REPRODUCING KERNEL SPLINES

IN THE

ANALYSIS OF MEDICAL DATA

by

Jan-Christian Lehtinen

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Mathematics and Natural Sciences of the University of Turku, for public criticism in the Auditorium of the Computer Science Department on June 10th, 2000, at 12 noon.

University of Turku
Department of Computer Science
Turku, Finland
2000

Supervised by

Professor Timo Järvi, Ph.D.
Department of Computer Science
University of Turku

and

Assistant Professor Timo Knuutila, Ph.D.
Department of Computer Science
University of Turku

Reviewed by

Professor Patrik Eklund, Ph.D.
Department of Computing Science
Umeå University

and

Researcher Petri Koistinen, Dr.Tech.
Rolf Nevanlinna Institute
University of Helsinki

ISBN 951-29-1722-X
ISSN 1239-1883
Painosalama Oy
Turku, Finland

Abstract

The aim of data mining is to find rules or functions that refine data. Many data mining techniques have difficulties in analyzing data with few observations and lots of missing values. The techniques presented in this thesis are especially suitable for the analysis of this type of data. The analysis in this thesis is of medical data mining problems.

We use *reproducing kernel* (RK) *splines* to fit a function of several variables to scattered, noisy data. We show how RK functions can be used to generate a RK Hilbert space for specific data. Matrix decompositions can be utilized in the calculation of RK spline fits while adaptively choosing multiple smoothing parameters with generalized cross validation (GCV). To alleviate the computational time requirements of GCV, we propose a *parallel version* of GCV.

Variable selection is an important aspect in nonparametric approximation. In medical data, the variables usually correlate with each other and therefore there is no optimal variable combination. We give a heuristic term selection algorithm, which minimizes the number of variables needed in approximation. The term selection is tested using both artificial and medical data sets. The experiments suggest that heuristic model selection should be favored for standard breadth-first selection.

To evaluate the use of RK splines in the analysis of clinical data, we use our techniques on several medical data sets. Serum LDL cholesterol is often estimated using a linear model. Our *model selection algorithm* discovered that the product of total cholesterol and triglycerides is significant and it should be included in the model. In the second experiment, we study nutrient intake in a low-cholesterol diet. In the third experiment, we study Down syndrome screening and compare the variables influencing the reference values of hCG and AFP hormones between populations. In the last experiment, we study whether pneumococcal vaccine given to the mother during the pregnancy would protect infants from infections. The experiments performed with RK splines support the hypothesis that vaccination provides some protection from Pnc infections.

RK splines can be used to analyze complex data with few observations and lots of missing values. In our case studies with RK splines, we found several phenomena that were not discovered with conventional techniques.

Acknowledgements

This study was carried out at the Department of Computer Science, University of Turku and at the Turku Centre for Computer Science during 1995-2000. I express my sincere thanks to all those who have generously contributed to the present work in many different ways.

I am indebted to my supervisor, Professor Timo Järvi, Head of the Department of Computer Science, who has provided me with excellent resources and continuous support enabling me to complete this study. I also wish to express my gratitude to Assistant Professor Timo Knuutila for his support in revising the manuscript of this thesis.

My warm thanks are due to Docent Jari Forsström. He originally guided me to the field of medical informatics. My sincere thanks go to Professor Olli Nevalainen and Professor Kari Ylinen for their valuable comments and suggestions for revising the manuscript.

This thesis was reviewed by Professor Patrik Eklund and Doctor Petri Koistinen. I am deeply grateful to the reviewers for their valuable comments on this thesis.

I am grateful to my colleagues Mehran Gomari, Harri Hakonen, Jaakko Järvi, and Antti Koski for positive criticism and numerous discussions during the past five years. It has been a privilege to work with them all this time.

I wish to acknowledge the medical partners of the research groups for a very fruitful cooperation. Especially, I would like to mention Professor Olli Simell, Harri Niinikoski, Hanna Nohynek, and Minna Räsänen.

I also would like to thank Gordon Alford for the spelling checking and correcting the grammar of this thesis. I acknowledge my gratitude to Professor Ralph-Johan Back, Head of the Turku Centre for Computer Science, for financial support. The Center for Scientific Computing has provided excellent computing resources during the years of the study.

Finally, I wish to thank my parents for their support and encouragement. Most of all, however, I owe the warmest thanks to Henna for her patience and understanding.

This study was financially supported by the Turku Centre for Computer Science and the National Technology Agency.

Contents

Chapter 1

Introduction

This thesis focuses on data mining. The techniques presented here are applicable in the analysis of very wide range of quantitative data. Our emphasis is on medical problems and therefore we often refer to *medical data*. The methods used are especially suitable for the analysis of data with few observations and many missing values.

In medical literature, artificial neural networks are sometimes considered as an alternative to statistical data analysis. However, many authors do not justify their use. Statistics has a long history that has formed certain practices for quantitative data analysis. Researchers of the applied fields are familiar with the concepts such as degrees of freedom, distribution of data, p-value, confidence interval, etc. In neural computation, these practices and concepts are still largely missing and neural networks are often used without the necessary knowledge of the theory of approximation. This sometimes leads the researcher to misinterpret the results.

One of the major differences between the statistical data analysis and neural networks is in measuring reliability. In neural computation, the reliability of the result is often expressed as a mean estimation error for some data set. However, mean errors are sensitive to the arrangements of the experiments. In medical statistics, authors often express the reliability of the clinical research results with the probability or confidence interval of the estimates. Therefore, it would be useful if the result of the neural computation could also be expressed using the same reliability measures. This would lessen the misuse of neural networks.

Bayesian techniques have gained popularity among neural network researchers within the past few years [Mac95b, Bis95]. The Bayesian approach combines traditional neural computation and an estimation of the statistical expectation. Bayesian techniques bring several new features to conventional neural computation. This approach allows different models to be compared using only the training data. For regression problems, it offers the possibility of determining confidence intervals. The relative importance of different inputs can be determined using automatic relevance determination [Mac94a, Mac95a]. For more properties of Bayesian techniques, see [Mac95b].

Reproducing kernel (RK) splines [Wah90, Gu96, Gu89] are less known than the Bayesian approach but they include similar features. RK splines have certain benefits over the Bayesian approach, one of them being the training process itself. RK splines do not use any network topology. This alleviates the training because there is no need to determine the parameters of the network topology. In addition, distributed matrix operations provide efficient parallelism for training of RK splines.

There have been a few experiments utilizing RK splines in the analysis of medical data. Diabetic retinopathy has been studied in [WWG+95a, WWG+95b, Wan94, Wan95b, WWG+93]. Wang has studied breast cancer in [Wan94] and Wang et al. circadian rhythms of the hormones ACTH and cortisol in [WGB97]. Gu has studied the time from HIV infection to AIDS diagnosis and survival time after heart transplant in [Gu97].

1.1 Aims of the Study

The aim of this study is to investigate the applicability of RK splines in the analysis of medical data. The parts of the study are:

- To describe RK splines and justify their use in the analysis of quantitative medical data.
- To develop RK splines especially from the view point of medical data analysis. The new features introduced in this study include among other things parallel computation of RK splines and automatic model space selection.
- To test and demonstrate the use of RK splines with real medical data.

Reproducing kernels have been known for a long time although their characteristic properties have only been applied since the beginning of the 20th century. Previously, reproducing kernels were called with names like "positive definite kernels", "positive Hermitian matrices" and "positive definite functions". The term "reproducing kernel" became standard in 1950's and in approximation theory reproducing kernels were introduced in 1970's. Unlike neural networks, which often aim at simulating the cognitive processes of brain, RK splines originate from the mathematical theory of approximation and generalization. Until now, very few authors have considered RK splines from the point of view of *machine learning*. The overall aim of this study is to discuss RK splines as a machine learning technique and to demonstrate the use of RK splines in data mining.

1.2 Overview of the Chapters

Chapter 2 deals with the problem of learning and generalization when using finite sets of data. We discuss the concept of RK and some properties of the reproducing kernel Hilbert space (RKHS). Further, we derive a training algorithm for RK splines.

In neural computation and many other machine learning techniques, the original training data is often divided into two subsets. One subset is used to train and the other one to test and adjust the training parameters. Chapter 3 deals with the generalized cross validation (GCV) technique which avoids the division of the training set. The computation of the GCV function requires $O(n^3)$ operations, where n is the number of training patterns. In Chapter 3, we introduce a parallel algorithm that calculates the GCV function efficiently.

Chapter 4 concentrates on automatic model space and variable selection. The terms of the RK spline model include univariate functions and functions of two or more variables. Model space selection aims at finding a set of these functions such that their sum approximates the relationship between the dependent and independent variables. Variable selection is a natural consequence of the model space selection. A variable is excluded from the model space if none of the terms of the model space contain the variable.

Chapter 5 demonstrates the use of RK splines in the analysis of medical data. The chapter is divided into four sections, each dealing with a different type of a medical problem. Section 5.1 focuses on low density lipoprotein (LDL) cholesterol estimation. High LDL cholesterol is a major risk factor for coronary heart disease. In order to calculate LDL cholesterol concentration, serum LDL needs to be separated from blood sample by ultracentrifugation, which is not cost-efficient in frequent use. Therefore, serum LDL concentration is usually estimated from other lipid values. Our model selection algorithm detected that by including the product of total cholesterol and triglycerides in the model the accuracy of the estimate increases. The discovery is statistically very significant ($p < 0.000001$).

High LDL cholesterol is usually due to unhealthy diet. In Section 5.2, we study nutrient intake in a low-cholesterol diet. The project aims at decreasing exposure of young children to known atherosclerosis risk factors. RK splines were used to study the effect of nutrient intake to serum cholesterol. One of the input variables was a ratio of two measurements (polyunsaturated fatty acids and saturated fatty acids) and RK splines were able to detect the nonlinearity of the ratio.

Down syndrome (DS) is the most common serious chromosome aberration in which affected individuals survive beyond infancy. Maternal age is known to be associated with the DS risk and therefore pregnant women over 35 years are routinely offered an opportunity of amniocentesis and caryotyping of fetal cells. However, two thirds of the DS cases are born to mothers less than 35 years. Maternal serum markers, such as alpha-fetoprotein (AFP) and human chronic gonadotropin (hCG) are usually used to estimate the DS risk of women under 35 years. Serum hCG and AFP decrease gradually with maternal weight because the same amount of hormones diffuses to larger volume. Both AFP and hCG are population dependent and therefore the reference AFP and hCG values need to be adjusted for each population separately. In Section 5.3, we study the differences of the reference AFP and hCG between populations. RK splines discovered a difference in the effect of maternal weight. In Taiwan the decrease rate of hormone concentration per kg is greater than in the UK. The finding is true for both hCG and AFP. Therefore, different weight correction should be used for British and Taiwanese women. RK splines also revealed a difference in the effect of gestational age between the UK and Taiwan.

Currently used pneumococcal (Pnc) vaccines give $50 - 80\%$ protection against serious Pnc infections. However, children under two years old are unable to produce antibodies against those Pnc bacteria that are common cause of childhood infections and therefore these vaccines do not protect young children. In Section 5.4, we study the influence of the conventional Pnc polysaccharide vaccine given during pregnancy and the potential of maternally transferred antibodies for the protection of infants in their first $3 - 4$ months of life against Pnc diseases. In 15 of the 18 experiments performed with RK splines, results support the hypothesis that vaccination provides some protection from Pnc infections.

The data set descriptions of the Friedewald data (Section 5.1), STRIP data (Section 5.2), Down syndrome data (Section 5.3) and Pneumococcal data (Section 5.4) are given in Appendix A.

Chapter 2

Theoretical Background of Reproducing Kernel (RK) Splines

The aim of this chapter is to provide an introduction to the theory of RK splines and demonstrate the use of RK splines with examples. The following chapters rely on the results derived in this chapter and therefore we encourage a reader new in the field of RK splines to read the chapter carefully. A reader familiar with RK splines may skip the chapter or read it only briefly.

Kimeldorf and Wahba introduced RK splines and showed the connection between RK splines and Bayes estimates in 1970 [KW70, KW71]. The formulas presented in the original paper were unsuitable to the computing capabilities of the day. Multivariate splines did not attract researchers in early 1970's and offshoots of these papers were often rejected by journals (see foreword of [Wah90]). In 1978 Wahba introduced the use of matrix decompositions in calculation of RK splines [Wah78]. This decreased the computational barrier of RK splines and enabled small experiments. The recent development of computers has made it feasible to use RK splines as an alternative to traditional approximation in the analysis of quantitative data.

RK splines have a number of features that make them especially favorable for medical data analysis. The duality between RK splines and Bayes estimates ensures sound theoretical background for approximation. One can also calculate confidence intervals to the RK spline estimates [Wah83].

There are often randomly or systematically located missing values in medical data. RK splines use a function decomposition in approximation, which alleviates the problems associated with missing values. For example, a two-variable function s can be presented in the form

$$s(x, y) = c + f(x) + g(y) + h(x, y),$$

where c is a constant, f and g are one-variable functions and h is a two-variable function. In the decomposition, functions f, g and h are considered as correction terms for the initial estimate

$$\tilde{s}_1(x, y) = c.$$

Following the same pattern, h is a correction to the estimate

$$\tilde{s}_2(x, y) = c + f(x) + g(y). \tag{2.1}$$

If either x or y is missing, we can approximate s by

$$\begin{aligned} \tilde{s}_3(x, y) &= c + g(y) \quad \text{or} \\ \tilde{s}_4(x, y) &= c + f(x). \end{aligned}$$

If both x and y are missing, we use the estimate $\tilde{s}_1(x, y) = c$. In the decomposition, there are no missing values and therefore the approach overcomes most of the problems associated with randomly located missing values. For a survey on additive models, see [BHT89, HT90, HT86].

The *curse of dimensionality* [Bis95] is one of the problems in nonparametric approximation. The input space becomes sparse, if we introduce new independent variables but keep the number of observations unchanged. The above function decomposition solves the curse of dimensionality elegantly. We can approximate $s(x, y)$ using, for example, decomposition (2.1). If the original $s(x, y)$ and our estimate $\tilde{s}_2(x, y)$ are close to each other, we have successfully replaced the original two-variable approximation problem with a set of one-variable approximation problems. Our experiments suggest that biological phenomena can often be explained with a function decomposition that has only a few two-variable terms in addition to the constant and one-variable terms.

2.1 Learning and Generalization

Mitchell proposes a perspective that machine learning involves searching a very large space of possible hypotheses to determine one that best fits the observed data and the prior knowledge held by the learner [Mit97]. Learning from examples can be considered in many cases to be equivalent to multivariate function approximation [GJP93]. The goal of learning is not to learn an exact representation of the training data itself, but to build a statistical model of the process which generated the data. The statistical model is usually constructed by selecting a set of parameters that minimize some error function. The choice of the error function depends on the approximation method and data that is analyzed. The mean square error (MSE) and its variants are the most used error functions in nonparametric approximation. In this section, we derive the MSE function and justify its use in nonparametric approximation. For theory of generalization in function approximation, see [PG89, NG94]

2.1.1 Mean Square Error (MSE)

Our aim is to predict target values from input values. Let matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a sequence of input vectors and $\mathbf{y} = (y_1, \ldots, y_n)^T$ the sequence of corresponding target values, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, for $i = 1, \ldots, n$. We suppose that pairs $(\mathbf{x}_i, y_i)$ are drawn independently from the same distribution. Let $p(\mathbf{x}_i, y_i)$ be the joint probability density of $\mathbf{x}_i$ and y_i, $p(\mathbf{x}_i)$ the unconditional density of $\mathbf{x}_i$, and $p(y_i|\mathbf{x}_i)$ the probability density of y_i given that $\mathbf{x}_i$ takes particular value. Using the assumption of independence, we may write the likelihood $\mathcal{L}$ of set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ as

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^{n} p(\mathbf{x}_i, y_i) \\ &= \prod_{i=1}^{n} p(y_i|\mathbf{x}_i) p(\mathbf{x}_i). \end{aligned}$$

We try to maximize $\mathcal{L}$. Maximizing the likelihood is equivalent to minimizing the negative logarithm of the likelihood

$$-\ln \mathcal{L} = -\sum_{i=1}^{n} \ln p(y_i|\mathbf{x}_i) - \sum_{i=1}^{n} \ln p(\mathbf{x}_i). \tag{2.2}$$

The unconditional density $p(\mathbf{x}_i)$ is independent of the target values. In this study we concentrate on techniques which do not utilize the unconditional densities $p(\mathbf{x}_i)$, $i = 1, \ldots, n$. Therefore in order to maximize $\mathcal{L}$ we need to minimize only the first part of (2.2), i.e.,

$$E_1 = -\sum_{i=1}^{n} \ln p(y_i|\mathbf{x}_i). \tag{2.3}$$

In function approximation it is assumed that the target values are given by some *continuous function* f, where

$$y_i = f(\mathbf{x}_i) + \epsilon_i. \tag{2.4}$$

Terms ϵ_i are called *noise* and they are expected to have a normal distribution with zero mean and standard deviation σ which does not depend on $\mathbf{X}$ or on i, i.e., $(\epsilon_1, \ldots, \epsilon_n)^T \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n\times n})$. The notation $\mathbf{I}$ stands for the identity matrix. The use of the normal distribution is usually motivated by referring to the *central limit theorem* which states that the probability distribution governing the sample mean approaches the normal distribution as $n \to \infty$ [Mit97, pp. 142]. Some other choice of noise distribution would lead to a different error function. The distribution of ϵ_i is given by the density function

$$p(\epsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{\epsilon_i^2}{2\sigma^2}). \tag{2.5}$$

Using (2.4) and (2.5), we see that the probability distribution of the target data is given by the density function

$$p(y_i|\mathbf{x}_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(y_i - f(\mathbf{x}_i))^2}{2\sigma^2}).$$

Now we can present (2.3) in the form

$$\begin{aligned} E_1 &= -\sum_{i=1}^{n} \ln\left[\frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(y_i - f(\mathbf{x}_i))^2}{2\sigma^2})\right] \\ &= \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 + n\ln\sigma\sqrt{2\pi}. \end{aligned}$$

The last term is independent of $f(\mathbf{x}_i)$ and therefore we do not need it in the error function minimization. Also the inverse of the variance in front of the sum can be discarded as constant. We obtain the familiar error function

$$\begin{aligned} E_2 &= \frac{1}{2}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 \\ &= \frac{1}{2}\|\mathbf{y} - \mathbf{f}(\mathbf{X})\|^2, \end{aligned}$$

where $\mathbf{f}(\mathbf{X}) = (f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n))^T$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. The error function is usually scaled with the number of patterns and presented in the form

$$E_3 = \frac{1}{2n}\|\mathbf{y} - \mathbf{f}(\mathbf{X})\|^2. \tag{2.6}$$

The error functions E_2 and E_3 have the minimum at the same point. The error function E_3 is called the *mean square error (MSE)*.

2.1.2 MSE with Infinite Training Data

The minimum of (2.6) considered as a function of f is obtained when the model function f interpolates the training data, i.e., $\mathbf{y} = \mathbf{f}(\mathbf{X})$. Our aim, however, is not to interpolate the training data but to construct a model function that predicts the target values for unseen input vectors. Consider the MSE with an infinite number of training patterns, i.e.,

$$\begin{aligned} E_4 &= \lim_{n\to\infty}\frac{1}{2n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 \\ &= \frac{1}{2}\int\int(y - f(\mathbf{x}))^2 p(\mathbf{x}, y)\,dy\,d\mathbf{x} \\ &= \frac{1}{2}\int\int(y - f(\mathbf{x}))^2 p(y|\mathbf{x})p(\mathbf{x})\,dy\,d\mathbf{x}. \end{aligned}$$

Theorem 2.1 *The MSE function for infinite training data can be presented in the form*

$$\begin{aligned} E_4 &= \frac{1}{2}\int [E(y|\mathbf{x}) - f(\mathbf{x})]^2 p(\mathbf{x})d\mathbf{x} \\ &+\frac{1}{2}\int [E(y^2|\mathbf{x}) - E(y|\mathbf{x})^2]p(\mathbf{x})dy, \end{aligned} \tag{2.7}$$

where $E(y|\mathbf{x})$ is the conditional expectation given by

$$E(y|\mathbf{x}) = \int yp(y|\mathbf{x})dy.$$

Proof. See [Bis95, p. 202]. □

The minimum of the error function (2.7) is obtained when $f(\mathbf{x}) = E(y|\mathbf{x})$, since the second term is independent of $f(\mathbf{x})$. The second term represents the noise in data and it determines the lower limit of the error.

2.1.3 MSE with Finite Training Data

The error function (2.7) is not very practical in its current form, because we usually have only a limited number of training patterns. We next discuss the case of finite training data.

Let $\mathcal{D}$ be a training set with n patterns. If $\mathcal{D}$ is finite, the function $f(\mathbf{x})$ depends on the choice of $\mathcal{D}$. We denote expectation over all sets $\mathcal{D}$ by $E_{\mathcal{D}}(\cdot)$. If we replace the term $[E(y|\mathbf{x}) - f(\mathbf{x})]^2$ in (2.7) with its data dependent expectation $E_{\mathcal{D}}[(E(y|\mathbf{x}) - f(\mathbf{x}))^2]$ and discard noise that is independent of our model function f, we can express the error function in the form

$$E_5 = \frac{1}{2}\int E_{\mathcal{D}}[(E(y|\mathbf{x}) - f(\mathbf{x}))^2]p(\mathbf{x})d\mathbf{x}.$$

Theorem 2.2 *The MSE function for finite training data can be presented in the form*

$$\begin{aligned} E_5 &= \frac{1}{2}\int [E(y|\mathbf{x}) - E_{\mathcal{D}}(f(\mathbf{x}))]^2 p(\mathbf{x})d\mathbf{x} \\ &+\frac{1}{2}\int E_{\mathcal{D}}[[E_{\mathcal{D}}(f(\mathbf{x})) - f(\mathbf{x})]^2]p(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{2.8}$$

Proof. See [Bis95, pp. 333-335]. □

The first term of (2.8) is called *bias error* and the second one *variance error*. The distinction between the bias error and the variance error is important. The bias error measures the extent to which $f(\mathbf{x})$ differs from its desired value, in this case $E(y|\mathbf{x})$. The more training data we have the smaller the bias error is. The variance error measures the extent to which $f(\mathbf{x})$ is sensitive to the choice of a particular $\mathcal{D}$.

Consider the extreme case where $y_i = f(\mathbf{x}_i)$ for $i = 1, \dots, n$. If f interpolates the training data the bias error disappears at the training points. In the other extreme, if f does not depend on selection of $\mathcal{D}$, then $E_{\mathcal{D}}(f(\mathbf{x})) = f(\mathbf{x})$ and the variance error disappears at the training points. Both extremes are undesirable when $\mathcal{D}$ is finite. The objective of training is to find a suitable balance between the bias and variance error.

2.1.4 Generalization

In order to avoid overfitting of the training data some restrictions to f are necessary. We include to (2.6) a penalty functional $J_\lambda(f)$ to control the tradeoff between the bias and variance error. The error function (2.6) takes the form

$$E = \frac{1}{2n}\|\mathbf{y} - \mathbf{f}(\mathbf{X})\|^2 + J_\lambda(f),$$

where vector λ contains parameters controlling the penalty functional.

