Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 211–225
http://www.ssca.org.in/journal

Testing with Cubic Smoothing Splines

Tapio Nummi[1], Jyrki Möttönen[2] and Jianxin Pan[3]
[1]*Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland*
[2]*Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland*
[3]*Faculty of Science and Technology, BNU-HKBU United International College, Zhuhai, Guangdong 519087, PR China*

Received: 28 May 2024; Revised: 3 August 2024; Accepted: 10 August 2024

Abstract

In this paper, we present some possible ways to perform estimation and testing for cubic smoothing splines. Special emphasis is placed on the analysis of correlated data, when using semi-parametric regression models (Schimek, 2000), and the so-called spline growth model (Nummi and Koskela, 2008; Nummi *et al.*, 2017), an extension of the basic growth curve model (Potthoff and Roy, 1964; Rao, 1965). Furthermore, practical applications in fields such as medicine and animal breeding are introduced, highlighting the versatility and efficacy of cubic smoothing splines in real-world applications.

Key words: Covariance structures; Eigenvalue decomposition; Growth curves; Semi-parametric regression.

AMS Subject Classifications: 62J05, 62J10

1. Introduction

In our paper, we specifically delve into the intricacies of cubic smoothing splines. One of the standout advantages inherent in smoothing splines is their adaptability, granting precise control over the delicate balance between interpolating data points and maintaining the overall smoothness of the curve. This control is facilitated by a smoothing parameter, empowering researchers to fine-tune the model for optimal performance. For statistical inference with smoothing splines and semi-parametric regression we can refer to the books by Eubank and Spiegelman (1990), Green and Silverman (1993), Ruppert *et al.* (2003), Wu and Zhang (2006), Harezlak *et al.* (2018) and Stasinopoulos *et al.* (2017), for example.

The notable flexibility of smoothing splines extends beyond their ability to capture intricate data patterns. They also boast a range of theoretical properties that significantly enhance their utility. In various scenarios, smoothing splines emerge as a compelling alternative to parametric models. This preference arises from the inherent challenge of justifying

Corresponding Author: Tapio Nummi
Email: tapio.nummi@tuni.fi

the choice of a parametric function, which often lacks a clear rationale or relies on a rough approximation of the true underlying function form.

Characterized by their high flexibility, splines offer an advantageous choice by providing a flexible and accurate approximation of the true function form. This is particularly valuable in situations where a clear parametric alternative may prove elusive or is based on a rough approximation. The limitations of parametric models become especially evident when testing different competing models against each other, as they typically also provide a limited set of possible alternative hypotheses. In contrast, cubic smoothing splines offer a very broad family of alternative model choices. When pitted against corresponding parametric models, they not only showcase their adaptability but also present a more comprehensive and versatile set of alternatives for a more robust model comparison. In this context papers by Speckman (1988), Eubank and Hart (1992), Azzalini and Bowman (1993), Cantoni and Hastie (2002), Härdle *et al.* (1998), Lin and Zhang (1999), Verbyla *et al.* (1999), Schimek (2000),Zhang and Lin (2003), Liu and Wang (2004), Nummi *et al.* (2011), and Nummi *et al.* (2013) serve as valuable references. This paper concentrates on the inference of cubic smoothing splines and semi-parametric regression. Our methods exhibits flexibility also in the sense that they apply also under correlated data, further extending its utility for testing growth curves (Koskela *et al.*, 2006; Nummi and Mesue, 2013; Nummi *et al.*, 2017), for example.

In Section 2.1, we present some methods used to estimate cubic smoothing splines and corresponding semi-parametric regression models. Subsequently, in Section 3, we elucidate techniques for accurately approximating the spline fit, and introduce a comprehensive set of hypotheses and tests relevant to semi-parametric regression models. Furthermore, we illustrate these methods with an example of medical testing, demonstrating their practical application potential. In Sections 4 and 5, we focus on estimation and testing in a spline growth model and its multivariate extension. These methods are illustrated with a practical application on animal breeding. In Section 6, some concluding remarks are provided.

2. Cubic smoothing splines and semi-parametric regression

2.1. Cubic smoothing splines

Consider the vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$, observed at measuring points $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ on the interval $[a, b]$, where $a < x_1 < x_2 < \cdots < x_n < b$. A cubic smoothing spline can be expressed as

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{g} = (g(x_1), g(x_2), \ldots, g(x_n))^\top$ represents a vector of the smooth, twice-differentiable curve $g(\cdot)$. The term $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top \sim N_n(\mathbf{0}, \sigma^2\mathbf{R})$ accounts for normally distributed errors, where $\mathbf{R}$ is a covariance matrix characterized by parameters within the vector $\boldsymbol{\theta}$.

The estimation of cubic smoothing splines $\mathbf{g}$ can be achieved through a penalized least squares criterion (PLS). This process commences by defining the roughness matrix $\mathbf{K} = \nabla\boldsymbol{\Delta}^{-1}\nabla^\top$, wherein the non-zero elements of the banded $n \times (n-2)$ matrix ∇ and the $(n-2) \times (n-2)$ matrix $\boldsymbol{\Delta}$ are given by

$$\nabla_{k,k} = \frac{1}{h_k}, \quad \nabla_{k+1,k} = -\left(\frac{1}{h_k} + \frac{1}{h_{k+1}}\right), \quad \nabla_{k+2,k} = \frac{1}{h_{k+1}}$$

and

$$\Delta_{k,k+1} = \Delta_{k+1,k} = \frac{h_{k+1}}{6}, \quad \Delta_{k,k} = \frac{h_k + h_{k+1}}{3},$$

where $k = 1, 2, \ldots, (n-2)$ and $h_j = x_{j+1} - x_j$, with $j = 1, 2, \ldots, (n-1)$. The penalized least squares criterion at points $x_1, x_2, \ldots, x_n$ is then expressed as

$$Q_1 = (\mathbf{y} - \mathbf{g})^\top \mathbf{R}^{-1}(\mathbf{y} - \mathbf{g}) + \alpha \mathbf{g}^\top \mathbf{K}\mathbf{g} \tag{2}$$

