vary as a smooth function of some chosen covariates, these models can model relationships that may not have been possible with simpler parametric models.

2. **Interpretability**: The estimated coefficient functions provide a clear picture of how the effects of covariates change with the variable of interest.

Cons:

1. **Estimation and Inference Complexity**: Both the estimation of the varying coefficient functions and the subsequent statistical inference are more complex and computationally demanding compared to parametric models.

2. **Overfitting**: Due to their flexibility, varying coefficient models can potentially overfit the data, that is fit the data ”too well” by capturing noise rather than the underlying trends.

3. **Potential interpretation challenges**: Although varying coefficient models offer a high level of interpretability, they can be challenging to interpret when the number of covariates is large or when there are complex interactions between covariates.

4. **Research Stage**: The varying coefficient models are newer and less studied compared to parametric models. The methodology is complex and further research is needed to make the varying coefficient models usable in broad research.

2. Estimation methods

Let us consider the longitudinal varying coefficient model in (1.6) discussed in Chapter 1. In this Chapter, we look closely at two methods of estimating the coefficient functions: the smoothing spline method and the polynomial spline method. They are closely related, the polynomial spline method can be considered a special case of the smoothing spline method.

Consider a longitudinal random sample of size $\sum_{i=1}^{n} n_i = N \in \mathbb{N}$

$$\left\{\left(Y_i\left(T_{il}\right), X_{i1}\left(T_{il}\right), \ldots, X_{ip}\left(T_{il}\right), T_{il}\right), i = 1, \ldots, n, l = 1, \ldots, n_i\right\}, \tag{2.1}$$

where each $(p+2)$-tuple $(Y_i(T_{il}), X_{i1}(T_{il}), \ldots, X_{ip}(T_{il}), T_{il})$ is a realization from the joint distribution $F_{(Y(T), X_1(T), \ldots, X_p(T), T)}$. In the given notation, T_{il} represents the l-th measurement time of the $i-th$ subject, $\mathbf{X}_i\left(T_{il}\right) = \left(X_{i1}\left(T_{il}\right), \ldots, X_{ip}\left(T_{il}\right)\right)^\top$ and $Y_i\left(T_{il}\right) = Y_{il}$ are the i-th subject's observed covariates and outcome measured at time T_{il}.

The data satisfy the varying coefficient model in (1.6) in the longitudinal setting, if for each subject i and measurement l we have:

$$Y_i(T_{il}) = X_{i1}(T_{il})\beta_1\left(T_{il}\right) + \cdots + X_{ip}(T_{il})\beta_p\left(T_{il}\right) + \varepsilon_i\left(T_{il}\right), \tag{2.2}$$

where $\left(\varepsilon_i\left(T_{il}\right), \ldots, \varepsilon_i\left(T_{in_i}\right)\right)^\top$ are realizations of the mean-zero random process $\varepsilon(t)$. The coefficient functions $\boldsymbol{\beta}\left(\cdot\right) = \left(\beta_1\left(\cdot\right), \ldots, \beta_p\left(\cdot\right)\right)^\top$ are unspecified. Our goal is to find functions $\hat{\boldsymbol{\beta}}\left(\cdot\right) = \left(\hat{\beta}_1\left(\cdot\right), \ldots, \hat{\beta}_p\left(\cdot\right)\right)^\top$ using the sample data which provide a reasonable approximation of the true underlying functions $\boldsymbol{\beta}$. We denote $\varepsilon_{il} = \varepsilon_i\left(T_{il}\right)$, $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i1}, \ldots, \varepsilon_{in_i}\right)^\top$, and $\boldsymbol{\varepsilon} = \left(\boldsymbol{\varepsilon}_1^\top, \ldots, \boldsymbol{\varepsilon}_n^\top\right)^\top$. The times of the measurements are random, all results in this thesis however hold also for deterministic times.

Denote by $\mathcal{C}^d$ the space of all smooth functions of order $d \in \mathbb{N}_0$. Theoretically, we could obtain the estimators $\hat{\boldsymbol{\beta}}$ as a smooth function of order d minimizing the (possibly weighted) sum of squared errors, i.e

$$\hat{\boldsymbol{\beta}}(\cdot) = \underset{\beta_j(\cdot) \in \mathcal{C}^d,\, j=1,\ldots,p}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{l=1}^{n_i} \left[Y_i(T_{il}) - \sum_{j=1}^{p} X_{ij}(T_{il})\beta_j(T_{il})\right]^2. \tag{2.3}$$

Estimating the varying coefficient model directly on the space of smooth functions is theoretically possible, but using spline approximation offers practical advantages. The space of all smooth functions is infinite-dimensional, which makes the optimization problem challenging to solve and computationally infeasible. Directly searching through this space can lead to overfitting, the resulting model can fit the training data well but perform poorly on new data. Additionally, solutions found in the space of smooth functions may lack interpretability, making it difficult to draw meaningful conclusions from the model. Spline approximation addresses these issues.

2.1 Polynomial spline approach

The first estimation method we consider is the polynomial spline method introduced by Huang et al. (2002) in the context of longitudinal data.

To estimate the coefficient functions, we approximate the unknown functions $\beta_j(\cdot)$ by some basis functions and then the estimate is obtained by minimizing the sum of squared errors. A common choice is to use B-splines.

2.1.1 B-spline basis

We start with an introduction of B-splines, as they are the key concept of the spline estimation methods. To be able to define B-splines, it is essential to introduce knot sequences.

Definition 3. *A knot sequence $\boldsymbol{\xi} \in \mathbb{R}^{M+2}$ of length $M+2$, $M \in \mathbb{N}$ is a non-decreasing sequence of real numbers, $\boldsymbol{\xi} := \{\xi_i\}_{i=1}^{M+2} = \{\xi_1, \dots, \xi_{M+2} : -\infty < \xi_1 \leq \xi_2 \leq \dots \leq \xi_{M+2} < \infty\}$. The elements of $\boldsymbol{\xi}$ are called knots. The knots are said to be distinct if $\xi_i \neq \xi_j$ for any $i \neq j$. For distinct knots, the set of M knots $\{\xi_2, \xi_3, \dots, \xi_{M+1}\}$ can be referred to as internal knots, while ξ_1 and ξ_{M+2} are the end points.*

We only consider distinct knots and refer to them simply as knots.

Provided that the number of internal knots $\boldsymbol{\xi}$, M, satisfies $M \geq d$, $d \in \mathbb{N}_0$, we can define B-spline basis of degree d over the knots $\boldsymbol{\xi}$.

According to De Boor (1972), we can define the B-spline basis (one of the possible basis of the space of the polynomial splines) as

Definition 4. *Let $M, d \in \mathbb{N}_0 : M \geq d$. A B-spline basis of degree $d \in \mathbb{N}_0$ with $M+2$ knots $\boldsymbol{\xi} = \{\xi_j,\, j = 1, \dots, M+2 : -\infty < \xi < \dots < \xi_{M+2} < \infty\}$ is a set of non-negative B-spline basis functions*

$$\mathcal{B}_{d,\boldsymbol{\xi}} = \{B_{j,d,\boldsymbol{\xi}}(\cdot),\, j = 1, \dots, M+d+1\}.$$

Each B-spline basis function $B_{j,d,\boldsymbol{\xi}}(\cdot)$ is defined recursively as follows

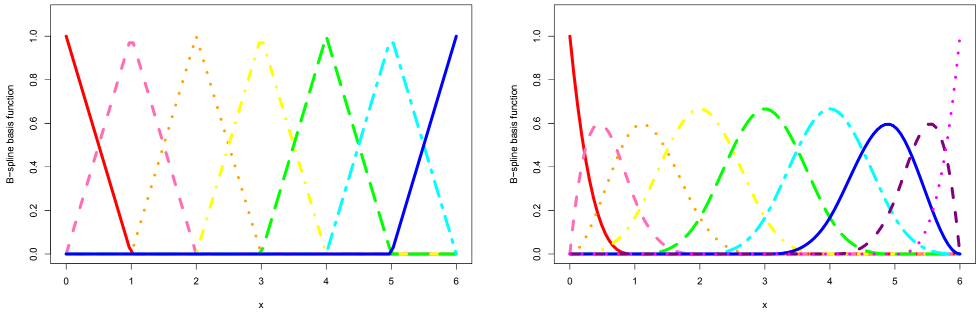
$$B_{j,0,\boldsymbol{\xi}}(x) = \begin{cases} 1 & \text{if } \xi_j \leq x < \xi_{j+1}, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$B_{j,k,\boldsymbol{\xi}}(x) = \frac{x - \xi_j}{\xi_{j+k} - \xi_j} B_{j,k-1,\boldsymbol{\xi}}(x) + \left(1 - \frac{x - \xi_j}{\xi_{j+k} - \xi_j}\right) B_{j+1,k-1,\boldsymbol{\xi}}(x),$$

for $k = 1, \dots, d$ and $j = 1, \dots, M+d+1$.

Roughly speaking, B-spline basis is a set of B-spline basis functions defined over a sequence of knots. Each B-spline basis function is a piecewise polynomial function of a given degree, and its value is non-zero only over a limited range defined by the knots. The sum of all B-spline basis functions at any given point is one. For more information about the B-splines see De Boor (1978).



(a) $d = 1$.

(b) $d = 3$.

Figure 2.1: B-spline basis functions for $d = 1$ and $d = 3$, and knots $\boldsymbol{\xi} = \{0, 1, 2, 3, 4, 5, 6\}$.

Any B-spline $B(\cdot)$ of degree $d \in \mathbb{N}_0$ with the knots $\{\xi_j,\, j = 1, \ldots, M+2\}$ can be written in terms of its B-spline basis as a linear combination of the basis B-spline functions

$$B(\cdot) = \sum_{j=1}^{M+d+1} \gamma_j B_j(\cdot),$$

for some $\gamma_j \in \mathbb{R}$, $j = 1, \ldots, M+d+1$. The B-spline functions of a given degree and knots form a linear space, the B-spline basis is the basis of the linear space.

It can be shown that any smooth function $f(\cdot)$ of degree d can be approximated on a closed interval $[a, b]$, $a \in \mathbb{R}, b \in \mathbb{R}, a < b$, by some B-spline function $b(\cdot)$ of degree d and sufficiently large number of knots, as

$$\sup_{z \in [a,b]} |f(z) - b(z)| \xrightarrow{M \to \infty} 0.$$

Let us assume that the observation times for all $i = 1, \ldots, n$, $l = 1, \ldots, n_i$ satisfy $T_{il} \in [a,\, b]$ and that the coefficient functions $\beta_j(\cdot)$ are $(d-1)$-times continuously differentiable, usually we take $d = 3$ (corresponds to cubic splines), or $d = 1$ (linear splines). Then we can approximate the coefficient functions using the B-spline basis functions of degree d, so that

$$\beta_j(t) \approx \sum_{k=1}^{K_j} \gamma_{jk} B_{jk}(t), \qquad K_j \in \mathbb{N}, \tag{2.4}$$

where $\{B_{jk}(\cdot), k = 1, \ldots, K_j\}$ is the B-spline basis from Definition 4 with a given degree and knots. The number $K_j \in \mathbb{N}$ represents the dimension of the corresponding B-spline linear space $\mathbb{G}_j$ (number of B-spline basis functions). As we know from Definition 4, K_j depends on the number of inner knots M_j and the B-spline degree d through the relation

$$K_j = M_j + d + 1.$$

To estimate the coefficient functions, we need to choose an appropriate B-spline basis $\mathcal{B}_j$ of a linear space $\mathbb{G}_j$, and estimate the parameters $\gamma_{j1}, \ldots, \gamma_{jK_j}$ by some $\hat{\gamma}_{j1}, \ldots, \hat{\gamma}_{jK_j}$ using the polynomial spline approach or the penalized spline approach (smoothing spline approach). The estimate is then given as

$$\hat{\beta}_j\left(r\right) = \sum_{k=1}^{K_j} \hat{\gamma}_{jk} B_{jk}(r). \tag{2.5}$$

2.1.2 Longitudinal model estimation

Suppose that smooth functions $\beta_j\left(\cdot\right)$ for $\jmath = 1, \ldots, p$ can be approximated by some B-spline function

$$\beta_j\left(\cdot\right) \approx \sum_{k=1}^{K_j} \gamma_{jk} B_{jk}(\cdot).$$

The approximation of $\boldsymbol{\beta}\left(\cdot\right)$ leads to the approximation of Model (2.2) as

$$\begin{aligned} Y_i(T_{il}) &= X_{i1}(T_{il}) \sum_{k=1}^{K_1} \gamma_{1k} B_{1k}(T_{il}) + \cdots + X_{ip}(T_{il}) \sum_{k=1}^{K_p} \gamma_{pk} B_{pk}(T_{il}) + \varepsilon_{il} \\ &= \sum_{j=1}^{p} \sum_{k=1}^{K_j} \gamma_{jk} B_{jk}(T_{il}) X_{ij}(T_{il}) + \varepsilon_{il}. \end{aligned}$$

To estimate the coefficient functions $\beta_1(\cdot), \ldots, \beta_p(\cdot)$, we need to estimate the coefficients $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma_1}^\top, \ldots, \boldsymbol{\gamma_p}^\top\right)^\top = \left(\gamma_{11}, \ldots, \gamma_{1K_1}, \ldots, \gamma_{p1}, \ldots, \gamma_{pK_p}\right)^\top$. That can be done by minimizing the sum of squared errors, which provides a natural measure of the goodness of fit.

Huang et al. (2002) suggested using a weighted sum of square errors to account for the within-subject correlation. Let $\hat{\boldsymbol{\gamma}} = \left(\hat{\boldsymbol{\gamma}}_1^\top, \ldots, \hat{\boldsymbol{\gamma}}_p^\top\right)^\top = \left(\hat{\gamma}_{11}, \ldots, \hat{\gamma}_{1K_1}, \ldots, \hat{\gamma}_{p1}, \ldots, \hat{\gamma}_{pK_p}\right)^\top$ be the estimated vector $\boldsymbol{\gamma}$. The estimate $\hat{\boldsymbol{\gamma}}$ of the vector $\boldsymbol{\gamma}$ can be calculated as

$$\hat{\boldsymbol{\gamma}} = \operatorname*{argmin}_{\boldsymbol{\gamma} \in \mathbb{R}^K} \sum_{i=1}^{n} w_i \sum_{l=1}^{n_i} \left(Y_i(T_{il}) - \sum_{j=1}^{p} X_{ij}(T_{il}) \sum_{k=1}^{K_j} \gamma_{jk} B_{jk}(T_{il}) \right)^2, \tag{2.6}$$

where $w_i \geq 0$, $\sum_{i=1}^{n} w_i = 1$ are some chosen weights for subjects $i = 1, \ldots, n$.

The choice of weights w_i is important, as it can greatly impact both the theoretical and practical properties of the estimators $\hat{\boldsymbol{\gamma}}$. An ideal choice of the weights might depend on the within-subject correlation, which is usually unknown. Two common choices of weights are:

1. $w_i = \frac{1}{nn_i}$, which assigns equal weight to each subject. This is recommended when the number of observations n_i for each subject i is similar.

2. $w_i = \frac{1}{N}$, which assigns equal weight to all observations. This can be used when the number of observations n_i for each subject i differs significantly.

It should be noted that if we relax the constraint $\sum_{i=1}^{n} w_i = 1$, alternative weights $w_i = \frac{1}{n_i}$ and $w_i = 1$ can be considered. These alternatives correspond to the two previously mentioned choices of weights, each scaled by a constant factor that is the same for all subjects. As a result, both sets of weights lead to the same solution for the minimization problem (2.6).

To find the solution to this minimization problem, we can rewrite (2.6) into a matrix notation. Let

$$\mathbb{B}(t) = \begin{pmatrix} B_{11}(t) & \dots & B_{1K_1}(t) & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & \dots & B_{p1}(t) & \dots & B_{pK_p}(t) \end{pmatrix}$$

be a $(p \times K)$ matrix. Further, let $\mathbb{W} = \operatorname{diag}(w_i, \dots, w_i)$, $\mathbb{U}_i = (\boldsymbol{U}_{i1}, \dots, \boldsymbol{U}_{in_i})^\top$, where $\boldsymbol{U}_{il}^\top = \boldsymbol{X}_i^\top(T_{il})\mathbb{B}(T_{il})$, $l = 1, \dots, n_i$, and the vector of observed responses of i-th subject $\boldsymbol{Y_i} = (Y(T_{i1}), \dots, Y(T_{in_i}))^\top$, then

$$\hat{\boldsymbol{\gamma}} = \underset{\gamma \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \mathbb{U}_i \boldsymbol{\gamma})^\top \mathbb{W}_i (\boldsymbol{Y}_i - \mathbb{U}_i \boldsymbol{\gamma}). \tag{2.7}$$

which resembles ordinary least square estimation in the case of linear regression models, and can be therefore solved analogously by using normal equations. Set

$$\begin{aligned} F(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} (\boldsymbol{Y}_i - \mathbb{U}_i \boldsymbol{\gamma})^\top \mathbb{W}_i (\boldsymbol{Y}_i - \mathbb{U}_i \boldsymbol{\gamma}) \\ &= \sum_{i=1}^{n} \left[\boldsymbol{Y}_i^\top \mathbb{W}_i \boldsymbol{Y}_i - \boldsymbol{Y}_i^\top \mathbb{W}_i \mathbb{U}_i \boldsymbol{\gamma} - \boldsymbol{\gamma}^\top \mathbb{U}_i^\top \mathbb{W}_i \boldsymbol{Y}_i + \boldsymbol{\gamma}^\top \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i \boldsymbol{\gamma} \right]. \end{aligned}$$

Let us calculate the derivative and set it to zero.

$$\frac{\partial F(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = -2 \sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \boldsymbol{Y}_i + 2 \sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i \boldsymbol{\gamma} = 0$$

By rearranging the terms we get the following system of linear equations

$$\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i \hat{\boldsymbol{\gamma}} = \sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \boldsymbol{Y}_i.$$

The solution can be analytically expressed by taking the inverse of the matrix $\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i$, provided the inverse exists. In Section 3.1 we list mild regularity conditions that guarantee its existence. The resulting estimate is

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i \right)^{-1} \sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \boldsymbol{Y}_i. \tag{2.8}$$

We focus on this estimator in this thesis, due to the nice asymptotic properties (consistency and asymptotic normality), shown in section 3.1. For completeness, let us also introduce the smoothing spline estimator.