The aim of the penalty term is to decrease the oscillation of f. A function is said to be *smoother* than another one if it oscillates less. When we study functions in the frequency domain we may say that a function is smoother than another one if it has less energy at high frequency. The general form of the penalty term in MSE minimization is

$$J_\lambda(f) = \int \frac{1}{G(\mathbf{s})}|F(\mathbf{s})|^2 d\mathbf{s}, \tag{2.9}$$

where $F(\mathbf{s})$ is the Fourier transform of f($\mathbf{x}$), see [GJP93] for details. We use symbols $\mathbf{x}$ and $\mathbf{s}$ to distinguish the original domain and the frequency domain. The term $1/G(\mathbf{s})$ is a high-pass filter. Function $G(\mathbf{s})$ can be any well defined positive, symmetric function that falls off to zero as $\|\mathbf{s}\| \to \infty$. By selecting an appropriate high-pass filter one can calculate with inverse Fourier transform the corresponding penalty functional for univariate and multivariate splines.

We demonstrate the use of (2.9) with two examples. In the first example we use the so-called *cubic splines*. Cubic splines are often used to interpolate one-dimensional data [Ver99]. In the cubic spline approximation the spline function does not interpolate the training points but passes them nearby.

Example 2.1 (Cubic Spline Penalty Term) *Let f be a one-dimensional differentiable continuous function. If we choose the high-pass filter $G(s) = 1/s^4$, we get the penalty functional*

$$\begin{aligned} J_\lambda(f) &= \lambda_1 \int_{\mathbb{R}} s^4 |F(s)|^2 ds \\ &= \lambda_1 \int_{\mathbb{R}} (f''(x))^2 dx, \end{aligned} \tag{2.10}$$

which is the penalty of one-dimensional cubic splines [Ver99].

Coefficient λ_1 determines the smoothness of the approximation. If $\lambda_1 \to 0$, f interpolates the training data. If $\lambda_1 \to \infty$, f is the linear regression line fitted to data. The tradeoff between the bias and variance error is controlled by selecting an appropriate λ_1. Using a similar derivation one can easily show that the high-pass filter of one-dimensional line interpolation is correspondingly $1/s^2$ [Ver99]. For the theory of multivariate cubic splines, see [Gu95].

In the second example we use *thin plate splines*. Thin plate spline approximation [Duc77] is one of the most used methods in two-dimensional interpolation and approximation on an irregularly spaced grid. The method produces a natural looking smooth surface and therefore thin plate splines are used especially in the approximation of two and three-dimensional objects.

Example 2.2 (Thin Plate Spline Penalty Term) *Let f be a two-dimensional differentiable continuous function. If we choose the high-pass filter $G(\mathbf{s}) = 1/\|\mathbf{s}\|^4$, we get the penalty functional*

$$\begin{aligned} J_\lambda(f) &= \lambda_1 \int_{\mathbb{R}^2} \|\mathbf{s}\|^4 |F(\mathbf{s})|^2 d\mathbf{s} \\ &= \lambda_1 \int_{\mathbb{R}^2} \left[\left(\frac{\partial^2 f}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2}\right)^2 \right] dx_1 dx_2, \end{aligned}$$

which is the penalty of two-dimensional thin plate splines [Ver99].

2.2 RK Functions

This far we have assumed that the model function f is continuous. Now we derive model functions that are applicable also for non-continuous index sets. For this purpose, we utilize reproducing kernel Hilbert spaces. In this section, we first discuss the RK functions on domains $[0,1]$, $[0,1]^d$ and $\{1, 2, \ldots, K\}$ and then we derive the RK spline training algorithm. The proofs of the theorems presented in this section are quite complex and therefore most of them are given only as references.

Consider a general index set $\mathcal{T}$, e.g.,

$$\begin{aligned} \mathcal{T} &= \mathbb{R}^d \\ \mathcal{T} &= [0,1] \\ \mathcal{T} &= \{(x,y,z) \mid x^2 + y^2 + z^2 = 1\}, \quad \text{(unit sphere)} \\ \mathcal{T} &= \{1, 2 \ldots, K\}. \end{aligned}$$

Reproducing kernel Hilbert space provides a framework for learning and generalization with any well defined index set. For example, a two-dimensional index $\mathcal{T} = [0,1] \times \{0,1,2\}$, which has both a continuous and a discrete variable, should not, in general, be modeled in $\mathbb{R}^2$.

Definition 2.1 (Reproducing Kernel Hilbert space, RKHS) *Let $s, t \in T$, where T is some set, and $K(s,t)$ be a function $T \times T \to \mathbb{R}$. The Hilbert space W of real-valued functions on T is called a* reproducing kernel Hilbert space *if it has the properties:*

i. $K(\cdot, t) \in W$ for every $t \in T$
ii. $\langle f(\cdot), K(\cdot, t)\rangle = f(t)$ for every $t \in T$ and $f \in W$

The function K is called a reproducing kernel.

Theorem 2.3 *Let K_1 be the reproducing kernel of Hilbert space W_1 and K_2 the reproducing kernel of W_2. The direct sum $W = W_1 \oplus W_2$ has the reproducing kernel $K = K_1 + K_2$ and the tensor product $W = W_1 \otimes W_2$ has the reproducing kernel $K = K_1 K_2$.*

Proof. See [Aro50]. □

Appendix B.1 contains the basics of Hilbert spaces.

2.2.1 RK Functions on $[0,1]$

The most commonly used one-dimensional index set in the RK spline approximation is $[0,1]$. Usually the continuous variables of the training set can be scaled to the interval $[0,1]$. If the variables cannot be scaled, we can use the index set $\mathbb{R}$, for example. In order to use the index set $[0,1]$, we need RKHS and RK for it.

Let $W_k^1[0,1]$ be a Hilbert space $\{f | f^{(i)}$ absolutely continuous on $[0,1], i = 0, \ldots, k-1, f^{(k)} \in \mathcal{L}_2[0,1]\}$, where the superscript in $W_k^1[0,1]$ denotes that the Hilbert space is a space of one-dimensional functions, and $\mathcal{L}_2[0,1]$ denotes the space of square integrable functions on the interval $[0,1]$. A function f is said to be absolutely continuous if it has a derivative whose indefinite integral is f.

Theorem 2.4 (RKHS on $[0,1]$) *The space $W_k^1[0,1]$ is an RKHS with the norm*

$$\|f\|^2_{W_k^1} = \sum_{i=0}^{k-1} \left(\int_0^1 f^{(i)}(x)dx\right)^2 + \int_0^1 (f^{(k)}(x))^2 dx. \tag{2.11}$$

Proof. See [CW79] or [Wah90, p. 129]. □

Theorem 2.5 *Let $h_i(x) = B_i(x)/i!$, where $B_i(x)$ is the i^{th} Bernoulli polynomial (Appendix B.2). The space $W_k^1[0,1]$ with the norm* (2.11) *can be represented as a direct sum of k orthogonal subspaces in the form*

$$\begin{aligned} W_k^1[0,1] &= \{h_0\} \oplus \{h_1\} \oplus \cdots \oplus \{h_{k-1}\} \oplus \mathcal{H}_1 \\ &= \mathcal{H}_0 \oplus \mathcal{H}_1 \end{aligned}$$

where $\{h_i\}$ is the subspace of one-dimensional functions spanned by h_i and $\mathcal{H}_1$ is the subspace orthogonal to $\mathcal{H}_0$ satisfying

$$\int_0^1 f^{(i)}(x)dx = 0, \quad i = 0, \ldots, k-1.$$

Proof. See [CW79]. □

The notation $\mathcal{H}_0$ denotes the *nonsmooth* RKHS, that is the space containing the terms that will not be penalized in the error function minimization and $\mathcal{H}_1$ denotes the *smooth* RKHS, that is the space containing the terms that will be penalized.

Theorem 2.6 (RK of RKHS on $[0,1]$) *The RK $R_0^i(s,t)$ of the RKHS $\{h_i\}$ is*

$$R_0^i(s,t) = h_i(s)h_i(t)$$

and the RK $R_1(s,t)$ of the RKHS $\mathcal{H}_1$ is

$$R_1(s,t) = h_k(s)h_k(t) + (-1)^{k-1}h_{2k}(|s-t|).$$

Proof. See [CW79]. □

Using theorems 2.5 and 2.6, the RK of $W_k^1[0,1]$ with the norm (2.11) can now be written as

$$\begin{aligned} R(s,t) &= \sum_{i=0}^{k-1} R_0^i(s,t) + R_1(s,t) \\ &= \sum_{i=0}^{k} h_i(s)h_i(t) + (-1)^{k-1}h_{2k}(|s-t|). \end{aligned}$$

Example 2.3 *Let $k = 1$. The RK of $W_1^1[0,1]$ is*

$$\begin{aligned} R(s,t) &= h_0(s)h_0(t) + h_1(s)h_1(t) + h_2(|s-t|) \\ &= 1 + (s-1/2)(h-1/2) + \frac{1}{2}(|s-t|^2 - |s-t| + 1/6) \\ &= \frac{1}{2}[s(s-1) + t(t-1) - |s-t| + \frac{8}{3}]. \end{aligned}$$

2.2.2 RK Functions on $[0,1]^d$

Before introducing the RKHS of general d-dimensional functions, we demonstrate the use of the RKHS on $[0,1]^2$. Consider two continuous variables $a, b \in [0,1]$. Using a similar norm as in Theorem 2.4, we can present the two variable RKHS $W_k^2[0,1]$ as a tensor product in the form

$$\begin{aligned} W_k^2[0,1] &= [\{h_0^a\} \oplus \{h_1^a\} \oplus \cdots \oplus \{h_{k-1}^a\} \oplus \mathcal{H}_1^a] \otimes \\ &\quad\ [\{h_0^b\} \oplus \{h_1^b\} \oplus \cdots \oplus \{h_{k-1}^b\} \oplus \mathcal{H}_1^b] \\ &= \mathcal{H}_0^{a,b} \oplus \mathcal{H}_1^{a,b}, \end{aligned}$$

where the space $\mathcal{H}_0^{a,b}$ is the direct sum of all subspaces that do not contain $\mathcal{H}_1^a$ or $\mathcal{H}_1^b$ and the space $\mathcal{H}_1^{a,b}$ is the direct sum of all other subspaces of W_k^2.

Example 2.4 *Let $k = 2$. The RKHS $W_2^2[0,1]$ can be presented in the form $\mathcal{H}_0^{a,b} \oplus \mathcal{H}_1^{a,b}$, where*

$$\begin{aligned} \mathcal{H}_0^{a,b} &= [\{1\}\otimes\{1\}] \oplus [\{1\}\otimes\{h_1^b\}] \oplus [\{h_1^a\}\otimes\{1\}] \oplus [\{h_1^a\}\otimes\{h_1^b\}] \\ &= \{1\} \oplus \{h_1^a\} \oplus \{h_1^b\} \oplus [\{h_1^a\}\otimes\{h_1^b\}] \\ &= \{1\} \oplus \{a-1/2\} \oplus \{b-1/2\} \oplus [\{a-1/2\}\otimes\{b-1/2\}] \end{aligned}$$

and

$$\begin{aligned} \mathcal{H}_1^{a,b} &= [\{1\}\otimes\mathcal{H}_1^b] \oplus [\{h_1^a\}\otimes\mathcal{H}_1^b] \oplus [\{1\}\otimes\mathcal{H}_1^a] \oplus [\{h_1^b\}\otimes\mathcal{H}_1^a] \oplus [\mathcal{H}_1^a\otimes\mathcal{H}_1^b] \\ &= \mathcal{H}_1^a \oplus \mathcal{H}_1^b \oplus [\{h_1^a\}\otimes\mathcal{H}_1^b] \oplus [\{h_1^b\}\otimes\mathcal{H}_1^a] \oplus [\mathcal{H}_1^a\otimes\mathcal{H}_1^b] \\ &= \mathcal{H}_1^a \oplus \mathcal{H}_1^b \oplus [\{a-1/2\}\otimes\mathcal{H}_1^b] \oplus [\{b-1/2\}\otimes\mathcal{H}_1^a] \oplus [\mathcal{H}_1^a\otimes\mathcal{H}_1^b]. \end{aligned}$$

The space $\mathcal{H}_0^{a,b}$ is thus a direct sum of four nonsmooth RKHSs and $\mathcal{H}_1^{a,b}$ is a direct sum of five smooth RKHSs. The smooth terms control the tradeoff between the bias and variance error.

Now consider the general case, $\mathbf{x} \in [0,1]^d$. The corresponding RKHS is

$$W_k^d[0,1] = \bigotimes_{i=1}^{d} [\{h_0^{x_i}\} \oplus \{h_1^{x_i}\} \oplus \cdots \oplus \{h_{k-1}^{x_i}\} \oplus \mathcal{H}_1^{x_i}] \tag{2.12}$$

The space $W_k^d[0,1]$ is a tensor product of d RKHSs. The total number of direct sums in (2.12) is $(k+1)^d$.

Definition 2.2 (Elementary Space) *We call the subspaces of the direct sum* (2.12) *elementary spaces. In* (2.12) *each elementary space is a tensor product of d subspaces. In the general case, the subspaces of the tensor product can be RKHSs of different index sets. For example, the RKHS of the index set $[0,1] \times \{0,1,2\}$ is a tensor product of the RKHSs of $[0,1]$ and $\{0,1,2\}$.*

The RK of $W_k^d[0,1]$ is the sum of the RKs of all elementary spaces of $W_k^d[0,1]$.

Example 2.5 *Consider one elementary space of* $W_2^4[0,1]$*, e.g.,*

$$\begin{aligned} W &= \{h_1^{x_1}\} \otimes \mathcal{H}_1^{x_2} \otimes \mathcal{H}_1^{x_3} \otimes \{h_1^{x_4}\} \\ &= \{x_1 - 1/2\} \otimes \mathcal{H}_1^{x_2} \otimes \mathcal{H}_1^{x_3} \otimes \{x_4 - 1/2\}. \end{aligned}$$

The space has both nonsmooth and smooth subspaces. Let $\mathbf{s}, \mathbf{t} \in [0,1]^4$*. The RK of* $W \subset \mathcal{H}_1$ *is*

$$\begin{aligned} R_1(\mathbf{s},\mathbf{t}) &= R_0^1(s_1,t_1)R_1(s_2,t_2)R_1(s_3,t_3)R_0^1(s_4,t_4) \\ &= h_1(s_1)h_1(t_1) \cdot [h_2(s_2)h_2(s_2) - h_4(|s_2 - t_2|)] \cdot \\ &\quad [h_2(s_3)h_2(s_3) - h_4(|s_3 - t_3|)] \cdot h_1(s_4)h_1(t_4). \end{aligned}$$

2.2.3 RK Functions on set $\{1,\dots,K\}$

In addition to continuous variables, data sets often contain also discrete variables. In this section, we discuss the RKHS of a set of size K. Let $\mathcal{T} = \{1,2,\dots,K\}$ and $W_{\{K\}}$ be a Hilbert space $\{f | f \in \mathbb{R}^K\}$. In order to use discrete variables in approximation we need the corresponding RKHS and RK. The next three theorems are analogous to theorems 2.4, 2.5 and 2.6.

Theorem 2.7 (RKHS on $\{1,2,\dots,K\}$) *The space* $W_{\{K\}}$ *is an RKHS with norm*

$$\|f\|^2_{W_{\{K\}}} = \mathbf{f}^T\mathbf{f} = \mathbf{f}^T(\mathbf{1}\mathbf{1}^T/K)\mathbf{f} + \mathbf{f}^T(\mathbf{I} - \mathbf{1}\mathbf{1}^T/K)\mathbf{f}, \tag{2.13}$$

where

$$\mathbf{f} = (f(1), f(2), \dots, f(K))^T.$$

The notation **1** *stands for the vector of* K *ones.*

Proof. See [Gu96]. □

Theorem 2.8 *The space* $W_{\{K\}}$ *can be presented in the form*

$$W_{\{K\}} = \mathcal{H}_0 \oplus \mathcal{H}_1,$$

where $\mathcal{H}_0$ *is the column space of* $(\mathbf{1}\mathbf{1}^T/K)$ *and* $\mathcal{H}_1$ *is the column space of* $(\mathbf{I} - \mathbf{1}\mathbf{1}^T/K)$*. The space* $\mathcal{H}_0$ *is a nonsmooth space and the space* $\mathcal{H}_1$ *is orthogonal to* $\mathcal{H}_0$*. A space generated by the columns of a matrix is called a column space.*

Proof. See [Gu96]. □

Theorem 2.9 (RK of RKHS on $\{1, 2, \ldots, K\}$) *Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be as in Theorem 2.8. The RK of $\mathcal{H}_0$ is*

$$R_0(s,t) = (\mathbf{1}\mathbf{1}^T/K)_{st}$$

and the RK of $\mathcal{H}_1$ is

$$R_1(s,t) = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/K)_{st}.$$

Proof. See [Gu96] or [Gu95]. □

The RK of $W_{\{K\}}$ with the norm (2.13) can be presented as

$$\begin{aligned} R(s,t) &= R_0(s,t) + R_1(s,t) \\ &= 1/K + (\mathbf{I} - \mathbf{1}\mathbf{1}^T/K)_{st}. \end{aligned}$$

Example 2.6 *Consider the RKHS $W_{\{2\}}$. The RK $R_0(s,t)$ of $W_{\{2\}}$ is $1/2$. The corresponding $R_1(s,t)$ is $1/2$ for $s = t$ and $-1/2$ for $s \neq t$*

2.3 Use of RK Splines in Approximation

We are now ready to utilize the RKHS in function approximation. Suppose that the RKHS $\mathcal{H}$ has a decomposition

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1, \tag{2.14}$$

where $\mathcal{H}_0$ is a finite dimensional space containing the terms that are not going to be penalized and $\mathcal{H}_1$ is the orthogonal space to $\mathcal{H}_0$ containing the terms that will be penalized.

2.3.1 RK Spline Function

We used earlier a penalty term that was an integral over $\mathbb{R}$. In order to use non-continuous index sets, we need to use a more general form of the penalty. From now on we present the error function

$$E = \frac{1}{n}\|\mathbf{y} - \mathbf{f}(\mathbf{X})\|^2 + J_\lambda(f)$$

in a more general form

$$E = \frac{1}{n}\|\mathbf{y} - \mathbf{f}(\mathbf{X})\|^2 + \|P_\lambda f\|^2,$$

where $P_\lambda f$ is an orthogonal projection of f onto $\mathcal{H}_1$ in $\mathcal{H}$. The projection is a penalty of oscillation.

Let

$$\mathcal{H}_1 = \bigoplus_{j=1}^{p} \mathcal{H}^j$$

and the corresponding square norm

$$\|P_\lambda f\|^2 = \sum_{j=1}^{p} \frac{1}{\lambda_j} \|P^j f\|^2,$$

where $\lambda = (\lambda_1, \lambda_2 \dots \lambda_p)$, $\lambda_j > 0$, $j = 1, \dots, p$, *is a parameter vector controlling the level of smoothing* [Wah90, p. 127]. In Section 2.2 we used norms without λ. Now each subspace $\mathcal{H}^j$ has its own smoothing parameter. Let the RK of $\mathcal{H}_1$ be $R_{1,\lambda}$ and the RK of $\mathcal{H}^j$ correspondingly be $\lambda_j R^j$. Now the RK of the penalized terms can be expressed as

$$R_{1,\lambda} = \sum_{j=1}^{p} \lambda_j R^j.$$

Theorem 2.10 (RK Spline) *Let $\mathcal{H}_0 = \{h_0\} \oplus \{h_1\} \oplus \cdots \oplus \{h_m\}$. The solution to the minimization problem*

$$\min_{f \in \mathcal{H}} \frac{1}{n} \|\mathbf{y} - \mathbf{f}(\mathbf{X})\|^2 + \|P_\lambda f\|^2 \tag{2.15}$$

for a fixed λ has the form

$$f(\mathbf{x}) = \sum_{r=1}^{m} d_r h_r(\mathbf{x}) + \sum_{i=1}^{n} c_i R_{1,\lambda}(\mathbf{x}, \mathbf{x}_i), \tag{2.16}$$

with the prerequisite that $\mathbf{x}_i$ are distinct, $i = 1, \dots, n$. The coefficients $\mathbf{c}$ and $\mathbf{d}$ have a unique solution if λ is a constant vector.

Proof. See [KW71] or [Wah90, pp. 11-12]. □

2.3.2 RK Spline Error Minimization

Our aim is to find $\mathbf{c}$ and $\mathbf{d}$ minimizing (2.15). Let $\mathbf{Q}_{\lambda\, n\times n}$ and $\mathbf{S}_{n\times m}$ be matrices with entries

$$\begin{aligned} (Q_\lambda)_{ik} &= R_{1,\lambda}(\mathbf{x}_i, \mathbf{x}_k) \\ S_{ir} &= h_r(\mathbf{x}_i). \end{aligned}$$

Then (2.16) can be presented as

$$\mathbf{f}(\mathbf{X}) = \mathbf{S}\mathbf{d} + \mathbf{Q}_\lambda \mathbf{c}$$

and the error function (2.15) correspondingly as

$$E = \frac{1}{n}\|\mathbf{y} - \mathbf{S}\mathbf{d} - \mathbf{Q}_\lambda \mathbf{c}\|^2 + \|P_\lambda f\|^2,$$

where the penalty can be extracted as

$$\begin{aligned}
\|P_\lambda f\|^2 &= \langle P_\lambda f, P_\lambda f\rangle \\
&= \langle \sum_{i=1}^{n} c_i R_{1,\lambda}(\cdot, \mathbf{x}_i), \sum_{j=1}^{n} c_j R_{1,\lambda}(\cdot, \mathbf{x}_j)\rangle \\
&= \sum_{i,j=1}^{n} c_i c_j R_{1,\lambda}(\mathbf{x}_i, \mathbf{x}_j) \\
&= \mathbf{c}^T \mathbf{Q}_\lambda \mathbf{c}.
\end{aligned}$$

So, in order to minimize (2.15) we are to find the vectors $\mathbf{c}$ and $\mathbf{d}$ minimizing the error function

$$E = \frac{1}{n}\|\mathbf{y} - \mathbf{S}\mathbf{d} - \mathbf{Q}_\lambda \mathbf{c}\|^2 + \mathbf{c}^T \mathbf{Q}_\lambda \mathbf{c} \tag{2.17}$$

for a given parameter vector λ.

Theorem 2.11 *The vectors* $\mathbf{c}$ *and* $\mathbf{d}$ *minimizing* (2.17) *can be solved from the system of equations*

$$\begin{aligned}
(\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{c} + \mathbf{S}\mathbf{d} &= \mathbf{y} \\
\mathbf{S}^T \mathbf{c} &= \mathbf{0}.
\end{aligned} \tag{2.18}$$

Proof. See [Wah78] or [Wah90, p. 12]. □

For any matrix $\mathbf{A}_{n\times m}$, $n \geq m$, that has linearly independent columns there exists a unique pair $\mathbf{Q}_{n\times n}$ and $\mathbf{R}_{n\times m}$, such that $\mathbf{A} = \mathbf{QR}$, $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is an upper triangular matrix. This factorization is called *QR-decomposition* [Fie86, p. 239]. To evaluate $\mathbf{c}$ and $\mathbf{d}$, consider the QR-decomposition

$$\mathbf{S} = \begin{pmatrix} \mathbf{F}_1 & \mathbf{F}_2 \end{pmatrix} \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}, \tag{2.19}$$

where $(\mathbf{F}_1\ \mathbf{F}_2)$ is orthogonal. In the decomposition, $\mathbf{F}_1$ is a $n \times m$ matrix, $\mathbf{F}_2$ a $n \times (n-m)$ matrix, $\mathbf{R}_{m\times m}$ an upper triangular matrix, and $\mathbf{0}_{(n-m)\times m}$ the zero matrix.