The minimum with a fixed positive smoothing parameter α is a cubic smoothing spline (*e.g.* Green and Silverman (1993))

$$\tilde{\mathbf{g}} = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{y} = \mathbf{S}_\alpha \mathbf{y}, \tag{3}$$

where we denote $\mathbf{H} = \mathbf{R}^{-1}$ and $\mathbf{S}_\alpha = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}$ is the so-called *smoother matrix*. It is easily seen that if the covariance matrix $\mathbf{R}$ satisfies the equation

$$\mathbf{R}\mathbf{K} = \mathbf{K}, \text{ or equivalently, } \mathbf{K} = \mathbf{H}\mathbf{K}, \tag{4}$$

the smoother matrix reduces to the form $\mathbf{S}_\alpha = (\mathbf{I} + \alpha\mathbf{K})^{-1}$. The resulting spline estimator in this case becomes as Nummi and Koskela (2008),

$$\hat{\mathbf{g}} = (\mathbf{I} + \alpha\mathbf{K})^{-1}\mathbf{y}. \tag{5}$$

It can be seen that this estimator does not depend on the covariance matrix $\mathbf{R}$. It is demonstrated in Nummi *et al.* (2011) that certain important covariance structures used in the analysis of repeated measures or longitudinal data satisfy condition (4). These structures include the uniform covariance structure $\mathbf{R} = \mathbf{I} + d^2\mathbf{1}\mathbf{1}^\top$ and the linear structure $\mathbf{R} = \mathbf{I} + \mathbf{X}\mathbf{D}\mathbf{X}^\top$, where $d^2 > 0$, $\mathbf{D}$ is positive definite, and $\mathbf{X} = (\mathbf{1}, \mathbf{x})$, for example. It is worth noting that in this scenario, when the smoothing parameter α is fixed, the estimated splines become simple linear functions of the observations $y_1, y_2, \ldots, y_n$, and further this offers also the possibility to use the methodology in the case of correlated data, which will be tackled in particular in Section 4.

2.2. Semi-parametric regression

The spline model in (1) seamlessly extends into a semi-parametric regression model

$$\mathbf{y} = \mathbf{U}\mathbf{b} + \mathbf{g} + \epsilon, \tag{6}$$

where $\mathbf{U}\mathbf{b}$ represents the linear component, with $\mathbf{U}$ being a full-rank $n \times k$ matrix of values of k explanatory variables (excluding the constant term), and $\mathbf{b}$ a k-vector of unknown parameters. Semi-parametric regression models have been considered in Nummi *et al.* (2013), Green and Silverman (1993), Schimek (2000), and Wu and Zhang (2006), for example. The PLS criterion for this case is expressed as

$$Q_2 = [\mathbf{y} - (\mathbf{U}\mathbf{b} + \mathbf{g})]^\top \mathbf{H}[\mathbf{y} - (\mathbf{U}\mathbf{b} + \mathbf{g})] + \alpha \mathbf{g}^\top \mathbf{K}\mathbf{g}. \tag{7}$$

Minimizing with respect to $\mathbf{b}$ and $\mathbf{g}$ leads to the following estimates (Green and Silverman, 1993)

$$\tilde{\mathbf{b}} = [\mathbf{U}^\top \mathbf{H}(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{U}]^{-1}\mathbf{U}^\top \mathbf{H}(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{y} \tag{8}$$

and

$$\tilde{\mathbf{g}} = \mathbf{S}_\alpha(\mathbf{y} - \mathbf{U}\tilde{\mathbf{b}}), \tag{9}$$

where $\mathbf{S}_\alpha = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}$. It can be shown that if the condition (4) holds, the estimates simplify to (Nummi *et al.*, 2013)

$$\hat{\mathbf{b}} = [\mathbf{U}^\top(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{U}]^{-1}\mathbf{U}^\top(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{y}, \tag{10}$$

where $\hat{\mathbf{g}} = \mathbf{S}_\alpha(\mathbf{y} - \mathbf{U}\hat{\mathbf{b}})$, $\mathbf{S}_\alpha = (\mathbf{I} + \alpha\mathbf{K})^{-1}$ and the fitted semi-parametric curve can be obtained as

$$\hat{\mu} = \mathbf{M}\mathbf{y}, \tag{11}$$

where $\mathbf{M} = \mathbf{S}_\alpha + \tilde{\mathbf{U}}[\tilde{\mathbf{U}}^\top\mathbf{U}]^{-1}\tilde{\mathbf{U}}^\top$ and $\tilde{\mathbf{U}} = (\mathbf{I} - \mathbf{S}_\alpha)\mathbf{U}$, respectively. It appears that, once the smoothing parameter α is fixed, the estimation process for both the cubic smoothing spline and the semi-parametric model becomes quite straightforward. In the upcoming chapter, we will delve into the methodologies employed for hypothesis testing.

3. Testing

3.1. Approximate fit

Testing in the context of cubic splines poses challenges, primarily because the smoother matrix inherently lacks the properties of a projector matrix. Consequently, established methods, such as those developed for linear models, do not seamlessly apply to cubic splines. Here we outline a few potential avenues and methodologies for conducting tests related to various hypotheses.

Our approach is centered around approximating the smoother matrix $\mathbf{S}_\alpha$ with a matrix possessing the properties of a projector matrix. This approximation not only yields a highly accurate representation of a cubic smoothing spline fit but also generates a cubic spline itself, as it is rooted in a linear combination of cubic splines (Nummi *et al.*, 2011). It can be demonstrated that $\mathbf{S}_\alpha$ can be decomposed as (see also Hastie (1996))

$$\mathbf{S}_\alpha = \mathbf{T}(\mathbf{I} + \alpha\mathbf{\Lambda})^{-1}\mathbf{T}^\top, \tag{12}$$

where the matrix of eigenvectors $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_n)$ can be directly calculated from the roughness matrix $\mathbf{K}$, and the eigenvalues $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_n)$ of $\mathbf{K}$ are interrelated with $\mathbf{S}_\alpha$ such that the eigenvalues of $\mathbf{S}_\alpha$ are given by $\gamma_j = 1/(1 + \alpha\lambda_i)$, indicating a reverse order of eigenvectors of $\mathbf{K}$ and $\mathbf{S}_\alpha$. Intriguingly, the sequence of eigenvectors of $\mathbf{S}_\alpha$ appears to increase in complexity like a sequence of orthogonal polynomials (see *e.g.*, Ruppert *et al.* (2003)), and the eigenvalues $\gamma_j \in (0, 1)$ show how much dumping is made for each $\mathbf{t}_j$ when the smoother is applied. We can effectively approximate $\mathbf{S}_\alpha$ by