2.2 Smoothing spline approach

The smoothing spline method, also known as the penalized spline method, is the second spline-based method we introduce. Introduced by Hastie and Tibshirani (1993) and further developed by Hoover et al. (1998), the method estimates the coefficient functions $\boldsymbol{\beta}$ by minimizing a criterion that combines the sum of squared errors and a penalty term that depends on the smoothness of the functions.

In fact, the polynomial spline approach, discussed in Section 2.1, is a special case of the smoothing spline method. In the polynomial spline approach, the penalty term is effectively set to zero, focusing the minimization solely on the sum of squared errors in (2.6), thus focusing only on the fit. On the other hand, the smoothing spline method balances both the fit and the smoothness by including a penalty term that penalizes the roughness of the coefficient functions.

2.2.1 Longitudinal model estimation

Suppose that the coefficients functions $\beta_1(\cdot), \ldots, \beta_p(\cdot)$ are twice continuously differentiable with bounded and square integrable second derivatives. For i.i.d. data, it is natural to estimate the functions $\beta_1(\cdot), \ldots, \beta_p(\cdot)$ by minimization of the penalized least squares. Hoover et al. (1998) suggested minimizing such criterion even in the case of repeated measurements, i.e. the estimates are obtained as

$$\begin{aligned}\hat{\boldsymbol{\gamma}} =& \underset{\boldsymbol{\gamma}\in\mathbb{R}^K}{\operatorname{argmin}} \left[\sum_{i=1}^{n}\sum_{l=1}^{n_i} \left(Y_i(T_{il}) - \sum_{j=1}^{p} X_{ij}(T_{il}) \sum_{k=1}^{K_j} \gamma_{jk} B_{jk}(T_{il}) \right)^2 \right. \\ & \left. + \sum_{j=1}^{p} \lambda_j \int_a^b \left(\sum_{k=1}^{K_j} \gamma_{jk} B''_{jk}(t) \right)^2 \mathrm{d}t \right], \qquad (2.9)\end{aligned}$$

where $\lambda_j \in \mathbb{R}_0^+$ are smoothing (tuning) parameters. That is, the estimation is conducted assuming a within-subject working independence. The first term of the penalized least squares measures the goodness of fit of the model to the data, while the second term penalizes the lack of smoothness of the coefficient functions based on their second derivatives. The parameters λ_j control the trade-off between the fit and the smoothness for each coefficient function. Larger values of λ_j lead to smoother functions, but can also increase the bias. If λ_j is very small, i.e. zero or close to zero, the penalty term is of no real importance and the minimization leads to the estimator with the best fit for the data given in Section 2.1.

To solve this optimalization problem, we need to rewrite (2.9) by using a matrix notation as

$$\begin{aligned}\boldsymbol{Y} &= \left(Y_1(T_{11}), \ldots, Y_1(T_{1n_1}), \ldots, Y_n(T_{n1}), \ldots, Y_n(T_{nn_n})\right)^\top, \\ \mathbb{D}_j &= \operatorname{diag}\left(X_{1j}(T_{11}), \ldots, X_{1j}(T_{1n_1}), \ldots, X_{nj}(T_{n1}), \ldots, X_{nj}(T_{nn_n})\right), \\ \boldsymbol{\Omega}_j &= \begin{pmatrix} \int_a^b B''_{j1}(t)B''_{j1}(t)\,\mathrm{d}t & \ldots & \int_a^b B''_{j1}(t)B''_{jK_j}(t)\,\mathrm{d}t \\ \vdots & \ddots & \vdots \\ \int_a^b B''_{jK_j}(t)B''_{j1}(t)\,\mathrm{d}t & \ldots & \int_a^b B''_{jK_j}(t)B''_{jK_j}(t)\,\mathrm{d}t \end{pmatrix}, \\ \mathbb{B}_j &= \begin{pmatrix} B_{j1}(T_{11}) & \ldots & B_{jK_j}(T_{11}) \\ \vdots & \ddots & \vdots \\ B_{j1}(T_{nn_n}) & \ldots & B_{jK_j}(T_{nn_n}) \end{pmatrix}.\end{aligned}$$

To minimize the expression (2.9), we need to solve the set of p equations

$$\frac{\partial}{\partial \boldsymbol{\gamma}_j} \left[\left\| \boldsymbol{Y} - \sum_{j=1}^{p} \mathbb{D}_j \mathbb{B}_j \boldsymbol{\gamma}_j \right\|_2^2 + \sum_{j=1}^{p} \lambda_j \boldsymbol{\gamma}_j^\top \boldsymbol{\Omega}_j \boldsymbol{\gamma}_j \right] = 0, \quad j = 1, \ldots, p. \qquad (2.10)$$

That corresponds to the set of equations

$$-2\mathbb{D}_j\mathbb{B}_j^\top\left(\mathbf{Y}-\sum_{k=1}^{p}\mathbb{D}_k\mathbb{B}_k\boldsymbol{\gamma}_k\right)+2\lambda_j\boldsymbol{\Omega}_j\boldsymbol{\gamma}_j=0,\quad j=1,\ldots,p. \tag{2.11}$$

Equation (2.11) can be reformulated through a series of algebraic manipulations.

$$\begin{aligned}
\mathbb{D}_j\mathbb{B}_j^\top\left(\mathbf{Y}-\sum_{k=1}^{p}\mathbb{D}_k\mathbb{B}_k\boldsymbol{\gamma}_k\right)&=\lambda_j\boldsymbol{\Omega}_j\boldsymbol{\gamma}_j,\\
\mathbb{B}_j^\top\mathbb{D}_j^2\mathbb{B}_j\boldsymbol{\gamma}_j+\lambda_j\boldsymbol{\Omega}_j\boldsymbol{\gamma}_j&=\mathbb{B}_j^\top\mathbb{D}_j\left(\mathbf{Y}-\sum_{k=1,k\neq j}^{p}\mathbb{D}_k\mathbb{B}_k\boldsymbol{\gamma}_k\right),\\
\left(\mathbb{B}_j^\top\mathbb{D}_j^2\mathbb{B}_j+\lambda_j\boldsymbol{\Omega}_j\right)\boldsymbol{\gamma}_j&=\mathbb{B}_j^\top\mathbb{D}_j\left(\mathbf{Y}-\sum_{k=1,k\neq j}^{p}\mathbb{D}_k\mathbb{B}_k\boldsymbol{\gamma}_k\right).
\end{aligned} \tag{2.12}$$

The solution can be obtained from (2.12) by solving a system of $K=\sum_{j=1}^{p}K_j$ equations. That can be solved using Gaussian elimination with computational time $O(K^3)$.

Hastie and Tibshirani (1993) suggested an alternative calculation requiring computational time only $O(K)$. For more details, see Hastie and Tibshirani (1993).

Alternatively, Chiang et al. (2001) proposed a different approach in the case of time invariant covariates, i.e. $X_{ij}(T_{il})=X_{ij}$ for all $i=1,\ldots,n$, $l=1,\ldots,n_i$, $j=1,\ldots,p$, minimizing a criterion for each $\beta_j(\cdot)$ separately. The usage of such models is somewhat limiting, as in practice, covariates often depend on the time of the collection. However, there are some examples of covariates used in longitudinal studies that are time-invariant, such as treatment, dosage of medication, or baseline values.

The idea is to consider a general model equation of the form

$$Y(t)=\boldsymbol{X}^\top\boldsymbol{\beta}(t)+\varepsilon(t). \tag{2.13}$$

Assuming the inverse of $\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\top$ exists, we can express $\boldsymbol{\beta}(t)$ as $\boldsymbol{\beta}(t)=(\mathsf{E}\,\boldsymbol{X}\boldsymbol{X})^{-1}(\mathsf{E}\,\boldsymbol{X}Y(t))$. Denote $e_{j,r}$ the (j,r)-th element of $\left(\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}$. The function $\beta_j(\cdot)$ can be expressed as

$$\beta_j(t)=\boldsymbol{e}_j^\top\left(\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}(\mathsf{E}\,\boldsymbol{X}Y(t))=\mathsf{E}\left(\sum_{r=1}^{p}e_{j,r}X_rY(t)\right).$$

The element $e_{j,r}$ does not depend on t, it can be estimated by using sample mean as

$$\hat{e}_{j,r}=\boldsymbol{e}_j^\top\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)^{-1}\boldsymbol{e}_r. \tag{2.14}$$

It follows that a reasonable estimator $\hat{\boldsymbol{\gamma}}_j$ can be obtained by penalized least squares as

$$
\begin{aligned}
\hat{\gamma}_j = \underset{\gamma \in \mathbb{R}^{K_j}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \sum_{l=1}^{n_i} \left[\sum_{r=1}^{p} e_{j,r} X_{ir} Y_i(T_{il}) - \sum_{k=1}^{K_j} \gamma_{jk} B_{jk}(T_{il}) \right]^2 \\
+ \lambda_j \int_a^b \left[\sum_{k=1}^{K_j} \gamma_{js} B''_{jk}(t) \right]^2 \mathrm{d}t, \qquad (2.15)
\end{aligned}
$$

the weights w_i are usually chosen in the same way as in Section 2.1.

In comparison with the minimization in (2.9), the minimization in (2.15) requires generally less computational time. On the other hand, there is no explicit solution to (2.15). The solution is a cubic spline that can be approximated by an equivalent kernel function. The derivation is however beyond the scope of this thesis, for more details see Chiang et al. (2001).

2.2.2 Selection of the knots and the smoothing parameters

The choice of the number of knots, the knot locations, and the smoothing parameter can have a large impact on the quality of the spline approximation. In general, the knots should be placed in regions of high curvature or where the function changes rapidly.

Selection of the knot locations

There are various methods available for selecting the placement of knots once the number of knots is chosen.

One simple method for selecting knot locations is to use equidistant knots. That is, the knots are spread evenly along the range of the time variable. This approach is easy to implement and can provide quite reasonable approximation for smooth functions. However, it may not capture well the local behaviour of the function.

Alternatively, we can choose the knot locations by considering the distribution of the effect-modifying variable. One approach is to use quantile-based knots, where the knots are placed at the quantiles of the distribution. This way we can avoid clustering of knots in regions with sparse data.

Another method is to use cross-validation or generalized cross-validation, which are techniques that select the knots by minimizing a criterion that balances the fit and smoothness of the function, see Section 2.2.2. This methods can provide an optimal trade-off between bias and variance of the spline, but may be computationally intensive as it requires refitting the model many times.

Selection of the number of knots

Selecting the optimal number of knots is crucial for obtaining an accurate estimate, as the number of knots controls the complexity of the spline functions and consequently also of the coefficient functions $\boldsymbol{\beta}(\cdot)$. A larger number of knots enables capturing more complex patterns in the data, while a smaller number of knots produce smoother and more stable estimates of the coefficient functions.

There are several ways how to select the optimal number of knots, e.g., it could be chosen arbitrary based on some available recommendations. It might be however more appropriate to choose the number of knots, possibly together with the locations, based on some data-driven way.

a. Cross-validation

Cross-validation is a resampling technique that can be used to choose the number of knots out of a given set $\mathcal{K}_p \subset \mathbb{R}_+^p$. The idea is to split the data into a training set and a validation set. The model is fit on the training data and evaluated on the validation data, and this process is repeated multiple times. One common method is the leave-one-out cross-validation (leave-one-out CV), where each observation serves as the validation set. The optimal number of knots, which can differ for each $j = 1, \dots, p$, is $\boldsymbol{K} = \big(K_1, \dots, K_p)^\top \in \mathcal{K}_p$ that minimizes the average prediction squared error, or alternatively the sum of squared prediction errors, for short SSPE.

Huang et al. (2002) defines the sum of squared prediction errors as

$$K_{(\mathrm{CV})}^{\mathrm{Long}} = \underset{\boldsymbol{K} \in \mathcal{K}_p}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{l=1}^{n_i} w_i \left[Y_i(T_{il}) - \sum_{j=1}^{p} X_{ij}(T_{il}) \hat{\beta}_j^{[-i]}(T_{il}) \right], \tag{2.16}$$

where $\hat{\boldsymbol{\beta}}^{[-i]}(\cdot) = (\hat{\beta}_1^{[-i]}, \dots, \hat{\beta}_p^{[-i]})^\top$ is the polynomial spline estimate of $\boldsymbol{\beta}(\cdot)$ computed with all measurements of the i-th subject deleted.

The procedure for the leave-one-out CV can be summarized as follows:

1. Choose $\mathcal{K}_p$, the range of possible numbers of knots to test.
2. For each number of knots $\boldsymbol{K}$ in the range $\mathcal{K}_p$
   (a) For each subject $i = 1, \dots, n$
      i. Remove the i-th subject from the dataset.
      ii. Fit the varying coefficient model using the remaining data with the current number of knots, resulting in $\hat{\boldsymbol{\beta}}^{[-i]}$.
      iii. Compute the prediction error for the left-out observations:
      $$Y_i(T_{il}) - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}^{[-i]}(T_{il}).$$
   (b) Calculate the sum given in (2.16) by summing the squared prediction errors.
3. Select the number of knots that minimizes the expression (2.16).

Similarly, the leave-one-out CV can be also used to choose the location of the knots (by minimizing SSPE over a set of possible locations) or choosing the location and the number of knots simultaneously. It should be pointed out that leave-one-out CV can be computationally quite expensive, as it requires fitting the model n times for each $\boldsymbol{K}$. A possible alternative is to use generalised cross-validation, see for example Ruppert (2002).

b. Information Criteria

Information criteria provide another method for choosing the number of knots. They balance the fit of the model (as measured by the residual sum of squares) against the complexity of the model (as measured by the number of parameters K). Two commonly used criteria are the Akaike Information Criterion AIC (Akaike, 1974) and the Bayesian Information Criterion BIC (Schwarz, 1978).

For the AIC, we select the number of knots $\boldsymbol{K} = (K_1, \ldots, K_p)$ that minimizes:

$$\begin{aligned}\boldsymbol{K}_{(\mathrm{AIC})} &= \underset{\boldsymbol{K}\in\mathcal{K}_p}{\operatorname{argmin}} \left[\log\left(\frac{\mathrm{RSS}}{N}\right) + \frac{2K}{N}\right] \qquad (2.17)\\ &= \underset{\boldsymbol{K}\in\mathcal{K}_p}{\operatorname{argmin}} \left[\log\left(\frac{1}{N}\sum_{i=1}^{n} w_i \sum_{l=1}^{n_i}\left(Y_i(T_{il}) - \sum_{j=1}^{p} X_{ij}(T_{il}) \sum_{k=1}^{K_j} \hat{\gamma}_{jk} B_{jk}(T_{il})\right)^2\right) + \frac{2K}{N}\right].\end{aligned}$$

For the BIC, we select the number of knots that minimizes:

$$\boldsymbol{K}_{(\mathrm{BIC})} = \underset{\boldsymbol{K}\in\mathcal{K}_p}{\operatorname{argmin}} \left[\log\left(\frac{\mathrm{RSS}}{N}\right) + \log(N)\frac{K}{N}\right]. \qquad (2.18)$$

For more details see Huang and Shen (2004).

Selection of the smoothing parameter

Chiang et al. (2001) recommends using the cross-validation method introduced in Section 2.2.2 the context of knot number selection also to choose the smoothing parameter out of a preselected range.

2.3 Other methods

Polynomial and smoothing spline estimation are global methods, meaning that these methods construct estimates considering the entire range of the covariate space. This global perspective often leads to computationally efficient solutions. However, their global nature may limit their flexibility to capture localized variations.

An alternative method is the local polynomial approach which uses kernel function to weigh the data points close to a specific point of interest. It then uses these weights to create a polynomial fit that effectively describes the local characteristics of the data (see Hastie and Tibshirani, 1993; Cai et al., 2000a). Another alternative is the local maximum likelihood approach that maximizes the likelihood function over local regions (see Cai et al., 2000a).

3. Statistical inference

In this chapter, we derive the asymptotic properties, especially consistence and asymptotic normality of the polynomial spline estimator of the coefficient functions in Model (2.2). The asymptotic results are used to construct pointwise asymptotic confidence intervals and asymptotic confidence bands for the unknown coefficient functions.