Theorem 2.12 *The solution to the system of equations* (2.18) *is*

$$\begin{aligned}
\mathbf{c} &= \mathbf{F}_2(\mathbf{F}_2^T(\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{F}_2)^{-1}\mathbf{F}_2^T \mathbf{y} \\
\mathbf{d} &= \mathbf{R}^{-1}\mathbf{F}_1^T(\mathbf{y} - (\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{c}).
\end{aligned}$$

Proof. The proof follows the guidelines of [Wah78] and [Wah90, pp. 12-13]. By multiplying (2.19) from left with $(\mathbf{F}_1\ \mathbf{F}_2)^T$, we get

$$\begin{pmatrix} \mathbf{F}_1^T \\ \mathbf{F}_2^T \end{pmatrix} \mathbf{S} = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix},$$

from which we see that $\mathbf{F}_2^T\mathbf{S} = \mathbf{0}_{m\times(n-m)}$ and $\mathbf{F}_1^T\mathbf{S} = \mathbf{R}_{m\times m}$. If we multiply (2.18) from left by $\mathbf{F}_2^T$ and replace $\mathbf{c}$ with $\mathbf{F}_2\mathbf{z}$ where $\mathbf{z}$ is a suitable vector (this can be done because $\mathbf{S}^T\mathbf{c} = \mathbf{0}$ and therefore $\mathbf{c}$ belongs to the column space of $\mathbf{F}_2$), we get

$$\begin{aligned} \mathbf{F}_2^T(\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{c} + \underbrace{\mathbf{F}_2^T\mathbf{S}}_{=\mathbf{0}}\,\mathbf{d} &= \mathbf{F}_2^T\mathbf{y} \\ \mathbf{F}_2^T(\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{F}_2\mathbf{z} &= \mathbf{F}_2^T\mathbf{y} \\ \mathbf{z} &= (\mathbf{F}_2^T(\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{F}_2)^{-1}\mathbf{F}_2^T\mathbf{y}. \end{aligned}$$

Now we get the explicit formula for $\mathbf{c}$,

$$\begin{aligned} \mathbf{c} &= \mathbf{F}_2\mathbf{z} \\ &= \mathbf{F}_2(\mathbf{F}_2^T(\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{F}_2)^{-1}\mathbf{F}_2^T\mathbf{y}. \end{aligned} \tag{2.20}$$

Vector $\mathbf{d}$ can be calculated by multiplying (2.18) from left with $\mathbf{F}_1^T$,

$$\begin{aligned} \mathbf{F}_1^T(\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{c} + \underbrace{\mathbf{F}_1^T\mathbf{S}}_{=\mathbf{R}}\,\mathbf{d} &= \mathbf{F}_1^T\mathbf{y} \\ \mathbf{d} &= \mathbf{R}^{-1}\mathbf{F}_1^T(\mathbf{y} - (\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{c}). \quad \square \end{aligned} \tag{2.21}$$

There is no closed form expression from which one could calculate the value of the parameter vector λ. In order to develop an efficient algorithm we need to study the relationship of $\mathbf{y}$ and $\mathbf{f}(\mathbf{X})$. Recall that

$$\mathbf{f}(\mathbf{X}) = \mathbf{S}\mathbf{d} + \mathbf{Q}_\lambda\mathbf{c}.$$

If we substitute this into (2.18) and replace $\mathbf{c}$ with (2.20), we get

$$\begin{aligned} \mathbf{f}(\mathbf{X}) &= \mathbf{y} - n\mathbf{I}\mathbf{c} \\ &= (\mathbf{I} - n\mathbf{F}_2(\mathbf{F}_2^T(\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{F}_2)^{-1}\mathbf{F}_2^T)\mathbf{y} \\ &= \mathbf{A}_\lambda\mathbf{y}, \end{aligned} \tag{2.22}$$

where $\mathbf{A}_\lambda$ is called a *smoothing matrix*. An iterative algorithm determining the value of the parameter vector λ will be described in Chapter 3. The algorithm is called *generalized cross validation* [CW79] and it utilizes the smoothing matrix above.

2.3.3 RK spline Example on $[0,1]$

The purpose of next example is to demonstrate the use of RK splines on a simple one-dimensional approximation problem. Consider the training pairs (x_i, y_i), where $i = 1, \ldots, 30$,

$$\begin{aligned}(x_1, \ldots, x_{15})^T &\sim \mathbf{0.25} + 0.05 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I}_{15\times 15}),\\ (x_{15}, \ldots, x_{30})^T &\sim \mathbf{0.75} + 0.05 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I}_{15\times 15}).\end{aligned}$$

The corresponding y-values are given by

$$y_i = \sin(7x_i^2) + \epsilon_i,$$

where $(\epsilon_1, \ldots, \epsilon_{30})^T \sim 0.1 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I}_{30\times 30})$. We pretend not to know the function that generates the data, i.e., $\sin(7x^2)$, but we try to estimate it based on the observed pairs (x_i, y_i). Figure 2.1 displays the randomized pairs.

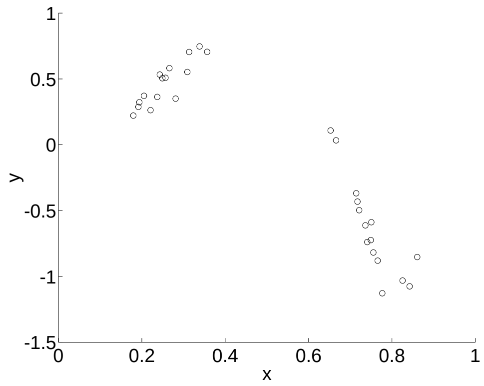


Figure 2.1: Observed pairs (x_i, y_i).

In this experiment, we can utilize the index set $\mathcal{T} = [0,1]$. We need to find a model f from the RKHS

$$W_2^1[0,1] = \{h_0\} \oplus \{h_1\} \oplus \mathcal{H}_1.$$

The RK of the RKHS $W_2^1[0,1]$ is $R_0(s,t) + R_1(s,t)$, where

$$\begin{aligned}R_0(s,t) &= h_0(s)h_0(t) + h_1(s)h_1(t)\\ &= 1 + (s-1/2)(t-1/2)\end{aligned}$$

and

$$\begin{aligned}R_1(s,t) &= h_2(s)h_2(t) - h_4(|s-t|)\\ &= 1 + (s-1/2)(t-1/2) + 1/16(s^2 - s + 1/6)(t^2 - t + 1/6)\\ &\quad -1/24((s-t)^4 - 2|s-t|^3 + (s-t)^2 - 1/30).\end{aligned}$$

Using (2.16), the function f can be written as

$$f(x) = d_0 h_0(x) + d_1 h_1(x) + \sum_1^{30} c_i \lambda_1 R_1(x, x_i),$$

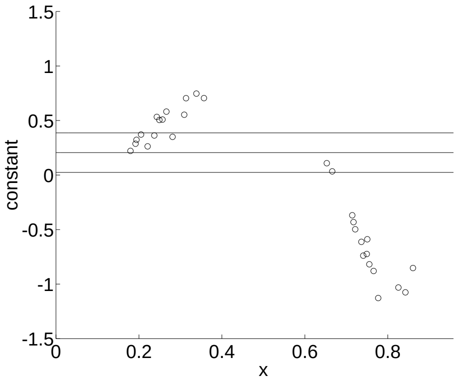
where vectors **c** and **d** are calculated as shown in (2.20) and (2.21). The smoothing coefficient λ_1 is determined using the generalized cross validation algorithm described in Chapter 3. In this example, the calculated value of λ_1 is $3.6053 \cdot 10^{-5}$.

Figure 2.2 shows the decomposition of the model function f. Figure (a) displays the constant d_0. The constant is unique because the integral over the domain $[0, 1]$ of all other terms equals to zero. Figure (b) shows the nonsmooth term $d_1(x - 1/2)$. It is a regression line that equals to zero at $x = 1/2$. Figure (c) shows the nonlinear term and Figure (d) the sum of all three terms with the 95% confidence interval (CI). The calculation of the CI is described in Chapter 3. Figure (d) also shows the original *sin* function. The original function is inside the 95% CI in all points of the domain $[0, 1]$. The CI is large on the areas where there are few or no training points, but narrow on the areas where there is a lot of training data.

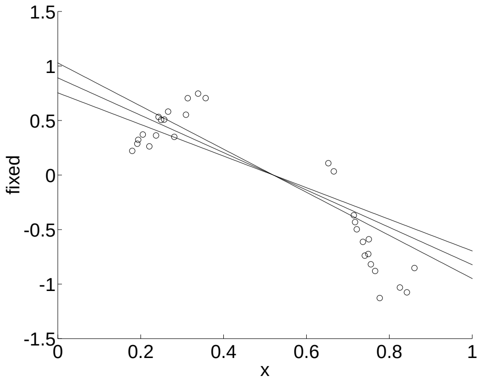
2.4 Further Reading

Shepard introduced bivariate metric interpolation in 1968 [She68, GW78]. The method was able to produce a surface from irregularly spaced data. A few years later Duchon presented multivariate thin plate spline interpolation [Duc77, Wah90, VW83]. Fast evaluation of these splines has been studied in [Pow93, LS94, BN92]. Franke has compared two-dimensional interpolation methods in [Fra82]. For multivariate thin plate splines, see [Wah90] and for other multivariate splines, see [GS94, Chu88, Buh90, BF96, Che86].

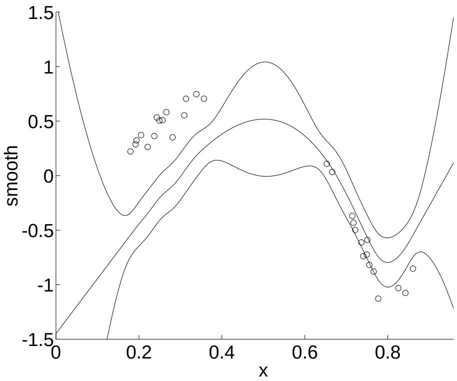
Guidelines on calculating RK functions to new domains can be found in [TA96, Wah90]. For example, global temperature estimation requires an RKHS for a sphere. Symbolic mathematics software systems, such as Macsyma, Mathematica or Maple are useful tools in calculation of new kernel functions. Precalculated RK functions for many common cases can be found, e.g., in [Wah90, TA96, Gir98, Att92, BV93].



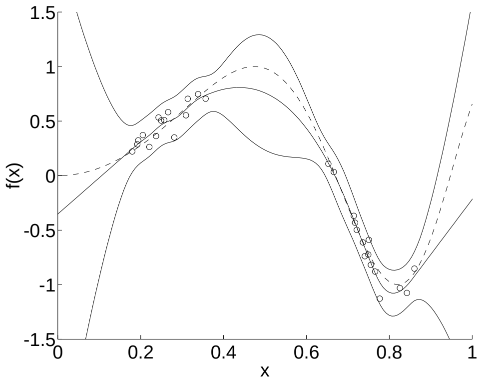
(a) Constant term $d_0 h_0(x)$.



(b) Nonsmooth term $d_1 h_1(x)$.



(c) Smooth term $\sum_{i=1}^{30} c_i \lambda_1 R_1(x, x_i)$.



(d) Model function $f(x)$.

Figure 2.2: Approximation of $sin(7x^2)$ on $[0, 1]$. Circles represent the training data, three solid lines the mean and component-wise 95% Bayesian confidence interval and the dashed line in Figure (d) the original sin-function.

Chapter 3

Generalized Cross Validation (GCV)

The goal of training is to build a statistical model f of the process that generated the training data. The function f is determined using the training data and the performance of the function is estimated comparing the output of the model to test data. The form of f is controlled by a parameter vector λ. Below we will denote the model function with f_λ to emphasize the fact that it depends on the control parameters. If we consider the linear regression and smooth interpolation of the training data as the two extremes of approximation, a good f_λ usually lies somewhere between them [Wah94]. Linear regression is easy to use but it lacks the expressive power of the nonlinear methods. A smooth interpolation of the training data obviously models the training data well but usually the performance with the test data is poor due to overfitting.

In the previous chapter, we used RK splines with a fixed parameter vector λ. The parameter vector determines the level of smoothing, that is how well the model function interpolates the training data. The aim of generalization is to determine a λ that minimizes the modeling error for the test data. In this chapter, we discuss an iterative *generalized cross validation* (GCV) algorithm [CW79] to determine λ.

Neural computation is often presented emphasizing the possibility to solve approximation problems in parallel. Few neural network implementations actually utilize parallelism because of its inefficiency. The nature of many training algorithms, including back-propagation, is such that the communication between the nodes will become a bottleneck in parallel implementations. The GCV algorithm is similar to the training algorithms used in neural computation. In this chapter, we introduce an efficient parallel GCV algorithm, which distributes calculation to several processors. The parallel GCV algorithm utilizes distributed matrix decompositions and it minimizes the communication between the processors. The purpose of the chapter is to show that RK splines can be calculated efficiently in parallel. The parallel GCV algorithms have not been published earlier.

Before describing GCV, we study ordinary *cross validation* (CV) [All74]. Let $\{f_\lambda^i \mid i = 1, \ldots, n\}$ be a set of model functions each trained with all training patterns excluding the pattern i. The aim of CV is to find the parameter vector λ that minimizes the CV function

$$C_\lambda = \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\lambda^i(\mathbf{x}_i))^2. \tag{3.1}$$

This form of the CV function is impractical because it requires n model functions f_λ^i. A common method to alleviate the computational cost is to construct only $m \ll n$ model functions each trained with $n-(n/m)$ patterns. This method is called m-fold cross validation [MST94].

We next describe, how the CV function (3.1) can be calculated utilizing the smoothing matrix of (2.22) to alleviate the computational complexity of generalization.

Definition 3.1 (Smoothing matrix) *The matrix*

$$\mathbf{A}_\lambda = (\mathbf{I} - n\mathbf{F}_2(\mathbf{F}_2^T(\mathbf{Q}_\lambda + n\mathbf{I})\mathbf{F}_2)^{-1}\mathbf{F}_2^T) \tag{3.2}$$

in (2.22) *is called a* smoothing matrix*.*

Theorem 3.1 (CV function) *Utilizing the smoothing matrix* (3.2) *the CV function* (3.1) *can be written in the form*

$$C_\lambda = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - f_\lambda(\mathbf{x}_i))^2}{(1 - a_{ii})^2}, \tag{3.3}$$

where a_{ii} are the diagonal elements of $\mathbf{A}_\lambda$.

Proof. See [CW79, Wah90]. □

3.1 GCV Function

GCV is a modification of CV. In GCV, the coefficients a_{ii} of the CV function (3.3) are replaced by their average $1/n \sum_{i=1}^{n} a_{ii}$. GCV has certain invariance properties that do not hold in the ordinary CV [Wah90]. Monte Carlo experiments on artificial data suggest that on average the GCV performs better than the CV [CW79, WW93].

Definition 3.2 (GCV function) *The function*

$$G_\lambda = \frac{\| (\mathbf{I} - \mathbf{A}_\lambda)\mathbf{y} \|^2}{\mathrm{trace}(\mathbf{I} - \mathbf{A}_\lambda)^2}. \tag{3.4}$$

is called the GCV function.

By substituting the average of the diagonal elements of $\mathbf{A}_\lambda$ to (3.3) and remembering that $\mathbf{f}_\lambda = \mathbf{A}_\lambda \mathbf{y}$ (as was shown in (2.22)) we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - f_\lambda(\mathbf{x}_i))^2}{(1 - 1/n\sum_{i=1}^{n} a_{ii})^2} &= \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - f_\lambda(\mathbf{x}_i))^2}{(1 - \frac{1}{n}\mathrm{trace}(\mathbf{A}_\lambda))^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_\lambda(\mathbf{x}_i))^2}{(\frac{1}{n}\mathrm{trace}(\mathbf{I} - \mathbf{A}_\lambda))^2} \\
&= \frac{n \parallel \mathbf{y} - \mathbf{f}_\lambda \parallel^2}{\mathrm{trace}(\mathbf{I} - \mathbf{A}_\lambda)^2} \\
&= \frac{n \parallel (\mathbf{I} - \mathbf{A}_\lambda)\mathbf{y} \parallel^2}{\mathrm{trace}(\mathbf{I} - \mathbf{A}_\lambda)^2} \\
&= nG_\lambda. \qquad (3.5)
\end{aligned}$$

The coefficient n is a constant and therefore it does not influence the minimum point of the function.

We use G_λ as the error function when determining the parameter vector λ of the RK splines. The evaluation of $\mathbf{A}_\lambda$ and therefore also G_λ requires $O(n^3)$ operations, where n is the number of training patterns. The parameter vector λ is determined using an iterative algorithm in which G_λ is evaluated repeatedly with different parameter values. The most efficient algorithms use either the Cholesky decomposition [Mei79] or tridiagonal Householder decomposition [GBCW89]. We use the former decomposition because it is easier to calculate in parallel. Several authors have also suggested heuristic GCV functions [WJGG94, WL95, Luo96b, Luo96a, XW95, Hut89] to speed up generalization.

3.1.1 Fast Evaluation of GCV

In order to minimize the evaluation time of the numerator of the GCV function (3.4), we rewrite it as

$$\parallel (\mathbf{I} - \mathbf{A}_\lambda)\mathbf{y} \parallel^2 = \left\| n\mathbf{F}_2(\mathbf{F}_2^T\mathbf{Q}_\lambda\mathbf{F}_2 + n\mathbf{I})^{-1}\mathbf{F}_2^T\mathbf{y} \right\|^2 . \qquad (3.6)$$

This can be done by substituting (3.2) in the place of $\mathbf{A}_\lambda$.

Lemma 3.2 *Matrix* $(\mathbf{F}_2^T\mathbf{Q}_\lambda\mathbf{F}_2 + n\mathbf{I})$ *has the Cholesky decomposition*

$$\mathbf{U}_\lambda^T\mathbf{U}_\lambda = \mathbf{F}_2^T\mathbf{Q}_\lambda\mathbf{F}_2 + n\mathbf{I}, \qquad (3.7)$$

where $\mathbf{U}_\lambda$ *is an invertible upper triangular matrix.*

Proof. A symmetric real matrix $\mathbf{A}$ is said to be *positive definite* if $\langle \mathbf{Ax}, \mathbf{x}\rangle > 0$ holds for any real vector $\mathbf{x} \neq 0$. A matrix has the *Cholesky decomposition* if and only if it is positive definite [Fie86, Chapter 2]. Therefore it is enough to show that $(\mathbf{F}_2^T\mathbf{Q}_\lambda\mathbf{F}_2 + n\mathbf{I})$ is positive definite.

i. $\mathbf{Q}_\lambda$ is a positive definite matrix (see [Wah90] for details).

ii. Using the observation (i) we know that for all $\mathbf{x} \neq 0$

$$\mathbf{x}^T(\mathbf{F}_2^T\mathbf{Q}_\lambda\mathbf{F}_2 + n\mathbf{I})\mathbf{x} = \mathbf{y}^T\mathbf{Q}_\lambda\mathbf{y} + n\mathbf{x}^T\mathbf{x} \geq 0,$$

where $\mathbf{y} = \mathbf{F}_2\mathbf{x}$. Therefore the matrix $(\mathbf{F}_2^T\mathbf{Q}_\lambda\mathbf{F}_2 + n\mathbf{I})$ is positive definite. □

Theorem 3.3 (Fast GCV function) *The GCV function* (3.4) *can be written in the form*

$$G_\lambda = \frac{n\left\| \mathbf{F}_2(\mathbf{U}_\lambda^{-1})^T\mathbf{U}_\lambda^{-1}\mathbf{F}_2^T\mathbf{y} \right\|^2}{(\sum_{i=1}^n \sum_{j=i}^n u_{ij}^2)^2}, \tag{3.8}$$

where u_{ij} are the elements of $\mathbf{U}_\lambda^{-1}$.

Proof. Using Lemma 3.2 and (3.6), the numerator of (3.4) can be rewritten as

$$\begin{aligned}
\| (\mathbf{I} - \mathbf{A}_\lambda)\mathbf{y} \|^2 &= n^2 \left\| \mathbf{F}_2(\mathbf{F}_2^T\mathbf{Q}_\lambda\mathbf{F}_2 + n\mathbf{I})^{-1}\mathbf{F}_2^T\mathbf{y} \right\|^2 \\
&= n^2 \left\| \mathbf{F}_2(\mathbf{U}_\lambda^T\mathbf{U}_\lambda)^{-1}\mathbf{F}_2^T\mathbf{y} \right\|^2 \\
&= n^2 \left\| \mathbf{F}_2(\mathbf{U}_\lambda^{-1})^T\mathbf{U}_\lambda^{-1}\mathbf{F}_2^T\mathbf{y} \right\|^2.
\end{aligned}$$

Due to its construction the matrix $\mathbf{F}_2$ is independent of the smoothing parameters. The inverted triangular matrix $\mathbf{U}_\lambda^{-1}$, however, depends on λ and therefore it has to be recalculated each time the parameters change. In the next section, we describe a parallel algorithm to calculate $\mathbf{U}_\lambda^{-1}$.

We have now done with the nominator of (3.4). The denominator of (3.4) can be calculated as

$$\begin{aligned}
\mathrm{trace}(\mathbf{I} - \mathbf{A}_\lambda)^2 &= n \cdot \mathrm{trace}(\mathbf{F}_2(\mathbf{U}_\lambda^{-1})^T\mathbf{U}_\lambda^{-1}\mathbf{F}_2^T)^2 \\
&= n \cdot \mathrm{trace}(\mathbf{F}_2^T\mathbf{F}_2(\mathbf{U}_\lambda^{-1})^T\mathbf{U}_\lambda^{-1})^2 \\
&= n \cdot \mathrm{trace}((\mathbf{U}_\lambda^{-1})^T\mathbf{U}_\lambda^{-1})^2 \\
&= n \, \| \mathbf{U}_\lambda^{-1} \|_F^4 \\
&= n(\sum_{i=1}^n \sum_{j=i}^n u_{ij}^2)^2,
\end{aligned}$$

where u_{ij} are the elements of $\mathbf{U}_\lambda^{-1}$. The Frobenius norm $\| \cdot \|_F$ has an equivalence $\|\mathbf{A}\|_F^2 = \mathrm{trace}(\mathbf{A}^T\mathbf{A})$, which has been utilized above. □

3.2 Confidence Interval of GCV

The reliability of the approximation is often the key issue in the medical data analysis. Some observations can be classified reliably by machine learning algorithms, but there are instances that require the professional evaluation of a physician. For this purpose it is useful to calculate the reliability of the model for observations.