$$\mathbf{M}_c = \mathbf{T}_c\mathbf{T}_c^\top, \tag{13}$$

where $\mathbf{T}_c = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_c)$ denotes the first c eigenvectors of $\mathbf{T}$, which can be chosen using modified generalized cross-validation criteria (Nummi and Mesue, 2013)

$$GCV_1(c) = \frac{\frac{1}{n}\sum_{i=i}^{n}[y_i - \bar{y}_i]^2}{(1 - \frac{c}{n})^2}, \tag{14}$$

where $\bar{y}_i$ is now computed using the formula (5) with $\mathbf{S}_\alpha$ replaced by $\mathbf{M}_c$, for instance. It was demonstrated in Nummi *et al.* (2011) that this yields a pretty good approximation especially if the number of effective degrees of freedom is not unreasonably large. Further decomposition of $\mathbf{T}_c$ $(c > 2)$ takes the form $\mathbf{T}_c = (\mathbf{T}_2, \mathbf{T}_{c-2})$, where $\mathbf{T}_2$ encompasses the first two eigenvectors, and $\mathbf{T}_{c-2}$ comprises the remaining eigenvectors. Note that we can take $\mathbf{T}_2 = (\mathbf{t}_1, \mathbf{t}_2)$, where $\mathbf{t}_1 = \mathbf{1}/\sqrt{n}$ and $\mathbf{t}_2$ is given by $t_{2i} = (x_i - \bar{x})/S_x^2$, where $\bar{x}$ is the mean of $x_i s$ and $S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ (Nummi *et al.*, 2011). It is easy to see that $\mathbf{t}_1$ and $\mathbf{t}_2$ span a straight-line model.

We can now approximate $\hat{\mathbf{g}}$ for model (1) by

$$\tilde{\mu} = \mathbf{M}_c\mathbf{y} = (\mathbf{M}_1 + \mathbf{M}_2)\mathbf{y}, \tag{15}$$

where $\mathbf{M}_1 = \mathbf{T}_2\mathbf{T}_2^\top$ and $\mathbf{M}_2 = \mathbf{T}_{c-2}\mathbf{T}_{c-2}^\top$ and further for the model (11) we have

$$\tilde{\mu} = \tilde{\mathbf{M}}\mathbf{y} = (\mathbf{M}_c + \mathbf{M}_3)\mathbf{y}, \tag{16}$$

where $\mathbf{M}_3 = \bar{\mathbf{U}}[\bar{\mathbf{U}}^\top\mathbf{U}]^{-1}\bar{\mathbf{U}}^\top$ and $\bar{\mathbf{U}} = (\mathbf{I} - \mathbf{M}_c)\mathbf{U}$, respectively.

3.2. Hypotheses and test statistics

Testing is based on sums of squares as defined in this paragraph. It is first noted that if we have the correlation model $\mathbf{R} = \mathbf{I} + \mathbf{X}\mathbf{D}\mathbf{X}^\top$, for example, we have $\tilde{\mathbf{M}}\mathbf{X}\mathbf{D}\mathbf{X}^\top = \mathbf{X}\mathbf{D}\mathbf{X}^\top$ and therefore

$$(\mathbf{I} - \tilde{\mathbf{M}})(\mathbf{I} + \mathbf{X}\mathbf{D}\mathbf{X}^\top)(\mathbf{I} - \tilde{\mathbf{M}}) = (\mathbf{I} - \tilde{\mathbf{M}}). \tag{17}$$

We can further note that, under normality and the assumed correlation model, the following relationships hold (Nummi *et al.*, 2013)

$$\sigma^{-2}\mathbf{y}^\top(\mathbf{I} - \tilde{\mathbf{M}})\mathbf{y} = \sigma^{-2}S_{min} \sim \chi^2_{n-c-k}. \tag{18}$$

Similarly, we can define

$$\sigma^{-2}\mathbf{y}^\top(\mathbf{I} - \mathbf{M}_c)\mathbf{y} = \sigma^{-2}S_{spl} \sim \chi^2_{n-c}, \tag{19}$$

$$\sigma^{-2}\mathbf{y}^\top(\mathbf{I} - \mathbf{M}_1)\mathbf{y} = \sigma^{-2}S_{lin} \sim \chi^2_{n-2} \tag{20}$$

and

$$\sigma^{-2}\mathbf{y}^\top(\mathbf{I} - \mathbf{P}_i)\mathbf{y} = \sigma^{-2}S_{reg,i} \sim \chi^2_{n-k-i}, i = 1, 2, \tag{21}$$

where $\mathbf{P}_i = \mathbf{U}_i(\mathbf{U}_i^\top\mathbf{U}_i)^{-1}\mathbf{U}_i$, where for $i = 1, \mathbf{U}_1 = (\mathbf{1}, \mathbf{U})$ and $i = 2, \mathbf{U}_2 = (\mathbf{X}, \mathbf{U})$, and where $\mathbf{X} = (\mathbf{1}, \mathbf{X})$, respectively. These sum-of-squares expressions can now be utilized for the hypothesis testing of different special cases of the basic semi-parametric model. We can now formulate a set of compelling hypotheses each designed to assess various aspects of the models introduced. Note that the tests introduced in this section are applicable also to correlated data, provided an appropriate form of covariance matrix is employed.

3.2.1. Test 1: Cubic smoothing spline

The first test introduced here aims to scrutinize whether the basic linear model is applicable when compared to the assumed cubic smoothing spline alternative (model (1)). The hypotheses are formulated as follows

$$\mathbf{H}_0\colon\ \mu = \mathbf{X}\mathbf{b}_2,$$

where $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$ and $\mathbf{b}_2$ is a vector of two regression coefficients. The alternative hypothesis is

$$\mathbf{H}_a\colon\ \mu = \mathbf{g},$$

where $\mathbf{g}$ represents the assumed spline model. Since $\mathbf{M}_c\mathbf{M}_1 = \mathbf{M}_1$ (columns $\mathbf{M}_1$ are in the span of $\mathbf{M}_c$) it is observed that $(\mathbf{I} - \mathbf{M}_c)(\mathbf{M}_c - \mathbf{M}_1) = \mathbf{0}$ and therefore S_{spl} and $S_{lin} - S_{spl}$ are independent and

$$F_1 = \frac{(S_{lin} - S_{spl})/(c-2)}{S_{spl}/(n-c)} \sim F(c-2, n-c). \tag{22}$$

Then observing a larger F_1 than quantile $F_{1-\alpha}(c-2, n-c)$ yields the rejection of the null hypothesis. It was shown in a power study of Nummi *et al.* (2011) that this test performed very well when compared to other alternatives.