3.1 Asymptotic properties

The asymptotic properties of the polynomial spline estimator from Section 2.1.2 for Model (2.2) were extensively studied by Huang et al. (2004). Here we summarize the findings. Firstly, let us define the necessary terminology.

Let $g : [a, b] \to \mathbb{R}$ be a function and let $\mathbb{G} = \{g_i(\cdot), g_i : [a, b] \to \mathbb{R},\ i \in I\}$ be some linear space of real-valued functions defined on the interval $[a, b]$, where I is an index set. Denote by

$$\operatorname{dist}(f, \mathbb{G}) = \inf_{g\in\mathbb{G}} \left(\sup_{x\in[a,b]} |f(x) - g(x)| \right) \tag{3.1}$$

the L^∞ distance of the real-valued function $f(\cdot)$ from the linear space of real-valued functions $\mathbb{G}$. Moreover, let $\|\cdot\|_{L^2}$ be the L^2 norm, i.e. for a real-valued function $g(\cdot)$

$$\|g\|_{L^2} = \left(\int_a^b g^2(t)\, \mathrm{d}t \right)^{\frac{1}{2}}. \tag{3.2}$$

For a vector function $\boldsymbol{g}(\cdot) = (g_1(\cdot), \dots, g_p(\cdot))^\top$, $g_j : [a, b] \to \mathbb{R}$, $j = 1, \dots, p$,

$$\|\boldsymbol{g}\|_{L^2} = \left(\sum_{j=1}^{p} \|g_j\|_{L^2}^2 \right)^{\frac{1}{2}}. \tag{3.3}$$

Assume that the times of measurements $T_{jl}, j = 1, \dots, n, l = 1, \dots, n_i$, are independent observations of a real-valued random variable T with support contained in the interval $[a, b]$, $a, b \in \mathbb{R}$. Furthermore, assume that the random variable T is independent of the random processes $\{Y(t), \boldsymbol{X}(t), t \in [a,\ b]\}$.

Without loss of generality, assume that the support of the random variable T, is contained in the interval $[0,\ 1]$. The results hold for the general case $[a,\ b]$ due to a possible transformation $T(b - a) + a$.

For real-valued vector functions $\boldsymbol{g}^{(1)}(\cdot) = (g_1^{(1)}(\cdot), \dots, g_p^{(1)}(\cdot))^\top$ and $\boldsymbol{g}^{(2)}(\cdot) = (g_1^{(2)}(\cdot), \dots, g_p^{(2)}(\cdot))^\top$ defined on $[0,\ 1]^p$, we define the theoretical inner product as

$$\langle \boldsymbol{g}^{(1)}, \boldsymbol{g}^{(2)} \rangle = \mathsf{E} \left[\left(\sum_{j=1}^{p} X_j(T) g_j^{(1)}(T) \right) \left(\sum_{j=1}^{p} X_j(T) g_j^{(2)}(T) \right) \right] \tag{3.4}$$

and the empirical inner product as

$$\langle \boldsymbol{g}^{(1)}, \boldsymbol{g}^{(2)} \rangle_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{l=1}^{n_i} \left(\sum_{j=1}^{p} X_{ij}(T_{il}) g_j^{(1)}(T_{il}) \right) \left(\sum_{j=1}^{p} X_{ij}(T_{il}) g_j^{(2)}(T_{il}) \right). \tag{3.5}$$

The expectation in (3.4) is taken with respect to the joint distribution of $(\boldsymbol{X}(T),\, T)$.

Denote the corresponding norm for the theoretical inner product as $\|\cdot\|$, i.e. $\|\boldsymbol{g}^{(1)}\|^2 = \langle \boldsymbol{g}^{(1)}, \boldsymbol{g}^{(1)}\rangle$, and for the empirical inner product as $\|\cdot\|_n$.

The results are proven using the weights $w_i = \frac{1}{n_i}$. Let us start with regularity conditions

(C1) The density $f_T(\cdot)$ of the random variable T is bounded away from 0 and infinity uniformly over $t \in [0,\, 1]$, i.e., there exist positive constants c_1 and c_2 such that
$$0 < c_1 \le f_T(t) \le c_2 < \infty \quad \text{for all}\, t \in [0,\, 1]\,.$$

(C2) The eigenvalues of $\boldsymbol{\Sigma}(t) = \mathsf{E}\left[(X_1(t), \ldots, X_p(t))(X_1(t), \ldots, X_p(t))^\top | T = t\right] = \mathsf{E}\left[\boldsymbol{X}(t)\boldsymbol{X}^\top(t) | T = t\right]$ are bounded away from 0 and infinity uniformly in $t \in [0,\, 1]$.

(C3) $|X_j(t)|$ for all $j = 1, \ldots, p$ and $\mathsf{E}\left[\varepsilon(t)^2\right]$ are bounded on $[0,\, 1]$.

(C4) $\limsup\limits_{n \longrightarrow \infty} \left(\dfrac{\max\limits_{j=1,\ldots,p} K_j}{\min\limits_{j=1,\ldots,p} K_j} \right) < \infty.$

(C5) The process $\{\varepsilon(t),\, t \in [0,\, 1]\}$ can be decomposed into a sum of two independent stochastic processes, an arbitrary mean zero process $\varepsilon_1(t)$ and a process $\varepsilon_2(t)$ of measurement errors that are independent at different time points and have mean zero and a constant variance σ^2.

From (C2), it follows that there exist positive constants c_1, c_2 such that for any vector function $\boldsymbol{g}(\cdot) = \left(g_1(\cdot), \ldots, g_p(\cdot)\right)^\top$, $g_j : [0,1] \to \mathbb{R}$, $j = 1, \ldots, p$ and all $t \in [0,1]$

$$c_1 \boldsymbol{g}^\top(t)\boldsymbol{g}(t) \le \lambda_p(t)\boldsymbol{g}^\top(t)\boldsymbol{g}(t) \le \boldsymbol{g}^\top(t)\boldsymbol{\Sigma}(t)\boldsymbol{g}(t) \le \lambda_1(t)\boldsymbol{g}^\top(t)\boldsymbol{g}(t) \le c_2 \boldsymbol{g}^\top(t)\boldsymbol{g}(t),$$

where $\lambda_1(t)$ and $\lambda_p(t)$ are the highest and lowest eigenvalues of the matrix $\boldsymbol{\Sigma}(t)$, respectively.

Under these regularity conditions, the polynomial spline estimator of Model (2.2) exists and is determined uniquely, as given by the following Lemma.

Lemma 1. *Suppose the conditions (C1)–(C4) hold. There are positive constants* C_1 *and* C_2 *such that all eigenvalues of* $\left(\frac{K_n}{n}\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)$ *are bounded almost surely. Consequently,* $\left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)$ *is invertible.*

Proof. See lemma A3 in Huang et al. (2004). □

Let us now repeat and extend the notation from Section 2.1. Recall that

$$\boldsymbol{U}_{il} = \left(X_{i1}(T_{il})B_{11}(T_{il}), \ldots, X_{i1}(T_{il})B_{1K_1}(T_{il}), \ldots, X_{ip}(T_{il})B_{p1}(T_{il}), \ldots, X_{ip}(T_{il})B_{pK_p}(T_{il})\right)$$

is a row vector of the matrix

$$\mathbb{U}_i = \begin{pmatrix} X_{i1}(T_{i1})B_{11}(T_{i1}) & \ldots & X_{ip}(T_{i1})B_{pK_p}(T_{i1}) \\ \vdots & \ddots & \vdots \\ X_{i1}(T_{in_i})B_{11}(T_{in_i}) & \ldots & X_{ip}(T_{in_i})B_{pK_p}(T_{in_i}) \end{pmatrix}.$$

Further, denote

$$\mathbb{U} = \begin{pmatrix} \mathbb{U}_1 \\ \mathbb{U}_2 \\ \vdots \\ \mathbb{U}_n \end{pmatrix}, \qquad \mathbb{W} = \begin{pmatrix} \mathbb{W}_1 & 0 & \cdots & 0 \\ 0 & \mathbb{W}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbb{W}_n \end{pmatrix},$$

where $\mathbb{W}_i$ are diagonal matrices with n_i on the diagonal. Then

$$\mathbb{U}^\top \mathbb{W} \mathbb{U} = \sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i$$

and the polynomial spline estimator is given as

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i \right)^{-1} \sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \boldsymbol{Y}_i = \left(\mathbb{U}^\top \mathbb{W} \mathbb{U} \right)^{-1} \sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \boldsymbol{Y}_i. \tag{3.6}$$

Denote

$$\mathcal{D} = \{(X_{i1}(T_{il}), \ldots, X_{ip}(T_{il}),\ T_{il}),\ i = 1, \ldots, n,\ l = 1, \ldots, n_i\}. \tag{3.7}$$

The variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$ conditioning on $\mathcal{D}$ is

$$\operatorname{Var}(\hat{\boldsymbol{\gamma}}|\mathcal{D}) = \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \right) \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i \right)^{-1}, \tag{3.8}$$

where for all $i = 1, \ldots, n$

$$\mathbb{V}_i = \operatorname{Var}(\boldsymbol{Y}_i|\mathcal{D}) = \begin{pmatrix} \operatorname{var}(\varepsilon_i(T_{i1})) & \cdots & \operatorname{cov}(\varepsilon_i(T_{i1}), \varepsilon_i(T_{in_i})) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}(\varepsilon_i(T_{in_i}), \varepsilon_i(T_{i1})) & \cdots & \operatorname{var}(\varepsilon_i(T_{in_i})) \end{pmatrix} \tag{3.9}$$

is the variance-covariance matrix of the error process. It follows that

$$\begin{aligned} \operatorname{Var}(\hat{\boldsymbol{\beta}}\,(t)\,|\mathcal{D}) &= \operatorname{Var}(\mathbb{B}(t)\hat{\boldsymbol{\gamma}}|\mathcal{D}) \\ &= \mathbb{B}(t) \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \right) \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i \right)^{-1} \mathbb{B}^\top(t). \end{aligned}$$

Let $\boldsymbol{e}_j \in \mathbb{R}^p$ be a vector with j-th element 1 and 0 elsewhere. The conditional variance of $\hat{\beta}_j(t)$ is

$$\operatorname{var}(\hat{\beta}_j(t)|\mathcal{D}) = \boldsymbol{e}_j^\top \operatorname{var}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D})\boldsymbol{e}_j, \quad j = 1, \ldots, p. \tag{3.10}$$

The only unknown quantity that needs to be estimated in order to use asymptotic properties for inference is the matrix $\mathbb{V}_i$ corresponding to the autocovariance function of the error process $\boldsymbol{\varepsilon} : C_\varepsilon(t, s),\ t \in [0, 1]\,,\ s \in [0, 1]\,.$ Huang et al. (2004) suggested using a tensor product spline on $[0, 1] \times [0, 1]$ to approximate $C_\varepsilon(t, s)$

$$C_\varepsilon(t, s) \approx \sum_{k=1}^{L} \sum_{l=1}^{L} \nu_{kl} B_k(t) B_l(s), \quad t, s \in [0, 1]\,,\ t \neq s,\ L \in \mathbb{N}, \tag{3.11}$$

where $\nu_{kl} = \nu_{lk}$ for $l = 1, \dots, L$, $k = 1, \dots, L$. The estimate is reasonable just for $t \neq s$. That is because in longitudinal settings, the autocovariance function $C_\varepsilon(t, s)$ might not be continuous at $t = s$, i.e

$$\lim_{s \to t} C_\varepsilon(t, s) \neq C_\varepsilon(t, t), \tag{3.12}$$

for details, see Diggle and Verbyla (1998). Therefore, we estimate $C_\varepsilon(t, t)$ separately by using spline approximation

$$C_\varepsilon(t, t) \approx \sum_{k=1}^{L} \upsilon_k B_k(t), \quad t \in [0, 1]\,. \tag{3.13}$$

The vector of coefficients $\boldsymbol{\nu} = (\nu_{11}, \dots, \nu_{lL}, \nu_{22}, \dots, \nu_{2L}, \dots, \nu_{LL})^\top$ of length $L(L+1)/2$ can be estimated as

$$\hat{\boldsymbol{\nu}} = \underset{\boldsymbol{\nu} \in \mathbb{R}^{L(L+1)/2}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{l=2}^{n_i} \sum_{l'=1}^{l-1} \left(\hat{\varepsilon}_{il} \hat{\varepsilon}_{il'} - \sum_{k=1}^{L} \sum_{r=1}^{L} \nu_{kl} B_k(T_{il}) B_r(T_{il'}) \right)^2 , \tag{3.14}$$

where $\hat{\varepsilon}_{il} = Y_{il} - \boldsymbol{X}_i^\top (T_{il})\, \hat{\boldsymbol{\beta}}\, (T_{il})$ are residuals of the model in (2.2).

Similarly, the vector $\boldsymbol{\upsilon} = (\upsilon_1, \dots, \upsilon_L)$ from the expression (3.13) can be estimated as

$$\hat{\boldsymbol{\upsilon}} = \underset{\boldsymbol{\upsilon} \in \mathbb{R}^{L}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{l=1}^{n_i} \left(\hat{\varepsilon}_{il}^2 - \sum_{k=1}^{L} \nu_k B_k(T_{il}) \right)^2 . \tag{3.15}$$

The approximation of both $C_\varepsilon(t, s)$ and $C_\varepsilon(t, t)$ relies on choosing an appropriate spline basis. Huang et al. (2004) suggested using 5−10 equidistant knots. Another option is to use data-driven way, such as Cross-validation, see Section 2.2.2. That could be however quite computationally intensive. In a numerical study in Chapter 4 we use 5 equidistant knots.

3.1.1 Consistency

Definition 5 (Asymptotic relations of sequences)**.** *Let $a = \{a_n\}_{n \in \mathbb{N}}$, $b = \{b_n\}_{n \in \mathbb{N}}$ be sequences of real numbers.*

1. *The sequences a and b are said to be asymptotically equivalent (denoted as $a_n \asymp b_n$), if the limit of the ratio of the corresponding terms a_n and b_n exists, i.e.*
$$\lim_{n \to \infty} \frac{a_n}{b_n} = c,$$
where $c \neq 0$ is a real constant.

2. *The sequence a is said to be asymptotically less than or equal to b (denoted as $a_n \lesssim b_n$), if there exists a real constant c such that $a_n \leq c\, b_n$ for all sufficiently large n.*

3. *The sequence a is said to be asymptotically greater than or equal to b (denoted as $a_n \gtrsim b_n$), if there exists a real constant c such that $a_n \geq c\, b_n$ for all sufficiently large n.*

There is a close relation between the norm $\|\cdot\|$ of a B-spline function and the L^2 norm of its B-spline coefficients used throughout the proofs given by the following Lemma. Set $K_n = \max\limits_{j=1,\dots,p} K_j$.

Lemma 2. *Suppose the conditions (C1) - (C4) hold. Let* $g_j(\cdot) = \sum_{k=1}^{K_j} \gamma_{jk} B_{jk}(\cdot)$ *be a B-spline on* $[0,1]$, $\boldsymbol{\gamma}_j = (\gamma_{j1},\dots,\gamma_{jK_j})^\top$ *for* $j = 1,\dots,p$, *and* $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top,\dots,\boldsymbol{\gamma}_p^\top)^\top$. *Set* $\boldsymbol{g}(\cdot) = (g_1(\cdot),\dots,g_p(\cdot))^\top$. *Then* $\|\boldsymbol{g}\|^2 \asymp \sum_{j=1}^p \|g_j\|_{L^2}^2 \asymp \frac{\|\boldsymbol{\gamma}\|_2^2}{K_n}$, *where* $\|\boldsymbol{\gamma}\|_2$ *denotes the Euclidean norm of the coefficients* $\boldsymbol{\gamma}$.