Theorem 3.4 (Point-wise conditional variance) *The conditional variance* $\mathrm{var}(f(x_i)|\mathbf{y})$ *can be calculated as*

$$\mathrm{var}(f(x_i)|\mathbf{y}) = \sigma^2 a_{ii},$$

where σ^2 *is the sampling variance estimate and* a_{ii} *are the diagonal elements of* $\mathbf{A}_\lambda$.

Proof. See [CW79, Wah83, GW92]. □

Using Theorem 3.4 we can express f with the point-wise Bayesian CI as

$$f(x_i) \pm z_{\alpha/2}\sqrt{\sigma^2 a_{ii}},.$$

where α is the required confidence level. The value of $z_{\alpha/2}$ can be obtained from any standard statistical handbook. See [Wah90, Gu96] for details.

3.3 Parallel GCV

We next describe a parallel algorithm to calculate G_λ of (3.8). The parallel Cholesky decomposition and triangular matrix inversion were performed using Scalable Linear Algebra PACKage (ScaLAPACK) [CDO$^+$94, DW97, CH91]. The ScaLAPACK is a general purpose linear algebra library which includes many parallel matrix operations.

In the ScaLAPACK, matrices are presented using a 2-dimensional *block cyclic distribution* [SH96]. The aim of the representation is to distribute the calculation to several processors as evenly as possible. The processors are arranged as a 2-dimensional grid and the matrices are distributed over the processor grid. Figure 3.1 shows an example of the block cyclic distribution.

The *block size* has a great influence on the performance and therefore it should be chosen carefully. A large block size distributes matrices unevenly and the processors on the upper left corner of the grid are busier than the processor of the lower right corner. On the other hand, a small block size distributes matrices evenly but the communication between the processors becomes a bottleneck. Guidelines on choosing an appropriate block size can be found in [CDO$^+$94].

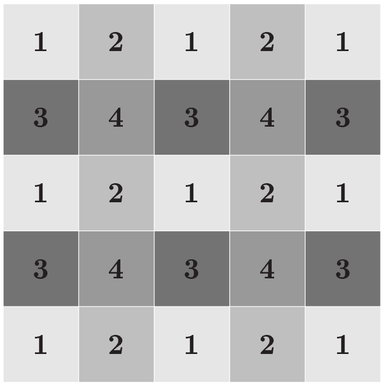


Figure 3.1: An example of a 2-dimensional block cyclic distribution of a matrix on a 2×2 processor grid. The matrix is divided into blocks of equal size starting from the upper left corner. The four gray levels illustrate the blocks owned by different processors. For example, if the block size is 50, the above figure illustrates the distribution of a 250×250 matrix to four processors.

We next describe an algorithm that evaluates expressions in parallel. We assume that the programs run in a computer system with shared memory, so that the parallel processes can access the memory of the parent process. In addition, the parallel processes have only a constant amount of local storage for the process identity and necessary local variables.

Algorithm 3.1 (Evaluation in Block Cyclic Distribution) *The algorithm calculates expressions in parallel using a 2-dimensional block cyclic distribution. The parameters i and j denote the row and column of a matrix block, S denotes the expressions to be evaluated. Each processor knows its position (pRow,pColumn) on the processor grid* $pMax \times pMax$. *The position is stored in the local memory of the process.*

function $evalBCD(i, j, S)$
if $(pRow = 1 + mod(i-1, pMax)) \wedge$
  $(pColumn = 1 + mod(j-1, pMax))$ **then** $evaluate(S)$
return

In algorithms we use the control word **pardo** to denote that the calculation is distributed over the processor grid. Algorithm 3.1 is an abstract presentation of calculation in a block cyclic distribution. In practice, parallel matrix operations are usually implemented using FORTRAN or C, and pseudo functions such as $evaluate(S)$ are not used. We use the notation $\mathbf{A}_{[ij]}$ to denote the matrix block $(1 + (i-1) * b : r)(1 + (j-1) * b : l)$ of $\mathbf{A}$, where $r = min(rank(\mathbf{A}), i * b)$, $l = min(rank(\mathbf{A}), j * b)$ and b is the block size.

3.3.1 Parallel Cholesky Decomposition

The most expensive operations in the calculation of G_λ are the Cholesky decomposition, triangular matrix inversion and matrix multiplication. All these operations can be performed in parallel. In this section, we discuss the Cholesky decomposition with the ScaLAPACK routines. Consider a positive definite matrix $\mathbf{A}$ that has a Cholesky decomposition $\mathbf{U}^T\mathbf{U}$, where $\mathbf{U}$ is an upper triangular matrix. The decomposition $\mathbf{A} = \mathbf{U}^T\mathbf{U}$ can be presented in a block form as

$$\begin{aligned} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{21}^T \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} &= \begin{pmatrix} \mathbf{U}_{11}^T & \mathbf{0} \\ \mathbf{U}_{21}^T & \mathbf{U}_{22}^T \end{pmatrix} \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{21} \\ \mathbf{0} & \mathbf{U}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{U}_{11}^T\mathbf{U}_{11} & \mathbf{U}_{11}^T\mathbf{U}_{21} \\ \mathbf{U}_{21}^T\mathbf{U}_{11} & \mathbf{U}_{21}^T\mathbf{U}_{21} + \mathbf{U}_{22}^T\mathbf{U}_{22} \end{pmatrix}, \end{aligned} \tag{3.9}$$

where $\mathbf{A}_{11}$, $\mathbf{A}_{22}$ are symmetric matrices, $\mathbf{U}_{11}$, $\mathbf{U}_{22}$ upper triangular matrices, and $\mathbf{A}_{21}$, $\mathbf{U}_{21}$ full matrices. The matrices $\mathbf{A}_{11}$ and $\mathbf{U}_{11}$ are selected so that their rank is the block size (unless rank($\mathbf{A}$) is less than the block size in which case no parallelism is required). From (3.9) we get

$$\begin{aligned} \mathbf{A}_{11} &= \mathbf{U}_{11}^T\mathbf{U}_{11} \\ \mathbf{A}_{21} &= \mathbf{U}_{21}^T\mathbf{U}_{11} \\ \mathbf{A}_{22} &= \mathbf{U}_{21}^T\mathbf{U}_{21} + \mathbf{U}_{22}^T\mathbf{U}_{22}. \end{aligned} \tag{3.10}$$

Consider the equations of (3.10) rewritten as

$$\begin{aligned} \mathbf{A}_{11} &= \mathbf{U}_{11}^T\mathbf{U}_{11} \\ \mathbf{U}_{21}^T &= \mathbf{A}_{21}\mathbf{U}_{11}^{-1} \\ \mathbf{A}_{22} - \mathbf{U}_{21}^T\mathbf{U}_{21} &= \mathbf{U}_{22}^T\mathbf{U}_{22}. \end{aligned} \tag{3.11}$$

The matrix $\mathbf{U}_{11}$ can be solved utilizing the standard Cholesky decomposition within a block [PTVF92] and $\mathbf{U}_{21}^T$ can be calculated from (3.11). The matrix $\mathbf{U}_{22}$ can be calculated using (3.9) on $\mathbf{A}_{22} - \mathbf{U}_{21}^T\mathbf{U}_{21}$.

We can distribute the calculation to several processors utilizing the block cyclic distribution. Algorithm 3.2 calculates the Cholesky decomposition in parallel. The core of the algorithm is executed only in one processor. The phases (i)-(iii) in the algorithm refer to the lines of (3.11). Figure 3.2

demonstrates one execution of the main loop of the algorithm.

Algorithm 3.2 (Parallel Cholesky decomposition) *The algorithm returns the upper triangular matrix* $\mathbf{U}$ *of the Cholesky decomposition* $\mathbf{A} = \mathbf{U}^T\mathbf{U}$. *The parameter* $\mathbf{A}$ *denotes a positive definite matrix and* b *the block size. The function* $cholesky(\mathbf{A}_{[ii]})$ *calculates* $\mathbf{U}_{[ii]}$ *of the standard Cholesky decomposition* $\mathbf{A}_{[ii]} = \mathbf{U}_{[ii]}^T\mathbf{U}_{[ii]}$ *without parallelism.*

> **function** $parallelCholesky(\mathbf{A}, b)$
> $m \leftarrow \lceil rank(\mathbf{A})/b \rceil$
> **for** $i \leftarrow 1$ **to** m **do**
> $\mathbf{A}_{[ii]} \leftarrow transpose(cholesky(\mathbf{A}_{[ii]}))$ % phase (i)
> $\mathbf{T} \leftarrow inverse(\mathbf{A}_{[ii]})$
>   **for** $j \leftarrow i+1$ **to** m **pardo** % phase (ii)
> $evalBCD(j, i,$ "$\mathbf{A}_{[ji]} \leftarrow \mathbf{A}_{[ji]}\mathbf{T}$")
>   **for** $j \leftarrow i+1$ **to** m **pardo** % phase (iii)
>   **for** $k \leftarrow i+1$ **to** m **pardo**
> $evalBCD(k, j,$ "$\mathbf{A}_{[kj]} \leftarrow \mathbf{A}_{[kj]} -$
> $transpose(\mathbf{A}_{[(i+1:m)i]})\mathbf{A}_{[(i+1:m)i]}$")
> **od**
> $return(\mathbf{A})$

3.3.2 Parallel Triangular Matrix Inversion

We also need an algorithm that calculates the parallel triangular matrix inversion used in G_λ of Theorem 3.3. In this section, we discuss the triangular matrix inversion with the ScaLAPACK routines. Consider a nonsingular upper triangular matrix $\mathbf{U}$ that should be inverted. In order to use a parallel algorithm, we need a block version of the matrix inversion. The matrix inversion $\mathbf{UV} = \mathbf{I}$ (i.e., $\mathbf{V} = \mathbf{U}^{-1}$) can be presented in a block form as

$$\begin{aligned} \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{21} \\ \mathbf{0} & \mathbf{U}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{21} \\ \mathbf{0} & \mathbf{V}_{22} \end{pmatrix} &= \begin{pmatrix} \mathbf{U}_{11}\mathbf{V}_{11} & \mathbf{U}_{11}\mathbf{V}_{21} + \mathbf{U}_{21}\mathbf{V}_{22} \\ \mathbf{0} & \mathbf{U}_{22}\mathbf{V}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \end{aligned} \tag{3.12}$$

where $\mathbf{U}_{11}$, $\mathbf{V}_{11}$, $\mathbf{U}_{22}$, $\mathbf{V}_{22}$ are upper triangular matrices and $\mathbf{U}_{21}$, $\mathbf{V}_{21}$ full matrices. The matrices $\mathbf{U}_{11}$ and $\mathbf{V}_{11}$ are selected so that their rank is the block size (unless rank($\mathbf{U}$) is less than the block size in which case no parallelism is required). From (3.12) we get

$$\begin{aligned} \mathbf{U}_{11}\mathbf{V}_{11} &= \mathbf{I} \\ \mathbf{U}_{22}\mathbf{V}_{22} &= \mathbf{I} \\ \mathbf{U}_{11}\mathbf{V}_{21} + \mathbf{U}_{21}\mathbf{V}_{22} &= \mathbf{0}. \end{aligned} \tag{3.13}$$

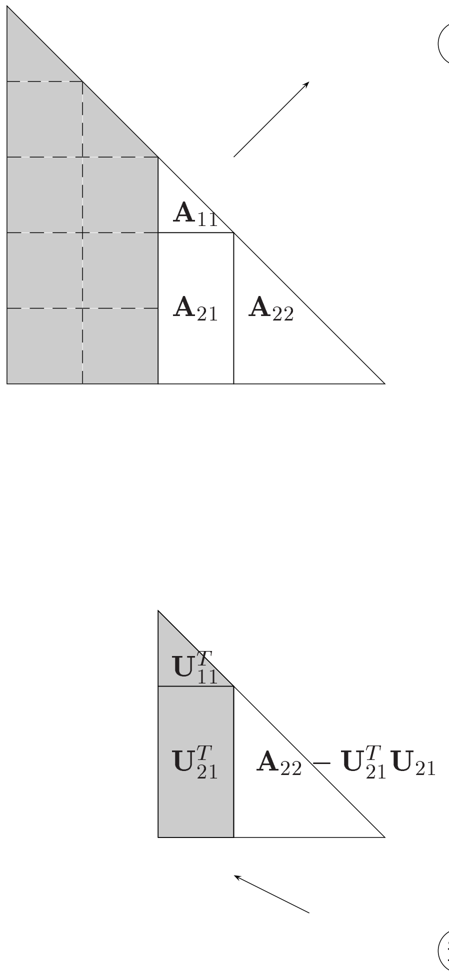


Figure 3.2: Demonstration of one iteration of the parallel Cholesky decomposition algorithm on a $5b \times 5b$ matrix using a 2×2 processor grid, where b is the block size. The matrix is symmetric and therefore only the lower triangular part of it will decomposed. The gray areas in the figure denote the parts of the matrix that have already been decomposed. Symbols (i), (ii) and (iii) refer to the corresponding phases of the Algorithm 3.2.

We can rewrite the equations of (3.13) as

$$\begin{aligned} \mathbf{V}_{11} &= \mathbf{U}_{11}^{-1} \\ \mathbf{V}_{22} &= \mathbf{U}_{22}^{-1} \\ \mathbf{V}_{21} &= -\mathbf{U}_{11}^{-1}\mathbf{U}_{21}\mathbf{U}_{22}^{-1}. \end{aligned} \tag{3.14}$$

The parallel algorithm calculating the triangular matrix inversion is similar to the one calculating the Cholesky decomposition. The phases (i) and (ii) in the algorithm refer to lines one and three in (3.14). Figure 3.3 demonstrates the use of the algorithm.

Algorithm 3.3 (Parallel triangular matrix inversion) *The algorithm inverts a given upper triangular matrix* $\mathbf{U}$ *in parallel. The parameter* b *denotes the block size. The function* $inverse(\mathbf{U}_{[ii]})$ *calculates the inverse of* $\mathbf{U}_{[ii]}$ *without parallelism.*

```
function parallelTriangularInverse(U, b)
m ← ⌈rank(U)/b⌉
for i ← 1 to m do
    U[ii] ← inverse(U[ii]) % phase (i)
od
for i ← m to 1 do
    for j ← i + 1 to m pardo % phase (ii)
        for k ← i + 1 to m pardo
            evalBCD(j, k, "T[jk] ← U[ik]U[kj]")
    for j ← i + 1 to m U[ij] ← 0
    for j ← i + 1 to m pardo
        evalBCD(i, j, " for k ← i + 1 to m do U[ij] ← U[ij] + T[kj]")
    for j ← i + 1 to m pardo
        evalBCD(i, j, "U[ij] ← U[ii]U[ij]")
od
return(U)
```

3.3.3 Parallel GCV Algorithm

The Algorithm 3.4 calculates the GCV function G_λ of Theorem 3.3 in parallel. The algorithm utilizes the parallel Cholesky decomposition and parallel triangular matrix inversion presented earlier. A parallel matrix multiplication is utilized in all matrix multiplications (see [DW97]). The number of floating point operations required per processor depends on the number of processors and the block size. For complexity analysis of ScaLAPACK

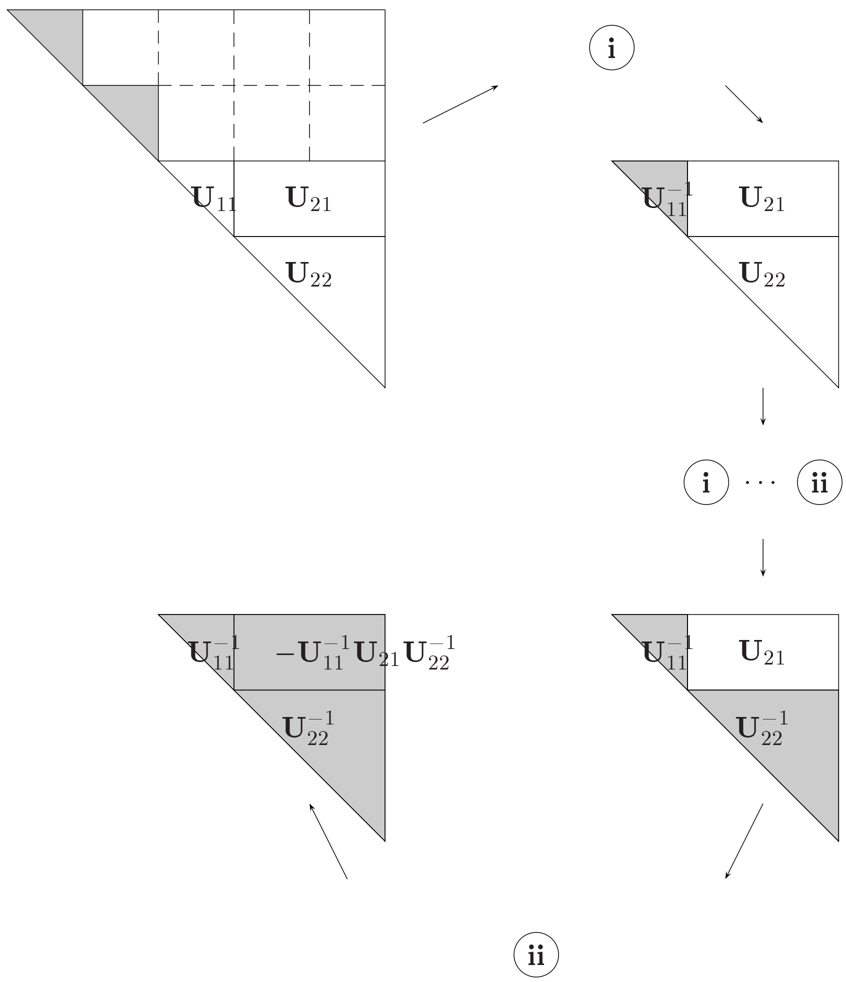


Figure 3.3: Demonstration of the parallel triangular matrix inversion algorithm on a $5b \times 5b$ matrix using a 2×2 processor grid, where b is the block size. Symbols (i) and (ii) refer to the corresponding phases of the Algorithm 3.3. The gray areas denote the parts of the matrix that have already been inverted.

routines, see e.g., [Dac99].

Algorithm 3.4 (Parallel GCV) *The algorithm calculates the GCV function* (3.8) *in parallel. The parameter b is the block size and the other parameters as in* (3.8)*.*

> **function** $parallelGCV(\mathbf{Q}_\lambda, \mathbf{F}_2, \mathbf{y}, b)$
> $n \leftarrow rank(\mathbf{Q}_\lambda)$
> % Calculate nominator
> $\mathbf{U} \leftarrow parallelTriangularInverse(parallelCholesky(\mathbf{F}_2^T\mathbf{Q}_\lambda\mathbf{F}_2 + n\mathbf{I}, b), b)$
> $\mathbf{C} \leftarrow \mathbf{U}\mathbf{F}_2^T\mathbf{y}$
> $\mathbf{C} \leftarrow \mathbf{U}^T\mathbf{C}$
> $\mathbf{C} \leftarrow \mathbf{F}_2\mathbf{C}$
> % Calculate denominator
> $d \leftarrow 0$
> **for** $i \leftarrow 1$ **to** n **do**
>   **for** $j \leftarrow i$ **to** n **do**
> $d \leftarrow d + u_{ij}^2$
> $return(n \cdot (norm(\mathbf{C})^2/d^2))$

3.4 Comparison of GCV and Parallel GCV

We performed two experiments to compare parallel GCV with ordinary GCV. The first experiment was performed with 4 processors and the second one with 16 processors. The runs were performed in a multiuser SGI mainframe maintained by the Finnish Center for Scientific Computing. The mainframe uses R8000 processors, which have a theoretical floating point rate of 300 Mflops.

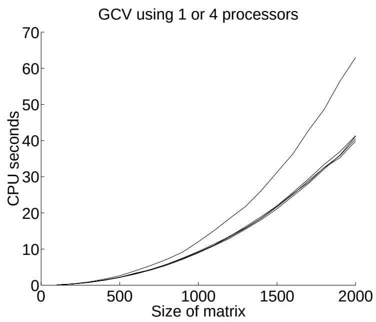
We used the GCV function

$$G_\lambda = n\frac{\left\|\mathbf{F}_2(\mathbf{U}_\lambda^{-1})^T\mathbf{U}_\lambda^{-1}\mathbf{F}_2^T\mathbf{y}\right\|^2}{(\sum_{i=1}^n\sum_{j=i}^n u_{ij}^2)^2},$$

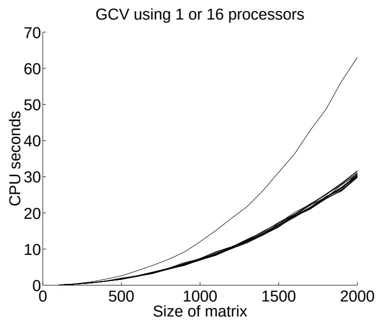
where

$$\mathbf{U}_\lambda^T\mathbf{U}_\lambda = \mathbf{F}_2^T\mathbf{Q}_\lambda\mathbf{F}_2 + n\mathbf{I}.$$

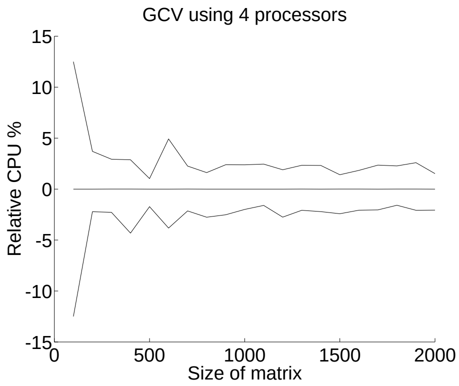
The matrix $\mathbf{F}_2$ was a random orthogonal matrix and $\mathbf{Q}_\lambda$ a random positive definite matrix. The runs were performed with the $\mathbf{Q}_\lambda$ and $\mathbf{F}_2$ of size $n \times n$, where $n = 100, 200, \ldots, 2000$. The block size was 50 in all experiments. Figure 3.4 shows the result of the comparisons. The calculation is distributed quite evenly between the processors. GCV with 16 processors was twice as fast as ordinary GCV for the matrix of size 2000×2000. The advantage of parallelism increases with the size of the matrices. Parallel GCV is not very cost efficient on sample sizes used in this study.



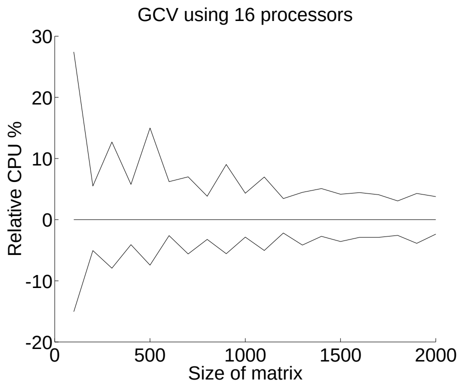
(a) Total CPU time per processor, 1 vs. 4 processors.



(b) Total CPU time per processor, 1 vs. 16 processors.



(c) Minimum and maximum CPU time per processor, 4 processors.



(d) Minimum and maximum CPU time per processor, 16 processors.

Figure 3.4: Parallel calculation of GCV. In Figures (a) and (b) the upper curve is the cost of the fast GCV with one processor. The lower curves are the costs of processors calculating the parallel GCV. In Figures (c) and (d) the lower curve is the minimum of the processors CPU time divided by the average CPU time and the upper curve is the corresponding maximum.