3.2.2. Tests 2: Semi-parametric model

A) Testing the significance of linear covariates in the full model

Suppose the full semi-parametric model may include a set of linear covariates, denoted as $\mathbf{U}$. We first test the significance of this set in the full model. The null hypothesis is

$$\mathbf{H}_0\colon\ \mu = \mathbf{g},$$

and the alternative hypothesis, a full semi-parametric model, is

$$\mathbf{H}_a\colon\ \mu = \mathbf{U}\mathbf{b} + \mathbf{g},$$

where $\mathbf{U}\mathbf{b}$ is a linear term and $\mathbf{g}$ is a smoothing spline term. Using similar arguments as before, we get

$$F_{2A} = \frac{(S_{spl} - S_{min})/k}{S_{min}/(n-k-c)} \sim F(k, n-k-c). \tag{23}$$

If the observed F_{2A} is larger than the critical value $F_{1-\alpha}(k, n-k-c)$, we reject the null hypothesis.

B) Assessing the fit of the model with linear model

This test evaluates whether the assumed linear model provides a better fit compared to a semi-parametric alternative. The hypotheses are defined as follows

$$\mathbf{H}_0\colon\ \mu = \mathbf{U}_{k+2}\mathbf{b}_{k+2},$$

where $\mathbf{U}_{k+2} = [\mathbf{X}, \mathbf{U}]$, $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$, and $\mathbf{b}_{k+2}$ is a vector of $k+2$ regression coefficients. The alternative hypothesis remains the same as in part A. The test statistic for this hypothesis becomes

$$F_{2B} = \frac{(S_{reg,2} - S_{min})/(c-2)}{S_{min}/(n-k-c)} \sim F(c-2, n-k-c). \tag{24}$$

If the observed F_{2B} exceeds the critical value $F_{1-\alpha}(c-2, n-k-c)$, we reject the null hypothesis. According to Nummi *et al.* (2013), the power of this test was investigated through a simulation study. The study found that estimating c from the observed data results in only a minimal loss of power compared to the scenario where c is known.

3.2.3. Test 3: Linear model

Ultimately, we can explore the need to include the variable $\mathbf{x}$, which was initially presumed to be a smooth term ($c > 2$), as a linear term alongside other linear terms within a full linear model. The hypotheses are formulated as

$$\mathbf{H}_0\colon\ \mu = \mathbf{U}_{k+1}\mathbf{b}_{k+1},$$

where $\mathbf{U}_{k+1} = [\mathbf{1}, \mathbf{U}]$, and $\mathbf{b}_{k+1}$ is a $k+1$ vector of regression coefficients. The alternative hypothesis is

$$\mathbf{H}_a\colon\ \mu = \mathbf{U}_{k+2}\mathbf{b}_{k+2},$$

where $\mathbf{U}_{k+2} = [\mathbf{X}, \mathbf{U}]$ and this can be tested as

$$F_3 = \frac{(S_{reg,1} - S_{reg,2})}{S_{reg,2}/(n-k-2)} \sim F(1, n-k-2). \tag{25}$$

Then observing a larger F_3 than quantile $F_{1-\alpha}(1, n-k-2)$ yields the rejection of the null hypothesis.

Example 1: PSA testing

As an illustration, we utilized part of the dataset gathered for the Finnprostate Study VII conducted by Professor Teuvo L. J. Tammela in Finland in 1990-2000 at Tampere University. The primary objective of this study was to examine individuals susceptible to prostate cancer. It is important to note that in this article, we will refrain from delving into the medical intricacies of the subject matter. Instead, our focus is solely on employing this dataset to exemplify the methodologies presented.

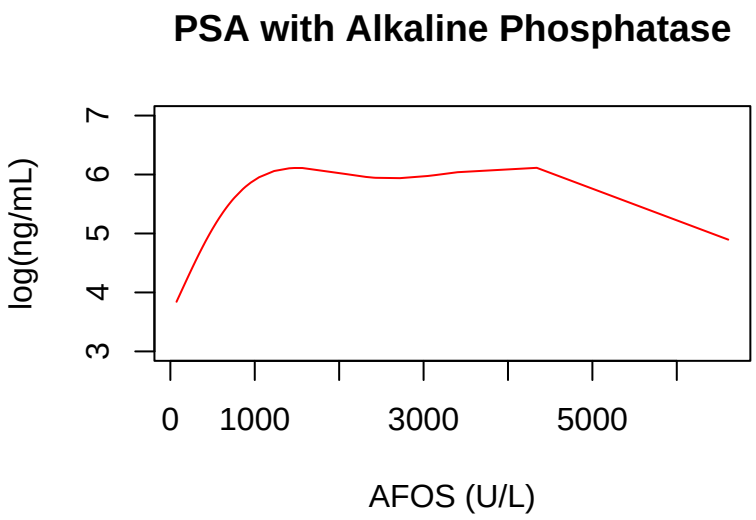

Figure 1: Plot of approximated spline fit for the values of prostate-specific antigen (log(ng/mL) as a function of values of alkaline phosphatase test (U/L).

In this instance, our examination of 537 individuals is centered on the variables Prostate-Specific Antigen (PSA, ng/mL), Body Mass Index (BMI, kg/m^2), Prostate Length (Length, cm), and Alkaline Phosphatase test (AFOS, U/L). The primary aim of our study is to construct a model for the log(PSA) value utilizing the variables AFOS, BMI, and Length. To commence, we explore the relationship between PSA and AFOS, assuming that a suitable spline model would best describe this connection. Employing the criteria $GCV^*(c)$, where c ranges from 1 to 6, we obtain the values 6.3152, 0.9808, 0.9042, 0.8791, 0.8755, and 0.8769. Consequently, our preferred choice is $c = 5$. It should be noted that for some measuring points $x_1, \ldots, x_n$, we have multiple values and therefore we need to replace the smoother matrix $\mathbf{S}_\alpha$ by

$$\mathbf{S}_\alpha = \mathbf{N}(\mathbf{N}^\top\mathbf{N} + \alpha\mathbf{K})^{-1}\mathbf{N}^\top, \tag{26}$$

where $\mathbf{N}$ is an incidence matrix of corresponding measuring times. The approximated spline fit is depicted in Figure 1. Upon subjecting this to a linear model test (Test 1), we obtain $F_1 = 22.45$, with the corresponding quantile $F_{0.95}(3, 532) = 2.622$. This unequivocally rejects the null hypothesis concerning a linear association.