Proof. From the independence of $\{\boldsymbol{X}(t),\ t \in [0,1]\}$ and T

$$\begin{aligned}\mathsf{E}\left[\left(\sum_{j=1}^p X_j(T) g_j(T)\right)^2\right] &= \int_0^1 \mathsf{E}\left[\left(\sum_{j=1}^p X_j(t) g_j(t)\right)^2 \middle| T = t\right] f_T(t)\,\mathrm{d}t \\ &= \int_0^1 \boldsymbol{g}^\top(t)\,\mathsf{E}\left[\boldsymbol{X}(t)\boldsymbol{X}^\top(t) | T = t\right]\boldsymbol{g}(t) f_T(t)\,\mathrm{d}t \\ &= \int_0^1 \boldsymbol{g}^\top(t)\boldsymbol{\Sigma}(t)\boldsymbol{g}(t) f_T(t)\,\mathrm{d}t. \qquad (3.16)\end{aligned}$$

By using conditions (C1) and (C2), it is easy to show that there are positive constants c_1, c_2, such that

$$c_1 \int_0^1 \boldsymbol{g}^\top(t)\boldsymbol{g}(t)\,\mathrm{d}t \le \int_0^1 \boldsymbol{g}^\top(t)\boldsymbol{\Sigma}(t)\boldsymbol{g}(t) f_T(t)\,\mathrm{d}t \ \le c_2 \int_0^1 \boldsymbol{g}^\top(t)\boldsymbol{g}(t)\,\mathrm{d}t. \qquad (3.17)$$

Hence

$$\|\boldsymbol{g}\|^2 = \mathsf{E}\left[\left(\sum_{j=1}^p X_j(T) g_j(T)\right)^2\right] \asymp \int_0^1 \boldsymbol{g}^\top(t)\boldsymbol{g}(t)\,\mathrm{d}t = \sum_{j=1}^p \|g_j\|_{L^2}^2. \qquad (3.18)$$

By the properties of B-splines, see De Boor (1978):

$$\|g_j\|_{L^2}^2 \asymp \|\boldsymbol{\gamma}_j\|_2^2 / K_j, \quad j = 1,\dots,p. \qquad (3.19)$$

By condition (C4) then

$$\|\boldsymbol{g}\|^2 \asymp \sum_{j=1}^p \|g_j\|_{L^2}^2 \asymp \frac{\|\boldsymbol{\gamma}\|_2^2}{K_n}. \qquad (3.20)$$

□

Definition 6 (Consistency in L2 norm). *An estimator* $\hat{\beta}_j(\cdot)$ *of* $\beta_j(\cdot)$ *is said to be consistent if*

$$\|\hat{\beta}_j - \beta_j\|_{L^2} = o_P(1),$$

where $o_P(1)$ *denotes convergence in probability to zero.*

Theorem 1 (Consistency). *Suppose that the conditions (C1)–(C4) are satisfied. If* $\lim\limits_{n\to\infty} K_n \frac{\log K_n}{n} = 0$, *then* $\|\hat{\beta}_j - \beta_j\|_{L^2}^2 = O_P\left(\frac{1}{n} + \frac{K_n}{n^2}\sum_{i=1}^n \frac{1}{n_i} + \rho_n^2\right)$, *where* $\rho_n = \max\limits_{j=1,\dots,p} \operatorname{dist}(\beta_j, \mathbb{G}_j)$. *If additionally* $\lim\limits_{n\to\infty} \rho_n = 0$, *then* $\hat{\beta}_j$, $j = 1,\dots,p$, *are consistent estimators of* β_j.

Proof. Set $\tilde{Y}_{il} = \boldsymbol{X}_i(T_{il})^\top \boldsymbol{\beta}(T_{il})$, $\tilde{\boldsymbol{Y}}_{\boldsymbol{i}} = (\tilde{Y}_{i1}, \dots, \tilde{Y}_{in_i})^\top$, $\tilde{\boldsymbol{Y}} = (\tilde{\boldsymbol{Y}}_1, \dots, \tilde{\boldsymbol{Y}}_n)^\top$, and $\tilde{\boldsymbol{\gamma}} = \mathsf{E}\,(\hat{\boldsymbol{\gamma}}|\mathcal{D}) = \sum_{i=1}^n \left(\mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \tilde{\boldsymbol{Y}}_i$, $\tilde{\boldsymbol{\beta}}(t) = \mathbb{B}(t)\tilde{\boldsymbol{\gamma}}$. Recall $\hat{\boldsymbol{\beta}}(t) = \mathbb{B}(t)\hat{\boldsymbol{\gamma}}$.

By triangle inequality

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{L^2}^2 \leq \|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L^2}^2 + \|\tilde{\boldsymbol{\beta}} - \boldsymbol{g}^*\|_{L^2}^2 + \|\boldsymbol{g}^* - \boldsymbol{\beta}\|_{L^2}^2, \tag{3.21}$$

where $\boldsymbol{g}^*(\cdot) = \left(g_1^*(\cdot), \dots, g_p^*(\cdot)\right)^\top$ is a vector B-spline function that is the best approximation of $\boldsymbol{\beta}(\cdot)$ on the space $\mathbb{G} = \mathbb{G}_1 \times \cdots \times \mathbb{G}_p$, i.e.

$$\|\boldsymbol{g}^* - \boldsymbol{\beta}\|_{L^2}^2 = \rho_n. \tag{3.22}$$

There is a vector $\boldsymbol{\gamma}^* = \left(\gamma_{11}^*, \dots, \gamma_{1K_1}^*, \dots, \gamma_{p1}^*, \dots, \gamma_{pK_p}^*\right)^\top$ of spline coefficients, such that $\boldsymbol{g}^*(t) = \mathbb{B}(t)\,\boldsymbol{\gamma}^* = \sum_{j=1}^p \sum_{k=1}^{K_j} \gamma_{jk}^* B_{jk}(t)$ for all $t \in [0,1]$.

Let us start with $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L^2}^2$. It follows from the spline properties, that

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L^2}^2 = \|\mathbb{B}(t)\,(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})\,\|_{L^2}^2 \asymp \frac{\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\|_2^2}{K_n}. \tag{3.23}$$

We can rewrite $\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\|_2^2$ by a series of algebraic operations and from Lemma 1

$$\begin{aligned}
\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\|_2^2 &= \left\|\left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i (\hat{\boldsymbol{Y}}_i - \tilde{\boldsymbol{Y}}_i)\right\|_2^2 \\
&= \left\|\left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \boldsymbol{\varepsilon}_i\right\|_2^2 = \left\|\left(\mathbb{U}^\top \mathbb{W} \mathbb{U}\right)^{-1} \mathbb{U}^\top \mathbb{W} \boldsymbol{\varepsilon}\right\|_2^2 \\
&= \frac{K_n^2}{n^2} \boldsymbol{\varepsilon}^\top \mathbb{W}^\top \mathbb{U} \left(\frac{K_n}{n} \mathbb{U}^\top \mathbb{W} \mathbb{U}\right)^{-1} \left(\frac{K_n}{n} \mathbb{U}^\top \mathbb{W} \mathbb{U}\right)^{-1} \mathbb{U}^\top \mathbb{W} \boldsymbol{\varepsilon} \\
&\asymp \frac{K_n^2}{n^2} \boldsymbol{\varepsilon}^\top \mathbb{W}^\top \mathbb{U} \mathbb{U}^\top \mathbb{W} \boldsymbol{\varepsilon} = \frac{K_n^2}{n^2} \left\|\mathbb{U}^\top \mathbb{W} \boldsymbol{\varepsilon}\right\|_2^2,
\end{aligned} \tag{3.24}$$

By utilizing the linearity of the expected value and taking the conditional expectation of (3.24), the expression can be simplified because of independence for different subjects $i = 1, \dots, n$ through elementwise representation as

$$\frac{K_n^2}{n^2} \left\|\mathbb{U}^\top \mathbb{W} \boldsymbol{\varepsilon}\right\|_2^2 = \frac{K_n^2}{n^2} \sum_{i=1}^n \frac{1}{n_i^2} \sum_{j=1}^p \sum_{k=1}^{K_j} \mathsf{E}\left[\left(\sum_{l=1}^{n_i} X_{ij}(T_{il}) B_{jk}(T_{il}) \varepsilon_{il}\right)^2 \middle| \mathcal{D}\right]. \tag{3.25}$$

For all $l = 1, \dots, n_i$, $l' = 1, \dots, n_i$, $l' \neq l$, by the properties of B-splines in Lemma 2 and condition (C3)

$$\mathsf{E}\left[\left(X_{ij}(T_{il}) B_{jk}(T_{il}) \varepsilon_{il}\right)^2 | \mathcal{D}\right] \leq \frac{C}{K_n},$$

$$\mathsf{E}\left[\left(X_{ij}(T_{il}) B_{jk}(T_{il}) \varepsilon_{il}\right)\left(X_{ij}(T_{il'}) B_{jk}(T_{il'}) \varepsilon_{il'}\right) | \mathcal{D}\right] \leq \frac{C}{K_n^2},$$

where C is a real constant. Therefore, $\mathsf{E}\left[\left(\sum_{l=1}^{n_i} X_{ij}(T_{il}) B_{jk}(T_{il}) \varepsilon_{il}\right)^2 | \mathcal{D}\right]$ can be expressed as a sum of n_i terms bounded by $\frac{C}{K_n}$ and $n_i^2 - n_i$ terms bounded by $\frac{C}{K_n^2}$.

We have

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|_{L^2}^2 \asymp \frac{\|\hat{\boldsymbol{\gamma}}-\tilde{\boldsymbol{\gamma}}\|_2^2}{K_n} &\asymp \frac{K_n}{n^2}\sum_{i=1}^{n}\frac{1}{n_i^2}\sum_{j=1}^{p}\sum_{k=1}^{K_j}\mathsf{E}\left[\left(\sum_{l=1}^{n_i}X_{ij}(T_{il})B_{jk}(T_{il})\varepsilon_{il}\right)^2 | \mathcal{D}\right]\\ &\lesssim \frac{K_n}{n^2}\sum_{i=1}^{n}\frac{1}{n_i^2}K_n\left(\frac{n_i}{K_n}+\frac{(n_i^2-n_i)}{K_n^2}\right) = O_P\left(\frac{K_n}{n^2}\sum_{i=1}^{n}\left[\frac{1}{n_i}+\frac{1}{K_n}\left(1-\frac{1}{n_i}\right)\right]\right).\end{aligned}$$

Let us now consider the second term from the inequality (3.21) $\|\tilde{\boldsymbol{\beta}}-\boldsymbol{g}^*\|_{L^2}^2$. Similarly, as for the first term $\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|_{L^2}^2$, by a series of algebraic operations it is easy to show that

$$\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\gamma}^*\|_{L^2}^2 = \|\mathbb{B}(t)\left(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\right)\|_{L^2}^2 \asymp \frac{\|\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2^2}{K_n}, \tag{3.26}$$

$$\begin{aligned}\|\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2^2 &= \left\|\left(\sum_{i=1}^{n}\mathbb{U}_i^\top\mathbb{W}_i\mathbb{U}_i\right)^{-1}\sum_{i=1}^{n}\mathbb{U}_i^\top\mathbb{W}_i(\tilde{\boldsymbol{Y}}_i-\mathbb{U}_i\boldsymbol{\gamma}^*)\right\|_2^2\\ &\asymp \frac{K_n^2}{n^2}\left\|\sum_{i=1}^{n}\mathbb{U}_i^\top\mathbb{W}_i\left(\tilde{\boldsymbol{Y}}_i-\mathbb{U}_i\boldsymbol{\gamma}^*\right)\right\|_2^2\\ &= \frac{K_n^2}{n^2}\sum_{j=1}^{p}\sum_{k=1}^{K_j}\left(\sum_{i=1}^{n}\frac{1}{n_i}\sum_{l=1}^{n_i}X_{ij}(T_{il})B_{jk}(T_{il})\left(\tilde{\boldsymbol{Y}}_i-\mathbb{U}_i\boldsymbol{\gamma}^*\right)_l\right)^2 \qquad (3.27)\end{aligned}$$

where $\left(\tilde{\boldsymbol{Y}}_i-\mathbb{U}_i\boldsymbol{\gamma}^*\right)_l = \boldsymbol{X}_i^\top(T_{il})\boldsymbol{\beta}(T_{il}) - \boldsymbol{X}_i^\top(T_{il})\mathbb{B}(T_{il})\boldsymbol{\gamma}^* = \boldsymbol{X}_i^\top(T_{il})(\boldsymbol{\beta}(T_{il}) - \mathbb{B}(T_{il})\boldsymbol{\gamma}^*)$ denotes the l-th term of $\left(\tilde{\boldsymbol{Y}}_i-\mathbb{U}_i\boldsymbol{\gamma}^*\right)$. From the condition (C3)

$$|\left(\tilde{\boldsymbol{Y}}_i-\mathbb{U}_i\boldsymbol{\gamma}^*\right)_l| = |\boldsymbol{X}_i^\top(T_{il})(\boldsymbol{\beta}(T_{il})-\mathbb{B}(T_{il})\boldsymbol{\gamma}^*)| \lesssim \rho_n, \tag{3.28}$$

hence

$$\begin{aligned}\|\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2^2 &\lesssim \frac{K_n^2\rho_n^2}{n^2}\sum_{j=1}^{p}\sum_{k=1}^{K_j}\left(\sum_{i=1}^{n}\frac{1}{n_i}\sum_{l=1}^{n_i}X_{ij}(T_{il})B_{jk}(T_{il})\right)^2\\ &\le K_n^2\rho_n^2\sum_{j=1}^{p}\sum_{k=1}^{K_j}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{l=1}^{n_i}X_{ij}^2(T_{il})B_{jk}(T_{il})\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{l=1}^{n_i}B_{jk}(T_{il})\right) \qquad (3.29)\\ &\lesssim K_n^3\rho_n^2\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{l=1}^{n_i}B_{jk}(T_{il})\right)^2, \qquad (3.30)\\ &\lesssim K_n\rho_n^2 \qquad (3.31)\end{aligned}$$

where the inequality (3.29) follows from the Cauchy-Schwartz inequality and the inequality (3.30) holds because of (C3). The inequality (3.31) holds because

$$\sup_{j,k}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{l=1}^{n_i}B_{jk}\left(T_{il}\right) \lesssim \frac{1}{K_n}, \tag{3.32}$$

for details, see Lemma A6 in Huang et al. (2004). The result follows. □

We have shown the consistence of the estimates $\hat{\boldsymbol{\beta}}$. Similarly, the proposed estimator of $\mathbb{V}_i$ is consistent (Huang et al., 2004).

3.1.2 Asymptotic normality

Theorem 2 (Asymptotic Normality)**.** *Let the conditions (C1) – (C5) hold and set $t \in [0,1]$. If $\lim_{n\to\infty} K_n \frac{\log K_n}{n} = 0$ and $\lim_{n\to\infty} K_n \frac{\max_i n_i}{n} = 0$, then*

$$\{\mathrm{var}(\hat{\boldsymbol{\beta}}(t)|\,\mathcal{D})\}^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}}(t) - \tilde{\boldsymbol{\beta}}(t)) \xrightarrow{d} \mathcal{N}(0, \mathbb{I}_p),$$

where $\tilde{\boldsymbol{\beta}}(t) = \mathsf{E}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D}) = \left(\tilde{\boldsymbol{\beta}}_1(t), \dots, \tilde{\boldsymbol{\beta}}_p(t)\right)^\top$. In particular,

$$\{\mathrm{var}(\hat{\boldsymbol{\beta}}_j(t)|\,\mathcal{D})\}^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}}_j(t) - \tilde{\boldsymbol{\beta}}_j(t)) \xrightarrow{d} \mathcal{N}(0,1) \quad \textit{for } j = 1, \dots, p.$$

Proof. Recall $\tilde{Y}_{il} = \boldsymbol{X}_i(T_{il})^\top \boldsymbol{\beta}(T_{il})$, $\tilde{\boldsymbol{Y}}_{\boldsymbol{i}} = (\tilde{Y}_{i1}, \dots, \tilde{Y}_{in_i})^\top$, $\tilde{\boldsymbol{Y}} = (\tilde{\boldsymbol{Y}}_1, \dots, \tilde{\boldsymbol{Y}}_n)^\top$, $\tilde{\boldsymbol{\gamma}} = \mathsf{E}(\hat{\boldsymbol{\gamma}}|\mathcal{D}) = \sum_{i=1}^n \left(\mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \tilde{\boldsymbol{Y}}_i$, $\tilde{\boldsymbol{\beta}}(t) = \mathbb{B}(t)\tilde{\boldsymbol{\gamma}}$, and $\hat{\boldsymbol{\beta}}(t) = \mathbb{B}(t)\hat{\boldsymbol{\gamma}}$.

It holds

$$\hat{\boldsymbol{\beta}}(t) - \tilde{\boldsymbol{\beta}}(t) = \mathbb{B}(t)\left(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\right),$$

so $\hat{\boldsymbol{\beta}}(t) - \tilde{\boldsymbol{\beta}}(t)$ can be considered a linear transformation of $\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}$ and it is sufficient to show

$$\{\mathrm{var}(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}|\,\mathcal{D})\}^{-\frac{1}{2}}\left(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\right) \xrightarrow{d} \mathcal{N}(0, \mathbb{I}_K). \tag{3.33}$$

By the Cramér-Wold device, it is sufficient to show that for any vector $\boldsymbol{c} \in \mathbb{R}^K$, $\boldsymbol{c} \neq (0, \dots, 0)^\top$

$$\frac{\boldsymbol{c}^\top(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})}{\sqrt{\mathrm{var}\left(\boldsymbol{c}^\top(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})|\,\mathcal{D}\right)}} \xrightarrow{d} \mathcal{N}(0,1).$$

The expressions can be rewritten as

$$\begin{aligned}
\boldsymbol{c}^\top(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}) &= \sum_{i=1}^n \boldsymbol{c}^\top \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{U}_i^\top \mathbb{W}_i \boldsymbol{\varepsilon}_i \\
&= \sum_{i=1}^n \boldsymbol{c}^\top \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i^{1/2} \mathbb{V}_i^{-1/2} \boldsymbol{\varepsilon}_i \\
&= \sum_{i=1}^n \boldsymbol{b}_i^\top \mathbb{V}_i^{-1/2} \boldsymbol{\varepsilon}_i = \sum_{i=1}^n \sqrt{\boldsymbol{b}_i^\top \boldsymbol{b}_i}\, \frac{1}{\sqrt{\boldsymbol{b}_i^\top \boldsymbol{b}_i}} \boldsymbol{b}_i^\top \mathbb{V}_i^{-1/2} \boldsymbol{\varepsilon}_i = \sum_{i=1}^n a_i \xi_i,
\end{aligned} \tag{3.34}$$

$$\mathrm{var}\left(\boldsymbol{c}^\top(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})\middle|\,\mathcal{D}\right) = \boldsymbol{c}^\top \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \boldsymbol{c}, \tag{3.35}$$

where

$$a_i = \sqrt{\boldsymbol{b}_i^\top \boldsymbol{b}_i} = \sqrt{\boldsymbol{c}^\top \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \boldsymbol{c}},$$

$$\boldsymbol{b}_i^\top = \boldsymbol{c}^\top \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i^{1/2}, \quad \xi_i = \frac{1}{\sqrt{\boldsymbol{b}_i^\top \boldsymbol{b}_i}} \boldsymbol{b}_i^\top \mathbb{V}_i^{-1/2} \varepsilon_i.$$

It holds that $a_i^2 = \boldsymbol{c}^\top \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \boldsymbol{c}$ and, conditioning on $\mathcal{D}$, ξ_i are independent random variables with

$$\mathsf{E}\left[\xi_i|\mathcal{D}\right] = 0, \quad \operatorname{var}\left[\xi_i|\mathcal{D}\right] = 1,$$

hence the random variables $a_i\xi_i$ have, conditionally on $\mathcal{D}$, mean zero and variance $a_i^2 = \boldsymbol{c}^\top \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \boldsymbol{c}$. It follows that $\mathsf{E}\left[a_i\xi_i|\mathcal{D}\right] = 0$, $\operatorname{var}\left[a_i\xi_i|\mathcal{D}\right] = a_i^2$. They are independent, but heteroscedastic. We use the Feller-Lindenberg central limit theorem to show the asymptotic normality.