3.5 Minimization of GCV Function

Any standard minimization algorithm, such as gradient descending, conjugate gradient descending or BFGS algorithm can be used to minimize the GCV function G_λ of (3.8). See [PTVF92, Chapter 10] for possible minimization algorithms. In each iteration step, the minimization algorithm selects a new set of parameters λ and calculates the corresponding G_λ. The procedure is repeated as long as the value of G_λ decreases. Guidelines on choosing good initial values for λ can be found in [Wah90]. We have used a stochastic gradient descending in all experiments.

3.6 Further Reading

The use of GCV with non-Gaussian data has been studied in [XW94, Xia96, Wan95a]. A modified version of GCV is required if there is a correlation in the random errors of the training data [Wan96], or if the data is combined from multiple sources [Gao93].

The bootstrap, Jackknife, cross validation and generalization of maximum likelihood (GML) are generalization methods similar to the GCV. Wang has compared the bootstrap CI with the Bayesian CI in [WW94]. The study suggests that the methods give similar results on average size samples, but on a small sample size the bootstrap CI appears to better than the Bayesian CI. The disadvantage of the bootstrap CIs is that computing them is very processor intensive. See [Efr83, Efr81] for the bootstrap CI. Efron has compared the bootstrap, Jackknife and CV in [EG83]. The experiments suggest that all three methods give similar results. Wahba has studied the GCV and GML in [Wah85]. The GML appears to have a modest edge over the GCV in some experiments although mixed results were obtained in other experiments.

D'Azevedo and Dongarra [DD97] and Choi et al. [CDO$^+$94] have reported good performance results in parallel Cholesky decomposition when the rank of the matrices is $5000-30000$. Dackland has studied the use of the parallel Cholesky decomposition with the ScaLAPACK libraries in [Dac99].

Chapter 4

Model Space Selection

Data mining is a technology for discovering useful information from large collections of data. The discovered information can be, for example, rules describing properties of the data, frequently occurring patterns or clustering of the objects in the data [Man97]. In this chapter, we amend RK splines with an automatic model space selection in order to improve the data mining properties of the method. In Section 2.3, we used RK splines with a predetermined model space. Several parametric and semiparametric methods, such as neural networks and genetic algorithms, have techniques for model selection. However, there are only a few model selection algorithms that can be used in nonparametric estimation. For RK splines there are no automatic model space selection algorithms that would utilize either smooth or non-additive elementary spaces (that is spaces spanned by the kernel functions of more than one variable). See Section 4.6 for currently used model space selection algorithms. In this chapter, we introduce a model space selection algorithm that utilizes both smooth and non-additive elementary spaces.

Definition 4.1 (Model Space) *We call a set of elementary spaces (Definition 2.2) of $\mathcal{H}$ a* model space*.*

The two extremes of the model space are the *constant model space* $\{1\}$ and the *full model space* $\mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ contains all nonsmooth elementary spaces and $\mathcal{H}_1$ all smooth elementary spaces.

There are two major motivations for automatic model space selection. The full model space of a high dimensional input space is complex and model space selection is needed to eliminate the unnecessary elementary spaces to alleviate the computational complexity. The second motivation is related to the interpretation of results. A selection procedure that can discard unnecessary elementary spaces and variables from the model space is of crucial help in interpreting the data.

The independent variables usually correlate with each other, which makes model space selection a challenging task. The number of elementary spaces in $\mathcal{H}$ is exponential with respect to the number of variables in data. Therefore, we cannot compare in practice all possible model spaces with each other, rather we need to use some heuristic methods in model space selection. For smooth elementary spaces, that is elementary spaces spanning $\mathcal{H}_1$, we also have to determine the smoothing parameters λ.

The simplest model space selection strategy is the manual one. We can select new candidate elementary spaces one by one from the full model space and test whether they should be included in the current model space. A new elementary space is included in the current model space if it decreases the error. The manual model space selection is impractical in the analysis of medical data. In automatic model space selection, stepwise techniques are often used. In *backward selection* the initial model space is often the full model space and in *forward selection* the constant model space.

4.1 Arrangement of Elementary Spaces

Consider a general RKHS

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1,$$

of (2.14) where $\mathcal{H}_0$ is spanned by the nonsmooth RK functions and $\mathcal{H}_1$ by the smooth RK functions. The result of model space selection is a model space that is a subspace of $\mathcal{H}$. It is possible to arrange the elementary spaces of $\mathcal{H}$ in such a way that the most potential candidates are likely to be tested in early stages of model space selection.

Let k be the number of variables in data. By reordering the elementary spaces of $\mathcal{H}$ we may express it as

$$\mathcal{H} = \{1\} \oplus \bigoplus_{i_1=1}^{k} \mathcal{H}^{i_1} \oplus \bigoplus_{1 \le i_1 < i_2 \le k} \mathcal{H}^{i_1,i_2} \oplus \cdots \oplus \bigoplus_{1 \le i_1 < i_2 \cdots i_k \le k} \mathcal{H}^{i_1,i_2,\cdots,i_k},$$

where $\mathcal{H}^{i_1,i_2,\cdots,i_r}$, $r \le k$, is the space of all r-ary functions of variables $x_{i_1}, x_{i_2}, \ldots, x_{i_r}$ in $\mathcal{H}$. The constant space $\{1\}$ denotes the space of constant functions.

Example 4.1 (Decomposition of $W_2^2[0,1]$) *Consider Example 2.4 in which we wrote the RKHS $W_2^2[0,1]$ in the form $\mathcal{H}_0^{a,b} \oplus \mathcal{H}_1^{a,b}$, where*

$$\begin{aligned}\mathcal{H}_0^{a,b} &= \{1\} \oplus \{h_1^a\} \oplus \{h_1^b\} \oplus [\{h_1^a\}\otimes\{h_1^b\}] \\ \mathcal{H}_1^{a,b} &= \mathcal{H}_1^a \oplus \mathcal{H}_1^b \oplus [\{h_1^a\}\otimes\mathcal{H}_1^b] \oplus [\{h_1^b\}\otimes\mathcal{H}_1^a] \oplus [\mathcal{H}_1^a\otimes\mathcal{H}_1^b].\end{aligned}$$

We can now rewrite $W_2^2[0,1]$ as

$$\{1\} \oplus \mathcal{H}^a \oplus \mathcal{H}^b \oplus \mathcal{H}^{a,b},$$

where

$$\begin{aligned}\mathcal{H}^a &= \{h_1^a\} \oplus \mathcal{H}_1^a \\ \mathcal{H}^b &= \{h_1^b\} \oplus \mathcal{H}_1^b \\ \mathcal{H}^{a,b} &= [\{h_1^a\}\otimes\{h_1^b\}] \oplus [\{h_1^a\}\otimes\mathcal{H}_1^b] \oplus [\{h_1^b\}\otimes\mathcal{H}_1^a] \oplus [\mathcal{H}_1^a\otimes\mathcal{H}_1^b].\end{aligned}$$

Example 4.2 (Decomposition of $\mathcal{H}^{a,b}$) *Consider a general two-variable RKHS $\mathcal{H}^{a,b}$. It can be written in the form*

$$\mathcal{H}^{a,b} = \{1\} \oplus \mathcal{H}^a \oplus \mathcal{H}^b \oplus \mathcal{H}^{a,b},$$

where

$$\begin{aligned}\mathcal{H}^a &= \mathcal{H}_0^a \oplus \mathcal{H}_1^a \\ \mathcal{H}^b &= \mathcal{H}_0^b \oplus \mathcal{H}_1^b \\ \mathcal{H}^{a,b} &= [\mathcal{H}_0^a\otimes\mathcal{H}_0^b] \oplus [\mathcal{H}_0^a\otimes\mathcal{H}_1^b] \oplus [\mathcal{H}_0^b\otimes\mathcal{H}_1^a] \oplus [\mathcal{H}_1^a\otimes\mathcal{H}_1^b].\end{aligned}$$

Below we will use abbreviation $\mathcal{H}_{x,y}^{a,b}$ for $\mathcal{H}_x^a\otimes\mathcal{H}_y^b$, where $x, y \in \{0,1\}$.

Definition 4.2 (Corresponding elementary spaces) *By the corresponding elementary spaces we mean the elementary spaces that have the RK functions of the same variables. For example, the corresponding smooth elementary spaces of $\mathcal{H}_{0,0}^{a,b} \in \mathcal{H}_0$ are $\mathcal{H}_{0,1}^{a,b}$, $\mathcal{H}_{1,0}^{a,b}$ and $\mathcal{H}_{1,1}^{a,b}$.*

4.1.1 Selection of Nonsmooth Elementary Spaces

We introduce two breadth-first search algorithms to select nonsmooth elementary spaces. The algorithms determine the order in which the elementary spaces of $\mathcal{H}_0$ are tested. The selection function, which we call *first-fit*, is a simple breadth-first search algorithm. The second algorithm, *best-fit*, is a greedy breadth-first search. It calculates correlations between the variables to determine the best order of the elementary spaces.

Algorithm 4.1 (Nonsmooth elementary space selection with first-fit) *The parameter $\mathcal{U}$ denotes the set of untested elementary spaces in $\mathcal{H}_0$. The function vars(h) determines the number of variables in the elementary space h. The dot in the argument list means that the argument is not used by the function.*

```
function nextNonSmoothSpace( · ,U)
min ← ∞
best ← ∅
for h ∈ U do
    if vars(h) < min
      min ← vars(h)
      best ← h
    fi
return(best)
```

Algorithm 4.2 (Nonsmooth elementary space selection with best-fit) *The parameter $\mathcal{V}$ denotes the validation data and $\mathcal{U}$ the set of untested elementary spaces in $\mathcal{H}_0$. The function corr(h,y) = $\{$ max corr(x,y) $|$ x is a variable of the elementary space h and the value of x is not missing$\}$ determines the maximum of the linear correlations between the independent variables of h and the dependent variable y of $\mathcal{V}$.*

```
function nextNonSmoothSpace(V,U)
min ← ∞
best ← ∅
for h ∈ U do
    if vars(h) < min
      min ← vars(h)
      best ← h
      max ← corr(h, y)
  else if (vars(h) = min) ∧ (corr(h, y) > max)
        max ← corr(h, y)
        best ← h
      fi
return(best)
```

It is not necessary to use validation data in elementary space selection. However, the use of the validation data usually improves the performance of the elementary space selection function. If the number of independent variables is large, the result of model space selection is likely to depend on the order in which the elementary spaces are included in the model space.

The best-fit is likely to be superior to the first-fit, because the elementary spaces whose variables correlate with the dependent variable are included in the model space earlier. A comparison of the best-fit and first-fit elementary space selections will be shown later in this chapter. In most of our experiments, the best-fit performs better than the first-fit.

4.1.2 Selection of Smooth Elementary spaces

We next describe the elementary space selection function that determines the order in which the elementary spaces of $\mathcal{H}_1$ are tested.

Algorithm 4.3 (Smooth elementary space selection) *The parameter $\mathcal{U}$ denotes the set of untested elementary spaces in $\mathcal{H}_1$, and $\mathcal{L} \in \mathcal{H}_0$ a elementary space limiting the search. The function corresponding$(h, \mathcal{L})$ determines whether the elementary spaces h and $\mathcal{L}$ are corresponding spaces according to Definition (4.2).*

```
function nextSmoothSpace(U, L)
best ← ∅
for h ∈ U do
    if corresponding(h, L)
        best ← h
    fi
return(best)
```

Note that for each elementary space of $\mathcal{H}_0 \setminus \{1\}$ there are one or more corresponding smooth elementary spaces in $\mathcal{H}_1$.

4.2 Comparison of Model Spaces

We now introduce the model test function that is used in both forward and backward selections to test whether a elementary space should be included in the current model space or not. The purpose of the model test function is to determine how well the current model space describes the validation data.

Algorithm 4.4 (Model test function) *The parameter $\mathcal{T}$ denotes the training data, $\mathcal{V}$ the validation data and $\mathcal{S}$ the model space to be tested.*

> **function** $testSpace(\mathcal{T}, \mathcal{V}, \mathcal{S})$
> *Calculate* λ *by GCV (3.8) using* $\mathcal{T}$ and $\mathcal{S}$
> *Calculate* $\mathbf{c}$ *of (2.20) using* $\mathcal{T}, \mathcal{S}$ *and* λ
> *Calculate* $\mathbf{d}$ *of (2.21) using* $\mathcal{T}, \mathcal{S}$ *and* λ
> $err \leftarrow 0$
> **for** $(\mathbf{x}, y) \in \mathcal{V}$ **do**
> $t \leftarrow f(\mathbf{x})$ *of (2.16) using* $\mathcal{S}, \lambda, \mathbf{c}$ *and* $\mathbf{d}$
> $err \leftarrow err + (y - t)^2$
> $return(err/|\mathcal{V}|)$

In the medical data analysis, there is often a practical concern on the importance of the individual elementary spaces. We use the relative importance of the elementary spaces to compare the first-fit and the best-fit model space selection algorithms.

Definition 4.3 (Relative importance of elementary spaces) *Let n be the number of elementary spaces in the RK spline model space $\mathcal{S}$. The relative importance $g(i)$ of the elementary spaces $i = 1, \ldots, n$ with respect to the training data $\mathcal{T}$ and the validation data $\mathcal{V}$ is given by*

$$g(i) = \frac{r(i)}{\sum_{i=1}^{n} r(i)}$$

where

$$r(i) = \frac{testSpace(\mathcal{T}, \mathcal{V}, \mathcal{S})}{testSpace(\mathcal{T}, \mathcal{V}, \mathcal{S} \setminus i)}.$$

4.3 Stepwise Model Space Selection

In this section, we present a heuristic algorithm for model space selection with RK splines. The procedure has two phases, forward and backward selection, which are executed in sequence.

4.3.1 Forward Selection

The forward selection step starts with the constant model space $\{1\}$. The current model space is updated by adding the most potential untested elementary space of $\mathcal{H}_0$ to the current model space. The "most potential elementary space" is determined by the heuristic elementary space selection function described in Section 4.1.1.

Algorithm 4.5 (Stepwise forward selection) *The parameter $\mathcal{T}$ denotes the training data, $\mathcal{V}$ the validation data, $\mathcal{H}$ the full model space and ϵ a small positive constant.*

$\underline{\textbf{function}}$ $forwardSelection(\mathcal{T}, \mathcal{V}, \mathcal{H}, \epsilon)$
$\mathcal{S} \leftarrow \{1\}$ % constant space
$\mathcal{U} \leftarrow \mathcal{H} \setminus \{1\}$
$\underline{\textbf{while}}$ $\mathcal{U} \neq \emptyset$ $\underline{\textbf{do}}$ % while there are untested elementary spaces
  $\mathcal{C} \leftarrow nextNonSmoothSpace(\mathcal{V}, \mathcal{U})$
  $\mathcal{C}_0 \leftarrow \mathcal{C}$
  $prune \leftarrow true$
  $\mathcal{P} \leftarrow \emptyset$
  $\underline{\textbf{while}}$ $\mathcal{C} \neq \emptyset$ $\underline{\textbf{do}}$ % while there are spaces corresponding to $\mathcal{C}_0$
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{C}$
    $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{C}$
    $\underline{\textbf{if}}$ $testSpace(\mathcal{T}, \mathcal{V}, \mathcal{S} \cup \mathcal{C}) < (1 - \epsilon) * testSpace(\mathcal{T}, \mathcal{V}, \mathcal{S})$
     $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{C}$ % found a new elementary space
     $prune \leftarrow false$
    $\underline{\textbf{fi}}$
    $\mathcal{C} \leftarrow nextSmoothSpace(\mathcal{U}, \mathcal{C}_0)$
  $\underline{\textbf{od}}$
  $\underline{\textbf{if}}$ $prune$ % prune more general elementary spaces
   $\underline{\textbf{for}}$ $p \in \mathcal{P}$ $\underline{\textbf{do}}$
    $\underline{\textbf{for}}$ $u \in \{\mathcal{X} \mid p \otimes \mathcal{X} \in \mathcal{U}\}$ $\underline{\textbf{do}}$
     $\mathcal{U} \leftarrow \mathcal{U} \setminus u$
    $\underline{\textbf{od}}$
   $\underline{\textbf{od}}$
  $\underline{\textbf{fi}}$
$\underline{\textbf{od}}$
$return(\mathcal{S})$

In our experiments, we have used $\epsilon = 0.001$, which means that a new elementary space is included in the current model space if it decreases the previous error at least by 0.1%.

The operation of the forward selection algorithm is simple. First it adds a new nonsmooth elementary space to the current model space. If this decreases the previous error, the new elementary space will be accepted to the current model space, otherwise it will be discarded. After this, all corresponding smooth elementary spaces are tested one by one and they are included in the current model space if they decrease the error. The number of corresponding smooth elementary spaces is usually low. Note that smooth elementary spaces can be included in the model space also without the corresponding nonsmooth elementary spaces. If neither the tested nonsmooth nor at least one of the corresponding smooth elementary spaces was included in the current model space, the more general elementary spaces are pruned from the set of untested elementary spaces. The procedure is repeated until the set of untested elementary spaces is empty.

4.3.2 Backward Selection

The next step in the model space construction is backward selection. Forward selection is likely to include unnecessary elementary spaces in the model space. Backward selection is used to simplify the model space. Backward selection does not improve the performance of the model space on unseen data but it discards unnecessary elementary spaces from it. The procedure should not be mixed up with postpruning that is often used in machine learning for generalization. In RK spline training, the GCV function (3.8) prevents overfitting.

Algorithm 4.6 (Stepwise backward selection) *The parameter $\mathcal{T}$ denotes the training data, $\mathcal{V}$ the validation data, $\mathcal{S}$ the initial model space and ϵ a small positive constant. We use the notation $\mathcal{S} = (\mathcal{S}^1, \dots, \mathcal{S}^{|\mathcal{S}|})$.*

function $backwardSelection(\mathcal{T}, \mathcal{V}, \mathcal{S}, \epsilon)$
$\mathcal{S} \leftarrow \mathcal{S} \setminus \{1\}$
while $\mathcal{S} \neq \emptyset$ **do**
    $k \leftarrow 1$
    for $i \leftarrow 2$ **to** $|\mathcal{S}|$ **do** % find the least important elementary space
        if $testSpace(\mathcal{T}, \mathcal{V}, \mathcal{S} \setminus \mathcal{S}^i) < testSpace(\mathcal{T}, \mathcal{V}, \mathcal{S} \setminus \mathcal{S}^k)$
            $k \leftarrow i$
        fi
    od
    if $testSpace(\mathcal{T}, \mathcal{V}, \mathcal{S} \setminus \mathcal{S}^k) < (1 + \epsilon) * testSpace(\mathcal{T}, \mathcal{V}, \mathcal{S})$
      then
            $\mathcal{S} \leftarrow \mathcal{S} \setminus \mathcal{S}^k$ % delete the least important elem. space
      else
            break
    fi
od
$return(\mathcal{S} \cup \{1\})$

We use the model space calculated by forward selection as the initial model space of backward selection. Before removing any elementary spaces from the initial model space, we test the relative importance of all elementary spaces. If the importance of the least important elementary space is less than some small positive constant ϵ, the elementary space is removed from the current model space. The procedure is repeated until the importance of all elementary spaces is greater than ϵ. We have used in our experiments the same ϵ as in forward selection, which means that a elementary space is excluded from the current model space if it decreases the error less than 0.1%.

4.4 Training and Test Algorithms

The next two algorithms summarize the RK spline training and test procedures presented piece by piece in chapters 2, 3 and 4. The training algorithm determines the model space, smoothing factors and coefficients of the RK spline for preselected training and validation sets. We have used Matlab 5.1 to implement the RK spline training and test algorithms.

Algorithm 4.7 (Training of RK splines) *The parameter $\mathcal{T}$ denotes the training data, $\mathcal{V}$ the validation data and ϵ a small positive constant, which determines the accuracy of automatic model space selection.*

> <u>**function**</u> $trainRK(\mathcal{T}, \mathcal{V}, \epsilon)$
> *Build $\mathcal{H}$ following the guidelines of Chapter 2*
> % Model space selection (Chapter 4)
> $\mathcal{S} \leftarrow$ *forwardSelection*$(\mathcal{T}, \mathcal{V}, \mathcal{H}, \epsilon)$
> $\mathcal{S} \leftarrow$ *backwardSelection*$(\mathcal{T}, \mathcal{V}, \mathcal{S}, \epsilon)$
> % GCV (Chapter 3)
> *Calculate λ from (3.8) using $\mathcal{T}$ and $\mathcal{S}$*
> % RK spline for fixed λ (Chapter 2)
> *Calculate $\mathbf{c}$ of (2.20) using $\mathcal{T}, \mathcal{S}$ and λ*
> *Calculate $\mathbf{d}$ of (2.21) using $\mathcal{T}, \mathcal{S}$ and λ*
> *return*$(\mathcal{S}, \lambda, \mathbf{c}, \mathbf{d})$

Model space selection, especially forward selection, is computationally expensive operation. The purpose of ϵ is to limit search space in model space selection. A small ϵ forces forward selection to test more elementary spaces than a large ϵ. In our experiments, we have used a small ϵ $(= 0.001)$. In large data mining tasks a larger ϵ may have to be used because of computational aspects.

Once the model has been built, it can be used in prediction. The RK spline test algorithm estimates the target value for a given input pattern.

Algorithm 4.8 (Testing of RK splines) *The parameter $\mathbf{x}$ denotes the test pattern, $\mathcal{S}, \lambda, \mathbf{c}$ and $\mathbf{d}$ the parameters returned by the training algorithm.*

> <u>**function**</u> $testRK(\mathbf{x}, \mathcal{S}, \lambda, \mathbf{c}, \mathbf{d})$
> $y \leftarrow f(\mathbf{x})$ *of (2.16) using $\mathcal{S}, \lambda, \mathbf{c}$ and $\mathbf{d}$*
> *return*(y)

4.5 Model Space Selection Experiments

The purpose of the following experiments is to compare the first-fit and best-fit model space construction techniques on small data sets. We tested the model space selection methods on five data sets. Three of the data sets were artificial and two from real medical experiments. We used the RKHS $W_2[0,1]$ for continuous variables and $W_{\{K\}}$ for discrete variables. In all experiments $\epsilon = 0.001$.

We now describe the notation used in the model space selection experiments. The experiments with artificial data were performed on training and validation sets of 100 and 200 observations. We present the RK spline model spaces of the artificial data sets in the form $\mathcal{F} = \mathcal{C} \oplus \mathcal{E}$, where $\mathcal{C}$ denotes the *correct elementary spaces* and $\mathcal{E}$ the *incorrect elementary spaces*. The distinction between the correct and incorrect elementary spaces can be made because the data is generated using an artificial function.

We show the result of the experiments with artificial data using the following notation. Consider, for example, some RK spline model space $\mathcal{C} \oplus \mathcal{E}$, where

$$\begin{aligned} \mathcal{C} &= \underset{66.6\%}{x_2} \oplus \underset{33.3\%}{\mathcal{H}_1^{x_3}} \\ \mathcal{E} &= \underset{0.1\%}{\mathcal{H}_{1,1}^{x_2,x_3}}, \end{aligned}$$

and x_2 is the abbreviation for $\mathcal{H}_0^{x_2}$. For simplicity, we do not display the constant space $\{1\}$ because we always include it in $\mathcal{C}$. In this example, the model space selection algorithm has detected two correct elementary spaces and one incorrect elementary space. The incorrect elementary spaces are the ones that should not have been included in the model space. The subscript percentages show the relative importance (Definition 4.3) of the elementary spaces.