Subsequently, we delve into semi-parametric model 6. Our preliminary analysis suggests that BMI, Length, and the interaction term Alkaline×OI can be utilized as explanatory variables in the $\mathbf{U}$-matrix, where OI is the obesity indicator ($OI = 1$ if $BMI > 30$, and 0 otherwise). Alkaline with $c = 5$ is used in (16) for model fitting and testing. Using the test statistic F_{2A}, we evaluated the significance of this set of covariates within the full semi-parametric model. The resulting value,$F_{2A} = 17.06$, exceeds the corresponding critical value $F_{0.95}(3, 529) = 2.62$, indicating clear significance. Additionally, the value of the test statistic F_{2B} is 28.63, which also surpasses the critical value $F_{0.95}(3, 529) = 2.62$. Consequently, the null hypothesis is firmly rejected, confirming that the model is semi-parametric rather than fully linear. Test 3 is not executed in this scenario, as it only becomes relevant if the null hypothesis from Test 2B is accepted.

4. Testing for growth data

In certain cases, growth modeling can be grounded in a theoretical framework, enabling the derivation of a parametric model for developmental processes. However, more frequently, such a theoretical foundation may be lacking, necessitating the adoption of alternative approximations. We found that cubic smoothing splines for many cases provide a well justified alternative since they quite accurately follow the true growth function. In the following, we outline the methodology for testing some relevant hypotheses when employing cubic smoothing splines to model the growth function.

The growth curve model of Potthoff and Roy (Potthoff and Roy, 1964) can be written as

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{A}^\top + \mathbf{E}, \tag{27}$$

where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ is the $q \times n$ matrix of independent $q \times 1$ response vectors, $\mathbf{T}$ is a $q \times p$ within-individual design matrix, $\mathbf{A}$ is an $n \times m$ between-individual design matrix, $\mathbf{B}$ is an unknown $p \times m$ parameter matrix to be estimated and $\mathbf{E}$ is a $q \times n$ matrix of random errors. It is assumed that the columns $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of $\mathbf{E}$ are independently normally distributed as $\mathbf{e}_i \sim N_q(\mathbf{0}, \mathbf{\Sigma}),\ i = 1, \ldots, n$.

We can write model (27) in a more general way by using cubic smoothing splines. Let

$$\mathbf{Y} = \mathbf{G}\mathbf{A}^\top + \mathbf{E}, \tag{28}$$

where $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m)$ is the matrix of smooth mean growth curves at time points $t_1, t_2, \ldots, t_q$. We further assume that $\mathbf{\Sigma}$ is a parsimonious covariance structure $\mathbf{\Sigma} = \sigma^2\mathbf{R}(\boldsymbol{\theta})$ with covariance parameters $\boldsymbol{\theta}$. Model (28) is referred to as the spline growth model (SGM). Note that we get the Potthoff and Roy model as a special case by setting $\mathbf{G} = \mathbf{T}\mathbf{B}$. The smooth solution for the matrix of mean growth curves $\mathbf{G}$ can be obtained by minimizing the following penalized least squares (PLS) objective function

$$Q = \mathrm{tr}[(\mathbf{Y} - \mathbf{G}\mathbf{A}^\top)^\top\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{G}\mathbf{A}^\top) + \alpha\mathbf{A}\mathbf{G}^\top\mathbf{K}\mathbf{G}\mathbf{A}^\top], \tag{29}$$

where α is a fixed smoothing parameter and $\mathbf{K}$ is the roughness matrix defined in Section 2.1. It can be easily seen that Q can be rewritten n the form

$$Q = \mathrm{tr}[(\mathbf{G}\mathbf{A}^\top - (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{Y})^\top(\mathbf{H} + \alpha\mathbf{K})(\mathbf{G}\mathbf{A}^\top - (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{Y})] + w, \tag{30}$$

where $\mathbf{H} = \mathbf{R}^{-1}$, $(\mathbf{H}+\alpha\mathbf{K})$ is a positive definite matrix and $w = \mathrm{tr}[\mathbf{Y}^\top\mathbf{H}^{-1}(\mathbf{H}+\alpha\mathbf{K})^{-1}\mathbf{H}^{-1}\mathbf{Y} - \mathbf{Y}^\top\mathbf{H}\mathbf{Y}]$ does not depend on $\mathbf{G}$. The function Q is minimized for fixed values of α and $\mathbf{H}$ when $\mathbf{G}\mathbf{A}^\top = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{Y}$. Multiplying both sides of the equation on the right by $\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}$ gives the spline estimator

$$\tilde{\mathbf{G}} = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}. \tag{31}$$

The estimator $\tilde{\mathbf{G}}$ has one drawback when thinking about practical applications. The matrix $\mathbf{H}$ is unknown, so it should be estimated from the data. However, in some special cases the estimator is simplified to a form that does not depend on the covariance matrix. Suppose that the matrix $\mathbf{H}$ fulfills the condition $\mathbf{K} = \mathbf{H}\mathbf{K}$ (or equivalently $\mathbf{R}$ fulfills the condition $\mathbf{K} = \mathbf{R}\mathbf{K}$). Then the spline estimator (31) simplifies to

$$\hat{\mathbf{G}} \;=\; (\mathbf{I}_q + \alpha\mathbf{K})^{-1}\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1} = \mathbf{S}\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \tag{32}$$

where $\mathbf{S} = (\mathbf{I}_q + \alpha\mathbf{K})^{-1}$ is the smoother matrix. The smoothing parameter α can be chosen by using the generalized cross-validation criteria

$$GCV_2(\alpha) = \frac{\frac{1}{nq}\mathrm{tr}[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^\top]}{(1-\frac{m\cdot edf}{nq})^2}, \tag{33}$$

where $\hat{\mathbf{Y}} = \hat{\mathbf{G}}\mathbf{A}^\top$ and $edf = \mathrm{tr}(\mathbf{S})$ is the effective degrees of freedom of the smoother matrix $\mathbf{S}$.