If the Lindenberg's condition is satisfied, then by the central limit theorem

$$\sum_{i=1}^n \frac{a_i\xi_i}{\sqrt{\sum_{i=1}^n a_i^2}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1). \tag{3.36}$$

Assume that

$$\frac{\max\limits_{i=1,\dots,n} a_i^2}{\sum_{i=1}^n a_i^2} \xrightarrow[n\to\infty]{} 0. \tag{3.37}$$

Define $s_n^2 = \sum_{i=1}^n a_i^2$, $m_n = \max a_i^2$.

The Lindeberg's condition states, that for every $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E}\left[a_i^2\xi_i^2 \mathbf{1}_{\{|a_i\xi_i|>\varepsilon s_n\}}|\mathcal{D}\right] = 0. \tag{3.38}$$

It holds, that

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathsf{E}\left[a_i^2\xi_i^2 \mathbb{1}_{\{|a_i\xi_i|\geq\varepsilon s_n\}}|\mathcal{D}\right] &= \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathsf{E}\left[a_i^2\xi_i^2 \mathbb{1}_{\left\{|\xi_i^2|\geq\frac{\varepsilon^2 s_n^2}{a_i^2}\right\}}\middle|\mathcal{D}\right] \\
&= \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathsf{E}\left[a_i^2\xi_i^2 \mathbb{1}_{\left\{|\xi_i^2|\geq\frac{\varepsilon^2 s_n^2}{a_i^2}\right\}}\middle|\mathcal{D}\right] \\
&\leq \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathsf{E}\left[a_i^2\xi_1^2 \mathbb{1}_{\left\{|\xi_1^2|\geq\frac{\varepsilon^2 s_n^2}{m_n}\right\}}\middle|\mathcal{D}\right] \\
&= \lim_{n\to\infty} \frac{1}{s_n^2} \mathsf{E}\left[s_n^2\xi_1^2 \mathbb{1}_{\left\{|\xi_1^2|\geq\frac{\varepsilon^2 s_n^2}{m_n}\right\}}\middle|\mathcal{D}\right] \\
&= \lim_{n\to\infty} \mathsf{E}\left[\xi_1^2 \mathbb{1}_{\left\{|\xi_1^2|\geq\frac{\varepsilon^2 s_n^2}{m_n}\right\}}\middle|\mathcal{D}\right].
\end{aligned} \tag{3.39}$$

Thus the condition (3.37) implies the Lindenberg's condition. From there (3.39) holds. It remains to show the condition (3.37).

Denote a vector $\boldsymbol{\lambda} = \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \boldsymbol{c} = \left(\boldsymbol{\lambda}_1^\top, \dots, \boldsymbol{\lambda}_p^\top\right)^\top$, where $\boldsymbol{\lambda}_j^\top = \left(\lambda_{j1}, \dots, \lambda_{jK_j}\right)$ for $j = 1, \dots, p$. Then

$$a_i^2 = \boldsymbol{\lambda}^\top \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \boldsymbol{\lambda}. \tag{3.40}$$

Let us start with the expression in the numerator of the Lindenberg's condition (3.37)

$$\boldsymbol{\lambda}^\top \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \boldsymbol{\lambda}.$$

It follows from condition (C3) that for any vector $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{n_i})^\top$ of length n_i

$$\boldsymbol{\psi}^\top \mathbb{V}_i \boldsymbol{\psi} = \mathsf{E}\left[\left(\sum_{l=1}^{n_i} \psi_l \varepsilon_i(T_{il})\right)^2\right] \leq \|\boldsymbol{\psi}\|_2^2 \sum_{l=1}^{n_i} \mathsf{E}\left(\varepsilon_i^2(T_{il})\right) \lesssim n_i \|\boldsymbol{\psi}\|_2^2, \tag{3.41}$$

hence by the Cauchy-Schwartz inequality, (C3), (C4), and the properties of the B-splines

$$\begin{aligned}
\boldsymbol{\lambda}^\top \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \boldsymbol{\lambda} &\lesssim n_i \boldsymbol{\lambda}^\top \mathbb{U}_i^\top \mathbb{W}_i \mathbb{W}_i \mathbb{U}_i \boldsymbol{\lambda} \\
&= \frac{1}{n_i} \boldsymbol{\lambda}^\top \mathbb{U}_i^\top \mathbb{U}_i \boldsymbol{\lambda} = \frac{1}{n_i} \sum_{l=1}^{n_i} \boldsymbol{\lambda}^\top \boldsymbol{U}_{il}^\top \boldsymbol{U}_{il} \boldsymbol{\lambda} \\
&= \frac{1}{n_i} \sum_{l=1}^{n_i} \left(\sum_{j=1}^{p} X_{ij}(T_{il}) \sum_{k=1}^{K_j} \lambda_{jk} B_{jk}(T_{il})\right)^2 \\
&\leq \frac{1}{n_i} \sum_{l=1}^{n_i} \left[\sum_{j=1}^{p} X_{ij}^2(T_{il}) \sum_{j=1}^{p} \left(\sum_{k=1}^{K_j} \lambda_{jk} B_{jk}(T_{il})\right)^2\right] && (3.42) \\
&\lesssim \sum_{j=1}^{p} \left\|\sum_{k=1}^{K_j} \lambda_{jk} B_{jk}\right\|_\infty^2 \lesssim K_n \sum_{j=1}^{p} \left\|\sum_{k=1}^{K_j} \lambda_{jk} B_{jk}\right\|^2 \lesssim \|\boldsymbol{\lambda}\|_2^2. && (3.43)
\end{aligned}$$

Let us now consider the denominator of (3.37). It follows from (C6) that $\mathbb{V}_i = \sigma^2 \mathbb{I}_{n_i} + \tilde{\mathbb{V}}_i$, where $\tilde{\mathbb{V}}_i$ is a positive semidefinite matrix. From there

$$\begin{aligned}
\boldsymbol{\lambda}^\top \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i\right) \boldsymbol{\lambda} &\geq \sigma^2 \boldsymbol{\lambda}^\top \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{W}_i \mathbb{U}_i\right) \boldsymbol{\lambda} \\
&= n\sigma^2 \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i^2} \sum_{l=1}^{n_i} \left(\sum_{j=1}^{p} X_{ij}(T_{il}) \sum_{k=1}^{K_j} \lambda_{jk} B_{jk}(T_{il})\right)^2\right] \\
&\geq \sigma^2 \min_{i=1,\ldots,n} \frac{n}{n_i} \left\|\sum_{k=1}^{K_j} \lambda_{jk} B_{jk}\right\|_n^2 \gtrsim \frac{n}{\min\limits_{i=1,\ldots,n} n_i} \left\|\sum_{k=1}^{K_j} \lambda_{jk} B_{jk}\right\|_n^2
\end{aligned} \tag{3.44}$$

It holds $\|\sum_{k=1}^{K_j} \lambda_{jk} B_{jk}\|_n^2 \asymp \|\sum_{k=1}^{K_j} \lambda_{jk} B_{jk}\|^2$ almost surely, for details, see Lemma A2 in Huang et al. (2004). By Lemma 2 we have $\|\sum_{k=1}^{K_j} \lambda_{jk} B_{jk}\|^2 \asymp \frac{\|\boldsymbol{\lambda}\|_2^2}{K_n}$, thus

$$\boldsymbol{\lambda}^\top \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i\right) \boldsymbol{\lambda} \gtrsim \frac{n}{\min\limits_{i=1,\ldots,n} n_i} \frac{1}{K_n} \|\boldsymbol{\lambda}\|_2^2. \tag{3.45}$$

From (3.43) and (3.45)

$$\frac{\max\limits_{i=1,\ldots,n} a_i^2}{\sum_{i=1}^{n} a_i^2} = \frac{\max\limits_{i=1,\ldots,n} \boldsymbol{\lambda}^\top \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i \boldsymbol{\lambda}}{\boldsymbol{\lambda}^\top \left(\sum_{i=1}^{n} \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i\right) \boldsymbol{\lambda}} \lesssim \max_{i=1,\ldots,n} n_i \frac{K_n}{n} \tag{3.46}$$

almost surely. The expression (3.37) follows from the assumption $\lim\limits_{n\to\infty} K_n \frac{\max_i n_i}{n} = 0$. □

Corollary 2.1. *Suppose $t_1, \ldots, t_m \in [0,1], m \in \mathbb{N}$ and let the assumptions of Theorem 2 be satisfied. Then for all $j = 1, \ldots, p$*

$$\mathbb{V}^{-\frac{1}{2}} \left(\hat{\beta}_j(t_1) - \tilde{\beta}_j(t_1), \ldots, \hat{\beta}_j(t_m) - \tilde{\beta}_j(t_m)\right)^\top \xrightarrow{d} \mathcal{N}_m(\mathbf{0}, \mathbb{I}_m),$$

where $\mathbb{V}$ is a $m \times m$ matrix with $(q,r)-th$ element equal to $\boldsymbol{e}_j^\top \mathbb{B}(t_q) \operatorname{Var}(\hat{\boldsymbol{\gamma}}|\mathcal{D}) \mathbb{B}^\top(t_r) \boldsymbol{e}_j$.

Proof. Recall $\hat{\beta}_j(t) = \boldsymbol{e}_j^\top \mathbb{B}(t) \hat{\boldsymbol{\gamma}}$, $\tilde{\beta}_j(t) = \boldsymbol{e}_j^\top \mathbb{B}(t) \tilde{\boldsymbol{\gamma}}$. From the proof of the Theorem 2,

$$\{\operatorname{Var}(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}|\, \mathcal{D})\}^{-\frac{1}{2}} (\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}) \xrightarrow{d} \mathcal{N}(0, \mathbb{I}_K).$$

Consequently,

$$\begin{pmatrix} \hat{\beta}_j(t_1) - \tilde{\beta}_j(t_1) \\ \vdots \\ \hat{\beta}_j(t_m) - \tilde{\beta}_j(t_m) \end{pmatrix} = \begin{pmatrix} \boldsymbol{e}_j^\top \mathbb{B}(t_1) \\ \vdots \\ \boldsymbol{e}_j^\top \mathbb{B}(t_m) \end{pmatrix} (\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}) \tag{3.47}$$

as a linear transformation of $(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})$ satisfies

$$\mathbb{V}^{-\frac{1}{2}} \begin{pmatrix} \hat{\beta}_j(t_1) - \tilde{\beta}_j(t_1) \\ \vdots \\ \hat{\beta}_j(t_m) - \tilde{\beta}_j(t_m) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \mathbb{I}_m), \tag{3.48}$$

where $\mathbb{V}$ is a $m \times m$ matrix

$$\begin{pmatrix} \boldsymbol{e}_j^\top \mathbb{B}(t_1) \\ \vdots \\ \boldsymbol{e}_j^\top \mathbb{B}(t_m) \end{pmatrix} \operatorname{Var}(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}|\, \mathcal{D}) \begin{pmatrix} \boldsymbol{e}_j^\top \mathbb{B}(t_1) \\ \vdots \\ \boldsymbol{e}_j^\top \mathbb{B}(t_m) \end{pmatrix}^\top. \tag{3.49}$$

□

We would like to replace $\tilde{\boldsymbol{\beta}}(\cdot)$ by $\boldsymbol{\beta}(\cdot)$ in theorem 2 and Corollary 2.1 to be able to make inference about $\boldsymbol{\beta}(\cdot)$, rather than $\tilde{\boldsymbol{\beta}}(\cdot)$ (we could however consider $\tilde{\boldsymbol{\beta}}(\cdot)$ the estimable part of $\boldsymbol{\beta}(\cdot)$). This replacement can be justified by the following theorems.

Theorem 3 (Bias)**.** *Suppose conditions (C1) – (C4) hold, and $\lim\limits_{n\to\infty} K_n \frac{\log K_n}{n} = 0$. Then $\sup\limits_{t\in[0,1]} |\tilde{\beta}_j(t) - \beta_j(t)| = O_P(\rho_n), j = 1, \ldots, p$.*

Proof. See Lemmas A9, A10, A11 in Huang et al. (2004). □

Under an additional assumption that β_j have bounded second derivatives, we have $\rho_n \asymp \frac{1}{K_n^2}$ (Schumaker, 2007).

Corollary 3.1. *Suppose conditions (C1) – (C5) hold,* $\lim\limits_{n\to\infty} K_n \frac{\log K_n}{n} = 0$, $\lim\limits_{n\to\infty} \frac{K_n^5}{n \max\limits_{i=1,\dots,n} n_i} = 0$, *and* $\beta_j(\cdot)$, $j = 1,\dots,p$, *have bounded second derivatives. Set* $m \in \mathbb{N}$, *and* $\boldsymbol{c} \in \mathbb{R}^m$, $\boldsymbol{c} \neq \mathbf{0}$, *then*

$$\sup_{t_1,\dots t_m \in [0,1]} \left| \left(\boldsymbol{c}^\top \mathbb{V} \boldsymbol{c}\right)^{-1/2} \left(\boldsymbol{c}^\top \left((\tilde{\beta}_j(t_1),\dots,\tilde{\beta}_j(t_m))^\top - (\beta_j(t_1),\dots,\beta_j(t_m))^\top \right)\right) \right| = o_p(1),$$

where $\mathbb{V}$ *is a variance-covariance matrix of* $((\hat{\beta}_j(t_1),\dots,\hat{\beta}_j(t_m))^\top$ *conditioning on* $\mathcal{D}$*. In particular* $\sup\limits_{t\in[0,1]} \left| \left(\operatorname{var}\left(\hat{\beta}_j(t)\,|\mathcal{D}\right)\right)^{-1/2} \left(\tilde{\beta}_j(t) - \beta_j(t)\right) \right| = o_p(1)$.

Proof.

$$\begin{aligned} \boldsymbol{c}^\top \mathbb{V} c &= \boldsymbol{c}^\top \mathbb{J} \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i\right) \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{J}^\top \boldsymbol{c} \\ &= \boldsymbol{\lambda}^\top \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{V}_i \mathbb{W}_i \mathbb{U}_i\right) \boldsymbol{\lambda}, \end{aligned}$$

for $\boldsymbol{\lambda} = \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{J}^\top \boldsymbol{c}$, where

$$\mathbb{J} = \begin{pmatrix} \boldsymbol{e}_j^\top \mathbb{B}(t_1) \\ \vdots \\ \boldsymbol{e}_j^\top \mathbb{B}(t_m) \end{pmatrix}.$$

From Inequality (3.45),

$$\boldsymbol{c}^\top \mathbb{V} c \gtrsim \frac{n}{\max\limits_{i=1,\dots,n} n_i} \frac{1}{K_n} \|\boldsymbol{\lambda}\|_2^2.$$

It holds

$$\|\boldsymbol{\lambda}\|_2^2 = \left\| \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{J}^\top \boldsymbol{c} \right\|_2^2 = \left\| \frac{K_n}{n} \frac{n}{K_n} \left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1} \mathbb{J}^\top \boldsymbol{c} \right\|_2^2.$$

According to Lemma 1, the matrix $\frac{n}{K_n}\left(\sum_{i=1}^n \mathbb{U}_i^\top \mathbb{W}_i \mathbb{U}_i\right)^{-1}$ has bounded eigenvalues, hence it is easy to see, e.g. by the spectral decomposition, that

$$\|\boldsymbol{\lambda}\|_2^2 \gtrsim \left\| \frac{K_n}{n} \mathbb{J}^\top \boldsymbol{c} \right\|_2^2 \gtrsim \frac{K_n^2}{n^2}.$$

The asymptotic upper bound for $\boldsymbol{c}^\top \mathbb{V} c$ is then

$$\frac{K_n}{n \max\limits_{i=1,\dots,n} n_i}.$$

Because $\beta_j(\cdot)$, $j = 1,\dots,p$, have bounded second derivatives, $\rho_n \asymp 1/K_n^2$, and from Theorem 3

$$\sup_{t_1,\dots,t_m \in [0,1]} \left| \boldsymbol{c}^\top \left((\tilde{\beta}_j(t_1),\dots,\tilde{\beta}_j(t_m))^\top - (\beta_j(t_1),\dots,\beta_j(t_m))^\top \right) \right| = O_P(1/K_n^2).$$

From there the statement of the theorem follows.