The model spaces of the medical data sets are presented as tables. The incorrect elementary spaces cannot be displayed because we do not know the original process that has generated the data. To alleviate the comparison of the selection methods we show the classification error for the validation data. In the model spaces of the medical data, the relative importance of the elementary spaces is displayed as table columns.

4.5.1 Modeling Noise

There is a risk that automatic model space selection will include noise terms in the RK spline model space. The next two experiments test the performance of the model space selection algorithms when modeling pure Gaussian noise. We used the function

$$f_0(\mathbf{x}) = 0 \cdot \sum_{i=1}^{10} x_i = 0.$$

The 10-dimensional input vectors were generated from the unit hypercube using uniform distribution. The corresponding target values were randomized from $f_0(\mathbf{x}) + \mathcal{N}(0, 1)$.

Experiment 4.1 (Noisy data, 100 observations) *The first-fit and best-fit produced the same RK spline model space*

$$\mathcal{E} = \underset{67.4\%}{x_2} \oplus \underset{32.7\%}{\mathcal{H}_1^{x_2}}. \tag{4.1}$$

Backward selection could not discard any elementary spaces from the model space.

Experiment 4.2 (Noisy data, 200 observations) *The first-fit and best-fit both generated the model space*

$$\mathcal{E} = \underset{29.6\%}{x_7} \oplus \underset{70.4\%}{\mathcal{H}_1^{x_7}}. \tag{4.2}$$

Backward selection discarded two incorrect elementary spaces from the model space.

The model space selection algorithms seem to be able to discard most of the Gaussian noise. However, backward selection was not successful because it discards elementary spaces based on their relative importance. In (4.1) and (4.2), the relative importance of all elementary spaces is greater than 29%. In true data there is usually at least one relatively noise-free variable and the above problem is unlikely to occur.

4.5.2 Modeling Additive Data

In the following experiments, we generated data using a true additive function. The purpose of the experiments was to test whether the model space selection algorithms are able to find the underlying model space and more importantly, do the selection algorithms include incorrect non-additive elementary spaces in the model space. We used the function

$$f_0(\mathbf{x}) = 0.1e^{4x_1} + 4/(1 + e^{-20(x_2 - 1/2)}) + 3x_3 + 2x_4 + x_5 + 0 \cdot \sum_{i=6}^{10} x_i$$

to generate the test data. Friedman has used the same function to test his MARS procedure [Fri91]. The additive function has two nonlinear terms, three linear terms and five noise terms. The $\mathbf{x}$-vectors were generated from the unit hypercube using the uniform distribution and the target values from $f_0(\mathbf{x}) + \mathcal{N}(0, 1)$.

Experiment 4.3 (Additive data, 100 observations) *The first-fit selection produced the model space*

$$\mathcal{C} = \underset{17.4\%}{x_1} \oplus \underset{47.2\%}{x_2} \oplus \underset{26.2\%}{x_3} \oplus \underset{4.0\%}{x_4}$$

$$\mathcal{E} = \underset{1.0\%}{\mathcal{H}_{1,0}^{x_3,x_{10}}} \oplus \underset{1.0\%}{\mathcal{H}_{0,1}^{x_4,x_{10}}} \oplus \underset{0.8\%}{x_6} \oplus \underset{1.4\%}{x_7} \oplus \underset{0.9\%}{\mathcal{H}_1^{x_7}} \oplus \underset{1.0\%}{x_{10}}$$

and the best-fit selection the model space

$$\mathcal{C} = \underset{17.0\%}{x_1} \oplus \underset{46.7\%}{x_2} \oplus \underset{26.3\%}{x_3} \oplus \underset{3.7\%}{x_4}$$

$$\mathcal{E} = \underset{1.1\%}{\mathcal{H}_{0,1}^{x_4,x_{10}}} \oplus \underset{0.5\%}{x_6} \oplus \underset{1.9\%}{x_7} \oplus \underset{1.1\%}{\mathcal{H}_1^{x_7}} \oplus \underset{1.6\%}{x_{10}} .$$

Backward selection discarded three incorrect elementary spaces from the first-fit model space and two from the best-fit model space.

The selection algorithms were unable to detect the nonlinearity of variables x_1 and x_2 and the algorithms did not include variable x_5 in the model spaces. Experiment 4.3 shows that the selection methods were unable to perform a successful model space selection on the additive data with 100 observations.

Experiment 4.4 (Additive data, 200 observations) *The first-fit and best-fit both generated the same model space*

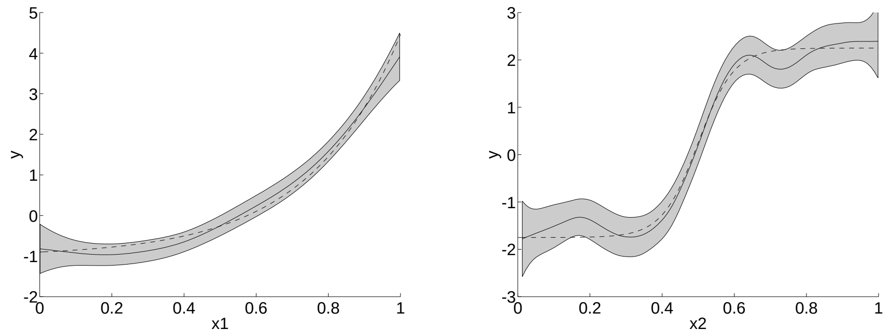
$$\mathcal{C} = \underset{17.1\%}{x_1} \oplus \underset{12.6\%}{\mathcal{H}_1^{x_1}} \oplus \underset{10.2\%}{x_2} \oplus \underset{14.8\%}{\mathcal{H}_1^{x_2}} \oplus \underset{30.2\%}{x_3} \oplus \underset{12.8\%}{x_4} \oplus \underset{2.5\%}{x_5} .$$

Backward selection discarded eight incorrect elementary spaces from the first-fit model space and five from the best-fit model space.

The model space has all the linear and nonlinear elementary spaces of the original function and no incorrect elementary spaces. Figure 4.1 shows the nonlinear elementary spaces in range $[0, 1]$. In RK spline model spaces, there is only one constant elementary space. Therefore the constants attached to particular terms of the original function cannot be extracted from the model. In Figure 4.1, the RK spline approximation is scaled with an artificial constant to obtain the same value domain as in the original function.

4.5.3 Modeling Non-Additive Data

In Experiment 4.4 our model space selection procedure did not detect false non-additive elementary spaces on additive data in which there were no



(a) Original function $0.1e^{4x_1}$ and its approximation $x_1 \oplus \mathcal{H}_1^{x_1}$.

(b) Original function $4/(1+e^{-20(x_2-1/2)})$ and its approximation $x_2 \oplus \mathcal{H}_1^{x_2}$.

Figure 4.1: Nonlinear elementary spaces of the additive data. The dotted line is the original function and solid line the RK spline approximation of it. The gray area is the pointwise 95% CI of the approximation.

interactions. Now it is time to focus on data that has interactions. We use the function

$$f_0(\mathbf{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - \frac{1}{2})^2 + 10x_4 + 5x_5 + 0 \cdot \sum_{i=6}^{10} x_i,$$

which has one nonlinear interaction of two variables, one nonlinear term, two linear terms and five noise terms [Fri91]. The $\mathbf{x}$-vectors were generated from the unit hypercube and the corresponding target values were $f_0(\mathbf{x})+\mathcal{N}(0,1)$.

Experiment 4.5 (Non-additive data, 100 observations) *The first-fit produced the model space*

$$\begin{aligned}
\mathcal{C} \;=\;& \underset{5.4\%}{x_1} \oplus \underset{2.1\%}{\mathcal{H}_1^{x_1}} \oplus \underset{1.0\%}{\mathcal{H}_{0,1}^{x_1,x_2}} \oplus \underset{1.0\%}{\mathcal{H}_{1,0}^{x_1,x_2}} \oplus \underset{3.9\%}{x_2} \oplus \\
& \underset{6.0\%}{\mathcal{H}_1^{x_2}} \oplus \underset{10.3\%}{\mathcal{H}_1^{x_3}} \oplus \underset{35.3\%}{x_4} \oplus \underset{16.7\%}{x_5} \\
\mathcal{E} \;=\;& \underset{1.0\%}{x_1x_3} \oplus \underset{1.0\%}{\mathcal{H}_{1,0}^{x_1,x_8}} \oplus \underset{1.0\%}{\mathcal{H}_{1,0,0}^{x_1,x_3,x_8}} \oplus \underset{1.0\%}{x_2x_3} \oplus \underset{1.1\%}{x_3} \oplus \\
& \underset{1.0\%}{\mathcal{H}_{1,1}^{x_3,x_6}} \oplus \underset{1.0\%}{\mathcal{H}_{0,1}^{x_3,x_8}} \oplus \underset{1.4\%}{x_6} \oplus \underset{1.4\%}{x_7} \oplus \underset{1.0\%}{x_7,x_8} \oplus \underset{0.8\%}{x_8}
\end{aligned}$$

and the best-fit generated the model space

$$\begin{aligned}
\mathcal{C} &= \underset{3.5\%}{x_1} \oplus \underset{1.5\%}{\mathcal{H}_1^{x_1}} \oplus \underset{5.9\%}{\mathcal{H}_{0,1}^{x_1,x_2}} \oplus \underset{2.1\%}{\mathcal{H}_{1,0}^{x_1,x_2}} \oplus \underset{2.8\%}{x_2} \oplus \\
&\quad \underset{8.6\%}{\mathcal{H}_1^{x_2}} \oplus \underset{15.7\%}{\mathcal{H}_1^{x_3}} \oplus \underset{40.0\%}{x_4} \oplus \underset{19.6\%}{x_5} \\
\mathcal{E} &= \underset{0.3\%}{x_1 x_3} \oplus \underset{0.1\%}{x_7} \, .
\end{aligned}$$

Backward selection discarded one incorrect elementary space from the first-fit model space and four from the best-fit model space.

There are 9 correct and 11 incorrect elementary spaces in the model space generated by the first-fit. Backward selection was unsuccessful in discarding the incorrect elementary spaces from the model space generated by forward selection. The best-fit generated the same correct elementary spaces as the first-fit, but only two incorrect elementary spaces. The reduction of the incorrect elementary spaces is remarkable considering that it was achieved just by reordering the elementary spaces of $\mathcal{H}_0$.

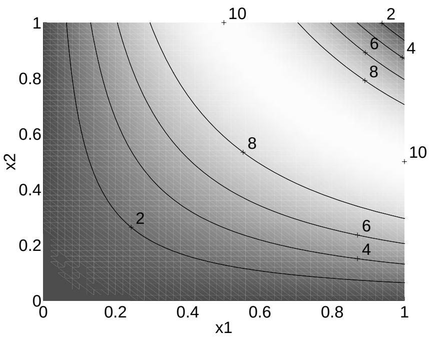
Experiment 4.6 (Non-additive data, 200 observations) *Both the first-fit and best-fit produced the model space*

$$\begin{aligned}
\mathcal{C} &= \underset{2.0\%}{x_1} \oplus \underset{11.7\%}{x_2} \oplus \underset{0.1\%}{x_1 x_2} \oplus \underset{4.5\%}{\mathcal{H}_1^{x_1}} \oplus \underset{7.7\%}{\mathcal{H}_1^{x_2}} \oplus \underset{2.1\%}{\mathcal{H}_{0,1}^{x_1,x_2}} \\
&\quad \oplus \underset{3.6\%}{\mathcal{H}_{1,0}^{x_1,x_2}} \oplus \underset{1.0\%}{\mathcal{H}_{1,1}^{x_1,x_2}} \oplus \underset{17.8\%}{\mathcal{H}_1^{x_3}} \oplus \underset{34.7\%}{x_4} \oplus \underset{14.8\%}{x_5} \, . \qquad (4.3)
\end{aligned}$$

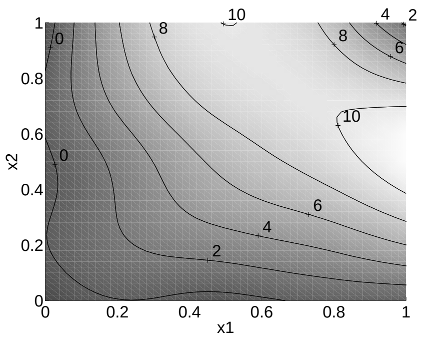
Backward selection discarded six incorrect elementary spaces from the first-fit model space. In the best-fit model space there were no incorrect elementary spaces after the forward selection.

Model space (4.3) has no incorrect elementary spaces. If we examine the first eight elementary spaces of the model space, we notice that there are all possible combinations of the smooth and nonsmooth elementary spaces of variables x_1 and x_2. The original function has the interaction $10\sin(\pi x_1 x_2)$. We have not found this complex interactions in our experiments with medical data. However, if such interactions exist, the model space selection algorithm has a potential to find the model space with the prerequisite that there is enough training data.

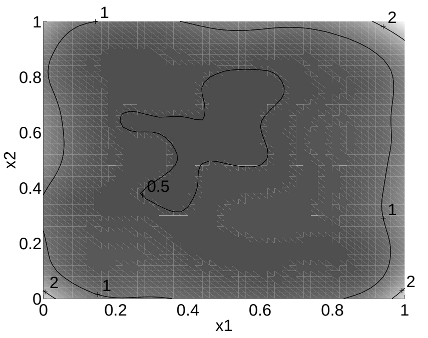
Figure 4.2 (a) shows a gray scale image of the original $10\sin(\pi x_1 x_2)$ and Figure 4.2 (b) its RK spline approximation, that has been trained with 200 observations. The RK spline approximation has been scaled with a constant



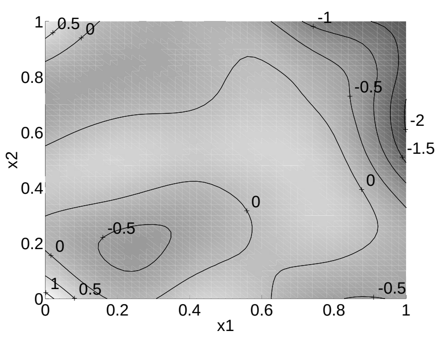
(a) Original function $10\sin(\pi x_1 x_2)$.



(b) RK spline approximation of $10\sin(\pi x_1 x_2)$.



(c) 95% CI of the approximation.



(d) Difference between the original function and its approximation.

Figure 4.2: The nonlinear interaction terms of the non-additive data. The model space was generated using 200 training and validation patterns.

to obtain the same domain range with the original function. In Figure 4.2 (c) there is the 95% CI of the approximation and in Figure 4.2 (d) the difference between the original function and the RK spline approximation. The original function lies inside the 95% CI of the approximation.

Figure 4.3 shows the original function of x_3 and its RK spline approximation. The linear regression of x_3 is a horizontal line, i.e., constant, which has been included in the constant elementary space of the model space. Therefore, there is no linear elementary space of x_3 in the RK spline model space (4.3).

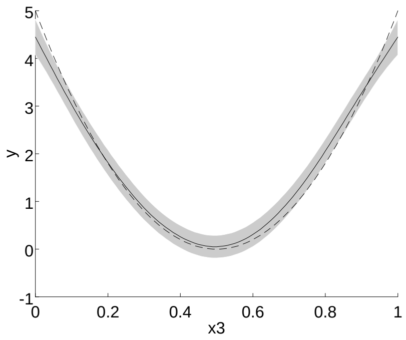


Figure 4.3: The nonlinear one-variable term of the non-additive data. The original function of x_3 is displayed as dotted line and its RK spline approximation as solid line. The gray area denotes the 95% CI of the approximation. The model was generated using 200 training and validation patterns.

4.5.4 Modeling Pima Indian Diabetes Data

We have demonstrated that the model space selection algorithms perform well on artificial data sets. The problem with any real life data is that we do not know the mathematical function of the phenomenon that has generated the data. Therefore, we cannot compare the RK spline model with the original function. To alleviate the comparison we show the classification error (% of incorrectly classified subjects) for the validation data.

In the following experiment, we use the Pima Indian diabetes data [MST94]. The data contains 768 observations, which are medical records of Pima Indian women at least 21 years of age. The problem with the data set is to predict whether a patient would test positive for diabetes given a number of physiological measurements and medical test results. See Appendix A.1 for a detailed data set description. We used 384 observations both in the training and validation sets. We assume that the relative proportion of the classes is approximately same in the training and validation set. An observation is classified as positive if the value of the regression function is greater than or equal to 0.5, otherwise it is classified as negative.

Experiment 4.7 (Pima Indian diabetes data) *Table 4.1 summarizes the result of model space selection.*

The both selection methods, first-fit and best-fit, result in 23% classification error on the validation set. Wahba et al. have studied the PIMA Indian diabetes data with RK splines in [WGWC93]. They suggest that from the total of eight variables of the data set, *times pregnant*, *plasma glucose*, *body mass index (BMI)* and *diabetes pedigree* are the ones that should be included in the RK spline model space. Table 4.1 shows the best of the four RK spline model spaces suggested in [WGWC93]. According to our analysis, *age* is one of the most important variables when predicting the diabetes of Pima Indians. Wahba et al. have used only nonsmooth elementary spaces in model space selection. Table 4.1 shows that *age* requires a smooth elementary space and its relative importance is 9% or 27% depending on the model space. In other respects, the results of the first-fit and best-fit model space selections are similar to the result of Wahba et al.

4.5.5 Modeling Heart Disease Data

The problem with the heart disease data is to predict the presence or absence of heart disease given the values of 13 attributes [MST94]. The data set contains 270 observations, from which 135 were used as training and 135 as validation patterns. See Appendix A.2 for a detailed data set description.

Experiment 4.8 (Heart disease data) *Table 4.2 summarizes the result of model space selection.*

Elementary space	first-fit	best-fit	Wahba
Age_1	27.0%	9.0%	-
Blood pressure_0	-	1.5%	-
BMI_0	7.4%	0.3%	29.0%
BMI_1	-	-	0.0%
$\text{Diabetes pedigree}_1$	7.6%	3.0%	-
Plasma glucose_0	-	18.6%	18.9%
Plasma glucose_1	1.4%	-	0.0%
Skin fold_0	0.3%	-	-
Serum insulin_0	-	1.5%	-
Times pregnant_0	13.5%	9.9%	6.6%
Times pregnant_1	-	-	0.0%
Blood pressure_0, Plasma glucose_0	-	11.8%	-
Blood pressure_0, Serum insulin_0	-	1.7%	-
BMI_0, Serum insulin_0	-	6.2%	-
BMI_0, Plasma glucose_0	-	-	2.5%
BMI_0, Plasma glucose_1	-	-	0.0%
BMI_1, Plasma glucose_0	-	-	0.0%
BMI_1, Plasma glucose_1	24.2%	36.5%	43.0%
Times pregnant_0, Skin fold_1	18.7%	-	-
Classification error of validation set	23.18%	22.92%	24.74%

Table 4.1: Pima Indian diabetes data. The percentages denote the relative importance of the elementary spaces (Definition 4.3). Subscript 0 denotes a nonsmooth and 1 a smooth elementary space.

Elementary space	first-fit	best-fit
Age_1	0.3%	-
Chest pain type_1	27.0%	10.8%
Exercise induced angina_1	1.6%	2.8%
Maximum heart rate_0	7.2%	1.3%
Number of major vessels_0	2.7%	15.8%
Number of major vessels_1	0.9%	8.3%
Old peak_0	0.8%	14.2%
Sex_0	2.4%	-
Sex_1	37.7%	11.7%
Slope of peak_1	5.1%	10.9%
Age_0, Maximum heart rate_0	2.7%	-
Age_0, Old peak_0	2.0%	-
Number of major vessels_1, Sex_1	5.8%	-
Number of major vessels_0, Serum cholesterol_0	0.6%	-
Number of major vessels_0, Thal_1	-	5.2%
Old peak_0, Serum cholesterol_0	3.4%	-
Old peak_0, Resting blood pressure_0	-	11.6%
Resting blood pressure_0,Slope of peak_1	-	7.6%
Classification error of validation set	14.81%	13.33%

Table 4.2: Heart Disease data. The percentages denote the relative importance of the elementary spaces (Definition 4.3). Subscript 0 denotes a nonsmooth elementary space and 1 a smooth elementary space.

The best-fit performed better than the first-fit. The first-fit included 15 elementary spaces in the model space and the best-fit only 11. Different model space construction techniques are likely to produce different model spaces when the number of variables is large. This is due to correlation between the independent variables. If there are several independent variables and few observations, best-fit selection should be used because it favors the elementary spaces whose variables have the greatest correlation between the dependent variable.

4.5.6 Conclusion of Modeling Experiments

We have performed six experiments on artificial data and two with real medical data to compare the first-fit and best-fit selection algorithms. Table 4.3 summarizes the numbers of correct and incorrect elementary spaces in the model spaces of the artificial data sets. In the experiments with 200 observations, the first-fit and best-fit both found all elementary spaces. However, there were significant differences in the experiments with 100 observations. The number of correct elementary spaces was independent of the selection algorithm, but the first-fit included more incorrect elementary spaces in the

Data	first-fit	best-fit
Noise 100	2	2
Noise 200	2	2
Additive 100	9	8
Additive 200	0	0
Non-additive 100	13	4
Non-additive 200	0	0

Table 4.3: Summary of the experiments with the artificial data sets. The columns of the table show the number of correct elementary spaces not found by the selection algorithm plus the number of incorrect elementary spaces.

model spaces than the best-fit. The difference was greatest in the model spaces of the non-additive data in which the first-fit included 11 incorrect elementary spaces and the best-fit only 2.

In the experiments with medical data, the best-fit selection gave smaller classification error on the validation data than the first-fit. We also found a new important variable (age) from Pima Indian diabetes data that had not been previously discovered with RK splines.

Comparison of the first-fit or best-fit model space selections does not justify a conclusion that one should be favored to another one. However, the noise reduction obtained suggests that the best-fit performs better when the number of observations is small. Therefore, in the experiments of Chapter 5 we use the best-fit selection to perform the RK spline model space construction.

4.6 Further Reading

Friedman and Silverman presented an additive model selection algorithm called TURBO in 1989 [FS89]. One year later, Hastie and Tibshirani published BRUTO, an improved version of TURBO [HT90]. BRUTO is based on adaptive backfitting and it uses a stepwise backward selection to delete unnecessary variables and base functions from the model. The strength of the method culminates on the error function the algorithm minimizes. It is similar to GCV, but the calculation requires only $O(n)$ operations, where n is the number of training patterns.

Friedman presented Multivariate Adaptive Regression Splines (MARS) in 1991 [Fri91]. The technique utilizes both backward and forward selection. Unlike TURBO or BRUTO, MARS is also able to determine the non-additive terms in model selection. Friedman has used a GCV like measure to test the performance of the models in [Fri91]. The method avoids the use of a separate validation set but requires a user-defined parameter to control the level of generalization.