As in Section 3, for testing we need to approximate the smoother matrix with a matrix that has the properties of a projection matrix. We can approximate the spline estimate (32) with

$$\breve{\mathbf{G}} \;=\; \mathbf{M}_c\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \tag{34}$$

where $\mathbf{M}_c = \mathbf{T}_c\mathbf{T}_c^\top$ and $\mathbf{T}_c$ contains the c first eigenvectors of the smoother matrix $\mathbf{S}$. The number of eigenvectors c can be easily estimated using a modified generalized cross-validation criterion obtained by replacing $\hat{\mathbf{Y}}$ and edf in formula (14) with $\breve{\mathbf{Y}} = \breve{\mathbf{G}}\mathbf{A}^\top$ and c, respectively. We can now approximate the fitted spline curves with

$$\breve{\mathbf{Y}} \;=\; \breve{\mathbf{G}}\mathbf{A}^\top = \mathbf{T}_c\mathbf{T}_c^\top\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top = \mathbf{T}_c\hat{\mathbf{\Omega}}\mathbf{A}^\top, \tag{35}$$

where $\hat{\mathbf{\Omega}} = \mathbf{T}_c^\top\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}$. The matrix $\hat{\mathbf{\Omega}}$ contains all the relevant information for testing mean curves and it is also an unbiased estimate of the parameter matrix $\mathbf{\Omega}$ of the statistical model $\mathbf{Y} = \mathbf{T}_c\mathbf{\Omega}\mathbf{A}^\top + \mathbf{E}$. Therefore, we will henceforth focus on testing linear hypotheses of the form

$$H_0 : \mathbf{C}\mathbf{\Omega}\mathbf{D} = \mathbf{0},$$

where $\mathbf{C}$ and $\mathbf{D}$ are known $\nu \times c$ and $m \times g$ matrices with ranks ν and g respectively. It is shown in Nummi and Mesue (2013) that testing can be based on

$$F = \frac{Q_*/\nu g}{\hat{\sigma}^2} \sim F[\nu g, n(q-c)], \tag{36}$$

where

$$Q_* = \mathrm{tr}([\mathbf{D}^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{D}]^{-1}[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]^\top[\mathbf{C}\mathbf{T}_c^\top\mathbf{R}\mathbf{T}_c\mathbf{C}^\top]^{-1}[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]) \tag{37}$$

and

$$\hat{\sigma}^2 = \frac{1}{n(q-c)}\mathrm{tr}[\mathbf{Y}^\top(\mathbf{I}_q - \mathbf{M}_c)\mathbf{Y}]. \tag{38}$$

In real-life applications, the matrix $\mathbf{R}$ contains parameters to be estimated and therefore the distribution of the F-statistic is only approximate. However, if we are only interested in progression in time we can drop the constant term by using $\mathbf{C} = [\mathbf{0}, \mathbf{I}_{c-1}]$, and if the uniform covariance model $\mathbf{R} = d^2\mathbf{1}_q\mathbf{1}_q^\top + \mathbf{I}_q$ is assumed, the test statistic Q_* simplifies to

$$Q_{**} = \mathrm{tr}([\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}][\mathbf{D}^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{D}]^{-1}[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]^\top). \tag{39}$$

It can be shown that the distribution of the test statistics Q_{**} is exact. This is an important result since the uniform covariance model is quite common and a good approximation in many situations. In Nummi and Mesue (2013) other kinds of situations are discussed, that give an exact version of the F-test introduced here.

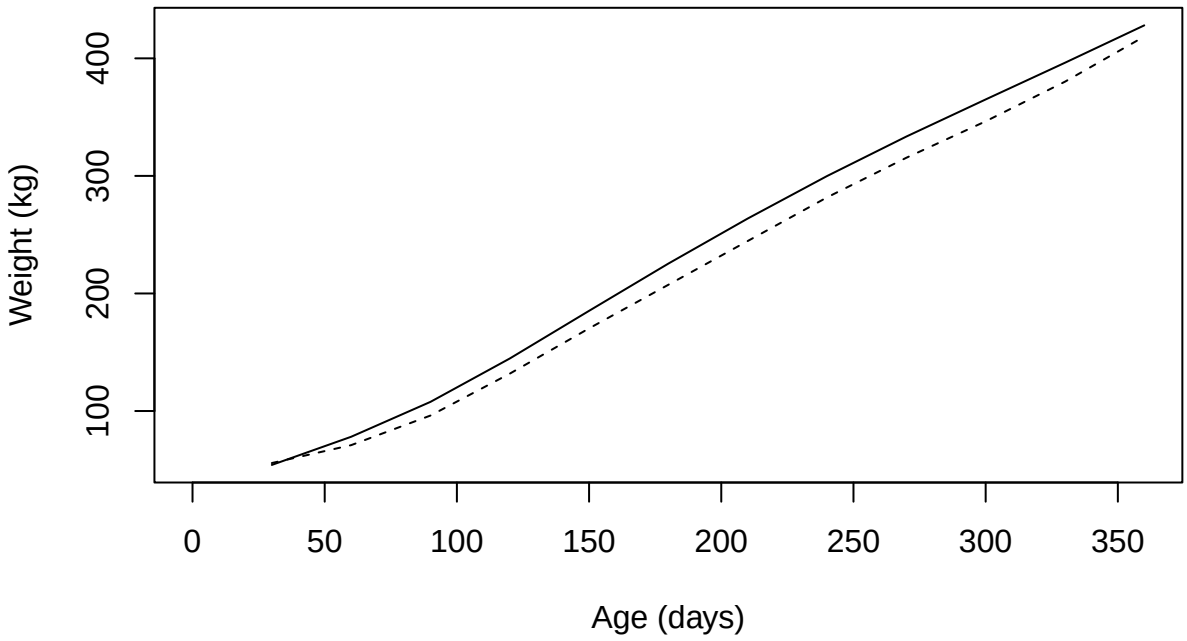

Figure 2: Plot of approximated spline fits for Finncattle bulls (solid curve) and Ayrshire bulls (dashed curve).