The special case follows by setting $\boldsymbol{c} = \boldsymbol{e}_1$. □

Corollary 3.2. *Let $t_1, \dots, t_m \in [0,1], m \in \mathbb{N}$ and let the assumptions of Corollary 3.1 be satisfied. Then for all $j = 1, \dots, p$*

$$\mathbb{V}^{-\frac{1}{2}} \left(\hat{\beta}_j(t_1) - \beta_j(t_1), \dots, \hat{\beta}_j(t_m) - \beta_j(t_m)\right)^\top \xrightarrow{d} \mathcal{N}_m(\mathbf{0}, \mathbb{I}_m),$$

where $\mathbb{V}$ is a $m \times m$ matrix with $(q,r)-th$ element equal to $\boldsymbol{e}_j^\top \mathbb{B}(t_q) \operatorname{Var}(\hat{\boldsymbol{\gamma}}|\mathcal{D}) \mathbb{B}^\top(t_r) \boldsymbol{e}_j$.

Proof.

$$\mathbb{V}^{-\frac{1}{2}} \begin{pmatrix} \hat{\beta}_j(t_1) - \beta_j(t_1) \\ \vdots \\ \hat{\beta}_j(t_m) - \beta_j(t_m) \end{pmatrix} = \mathbb{V}^{-\frac{1}{2}} \begin{pmatrix} \hat{\beta}_j(t_1) - \tilde{\beta}_j(t_1) \\ \vdots \\ \hat{\beta}_j(t_m) - \tilde{\beta}_j(t_m) \end{pmatrix} + \mathbb{V}^{-\frac{1}{2}} \begin{pmatrix} \tilde{\beta}_j(t_1) - \beta_j(t_1) \\ \vdots \\ \tilde{\beta}_j(t_m) - \beta_j(t_m) \end{pmatrix},$$

the result follows from Corollary 2.1 and from Corollary 3.1. □

As a result, Theorems 3.1 and 3.2 can be used to construct asymptotic confidence intervals and asymptotic confidence bands for the coefficient functions $\beta_j(\cdot)$.

3.2 Confidence intervals and bands

3.2.1 Confidence intervals

From Theorem 2 for $j = 1, \dots, p$, and $t \in [0, 1]$

$$\operatorname{var}(\hat{\beta}_j(t)|\mathcal{D})^{-\frac{1}{2}} \left(\hat{\beta}_j(t) - \mathsf{E}(\hat{\beta}_j(t)|\mathcal{D})\right) \xrightarrow[n\to\infty]{P} \mathcal{N}(0,1), \tag{3.50}$$

where $\mathsf{E}(\hat{\beta}_j(t)|\mathcal{D})$ and $\operatorname{var}(\hat{\beta}_j(t)|\mathcal{D})$ are the mean and variance of $\hat{\beta}_j(t)$ conditioning on $\mathcal{D}$. For simplicity, we write just $\mathsf{E}(\hat{\beta}_j(t))$ and $\operatorname{var}(\hat{\beta}_j(t))$.

Suppose that there is an estimate $\widehat{\operatorname{var}}(\hat{\beta}_j(t))$ of $\operatorname{var}(\hat{\beta}_j(t))$ such that

$$\frac{\widehat{\operatorname{var}}(\hat{\beta}_j(t))}{\operatorname{var}(\hat{\beta}_j(t))} \xrightarrow[n\to\infty]{P} 1. \tag{3.51}$$

It follows from the Cramer-Slutsky theorem, that

$$\left\{\widehat{\operatorname{var}}(\hat{\beta}_j(t))\right\}^{-\frac{1}{2}} (\hat{\beta}_j(t) - \mathsf{E}(\hat{\beta}_j(t))) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1), \tag{3.52}$$

which gives us an approximate $(1-\alpha)\%$ asymptotic confidence interval for $\mathsf{E}(\hat{\beta}_j(t))$ with end points

$$\hat{\beta}_j(t) \pm z_{1-\alpha/2}\sqrt{\widehat{\operatorname{var}}(\hat{\beta}_j(t))}, \tag{3.53}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

If the conditions of Corollary 3.1 are satisfied, then by the Cramer-Slutsky theorem

$$\left\{\widehat{\operatorname{var}}(\hat{\beta}_j(t))\right\}^{-\frac{1}{2}} (\hat{\beta}_j(t) - \mathsf{E}(\hat{\beta}_j(t)) + \mathsf{E}(\hat{\beta}_j(t)) - \beta_j(t)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1) \tag{3.54}$$

and (3.53) is also a $(1-\alpha)\%$ asymptotic confidence interval for $\beta_j(t)$.

For completeness, let us also include simultaneous confidence intervals for all linear combinations of β_j at multiple time points. For $t_1, \ldots, t_m \in [0,1]$, $m \in \mathbb{N}$, denote $\boldsymbol{\beta}_j^m = (\beta_j(t_1), \ldots, \beta_j(t_m))^\top$ and $\hat{\boldsymbol{\beta}}_j^m = \left(\hat{\beta}_j(t_1), \ldots, \hat{\beta}_j(t_m)\right)^\top$. From Corollary 3.2 we know

$$\left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)^\top \mathbb{V}^{-1} \left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right) \xrightarrow{d} \chi_m^2 \quad \text{for}\, n \to \infty.$$

Let $\hat{\mathbb{V}}$ be a consistent estimator of $\mathbb{V}$. Then from the Cramer-Slutsky theorem

$$\left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)^\top \hat{\mathbb{V}}^{-1} \left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right) \xrightarrow{d} \chi_m^2. \tag{3.55}$$

It follows that the $(1-\alpha)$ confidence region for $\boldsymbol{\beta}_j^m$ is

$$\left\{\boldsymbol{\beta}_j^m \in \mathbb{R}^m : \left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)^\top \hat{\mathbb{V}}^{-1} \left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right) \leq \chi_m^2(1-\alpha)\right\}. \tag{3.56}$$

Additionally, for any vector $\boldsymbol{c} \in \mathbb{R}^m$, the confidence interval for the linear combination $\boldsymbol{c}^\top \boldsymbol{\beta}_j^m$ is

$$t(\boldsymbol{c}) = \left\{\boldsymbol{c}^\top \boldsymbol{\beta}_j^m : \boldsymbol{\beta}_j^m \in \mathbb{R}^m, \frac{\left(\boldsymbol{c}^\top \left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)\right)^2}{\boldsymbol{c}^\top \hat{\mathbb{V}} \boldsymbol{c}} \leq \chi_1^2(1-\alpha)\right\}, \tag{3.57}$$

where $\chi_1^2(1-\alpha)$ denotes $(1-\alpha)-$quantile of χ_1^2 distribution. We can construct simultaneous confidence intervals for all $\boldsymbol{c}^\top \boldsymbol{\beta}_j^m$ by calculating $\max\limits_{\boldsymbol{c} \in \mathbb{R}^m} t(\boldsymbol{c})$.

$$\begin{aligned}
\max_{\boldsymbol{c} \in \mathbb{R}^m} t(\boldsymbol{c}) &= \max_{\boldsymbol{c} \in \mathbb{R}^m} \frac{\left(\boldsymbol{c}^\top \left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)\right)\left(\boldsymbol{c}^\top \left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)\right)^\top}{\boldsymbol{c}^\top \hat{\mathbb{V}} \boldsymbol{c}} \\
&= \operatorname{Tr}\left(\hat{\mathbb{V}}^{-1}\left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)\left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)^\top\right) \qquad (3.58) \\
&= \operatorname{Tr}\left(\left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)^\top \hat{\mathbb{V}}^{-1}\left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)\right) \\
&= \left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)^\top \hat{\mathbb{V}}^{-1}\left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right), \qquad (3.59)
\end{aligned}$$

which has asymptotically χ_m^2 distribution. The equality (3.58) holds as a maximum of quadratic form (Theorem 2.5 Härdle and Simar, 2019), and because $\hat{\mathbb{V}}^{-1}\left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)\left(\hat{\boldsymbol{\beta}}_j^m - \boldsymbol{\beta}_j^m\right)^\top$ is a matrix of rank 1, the trace is equal to its only eigenvalue. It follows that the simultaneous confidence intervals for all $\boldsymbol{c}^\top \boldsymbol{\beta}_j^m$ (with simultaneous coverage level of $1-\alpha$) are

$$\left(\boldsymbol{c}^\top \hat{\boldsymbol{\beta}}_j^m - \sqrt{\chi_m^2(1-\alpha)\, \boldsymbol{c}^\top \hat{\mathbb{V}} \boldsymbol{c}}, \boldsymbol{c}^\top \hat{\boldsymbol{\beta}}_j^m + \sqrt{\chi_m^2(1-\alpha)\, \boldsymbol{c}^\top \hat{\mathbb{V}} \boldsymbol{c}}\right). \tag{3.60}$$

As an example we can construct the confidence intervals for each $\beta(t_k)$, $k = 1, \ldots, m$, by setting $\boldsymbol{c} = \boldsymbol{e}_k$

$$\left(\hat{\beta}_j(t_k) - \sqrt{\chi_m^2(1-\alpha)\, \widehat{\operatorname{var}}(\hat{\beta}_j(t))}, \hat{\beta}_j(t_k) + \sqrt{\chi_m^2(1-\alpha)\, \widehat{\operatorname{var}}(\hat{\beta}_j(t))}\right). \tag{3.61}$$

As we can see, the distinguishing factor between the interval in question and the previously proposed interval in (3.53) lies in the choice of the quantile. In general, the coverage of (3.61) is quite conservative, and this method will provide wider confidence intervals then the Bonferroni adjustment for each individual confidence interval for $\beta_j(t_k)$, see Section 3.2.2. However, suppose we want to analyze the differing impact of a covariate (such as a dosage of a medication) on a particular outcome (like blood pressure) at varying time points, e.g. differing by a particular time unit. In this case, we could express the pairwise difference in these effects by a linear combination of the coefficient functions at those times. (3.60) gives an easy way to construct many of such intervals that hold simultaneously.

Pointwise confidence intervals are not always sufficient in application. This is because a $(1-\alpha)\%$ pointwise confidence interval $(L_{\alpha/2}(t),\, U_{\alpha/2}(t))$ only guarantees that

$$\mathsf{P}\left(L_{\alpha/2}(t) \leq \beta_j(t) \leq U_{\alpha/2}(t)\right) = 1 \tag{3.62}$$

for some given $t \in [0,1]$, which does not imply

$$\mathsf{P}\left(L_{\alpha/2}(t) \leq \beta_j(t) \leq U_{\alpha/2}(t) \text{ for any } t \in [0,1]\right) = 1-\alpha. \tag{3.63}$$

More interesting are simultaneous confidence bands.

3.2.2 Confidence bands

We might use the pointwise confidence intervals to construct simultaneous confidence bands by adjusting the level of significance in a Bonferroni-like way as proposed by Knafl et al. (1985). The idea is to construct pointwise confidence intervals on a fine grid of $[0,1]$ while adjusting the confidence level (for example by the Bonferroni method) to obtain simultaneous confidence band and bridge the gaps between the points by imposing some smoothness conditions on the coefficient function. Huang et al. (2004) suggested calculating the confidence bands in the following manner. Partition the interval $[0,1]$ using $M+1,\ M \in \mathbb{N}$, equidistant points $0 = \zeta_1 < \cdots < \zeta_{M+1} = 1$ and consider the set of asymptotic confidence intervals $\left(L_{\alpha/(2(M+1))}\left(\zeta_r\right), U_{\alpha/(2(M+1))}\left(\zeta_r\right)\right)$ for $r = 1, \ldots, M+1$ as given by (3.53). Let $L^I_{\alpha/(2(M+1))}(t)$, $U^I_{\alpha/(2(M+1))}(t)$, and $\mathsf{E}^I(\hat{\beta}_j(t))$, $\zeta_r \leq t \leq \zeta_{r+1}$, be the linear interpolation of $L_{\alpha/(2(M+1))}\left(\zeta_r\right)$ and $L_{\alpha/(2(M+1))}\left(\zeta_{r+1}\right)$, $U_{\alpha/(2(M+1))}\left(\zeta_r\right)$ and $U_{\alpha/(2(M+1))}\left(\zeta_{r+1}\right)$, and $\mathsf{E}\,(\hat{\beta}_j(\zeta_r))$ and $\mathsf{E}\,(\hat{\beta}_j(\zeta_{r+1}))$, respectively. For example,

$$\mathsf{E}^I\,\hat{\beta}_j(t) = M(\zeta_{r+1}-t)\,\mathsf{E}\,\hat{\beta}_j(\zeta_r) + M(t-\zeta_r)\,\mathsf{E}\,\hat{\beta}_j(\zeta_{r+1}). \tag{3.64}$$

Let us first look at the confidence bands for $\mathsf{E}\,(\hat{\beta}_j(t))$. Assume that either

$$\sup_{t\in[0,\,1]} |\{\mathsf{E}\,(\hat{\beta}_j(t))\}'| \leq c_1 \quad \text{for some known positive constant } c_1 \tag{3.65}$$

or

$$\sup_{t\in[0,\,1]} |\{\mathsf{E}\,(\hat{\beta}_j(t))\}''| \leq c_2 \quad \text{for some known positive constant } c_2. \tag{3.66}$$

Now by assuming condition (3.66) by using the Taylor theorem it is easy to show

$$\left|\mathsf{E}\,\hat{\beta}_j(t) - \mathsf{E}^{\mathrm{I}}\,\hat{\beta}_j(t)\right| \leq \frac{c_2}{2}\left(\zeta_{r+1}-t\right)\left(t-\zeta_r\right). \tag{3.67}$$

Similarly, in the case of (3.65), it can be shown that

$$\left|\mathsf{E}\,\hat{\beta}_j(t) - \mathsf{E}^{\mathrm{I}}\,\hat{\beta}_j(t)\right| \leq 2c_1 M\left(\zeta_{r+1} - t\right)\left(t - \zeta_r\right). \tag{3.68}$$

$\left(L^{\mathrm{I}}_{\alpha/(2(M+1))}\left(t\right), U^{\mathrm{I}}_{\alpha/(2(M+1))}\left(t\right)\right)$ is an approximate $(1-\alpha)$ confidence band for $\mathsf{E}^{\mathrm{I}}\,\hat{\beta}_j(t)$. By adjusting the width we get an approximate $(1-\alpha)$ confidence band for $\mathsf{E}\,\hat{\beta}_j(t)$

$$\left(L^{\mathrm{I}}_{\alpha/(2(M+1))}\left(t\right) - 2Mc_1\left(\zeta_{r+1} - t\right)\left(t - \zeta_r\right), U^{\mathrm{I}}_{\alpha/(2(M+1))}\left(t\right) + 2Mc_1\left(\zeta_{r+1} - t\right)\left(t - \zeta_r\right)\right) \tag{3.69}$$

in the case of (3.65), or

$$\left(L^{\mathrm{I}}_{\alpha/(2(M+1))}\left(t\right) - \frac{c_2}{2}\left(\zeta_{r+1} - t\right)\left(t - \zeta_r\right), U^{\mathrm{I}}_{\alpha/(2(M+1))}\left(t\right) + \frac{c_2}{2}\left(\zeta_{r+1} - t\right)\left(t - \zeta_r\right)\right) \tag{3.70}$$

in the case of (3.66). The confidence bands are confidence bands also for $\beta_j\left(\cdot\right)$, if the assumptions of Corollary 3.1 hold and if the conditions (3.65), resp. (3.66) hold for $\beta_j\left(\cdot\right)$, i.e

$$\sup_{t\in[0,\,1]} |\{\beta_j(t)\}^{'}| \leq c_1 \quad \text{for some known positive constant}\, c_1 \tag{3.71}$$

or

$$\sup_{t\in[0,\,1]} |\{\beta_j(t)\}^{''}| \leq c_2 \quad \text{for some known positive constant}\, c_2. \tag{3.72}$$

The issue is how to choose the number of intervals M, as there are no available guidelines, and the constants c_1, resp. c_2. The values c_1 and c_2 should be chosen carefully based on an expert's opinion. Another issue is the choice of the significance level of each interval. The suggested Bonferroni adjustment has an advantage of being simple to implement as it does not require any information about the correlation. However, it could lead to too conservative bands, especially for large M.

We investigate the usage of another confidence level adjustment, based on the inclusion-exclusion principle, taking the correlation structure into account. This way we should get narrower confidence intervals and consequently also confidence bands. On the other hand, the following procedure relies heavily on a consistent estimator of the covariance structure, therefore the confidence bad might be too narrow for data with insufficient n.

The inclusion-exclusion principle is a fundamental concept in combinatorics and probability theory. It provides a way to calculate the probability of the union of multiple events.