Chen has studied a stepwise RK spline model space selection in [Che87]. Wahba et al. have used so called GLIM algorithm in RK spline model space selection in [WGWC93]. The GLIM algorithm, however, cannot utilize the non-additive elementary spaces.

Chapter 5

Applications with Medical Data

In this chapter, we use RK splines to refine medical data. Although RK splines require no user defined parameters, there are several other factors that should be taken into account when using the method. For example, outliers in the training data may influence the result of the approximation. The purpose of this chapter is to show how RK splines can be used to analyze medical data and how the results should be interpreted.

There have been only a few experiments utilizing RK splines in the analysis of medical data. Diabetic retinopathy has been studied in [WWG$^+$95a, WWG$^+$95b, Wan94, Wan95b, WWG$^+$93]. Wang has studied breast cancer in [Wan94] and Wang et al. circadian rhythms of the hormones ACTH and cortisol in [WGB97]. Gu has studied the time from HIV infection to AIDS diagnosis and survival time after heart transplant in [Gu97].

The chapter is divided into four sections, each dealing with a separate type of a medical problem. Section 5.1 focuses on low density lipoprotein (LDL) cholesterol estimation. The study has been made in co-operation with the Medical Informatics Research Centre in Turku (MIRCIT) headed by Jari Forsström at the time of writing. In Section 5.2, we investigate nutrient intake in a low-cholesterol diet. The study is a part of the STRIP baby project conducted by Olli Simell in the Turku University Central Hospital. Section 5.3 deals with the differences of the reference AFP and hCG between populations. The study is a part of the Down syndrome project of the MIRCIT. Finally, in Section 5.4 we study the influence of the conventional Pnc polysaccharide vaccine against childhood pneumonia and other Pnc diseases. Section 5.4 is prepared with Hanna Nohynek from the ARIVAC project of the National Public Health Institute in Finland and the Research Institute for Tropical Medicine in the Philippines. At the end of each section, there is a short conclusion describing the main results from the medical and methodological viewpoints.

We perform automatic model space selection whenever there is sufficient data. The original data set is divided into two subsets. The first subset, that is the training set, is used to train the RK spline model and the other subset, that is the validation set, to compare different model spaces during the forward and backward selections. If the number of observations is too low for automatic model space selection a full additive model space is assumed. The full additive model space is the most general model space in which there are no interactions of two or more variables. For example, a data set with the independent variables a and b is modeled using the RK spline model space

$$\mathcal{F} = a \oplus \mathcal{H}_1^a \oplus b \oplus \mathcal{H}_1^b$$

if there is not enough data for automatic model space selection.

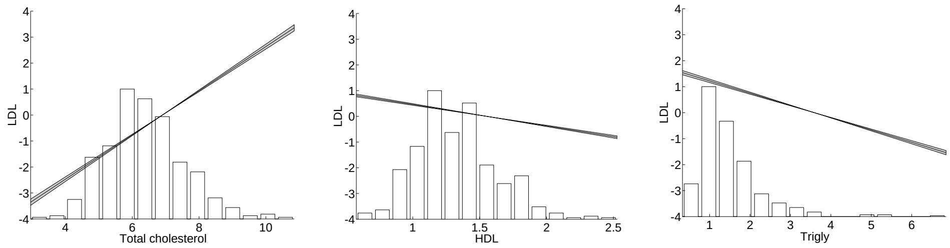
In Chapter 2 we justified the use of the MSE in function approximation. In this chapter, we use the MSE on the validation data to compare the performance of different methods. The validation data is not used to training RK splines, but only to select the RK spline model space. The use of the validation data in comparisons is justified because the backward selection algorithm of Chapter 4 does not decrease the MSE on the validation data. We use the RKHS $W_2[0,1]$ for continuous variables and $W_{\{K\}}$ for discrete variables. Below, we will use the word *model* to refer to RK spline model space.

5.1 Serum LDL Cholesterol

Several studies suggest that increased serum cholesterol is a major risk factor for coronary heart disease [FMH$^+$90, KSKG92, NW92]. *Low density lipoproteins* (LDL) and *high density lipoproteins* (HDL) are the two main types of lipoproteins that transport cholesterol. LDL is responsible for depositing cholesterol on the artery walls. HDL carries cholesterol away from body cells and tissues to the liver for removal from the body. Patients at the risk of coronary disease should be measured for serum LDL and HDL. Serum HDL is inexpensive to measure, but serum LDL needs to be separated by ultracentrifugation, which is not cost efficient in frequent use. Therefore, serum LDL concentration (mmol/l) is usually estimated using Friedewald formula

$$LDL = TC - HDL - Trigly/2.2, \tag{5.1}$$

where TC denotes *total cholesterol* and Trigly *triglycerides* [FLF72]. The formula is usually not used to estimate LDL of patients with Trigly greater than 4 mmol/l. In this section, we study if RK splines can estimate LDL more accurately than the Friedewald formula. The influence of Trigly is population dependent and therefore the divisor of Trigly in (5.1) may vary between the populations.



(a) TC (0.85 ± 0.031).

(b) HDL (-0.83 ± 0.055).

(c) Trigly (-0.47 ± 0.025).

Figure 5.1: Linear terms of the LDL estimation obtained with the RK spline model (5.2). Vertical bars represent histogram of the measured values of the variable. The numbers in parentheses show the coefficient of the term with 95% CI.

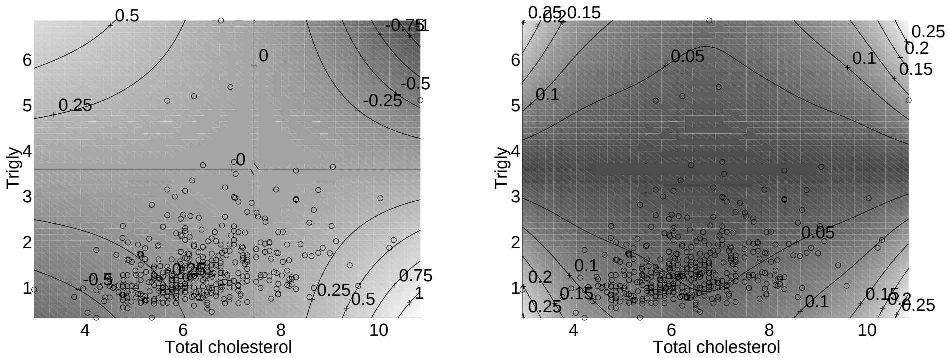
5.1.1 Comparison of LDL Models

The Friedewald data comprises 809 observations, which were randomized to a training set (405 observations) and a validation set (404 observations). Each subject was measured for TC, HDL, Trigly and LDL. There are no missing values in the data set. See Appendix A.3 for a detailed data set description.

Experiment 5.1 (LDL, full data) *In the first experiment, 405 training and 404 validation patterns were used. Automatic model selection generated the RK spline model*

$$LDL = TC \oplus HDL \oplus Trigly \oplus TC \cdot Trigly \oplus \mathcal{H}_{1,0}^{TC,Trigly}. \qquad (5.2)$$

In addition to linear TC, HDL and Trigly terms, which are present also in the original Friedewald formula, the model (5.2) has a linear and nonlinear interaction of TC and Trigly. Figure 5.1 shows the linear terms of the model with 95% CI. Figure 5.2 (a) shows the interaction of TC and Trigly, and Figure 5.2 (b) the corresponding 95% CI. Figure 5.3 shows the approximated LDL vs. observed LDL.



(a) Nonlinear interaction of TC and Trigly.

(b) The 95% CI of the interaction of TC and Trigly.

Figure 5.2: Interaction of the model (5.2). The figure shows the sum of the linear and nonlinear interaction.

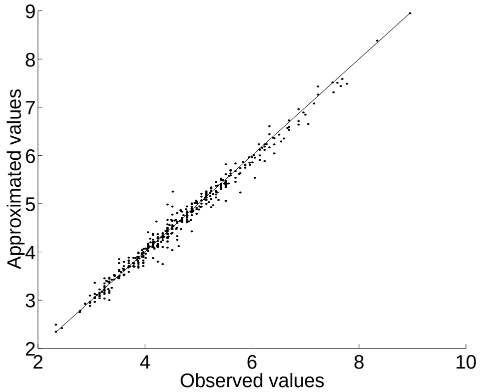


Figure 5.3: The observed LDL vs. approximated LDL using the RK spline model (5.2) and 404 validation patterns. The diagonal line is to help reader to compare the observed values with the estimated ones.

	MSE	MSE %
Linear regression	0.02059	–
RK spline model (5.2)	0.02021	−1.85
RK spline model (5.3)	0.02025	−1.65
Friedewald	0.05178	151.49
Friedewald with *Trigly*/2.7	0.02570	24.79

Table 5.1: Comparison of the models. The MSE was calculated using 398 validation patterns (Trigly $\leq$ 4 mmol/l). The MSE % column denotes the performance increase (-) or decrease (+) to the linear regression. The RK spline model (5.2) was trained with 405 patterns and (5.3) with 400 patterns (Trigly $\leq$ 4 mmol/l).

We performed a separate test using the same training and validation sets as in the Experiment 5.1, but excluding subjects with Trigly greater than 4 mmol/l. There were 400 such training and 398 validation patterns.

Experiment 5.2 (LDL, Trigly $\leq$ 4 mmol/l) *The model selection produced the model*

$$LDL = TC \oplus HDL \oplus Trigly \oplus TC \cdot Trigly, \tag{5.3}$$

which differs from the model (5.2) *only by the interaction term.*

Table 5.1 summarizes the result of the comparison of the linear regression, RK spline model (5.2), RK spline model (5.3), Friedewald formula, and population corrected Friedewald formula. Trigly is population dependent and therefore population correction improves the performance of the Friedewald formula. Population correction is performed by replacing the original divisor of Trigly in (5.1) with a population specific one. In our sample, divisor 2.7 minimizes the MSE of the Friedewald formula. The coefficients of the models were determined using the training data and corresponding MSEs were calculated using the validation data.

5.1.2 Conclusion of LDL Estimation

Medical conclusion: The performance of the original Friedewald formula and its population corrected version was poor. The best performance was obtained using the RK spline models (5.2) and (5.3). The model (5.3) has no smooth terms and therefore it is same as the linear regression of TC, HDL, Trigly and TC $\cdot$ Trigly. The term TC $\cdot$ Trigly is statistically very significant ($p < 0.000001$).

In daily clinical practice the original Friedewald formula is a good estimator of LDL on patients whose Trigly is less than 4 mmol/l. The population corrected version of the Friedewald formula should be favored in clinical research. In our data, the MSE decreased by 50% when the original

Friedewald formula was replaced by the population corrected one. The best performance in LDL estimation was obtained using RK splines. In addition to TC, HDL and Trigly, the interaction of TC and Trigly provided to be significant.

Methodological conclusion: The RK spline models (5.2) and (5.3) were more accurate than the other models used in LDL estimation. We did not detect nonlinear dependencies between LDL and dependent variables TC, HDL and Trigly on subjects with Trigly less than 4 mmol/l. This suggests that linear regression with the product of TC and Trigly can be used to estimate LDL. The results obtained with RK splines indicate that there is no reason to use nonlinear methods, such as neural networks, in LDL estimation because linear models are accurate enough.

5.2 Nutrients in Low-Cholesterol Diet

Clinical trials in adult populations have shown that lowering cholesterol levels reduces coronary heart disease (CHD) risk [Mac94b]. Because saturated fatty acids raise blood cholesterol levels, a major emphasis in food counseling is placed on reducing saturated fat intake. There are no long-term studies of the relationship of blood cholesterol levels measured in childhood to CHD in later life. The influence of diet on cholesterol has been studied, e.g., in [FMM+89, TDR92, HFK95, WJ97].

Total blood cholesterol levels in children vary geographically. In countries such as the Philippines, Italy and Ghana, saturated fat constitutes less than 10% dietary intake, and the serum cholesterol level in boys 8 to 9 years of age is generally below 4.14 mmol/l. In boys from countries such as the Netherlands, Finland and the United States, the saturated fat intake varies from 13.5% to 17.7% of daily energy intake, and serum cholesterol levels are generally above 4.14 mmol/l [Com98]. In this section, we use RK splines to study nutrient intake of children in a low-cholesterol diet.

5.2.1 STRIP project

The STRIP project aims at decreasing exposure of young children to known atherosclerosis risk factors [Lap95, LVJ+95]. The project was launched in 1990 and 1062 Finnish infants at the age of seven months were randomized to intervention and control groups. The parents of intervention infants received individualized dietary and life style counseling twice a year. Parents of the control group received the same counseling as currently given at well-baby clinics in Finland. The project will continue though the first 20 years of life.

The focus of this section is to study nutrient intake of children at age of seven years. In the STRIP project, this age group comprises 307 intervention and 323 control children. Families of the intervention group received dietary counseling at 1-3 month intervals during the first two years and twice a year after that. Families of the control group were seen twice a year from the beginning of the trial. At each visit, a pediatrician examined the child and evaluated the development of the child. A nutritionist examined the child's dietary history that was based on 3-4 day food records. Intervention families were encouraged to use low-fat milk products and to increase the use of vegetables, fruits and whole meat products. Controls were encouraged to use milk with at least 1.9% fat after breast-feeding phase. The food records were entered in a program, which calculated the child's daily nutrient intake.

5.2.2 Modeling Nutrient Intake

We use the abbreviation SAFA for *saturated fatty acids*, PUFA for *poly-unsaturated fatty acids*, MUFA for *mono-unsaturated fatty acids*, P/S for ratio PUFA/SAFA, TC for *total cholesterol*, HDL for *high density lipoproteins*, LDL for *low density lipoproteins* and Trigly for *triglycerides*.

We used RK splines without automatic model selection to analyze the influence of fat, SAFA, PUFA, MUFA, P/S, carbohydrate and fiber to TC, HDL, LDL and Trigly. The ratio P/S is used as an input variable because it is generally known to correlate with lipid values. See Appendix A.4 for a detailed description of the STRIP data.

Experiment 5.3 (Influence of Nutrients on Lipids) *We used all 630 observations as training data and four RK spline models (one for each lipid)*

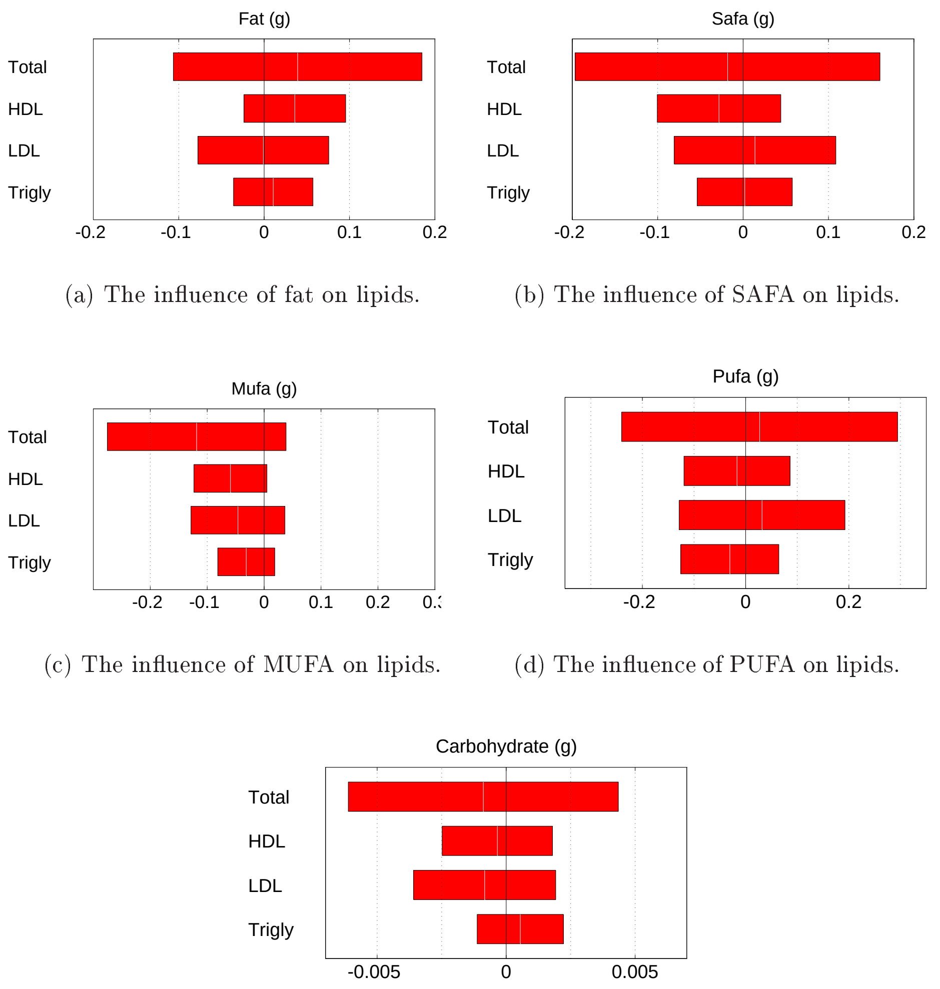
$$\begin{aligned} Lipid \quad = \quad & Fat \oplus \mathcal{H}_1^{Fat} \oplus \\ & SAFA \oplus \mathcal{H}_1^{SAFA} \oplus \\ & PUFA \oplus \mathcal{H}_1^{PUFA} \oplus \\ & MUFA \oplus \mathcal{H}_1^{MUFA} \oplus \\ & P/S \oplus \mathcal{H}_1^{P/S} \oplus \\ & Carbohydrate \oplus \mathcal{H}_1^{Carbohydrate} \oplus \\ & Fiber \oplus \mathcal{H}_1^{Fiber}, \end{aligned} \tag{5.4}$$

where Lipid is either TC, HDL, LDL or Trigly.

Low HDL and high TC, LDL and Trigly are risk factors of CHD. Therefore, we are looking for nutrients that increase HDL but decrease the other three lipid values. Figure 5.4 shows the effect of fat, SAFA, MUFA, PUFA and carbohydrate. Figures 5.5 (a)-(d) display the influence of P/S and Figures 5.5 (e)-(h) the influence of fiber.

5.2.3 Conclusion of Nutrient Study

Medical conclusion: The results obtained with RK splines suggest that MUFA decreases all lipids. Therefore possibly more emphasis in food counseling should be placed in the use of mono-unsaturated fatty acids. The minimum lipid values are achieved when the P/S ratio is close to 1/2. Usually food counseling aims at P/S ratio 1/1.
Methodological conclusion: There are only linear dependencies between the independent variables (excluding the P/S ratio) and the dependent variable in the STRIP data. Therefore, the performance of RK splines does not differ remarkably from linear regression. The only notable discovery with RK splines was the nonlinearity of the P/S ratio.



(a) The influence of fat on lipids.

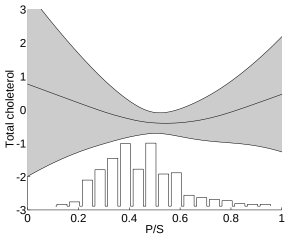
(b) The influence of SAFA on lipids.

(c) The influence of MUFA on lipids.

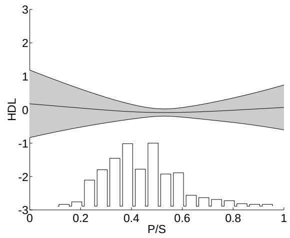
(d) The influence of PUFA on lipids.

(e) The influence of carbohydrate on lipids.

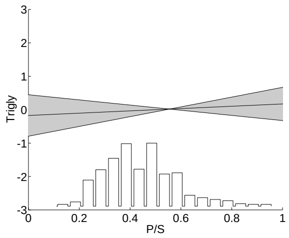
Figure 5.4: All fats and carbohydrate have a linear effect on lipids. The bars denote the 95% CI of the coefficients. A negative coefficient indicates that the corresponding lipid value decreases as a function of the nutrient and a positive coefficient indicates the opposite.



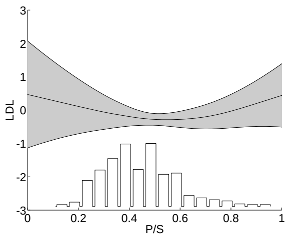
(a) The influence of P/S to TC.



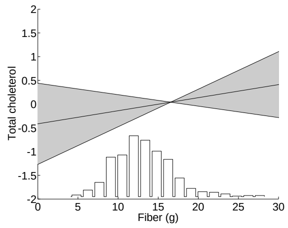
(b) The influence of P/S to HDL.



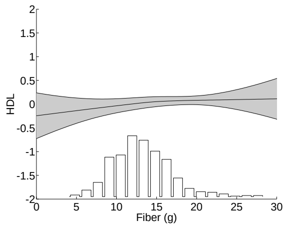
(c) The influence of P/S to Trigly.



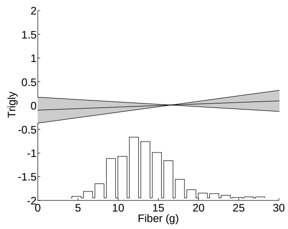
(d) The influence of P/S to LDL.



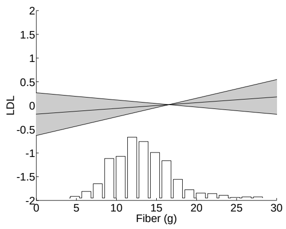
(e) The influence of fiber to TC.



(f) The influence of fiber to HDL.



(g) The influence of fiber to Trigly.



(h) The influence of fiber to LDL.

Figure 5.5: The influence of P/S and fiber on lipids. The P/S ratio and fiber are the only variables in the model (5.4) that have a nonlinear effect on lipids.

5.3 AFP and hCG in Down Syndrome (DS) Screening

Down syndrome (DS), or trisomy 21, is the most common serious chromosome aberration in which affected individuals survive beyond infancy. Maternal age is known to be associated with the DS risk and therefore in developed countries pregnant women over 35 years are routinely offered the opportunity of amniocentesis and caryotyping of fetal cells. However, two thirds of DS cases are born to mothers less than 35 years. Maternal serum markers are usually used to estimate the DS risk of women less than 35 years. Serum alpha-fetoprotein (AFP) was the first such marker identified. It has been shown to have an inverse correlation with the DS risk [LBJ95]. Another important marker is human chorion gonadotropin (hCG), which has a positive correlation with the DS [CJI$^+$97]. The other serum markers frequently used in the DS risk estimation are unconjugated estriol (uE3), free-alpha hCG, free-beta hCG and dimeric inhibin A [WKHM97]. In this study, we use only total hCG and AFP.

Serum hCG and AFP are usually measured at 14 to 16 weeks of pregnancy. Detected hormone values are compared to reference values and amniocentesis is performed on women whose detected values are not within safe interval. Wald et al. have reported that about 60% of DS cases can be detected this way when 5% false positive rate is used [WKHM97].

In the second trimester of pregnancy, AFP increases and hCG decreases. There are, however, several factors that may influence hCG and AFP levels. Therefore, the detected hormone values need to be adjusted to minimize the influence of disturbance before the values are compared to the reference values. Gestational age and maternal weight are the major factors influencing hCG and AFP. Conventionally, the detected hormone values are first linearly weight corrected and then compared with a reference table that has an entry for different gestational ages [HHSK98].