Example 2: Testing bulls at a research station in Finland

In this example, we present our methodology using a subset of data pertaining to 2712 bulls tested at an experimental station in Finland during the years 1965 to 1977. The original dataset comprised three breeds: Ayrshire, Finncattle, and Frisian. However, for the purposes of this illustration, we focused on a specific subset consisting of 208 bulls born in 1966, with 168 Ayrshire and 40 Finncattle bulls. The bulls underwent regular weighing, conducted every 30 days starting from the age of 30 days. For more comprehensive details, see the references Lindström and Maijala (1970) and Liski (1987).

To set up the spline growth model the between-individual design matrix $\mathbf{A}$ was defined as follows. For the Finncattle bulls, the rows of $\mathbf{A}$ are $(1, 0)$ and for the Ayrshire bulls the rows of $\mathbf{A}$ are $(0, 1)$. Using the generalized cross-validation criteria (33), we got the smoothing parameter $\alpha = 4142$. The number of eigenvectors c was then estimated using the modified generalized cross-validation criteria (33). The function $GCV_2(c)$ was minimized at $c = 7$. Figure 2 gives the approximated spline fits for the Finncattle bulls (solid curve) and the Ayrshire bulls (dashed curve).

To test if the progression is the same in both groups, we used the 6×7 matrix $\mathbf{C} = (\mathbf{0}, \mathbf{I}_6)$ and 2×1 vector $\mathbf{D} = (1, -1)^\top$. The value of the F-test statistic is

$$F = 102.1803,$$

which gives the P-value $\mathbb{P}(F_{6,1040} \geq 102.1803) \approx 0$. Therefore, the null hypothesis of equal progression of the response variable in the two test groups (Finncattle and Ayrshire) is clearly rejected. We also calculated the P-value of the permutation test. We randomly permuted the rows of matrix $\mathbf{A}$ and re-calculated the value of the F-statistic using the permuted matrix $\mathbf{A}$. After permuting $\mathbf{A}$ and re-calculating F-statistic $N = 100,000$ times, we got the estimated permutation test P-value

$$\frac{\#\{F_i \geq 102.1803\}}{N} = 0.00086.$$

Therefore, it can be affirmed that testing of the growth curves against each other can be readily implemented also using computational methods.

5. Testing in the multivariate spline growth model

The testing of the spline growth model can be generalized straightforwardly to a multivariate response case. The multivariate spline growth curve model can be written as

$$\mathbf{Y} = \mathbf{G}\mathbf{A}^\top + \mathbf{E}, \tag{40}$$

where

$$\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) = \begin{pmatrix} \mathbf{y}_{11} & \mathbf{y}_{21} & \cdots & \mathbf{y}_{n1} \\ \mathbf{y}_{12} & \mathbf{y}_{22} & \cdots & \mathbf{y}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_{1s} & \mathbf{y}_{2s} & \cdots & \mathbf{y}_{ns} \end{pmatrix}$$

is a $qs \times n$ matrix of the vectors of measurements of s responses and

$$\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m) = \begin{pmatrix} \mathbf{g}_{11} & \mathbf{g}_{21} & \cdots & \mathbf{g}_{m1} \\ \mathbf{g}_{12} & \mathbf{g}_{22} & \cdots & \mathbf{g}_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{1s} & \mathbf{g}_{2s} & \cdots & \mathbf{g}_{ms} \end{pmatrix}$$

is the corresponding $qs \times m$ matrix of smooth mean curves. See Nummi *et al.* (2017) for more details. For the covariance matrix $\mathbf{R}$ we can take, for example, a multivariate version of the uniform structure

$$\begin{aligned} \mathbf{R} &= (\mathbf{I}_s \otimes \mathbf{1}_q)\mathbf{D}(\mathbf{I}_s \otimes \mathbf{1}_q)^\top + \mathbf{I}_{qs} \\ &= \begin{pmatrix} d_1^2\mathbf{1}_q\mathbf{1}_q^\top + \mathbf{I}_q & d_{12}\mathbf{1}_q\mathbf{1}_q^\top & \cdots & d_{1s}\mathbf{1}_q\mathbf{1}_q^\top \\ d_{21}\mathbf{1}_q\mathbf{1}_q^\top & d_2^2\mathbf{1}_q\mathbf{1}_q^\top + \mathbf{I}_q & \cdots & d_{2s}\mathbf{1}_q\mathbf{1}_q^\top \\ \vdots & \vdots & \ddots & \vdots \\ d_{s1}\mathbf{1}_q\mathbf{1}_q^\top & d_{s2}\mathbf{1}_q\mathbf{1}_q^\top & \cdots & d_s^2\mathbf{1}_q\mathbf{1}_q^\top + \mathbf{I}_q \end{pmatrix}. \end{aligned} \tag{41}$$

If we now define the roughness part of the fitting criteria as

$$\mathbf{K}_s = \mathbf{W} \otimes \mathbf{K},$$

where $\mathbf{W} = \mathrm{diag}(\alpha_1, \ldots, \alpha_s)$ is a diagonal matrix of smoothing parameters $\alpha_1, \ldots, \alpha_s$ and $\mathbf{K}$ is the roughness matrix computed using the time points $t_1, \ldots, t_q$, then the roughness matrix $\mathbf{K}_s$ meets the condition

$$\mathbf{R}\mathbf{K}_s = \mathbf{K}_s \tag{42}$$

and the unweighted spline estimator becomes

$$\begin{aligned}\hat{\mathbf{G}} &= (\mathbf{I}_{qs} + \mathbf{W} \otimes \mathbf{K})^{-1}\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1} \\ &= \begin{pmatrix} \mathbf{S}(\alpha_1) & \mathbf{O} & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \mathbf{S}(\alpha_2) & \mathbf{O} & \dots & \mathbf{O} \\ \vdots & & \ddots & & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \dots & \mathbf{S}(\alpha_s) \end{pmatrix} \mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \end{aligned} \tag{43}$$

where $\mathbf{S}(\alpha_j) = (\mathbf{I}_q + \alpha_j\mathbf{K})^{-1}$, for $j = 1, \dots, s$. If we use the approximation technique introduced earlier we get

$$\hat{\mathbf{G}} = \begin{pmatrix} \mathbf{M}_{\bullet 1}\mathbf{M}_{\bullet 1}^\top & \mathbf{O} & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \mathbf{M}_{\bullet 2}\mathbf{M}_{\bullet 2}^\top & \mathbf{O} & \dots & \mathbf{O} \\ \vdots & \vdots & \ddots & & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \dots & \mathbf{M}_{\bullet s}\mathbf{M}_{\bullet s}^\top \end{pmatrix} \mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \tag{44}$$