For each ζ_r, where $r = 1, 2, ..., M+1$, let A_r denote the event that the confidence interval at ζ_r does not cover the true coefficient function, i.e., either the upper limit of the confidence interval is below the true coefficient function, or the lower limit is above the true coefficient function at ζ_r. Formally, let $(L_{\alpha/2}(\zeta_r), U_{\alpha/2}(\zeta_r))$ be the confidence interval at ζ_r given in (3.53), then

$$A_r(\alpha) = \{\omega : \beta_r(\zeta_r) < L_{\alpha/2}(\zeta_r;\omega) \text{ or } \beta_r(\zeta_r) > U_{\alpha/2}(\zeta_r;\omega)\} \tag{3.73}$$

where ω belongs to the underlying probability space Ω. For simplicity, we write just A_r instead of $A_r(\alpha)$.

By the principle of inclusion and exclusion

$$\mathsf{P}(\bigcup_{r=1}^{M+1} A_r) = \sum_{r=1}^{M+1} \mathsf{P}(A_r) - \sum_{1\leq i<j\leq M+1} \mathsf{P}(A_i \cap A_j) + \sum_{1\leq i<j<k\leq M+1} \mathsf{P}(A_i \cap A_j \cap A_k)$$
$$- \cdots + (-1)^{M+1}\,\mathsf{P}(A_1 \cap A_2 \cap \cdots \cap A_{M+1}). \tag{3.74}$$

By considering the intersections just up to order $R \in \mathbb{N}, R \leq M$, we either overestimate or underestimate the probability of union. For odd R from Bonferroni (1936)

$$\mathsf{P}\left(\bigcup_{r=1}^{M+1} A_r\right) \leq \sum_{r=1}^{M+1} \mathsf{P}(A_r) - \sum_{1\leq i,j\leq M+1 \;\; i\neq j} \mathsf{P}(A_i \cap A_j) \tag{3.75}$$
$$+ \ldots + (-1)^k \sum_{1\leq i_1<i_2<\ldots<i_k\leq M+1} \mathsf{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}). \tag{3.76}$$

For $R = 1$ it is the well-known Bonferroni inequality (also known as Boole's inequality) which justifies the Bonferroni multiple-testing adjustment. Problematic is that the number of items (in our case integrals) one has to calculate in order to use these inequalities is $\sum_{r=1}^{R} \binom{M+1}{r}$, which is even for $R = 3$ at least for potentially reasonable choices of M quite large. Alternative formula was developed for $R = 2$ by Hunter (1976)

$$\mathsf{P}(\bigcup_{r=1}^{M+1} A_r) \leq \sum_{r=1}^{M+1} \mathsf{P}(A_r) - \sum_{e_{ij}\in T} \mathsf{P}(A_i \cap A_j), \tag{3.77}$$

where T is any spanning tree with vertices $A_1, \ldots, A_{M+1}$ and $e_{ij} \in T$ means that the vertex A_i is connected with the vertex A_j by an edge e_{ij} in T. Hoover (1990) expanded on the idea and provided extension for $R > 2$

$$\mathsf{P}(\bigcup_{r=1}^{M+1} A_r) \leq \mathsf{P}(\bigcup_{r=1}^{R} A_r) + \sum_{r=R+1}^{M+1} \mathsf{P}\left(A_r \bigcap_{i_1<\cdots<i_{k-1}} \left(A_{i_1} \cup \cdots \cup A_{i_{k-1}}\right)^C\right), \tag{3.78}$$

where $i_1, \ldots, i_{R-1} \in S_r$ for S_r set with $R-1$ elements from $\{1, 2, \ldots, r-1\}$.

It is easy to show that for $R = 1$ it is again the standard Bonferroni inequality, and for $R = 2$ it corresponds to (3.77). Inequality (3.78) is particularly appealing as it consists of calculating only $M - R$ probabilities, each based on just R events. Inequality (3.78) holds for any sets S_r, and the optimal choice vary depending on the situation. We suggest to use $S_r = \{r-1, r-2, \ldots, r-R+1\}$.

Let us now demonstrate how to use (3.78) for $R = 3$ in our context. Recall, that from Corollary 3.2 are $\left(\hat{\beta}_j(\xi_1), \ldots, \hat{\beta}_j(\xi_1)\right)^\top$ asymptotically normally distributed with the mean $(\beta_j(\xi_1), \ldots, \beta_j(\xi_{M+1}))^\top$ and variance matrix $\mathbb{V}$. Suppose we have a consistent estimator of the variance matrix $\hat{\mathbb{V}}$. The goal is to find a confidence level $\tilde{\alpha}$, such that

$$\mathsf{P}(\bigcup_{r=1}^{M+1} A_r) = \alpha \tag{3.79}$$

for $A_r = A_r(\tilde{\alpha})$. Further, set $\hat{\mathbb{V}}_{ijk}$ the submatrix of $\hat{\mathbb{V}}$ corresponding to the terms $\hat{\beta}_j(\xi_i)$, $\hat{\beta}_j(\xi_j)$, and $\hat{\beta}_j(\xi_k)$, additionally $\hat{\sigma}_i = \boldsymbol{e}_i^\top \hat{\mathbb{V}} \boldsymbol{e}_i$ is the variance of $\hat{\beta}_j(\xi_i)$.

Set $u = z_{1-\tilde{\alpha}/2}$. Inequality (3.78) is then

$$\mathsf{P}(\bigcup_{r=1}^{M+1} A_r) \leq \mathsf{P}(A_1 \cup A_2 \cup A_3) + \sum_{r=4}^{M+1} \mathsf{P}\left(A_r \cap (A_{r-1} \cup A_{r-2})^C\right), \tag{3.80}$$

where

$$\begin{aligned}\mathsf{P}(A_1 \cup A_2 \cup A_3) = \mathsf{P}(A_1) + \mathsf{P}(A_2) + \mathsf{P}(A_3) - \mathsf{P}(A_1 \cap A_2) - \mathsf{P}(A_2 \cap A_3) \\ - \mathsf{P}(A_1 \cap A_3) + \mathsf{P}(A_1 \cap A_2 \cap A_3),\end{aligned}$$

and we approximate

$$\begin{aligned}&\mathsf{P}\left(A_r \cap (A_{r-1} \cup A_{r-2})^C\right) \\ &= 2\int_{\hat{\sigma}_r u}^{\infty} \int_{-\hat{\sigma}_{r-1}u}^{\hat{\sigma}_{r-1}u} \int_{-\hat{\sigma}_{r-2}u}^{\hat{\sigma}_{r-2}u} \frac{1}{\sqrt{(2\pi)^3 |\hat{V}_{r(r-1)(r-2)}|}} \exp\left(\frac{-1}{2}\boldsymbol{x}^\top \hat{V}_{r(r-1)(r-2)}^{-1} \boldsymbol{x}\right) dx_3 dx_2 dx_1.\end{aligned}$$

$\mathsf{P}(A_1 \cap A_2), \mathsf{P}(A_2 \cap A_3), \mathsf{P}(A_1 \cap A_3), \mathsf{P}(A_1 \cap A_2 \cap A_3)$ can be rewritten similarly as sums of integrals with bounds depending on the standard deviations and the quantile u (in all of the cases we integrate over region, in which all the intervals do not cover the true coefficient function). It should be pointed out that the correlation structure is unknown, and in smaller samples, the actual coverage probability can deviate from the nominal level due to its estimation.

We propose the following procedure: choose u such that (3.79) holds, and $\tilde{\alpha}$ so that the confidence level $\tilde{\alpha}$ satisfies $u = z_{1-\tilde{\alpha}/2}$ and consequently $\tilde{\alpha} = 2 - 2\Phi(u)$, where Φ denotes the cumulative distribution function of the standard normal distribution.

The confidence band is then constructed analogously as when using the Bonferroni correction. The only difference is that the confidence level for each of the $M + 1$ confidence intervals is taken as $\tilde{\alpha}$ instead of $\alpha/(M + 1)$. It holds $\tilde{\alpha} \geq \alpha/(M + 1)$ resulting in narrower confidence bands. However, the construction is more complex and relies on a good estimate of the covariance matrix. In Chapter 4 we compare both procedures for various choices of $\beta_j(\cdot)$ and M to see whether the adjusted procedure is indeed considerably less conservative, or not, and if it could be an alternative to Bonferroni adjustment at least for large datasets.

3.2.3 Usage for hypothesis testing

Once the covariate functions are estimated, we would like to know: (1) if the estimated unknown coefficient functions are statistically significant in the fitted model, (2) if the coefficient functions are really varying, and (3) if the estimated unknown coefficient functions could be expressed in a certain parametric form. That corresponds to the following hypotheses:

$$H_{0j}: \forall\, t \in [0,1] : \beta_j(t) = \beta_{j,0}(t, \boldsymbol{\omega}_{\boldsymbol{j}}^\top) \quad \text{vs.} \quad H_{1j}: \exists t \in [0,1] : \beta_j(t) \neq \beta_{j,0}(t, \boldsymbol{\omega}_{\boldsymbol{j}}^\top),$$

where $\beta_{j,0}(\cdot, \boldsymbol{\omega}_{\boldsymbol{j}}^\top)$ represents the parametric function of interest depending on a vector of parameters $\boldsymbol{\omega}_j$. As a special case to test (2), set $\beta_{j,0}(\cdot, \boldsymbol{\omega}_{\boldsymbol{j}}^\top) \equiv \beta_{j,0}$, where $\beta_{j,0}$ is an unknown real constant. To test (1), set $\beta_{j,0}(\cdot, \boldsymbol{\omega}_{\boldsymbol{j}}^\top) \equiv 0$.

We can test the hypotheses (3.2.3) by using confidence bands. If all necessary assumptions are satisfied, see Theorem 3, we have pointwise normality, and can construct asymptotic confidence intervals for $\beta_j\left(t\right), t \in [0,1]$. Then we can construct simultaneous confidence band for the coefficient function $\beta_j\left(\right)$ by methods discussed in Section 3.2.2.

If the entire function from $\beta_{j,0}(t, \boldsymbol{\omega}_{\boldsymbol{j}}^{\top})$ falls within the $100(1-\alpha)\%$ confidence band, we cannot reject the null hypothesis H_{0j}. On the other hand, if any part of the parametric function lies outside of this confidence band, we reject the null hypothesis. It should be however again pointed out that such confidence bands tend could be conservative by using the Bonferroni correction (have a low power), or be too narrow when considering the Hoover's inequality correction, especially for small datasets.

4. Simulation study

In the last Chapter, we compare the performance of the different types of approximate confidence bands based on the asymptotic normality approach from Section 3.2.2 by conducting a small numerical study. All computations were performed using the R software (R Core Team, 2022).

4.1 General framework

Let us consider the longitudinal varying-coefficient model with a single covariate and coefficient function, i.e the case

$$Y(t) = \beta(t)X(t) + \varepsilon(t), \tag{4.1}$$

where $\varepsilon(t)$ is a zero-mean error process.

We will investigate the coverage of the confidence bands using the Monte Carlo simulation in various scenarios.

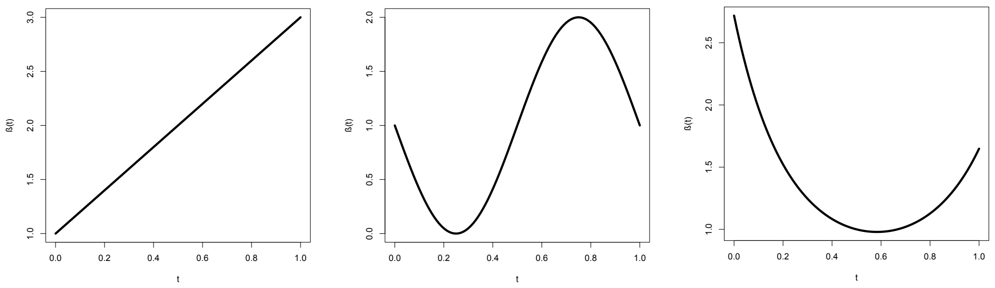
Data satisfying the model (4.1) are generated using the following mechanism. We consider $n = 100$, 200, 300, or 500 subjects, indexed as $i = 1, \ldots, n$. For each subject, we choose unequal number of observations $n_i \sim \text{Po}\,(8)$. If for some $i : n_i < 5$, we set $n_i = 5$, and if $n_i > 11$, we set $n_i = 11$ so that each subject has $5 - 11$ observations. All time points are generated as independent realisations of $T \sim \text{Unif}\,([0, 1])$ and ordered for each subject. The errors $\boldsymbol{\varepsilon}_i$ for each subject are independent samples from a zero-mean process ε, such that $\varepsilon_i(T_{il}) = Z_i(T_{il}) + \zeta_i(T_{il})$, where Z_i is a realization of a stationary Gaussian process with an autovocariance function $C_\varepsilon(s, t) = 2\exp(|s - t|^2/2)$ representing within-subject correlation and ζ_i is an independent realizations of $\mathcal{N}(0, 1)$ representing a measurement error. The covariates are chosen to be time invariant, generated $X_i \sim \mathcal{N}(5, 2)$ and $X_i(T_{il}) = X_i$ for all $l = 1, \ldots, n_i$.

We consider three different options for the coefficient function $\beta(t)$

1. $\beta^{\mathrm{A}}(t) = 1 + 2t$,
2. $\beta^{\mathrm{B}}(t) = 1 - \sin(2\pi t)$,
3. $\beta^{\mathrm{C}}(t) = e^{3(t-0.5)^2 + \frac{t-0.5}{2}}$.

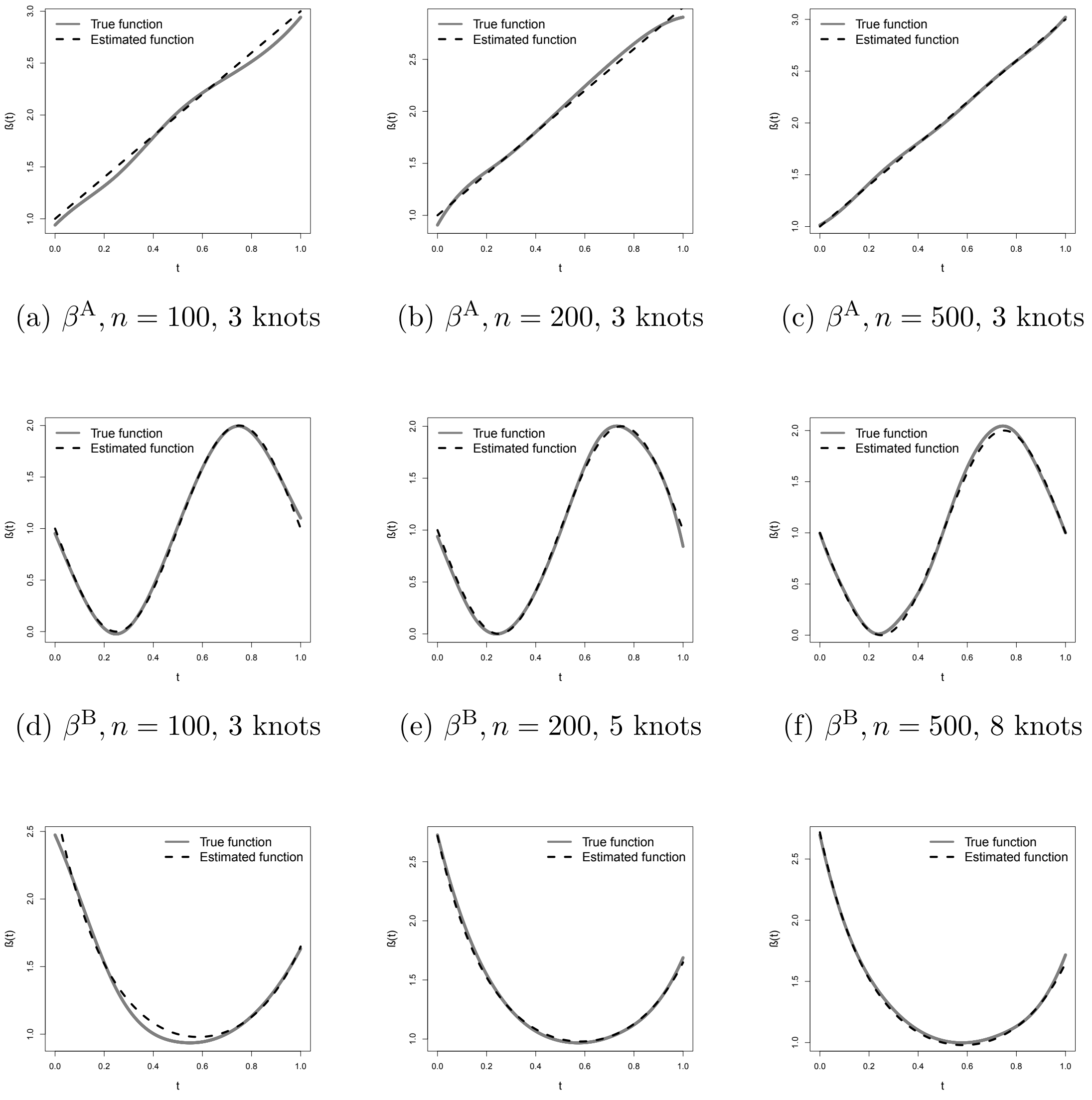
As we can see from Figure 4.1, β^{A} represents a linear trend, β^{B} is a sinusoid often encountered in natural sciences, and β^{C} represents a steep initial decline that slows down and reaches its minimum at $t = 0.5$, after which it increases, however at a slower pace then it decreased.