where $\mathbf{M}_{\bullet j}\mathbf{M}_{\bullet j}^\top = \mathbf{P}_j$ is an approximation matrix for the jth variable. Note that the dimensions needed can be estimated using the generalized cross-validation criteria introduced in 33. A straightforward generalization of the earlier considerations gives us an estimator

$$\hat{\mathbf{\Omega}} = \mathbf{M}_\bullet^\top\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \tag{45}$$

where $\mathbf{M}_\bullet = \text{diag}(\mathbf{M}_{\bullet 1}, \mathbf{M}_{\bullet 2}, \dots, \mathbf{M}_{\bullet s})$, of the multivariate growth curve model

$$\mathbf{Y} = \mathbf{M}_\bullet\Omega\mathbf{A}^\top. \tag{46}$$

Testing can be based on the linear hypothesis

$$H_0 : \mathbf{C}\mathbf{\Omega}\mathbf{D} = \mathbf{0},$$

where $\mathbf{C}$ and $\mathbf{D}$ are known $\nu \times c$ and $m \times g$ matrices with ranks ν and g, respectively, with

$$F = \frac{Q_*/\nu g}{\hat{\sigma}^2} \sim F[\nu g, n(sq - c_{tot})], \tag{47}$$

where $c_{tot} = c_1 + \cdots + c_s$ and

$$Q_* = \text{tr}\{[\mathbf{D}^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{D}]^{-1}[\mathbf{C}\hat{\Omega}\mathbf{D}]^\top[\mathbf{C}\mathbf{M}_\bullet^\top\mathbf{R}\mathbf{M}_\bullet\mathbf{C}^\top]^{-1}[\mathbf{C}\hat{\Omega}\mathbf{D}]\} \tag{48}$$

and

$$\hat{\sigma}^2 = \sum_{l=1}^{s} \frac{1}{n(q - c_l)}\text{tr}[\mathbf{Y}_l^\top(\mathbf{I}_q - \mathbf{P}_l)\mathbf{Y}_l]. \tag{49}$$

If we are interested in testing the equality of the progression of spline curves, then we can choose

$$\mathbf{C} = \text{diag}([\mathbf{0}, \mathbf{I}_{c_1 - 1}], \dots, [\mathbf{0}, \mathbf{I}_{c_s - 1}]) \quad \text{and} \quad \mathbf{D} = [\mathbf{1}_{m-1}, -\mathbf{I}_{m-1}]^\top$$

and, furthermore, if we assume that the covariance matrix has a uniform structure (41), the test statistic simplifies to the form

$$Q_{**} = \text{tr}\{[\mathbf{C}\hat{\Omega}\mathbf{D}][\mathbf{D}^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{D}]^{-1}[\mathbf{C}\hat{\Omega}\mathbf{D}]^\top\}, \tag{50}$$

which does not depend on the covariance matrix $\mathbf{R}$. The F statistic is then distributed as $F[df_1, df_2]$ with degrees of freedoms $df_1 = (c_{tot} - s)(m - 1)$ and $df_2 = n(sq - c_{tot})$.

6. Concluding remarks

In this paper, we explored various methodologies for estimating and testing cubic smoothing splines. We place particular emphasis on analyzing correlated data within semiparametric regression models, as well as the spline growth model, an extension of the basic growth curve model. Additionally, we introduced practical applications including medicine and animal breeding. These examples underscore the versatility and effectiveness of cubic smoothing splines in real-world scenarios.

Acknowledgements

We are deeply honored to have had the opportunity to contribute our publication to the special issue dedicated to C. R. Rao, one of the most esteemed statisticians in history. This is an immense privilege for us. We would also like to that the anonymous referee for the comments that led to improvements of the paper.

References

Azzalini, A. and Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**, 549–557.

Cantoni, E. and Hastie, T. (2002). Degrees-of-freedom tests for smoothing splines. *Biometrika*, **89**, 251–263.

Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *The Annals of Statistics*, **1**, 1412–1425.

Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association*, **85**, 387–392.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC, London, 1st ed. edition.

Härdle, W., Mammen, E., and Müller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, **93**, 1461–1474.

Harezlak, J., Ruppert, D., and Wand, M. P. (2018). *Semiparametric Regression With R*, volume 109. Springer.

Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 379–396.

Koskela, L., Nummi, T., Wenzel, S., and Kivinen, V. P. (2006). On the analysis of cubic smoothing spline-based stem curve prediction for forest harvesters. *Canadian Journal of Forest Research*, **36**, 2909–2919.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **61**, 381–400.

Lindström, U. and Maijala, K. (1970). Evaluation of performance test results for a.i. bulls. *Acta Agriculturae Scandinavica*, **20**, 207–218.

Liski, E. P. (1987). A growth curve analysis for bulls tested at station. *Biometrical Journal*, **29**, 331–343.

Liu, A. and Wang, Y. (2004). Hypothesis testing in smoothing spline models. *Journal of Statistical Computation and Simulation*, **74**, 581–597.

Nummi, T. and Koskela, L. (2008). Analysis of growth curve data by using cubic smoothing splines. *Journal of Applied Statistics*, **35**, 681–691.

Nummi, T. and Mesue, N. (2013). Testing of growth curves with cubic smoothing splines. In Dasgupta, R., editor, *Advances in Growth Curve Models*, pages 49–59, New York, NY. Springer.

Nummi, T., Möttönen, J., and Tuomisto, M. T. (2017). Testing of multivariate spline growth model. In Chen, D. G., Jin, Z., Li, G., Li, Y., Liu, A., and Zhao, Y., editors, *New Advances in Statistics and Data Science*, pages 75–85. Springer International Publishing, Cham.

Nummi, T., Pan, J., and Mesue, N. (2013). Testing linearity in semiparametric regression models. *Statistics and Its Interface*, **6**, 3–8.

Nummi, T., Pan, J., Siren, T., and Liu, K. (2011). Testing for cubic smoothing splines under dependent data. *Biometrics*, **67**, 871–875.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447–458.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York.

Schimek, M. G. (2000). Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference*, **91**, 525–540.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **50**, 413–436.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **48**, 269–311.

Wu, L. and Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*. Wiley, Hoboken, New Jersey.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, **4**, 57–74.