The coefficient functions are estimated using the polynomial spline estimator introduced in Section 2.1 with weights $\frac{1}{n_i}$. We use cubic splines ($d = 3$) and equidistant knots. The number of knots is chosen by the AIC criterion (2.17) due to the computational complexity of the leave-one-out CV (2.16), the number of inner knots for the estimation of the covariance function, and separately variance function of $\varepsilon(t)$ is chosen as 5, based on the recommendation of $5 - 10$ knots by Huang et al. (2004).



(a) $\beta^{\mathrm{A}}(t) = 1 + 2t$ (b) $\beta^{\mathrm{B}}(t) = 1 - \sin(2\pi t)$ (c) $\beta^{\mathrm{C}}(t) = e^{3(t-0.5)^2 + \frac{t-0.5}{2}}$

Figure 4.1: Graphs of the considered coefficient functions β^{A}, β^{B}, β^{C}.



(a) $\beta^{\mathrm{A}}, n = 100$, 3 knots (b) $\beta^{\mathrm{A}}, n = 200$, 3 knots (c) $\beta^{\mathrm{A}}, n = 500$, 3 knots

(d) $\beta^{\mathrm{B}}, n = 100$, 3 knots (e) $\beta^{\mathrm{B}}, n = 200$, 5 knots (f) $\beta^{\mathrm{B}}, n = 500$, 8 knots

(g) $\beta^{\mathrm{C}}, n = 100$, 3 knots (h) $\beta^{\mathrm{C}}, n = 200$, 3 knots (i) $\beta^{\mathrm{C}}, n = 500$, 3 knots

Figure 4.2: Polynomial spline estimates for $n = 100$, 200, 500, number of inner knots chosen through leave-one-out CV considering $3 - 8$ inner knots.

For all of these scenarios, we will calculate the empirical coverage probabilities of two types of confidence bands based on 1000 replications at the significance level 0.95 with $M = 50, 100, 250$ grid points used for the construction of the confidence intervals.

Let us construct the asymptotic confidence bands in (3.68). We consider two methods for choosing the confidence level of the $M + 1$ confidence intervals, the first one is based on the classical Bonferroni adjustment, we will refer to such confidence band as Bonferroni-adjusted band. The second one is based on the inclusion-exclusion principle, specifically on Inequality (3.78), we will refer to this confidence band as PIE-adjusted band. Recall, that the formula for the asymptotic confidence band for a coefficient function with bounded first derivatives (under some regularity conditions) is

$$\left(L^{\mathrm{I}}_{\alpha/(2(M+1))}(t) - 2Mc_1\left(\zeta_{r+1} - t\right)\left(t - \zeta_r\right), U^{\mathrm{I}}_{\alpha/(2(M+1))}(t) + 2Mc_1\left(\zeta_{r+1} - t\right)\left(t - \zeta_r\right)\right).$$

The formula depends on the choice of M and c_1, where c_1 is given in (3.71). The ideal choice of M is unknown, we consider 3 cases to compare the results. It should be pointed out that too small M (possibly $M = 50$) might not be the best choice, as the confidence band is constructed by fitting a parabola between two consecutive ends of confidence intervals and rather small M can provide quite high parabolas. The value c_1 is a constant. It should be evaluated based on an expert's opinion or past experience in the prospective field. We chose quite conservative values of c_1 : 8 for β^A, 12 for β^B, and 7 for β^C, in alignment with the available literature. The coverage was checked for 1001 grid points on $[0\,;\,1]$.

4.2 Results

Let us first asses the performance of the estimators. For that we define the square root of average squared error of an estimator $\hat{\beta}(\cdot)$ (RASE)

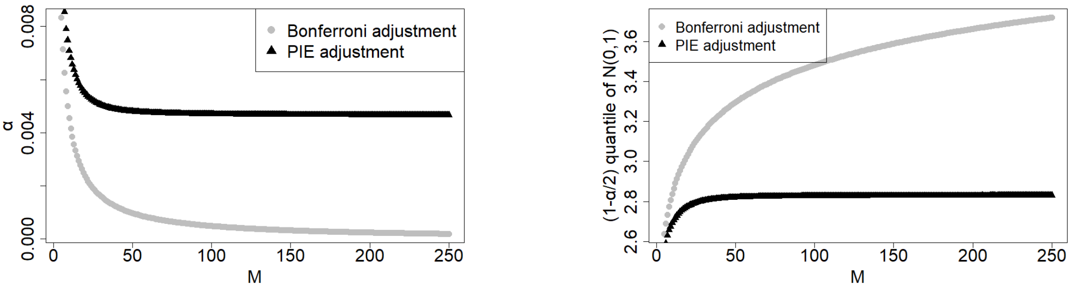
$$\mathrm{RASE} = \left[\frac{1}{N}\sum_{i=1}^{n}\sum_{l=1}^{n_i}\left(\hat{\beta}(T_{il}) - \beta(T_{il})\right)^2\right]^{1/2}. \tag{4.2}$$

Table 4.1 contains RASE estimators for all 12 cases.

	Leave-one-out CV			
	100	200	300	500
β^A	0.036 (0.012)	0.025 (0.07)	0.020 (0.007)	0.015 (0.005)
β^B	0.037 (0.012)	0.027 (0.09)	0.021 (0.006)	0.018 (0.006)
β^C	0.038 (0.012)	0.025 (0.08)	0.021 (0.007)	0.016 (0.006)
	AIC criterion			
	100	200	300	500
β^A	0.036 (0.013)	0.025 (0.008)	0.021 (0.007)	0.016 (0.006)
β^B	0.039 (0.012)	0.026 (0.009)	0.022 (0.008)	0.018 (0.006)
β^C	0.038 (0.014)	0.026 (0.008)	0.022 (0.006)	0.016 (0.005)

Table 4.1: Comparison of RASE mean(sd) for the leave-one-out cross validation and the Akaike criterion for subjects $n = 100$, $n = 200$, $n = 300$, $n = 500$, and the coefficient functions β^A, β^B, β^C, based 100 replications.

The two modifications of constructing the confidence bands, Bonferroni-adjustment and PIE-adjustment, differ only in the choice of the confidence levels of the confidence intervals α. The confidence levels α and the corresponding quantiles $u_{1-\frac{\alpha}{2}}$ depending on the number of intervals M are plotted on Figure 4.3. The Bonferroni-adjusted level of confidence is lower then the PIE-adjusted level of confidence. Additionally, with increasing number of points, the Bonferroni-adjusted level of significance continues to decrease much faster then the PIE-adjusted confidence level.



(a) Confidence levels.

(b) Quantiles.

Figure 4.3: An example of dependency of the confidence levels and quantiles on the number of confidence intervals M for $n = 100$, and β^A.

	Bonferroni correction			PIE Adjustment		
	M=50	M=100	M=250	M=50	M=100	M=250
β^A						
n=100	0.979	0.981	0.984	0.943	0.924	0.907
n=200	0.981	0.979	0.993	0.950	0.946	0.941
n=300	0.984	0.989	0.992	0.951	0.95	0.945
n=500	0.991	0.992	0.997	0.968	0.959	0.950
β^B						
n=100	0.977	0.982	0.992	0.931	0.933	0.924
n=200	0.988	0.988	0.992	0.960	0.955	0.939
n=300	0.990	0.991	0.992	0.962	0.959	0.950
n=500	0.987	0.993	0.993	0.967	0.964	0.953
β^C						
n=100	0.972	0.980	0.988	0.922	0.917	0.923
n=200	0.991	0.983	0.994	0.951	0.931	0.932
n=300	0.950	0.980	0.994	0.972	0.954	0.943
n=500	0.979	0.991	0.996	0.956	0.958	0.949

Table 4.2: An empirical coverage of asymptotic confidence bands based on the Bonferroni adjustment and PIE adjustment for $M = 50, 100, 250$, $n = 100, 200, 300, 500$ for coefficient functions β^A, β^B, β^C.

4.3 Discussion

From Table 4.1, the estimates get closer to the true value, as the sample size increases, which is in line with the consistence result from Section 3.1.1. We can see that the leave-one-out CV slightly outperforms the Akaike criterion, though it should be again noted that the leave-one-out CV is very computationally expensive. It seems that Akaike criterion in comparison performs rather well in our case with much less computational demands.

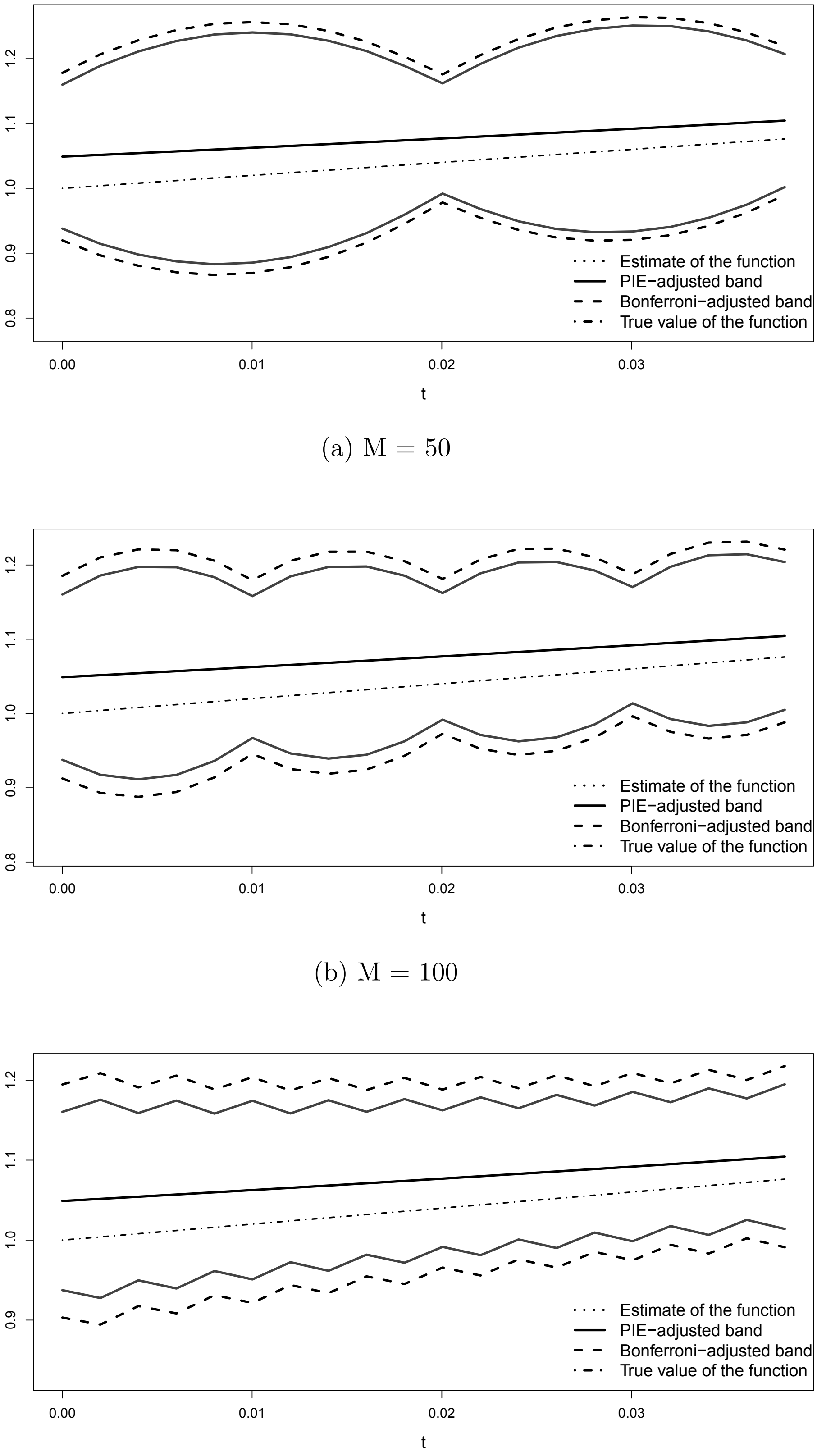
The results of the simulation study can be found in Table 4.2. We see that the Bonferroni-adjusted confidence bands are indeed very conservative. The PIE-adjusted confidence bands are much narrower. For insufficient sample size ($n = 100$, sometimes $n = 200$, $n = 300$) the PIE-adjusted bands do not maintain the confidence level 0.05. That is because the construction relies heavily on asymptotic results (consistency of both the estimate and the covariance structure).

The PIE-adjusted bands appear to be a quite reliable alternative to the Bonferroni-adjusted bands for larger datasets. The usage of PIE-adjusted bands for hypothesis testing as discussed in Section 3.2.3 would lead to higher power. One should be however careful and for small datasets use the Bonferroni-adjusted bands.

We considered three choices of M as the ideal choice is unknown. For the Bonferroni-adjusted bands the coverage probabilities tend to increase as M increases. That is because the bands gets wider (at least at the grid points where the confidence intervals were estimated) as the quantiles get larger, see Figure 4.3b. On the other hand, for the PIE-adjusted bands the coverage seems to decrease with increasing number of M. That is due to the fact that the quantiles increase at a much lower rate. The PIE-adjusted confidence bands were chosen such that the asymptotic simultaneous coverage of the confidence intervals is as close as possible to the threshold 0.05. It could be reasonable to slightly decrease the threshold and consider a more conservative construction, though still producing narrower confidence bands then the Bonferroni-adjusted bands.

It should be pointed out that the results of our simulation study are influenced by the choice of c_1, which is in practice an unknown value that needs to be carefully chosen. By taking a lower value of c_1, the confidence bands (the adjustments between the confidence intervals) will be slightly narrower and thus the empirical coverages in Table 4.2 will be slightly smaller. The Bonferroni-adjusted bands will presumably still be conservative, however, we might need greater sample size to maintain the level of significance in the case of PIE-adjusted confidence bands.

The estimation of c_1 is an important issue that has not been addressed properly, based on this small simulation study it seems reasonable to consider a slightly conservative estimate of c_1 together with the PIE-adjusted confidence bands for reasonably sized datasets. The Bonferroni-adjusted bands should be used in case of a small to moderate datasets, the choice of c_1 should be sensible.



(a) M = 50

(b) M = 100

(c) M = 250

Figure 4.4: Bonferroni-adjusted confidence bands, PIE-adjusted confidence bands, true value and estimated value of β^A based on a sample of size $n = 100$, $M = 50, 100, 250$, drawn by the values of 41 grid points on interval $[0\,;\,0.04]$.

Conclusion

In this thesis we introduced longitudinal varying coefficient models and two closely related spline estimation methods. In the third chapter, we derived asymptotic normality and consistence for the polynomial spline estimator under some mild regularity conditions. We showed how the asymptotic properties can be used to construct asymptotic confidence bands by taking pointwise confidence intervals on some grid, interpolating between the end points and slightly widening the band at the interpolated points to account for the interpolation.

An important question is how to choose the confidence level of the pointwise intervals. An appealing choice is to adjust the confidence levels using the Bonferroni correction. However, especially for high number of grid points, such bands might be very conservative. We described a possible modification based on Holder's inequalities which leads to a lower confidence level than the Bonferroni correction, but relies heavily on the precision of approximation.

In the fourth chapter, we investigated the performance of the confidence bands under various scenarios. We saw that the Bonferroni-adjusted band was indeed very conservative in all situations. We also saw that the adjustment of the confidence level based on Holder's inequalities showed some promising results. However, the usage requires a careful consideration, taking into account the complicatedness of the covariance structure, the sample size, and the prior information about the first or second derivative.

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

C. Brunsdon, A. S. Fotheringham, and M. E. Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298, 1996.

Z. Cai, J. Fan, and R. Li. Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451): 888–902, 2000a.

Z. Cai, J. Fan, and Q. Yao. Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451):941–956, 2000b.

C.-T. Chiang, J. A. Rice, and C. O. Wu. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96(454):605–619, 2001.

C. De Boor. On calculating with b-splines. *Journal of Approximation theory*, 6 (1):50–62, 1972.

C. De Boor. *A practical guide to splines*, volume 27. Springer, New York, 1978.

P. J. Diggle and A. P. Verbyla. Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, 54(2):401–415, 1998.

L. D. Fisher and D. Y. Lin. Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157, 1999.

J. W. Hardin and J. M. Hilbe. *Generalized estimating equations.* CRC Press, 2003.

W. K. Härdle and L. Simar. *Applied multivariate statistical analysis.* Springer, 2019.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1 (3):297–310, 08 1986.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993.

D. R. Hoover. Subset complement addition upper bounds-an improved inclusion-exclusion method. *Journal of statistical planning and inference*, 24(2):195–202, 1990.

D. R. Hoover, J. A. Rice, C. O. Wu, and L.-P. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 1998.

J. Z. Huang and H. Shen. Functional coefficient regression models for non-linear time series: a polynomial spline approach. *Scandinavian journal of statistics*, 31(4):515–534, 2004.

J. Z. Huang, C. O. Wu, and L. Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1): 111–128, 2002.

J. Z. Huang, C. O. Wu, and L. Zhou. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14(3): 763–788, 2004.

D. Hunter. An upper bound for the probability of a union. *Journal of Applied Probability*, 13(3):597–603, 1976.

G. Knafl, J. Sacks, and D. Ylvisaker. Confidence bands for regression functions. *Journal of the American Statistical Association*, 80(391):683–691, 1985.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

D. Ruppert. Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, 11(4):735–757, 2002.

L. Schumaker. *Spline functions: basic theory*. Cambridge university press, 2007.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.