The main purpose of this section is to search differences in reference values of hCG and AFP between populations using RK splines. Muller et al. have reported that race influences hCG and AFP values [MBP$^+$94]. The second objective is to search for new factors that influence gestational hormones. For example, insulin-dependent diabetes [PKH94, WDCC94] and smoking [PKH$^+$93, BHSB$^+$93] affect hCG and AFP.

5.3.1 Comparison of Finnish Cities

The purpose of the following experiments is to study whether there are differences in reference hCG and AFP between Finnish cities Helsinki, Tampere, Turku and Oulu on healthy pregnant women. Serum hCG and AFP concentrations were available from women with singleton pregnancies without DS or neural tube defects and without insulin-dependent diabetes mellitus.

The variables *gestational age*, *maternal weight*, *number of previous births* and *smoking* were used to estimate hCG and AFP. The Helsinki data set comprises 487 observations, Tampere 314, Turku 1105, and Oulu 782. There are missing values in all data sets. In addition, the Helsinki data set has no hCG. See Appendix A.5 for a detailed description of the data sets.

Automatic model selection on the largest data set, that is the Turku data, produced the models

$$\begin{aligned} hCG &= \underset{5.5\%}{Weight} \oplus \underset{4.5\%}{Smoking} \oplus \underset{90.0\%}{\mathcal{H}_1^{Gestation}} \\ AFP &= \underset{23.0\%}{Weight} \oplus \underset{77.0\%}{\mathcal{H}_1^{Gestation}}. \end{aligned} \tag{5.5}$$

There are no interaction terms in the above models. The number of observations is low in the data sets and therefore we use full additive models without automatic model selection. The full additive models are more general than the models (5.5).

Experiment 5.4 (Finnish DS comparison) *In comparison of the Helsinki, Tampere, Turku and Oulu data we used the model*

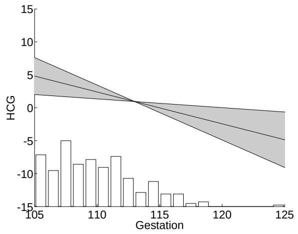
$$\begin{aligned} Marker = \; & Gestation \oplus \mathcal{H}_1^{Gestation} \oplus \\ & Weight \oplus \mathcal{H}_1^{Weight} \oplus \\ & Births \oplus \mathcal{H}_1^{Births} \oplus \\ & Smoking \oplus \mathcal{H}_1^{Smoking}, \end{aligned}$$

where Marker is either hCG or AFP.

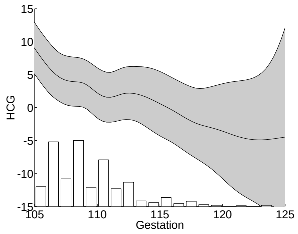
Gestational age. Gestational age is the major factor influencing hCG and AFP. Figures 5.6 (a)-(c) display hCG as function of gestational age. There are no significant differences in gestational age between the three hCG data sets.

Figures 5.7 (a)-(e) display AFP as function of gestational age. Figure 5.7 (b) shows that the variance of the error is not constant with respect to gestational age in the Helsinki data and therefore the observations with gestational age greater than 125 days were removed. Helsinki data has a nonlinear AFP, but the finding is not significant. The experiment shows that RK splines are sensitive to outliers in the data. In general, when a smoothing coefficient $\lambda_i \to 0$ there are usually outliers in the data.

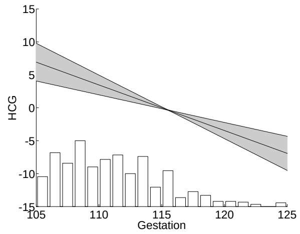
Maternal Weight. Figures 5.6 (d)-(f) show that Tampere and Oulu both have linear hCG models for maternal weight. Turku, however, has a nonlinear model. The Turku model suggests that for women over 100 kg, a nonlinear weight correction should be used instead of a linear one. In the Turku data, there were five women over 100 kg whereas in the Tampere and Oulu data there were only two such women in each. Comparison of the



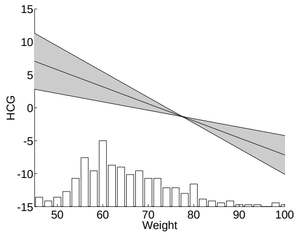
(a) Gestational age in Tampere.



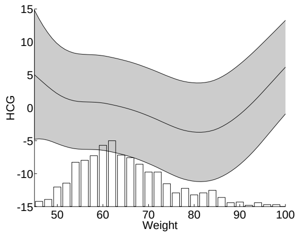
(b) Gestational age in Turku.



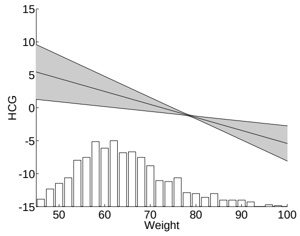
(c) Gestational age in Oulu.



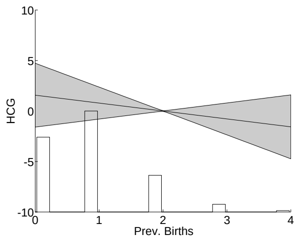
(d) Maternal weight in Tampere.



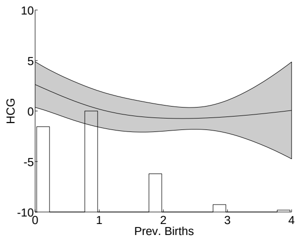
(e) Maternal weight in Turku.



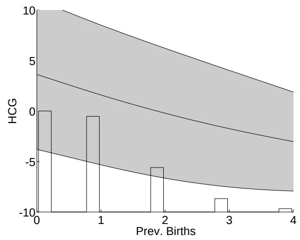
(f) Maternal weight in Oulu.



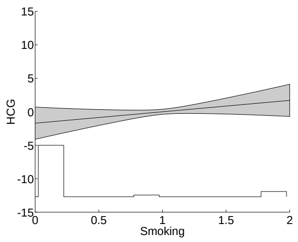
(g) Previous births in Tampere.



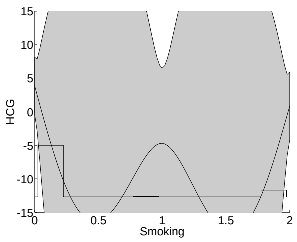
(h) Previous births in Turku.



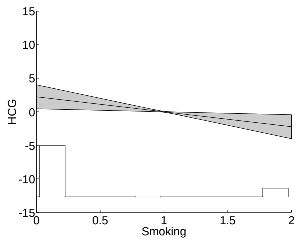
(i) Previous births in Oulu.



(j) Smoking in Tampere.

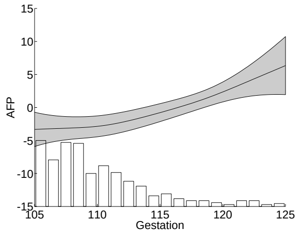


(k) Smoking in Turku.

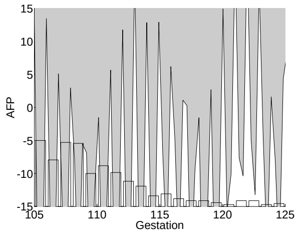


(l) Smoking in Oulu.

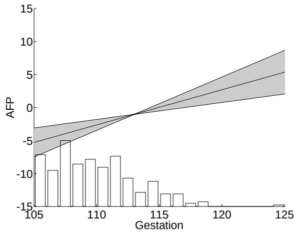
Figure 5.6: hCG in Finnish DS comparison.



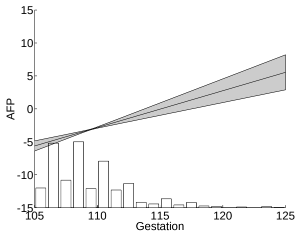
(a) Gestational age in Helsinki (Gestational age $\leq$ 125 days in training data).



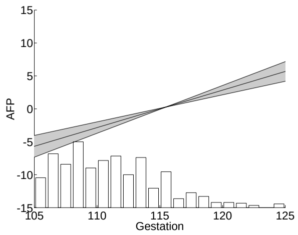
(b) Gestational age in Helsinki (full training data).



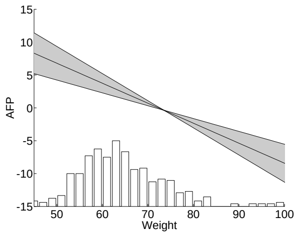
(c) Gestational age in Tampere.



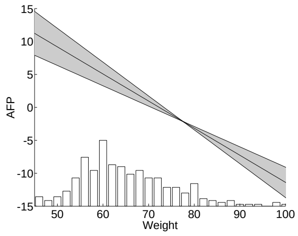
(d) Gestational age in Turku.



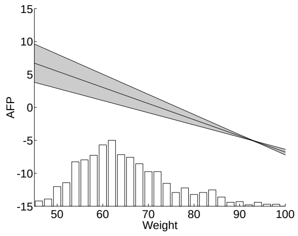
(e) Gestational age in Oulu.



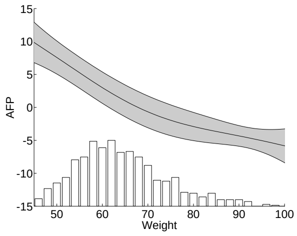
(f) Maternal weight in Helsinki.



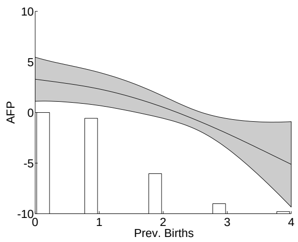
(g) Maternal weight in Tampere.



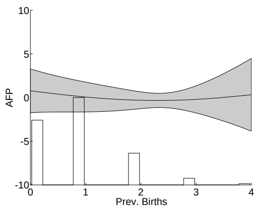
(h) Maternal weight in Turku.



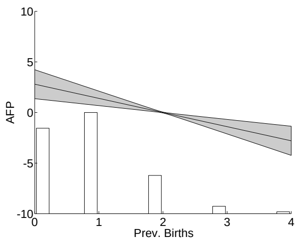
(i) Maternal weight in Oulu.



(j) Previous births in Helsinki.



(k) Previous births in Tampere.



(l) Previous births in Turku.

Figure 5.7: AFP in Finnish DS comparison.

Turku model with the Tampere and Oulu models is not justified for women over 100 kg because of the small amount of data.

The AFP models of maternal weight are similar in all four data sets. The greatest difference is between the Turku and Tampere data. All models are linear, except the Oulu model which has a weak nonlinear term. Figures 5.7 (f)-(i) show the AFP models.

Number of Previous Births. Wald and Watt have reported hCG levels to decrease by 3.1% per previous birth [WW96]. We could not detect the phenomenon as Figures 5.6 (g)-(i) show. All three experiments (Tampere, Turku and Oulu) indicate that hCG decreases per each previous birth, but the CI is too large to make any analytical conclusions.

In Helsinki and Turku AFP decreases gradually with the number of previous births. In Turku the decreases is 1.4 kU/l per previous birth (95% CI $[0.7, 2.1]$ kU/l). Figures 5.7 (j)-(l) show AFP as the function of previous births.

Smoking. Smoking has been reported to decrease hCG by $21 - 23\%$ [PKH+93, BHSB+93, BPL+89]. In our data sets, there were three smoking categories (0 = nonsmoking, 1 = smoked during the first trimester, 2 = smoking). The categories were modeled using a continuous model function. Figures 5.6 (j)-(l) show the influence of smoking on hCG. In Oulu, serum hCG decreases by 2.2 kU/l per smoking category (with 95% CI [0.4,4] kU/l). GCV tends to oversmooth insignificant terms, as can be seen in Figure 5.6 (k). Usually automatic model selection discards such terms, but in this analysis all one-variable terms were included in the RK spline model because there was not enough data to perform a separate model selection for each data set. We performed an unpaired t-test to nonsmoking (group 0) and smoking women (group 2). The average hCG of women who smoked was 12% less in Turku (p=0.11), 19% less in Oulu (p=0.042) and 13% greater in Tampere (p=0.35) compared with hCG of nonsmokers. The result confirms our findings using RK splines.

In Helsinki, Tampere and Turku, GCV failed to find a suitable hCG model due to oversmoothing. In Oulu, a statistically insignificant nonlinear hCG model was detected.

5.3.2 Comparison of Taiwan and the UK

In this section, the objective is to identify race-ethnicity-specific differences in hCG and AFP between Taiwanese and British women. In our data set, there were 14 003 Taiwanese and 20 232 British women. Each subject was measured for *maternal age*, *maternal weight*, *gestational age*, *hCG* and *AFP*. See Appendix A.5 for a detailed data set description.

We performed model selection separately on the Taiwanese and British data using 1000 training and 9000 validation patterns that were selected randomly from the original data sets.

Experiment 5.5 (Taiwanese and British hCG) *In hCG estimation, both the Taiwanese and British data produced the same RK spline model*

$$hCG = Weight \oplus Gestation \oplus \mathcal{H}_1^{Gestation}. \tag{5.6}$$

Maternal age was insignificant in the estimation of hCG and therefore the variable was excluded from the model. Maternal weight has a linear influence on hCG and gestational age a nonlinear influence. Figures 5.8 (a)-(d) show maternal weight and gestational age with the 95% CI. In maternal weight, there are some differences between the populations. The mean weight in Taiwan is 55 kg and in the UK 66 kg. The coefficient of the linear weight term in Taiwan is -0.21 ($[-0.14, -0.28]$ with 95% CI) and in the UK -0.12 ($[-0.08, -0.16]$).

There is a clear disparity between the populations in the models of gestational age. The UK has a nonlinear model that could apparently also be modeled with a second order parametric model. Taiwan has a complex nonlinear model, which at first sight seems to be due to oversmoothing. However, if we examine the frequency distribution of gestational age, we notice poles every seven days. There is a corresponding variation in the CI at seven days interval. The more there are training samples the narrower the CI is. Due to the cyclic frequency distribution of the training data, gestational age of Taiwan has an exceptional model. The true biological model of gestational age is probably close to the second order polynomial as the UK model suggest. However, the Taiwanese data is scattered in such a way that this conclusion is not justified from the training data. It is notable that the second order polynomial can still be fitted inside the 95% CI of gestational age in the Taiwanese model. RK splines with 1000 training patterns did not detect significant differences between the populations in hCG estimation.

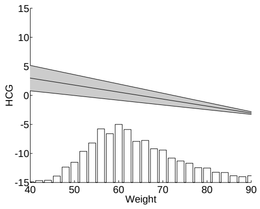
Figures 5.8 (a)-(d) show that the RK spline model (5.6) resembles the parametric model

$$hCG = a + b \cdot Weight + c \cdot Gestation + d \cdot Gestation^2. \tag{5.7}$$

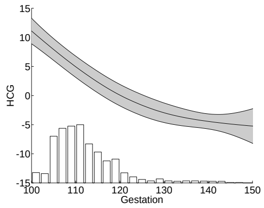
We used the model (5.7) and linear regression to determine whether differences exists or not. Instead of using only 1000 patterns, we used all available patterns from both populations. The only significant difference with 99% CI was in maternal weight. In Taiwan, every kilogram decreases hCG by 0.200 kU/l ($[0.165, 0.236]$ with 99% CI) and in the UK, 0.146 kU/l ($[0.129, 0.163]$). Therefore, a different weight correction should be performed for British and Taiwanese women.

Experiment 5.6 (Taiwanese and British AFP) *In AFP estimation, the Taiwanese data produced the RK spline model*

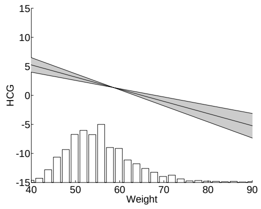
$$AFP = Weight \oplus Gestation$$



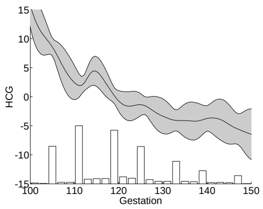
(a) hCG and Maternal weight in the UK.



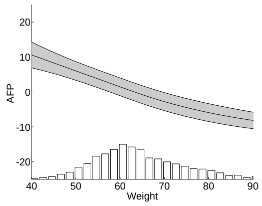
(b) hCG and Gestational age in the UK.



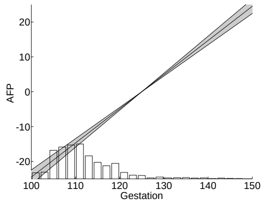
(c) hCG and Maternal weight in Taiwan.



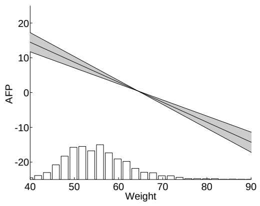
(d) hCG and Gestational age in Taiwan.



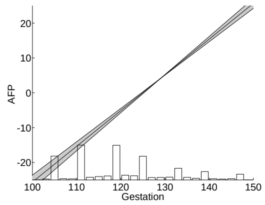
(e) AFP and Maternal weight in the UK.



(f) AFP and Gestational age in the UK.



(g) AFP and Maternal weight in Taiwan.



(h) AFP and Gestational age in Taiwan.

Figure 5.8: hCG and AFP models in Taiwanese and British DS comparison.

and the British data the model

$$AFP = Weight \oplus Gestation \oplus \mathcal{H}_1^{Weight}.$$

Figures 5.8 (e)-(h) show the AFP models of the UK and Taiwan. There is a weak nonlinear term in maternal weight in the UK whereas in Taiwan the term is fully linear. It should be kept in mind that there are few women over 75 kg in the Taiwanese data and the linearity may be due to this. In the effect of gestational age there are no significant differences between the UK and Taiwan.

Automatic model selection did not find any interaction terms and because the nonlinearity of the weight term in the UK model was weak, we performed a linear regression analysis on the full data using the model

$$AFP = a + b \cdot Weight + c \cdot Gestation.$$

The only significant difference with 99% CI was in maternal weight. In the UK every kilogram decreases AFP by 0.353 kU/l ($[0.332, 0.374]$ with 99% CI) and in Taiwan by 0.603 kU/l ($[0.550, 0.657]$). The difference is significant and therefore different weight correction should be used for British and Taiwanese women.

5.3.3 Conclusion of DS Screening

Medical conclusion: There are two discoveries in the Finnish data that deserve a more detailed study. The first one is the difference between Tampere and Turku on how maternal weight influences AFP. The difference is not statistically significant, but the Tampere data has only 314 samples, which is less than one third of the samples in the Turku data. More data from Tampere is required to determine whether the difference is significant or not. The second finding concerns the effect of weight on hCG. The Turku model suggests that for women over 100 kg a nonlinear weight correction should be used instead of a linear one.

In comparison of the British and Taiwanese populations, the only significant difference was in the effect of maternal weight. Serum hCG and AFP decrease gradually with weight because the same amount of hormones diffuses to larger volume. In Taiwan the decrease rate per kg is greater than in the UK. The finding is true for both hCG and AFP. Therefore, different weight correction coefficients should be used for British and Taiwanese women.

Methodological conclusion: An RK spline model usually contains both smooth and nonsmooth terms. The terms are scaled in such a way that their integral (if the variable is continuous) over the training domain is zero. This side condition guarantees a unique solution to the coefficients of RK splines. The differences in smooth terms between the populations have to be interpreted carefully. The differences in smooth terms propagate to constant terms because of the scaling and therefore if populations have different smooth terms a comparison of the constant terms is usually not justified.

The effect of gestational age on hCG in Helsinki data shows that *RK splines are vulnerable to outliers in the data*. However, the outliers can be detected easily with RK splines because of oversmoothing. In the case of oversmoothing, at least one of the smoothing coefficients $\lambda_i \to 0$. If RK splines are used for preliminary data analysis, the sensitivity of the method can be utilized in detecting whether there are outliers in the training data.

5.4 Carriage of Streptococcus Pneumoniae

Pneumonia kills about 3 million children every year in developing countries. The majority of these children die in their infancy. The most fatal pneumonia is caused by Haemophilus influenzae or Streptococcus pneumoniae [Sha95]. In most developed countries, children are routinely vaccinated against invasive infections caused by Haemophilus influenzae type b (Hib) bacterium. The Hib vaccine gives 94-100% protection against serious Hib diseases, such as meningitis and sepsis. Currently used 23-valent pneumococcal (Pnc) vaccine gives 50-80% protection against serious Pnc infections. However, children less than two years old are unable to produce antibodies against those Pnc bacteria that are common cause of childhood infections and therefore the vaccine does not protect young children. The carriage of Pnc in infancy has been studied in [GMG+89, MCFH92, CNS+95, LEOB+96].

In this section, we use RK splines to study the influence of the conventional Pnc polysaccharide vaccine given during pregnancy and the potential of maternally transferred antibodies for protection of infants in their first $3-4$ months of life against Pnc diseases.

5.4.1 ARIVAC project

The National Public Health Institute in Finland and the Research Institute for Tropical Medicine in the Philippines commenced in the early 1980's research collaboration for the evaluation of the Pnc vaccine in the prevention of childhood pneumonia in the Philippines. The name of the project is ARIVAC (Acute Respiratory Infections VACcine).

The study was a randomized, controlled, one-way blinded intervention study, where 108 mothers (intervention group) received Hib, Pnc and tetanus toxoid (TT) vaccines, 53 mothers (control group) only TT vaccine as is the practice in the Philippines. The majority of mothers were vaccinated in the second trimester of pregnancy. In both groups, infants received Hib and Pnc vaccines. See Appendix A.6 for details of the data set.

5.4.2 Comparison of Serotypes

There are 90 different serotypes of Pnc. Each serotype induces its own antibody and therefore immunity against one serotype usually does not protect from infections caused by the other serotypes. A nasopharyngeal swab was obtained from all children at the age of 6, 10, 14, 18 weeks, 9 and 10 months. Figure 5.9 shows Pnc carriage of each age group. In all age groups, Pnc carriage is more prevalent among controls than interventions.

Nasal swab samples were serotyped using a chessboard approach described in [Sør93]. Figure 5.10 shows Pnc carriage for ten most common

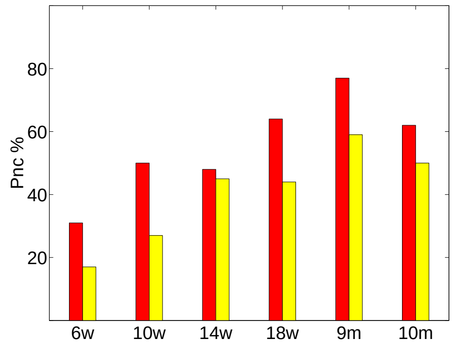


Figure 5.9: Pnc carriage of the age groups. The dark bars represent controls and the light bars interventions. Labels 6w-10m denote the age of the children from 6 weeks to 10 months.

serotypes at each age group. The rest of the serotypes are grouped by label *other*.

The 23-valent polysaccharide Pnc vaccine includes the 23 most common Pnc serotypes. From these serotypes, six were chosen for closer study and the antibodies of these six serotypes were measured from infants at the age of 6, 10, 14, 18 weeks, 9 and 10 months. We use labels *I6w*, *I10w*, *I14w*, *I18w*, *I9m* and *I10m* to denote the antibodies measured from infants. The antibodies were also measured from mothers before vaccination *(Mpre)*, one month after the vaccination *(Mpost)*, and from cord blood at the delivery *(Mcord)*. From the six serotypes selected for closer study, serotypes 5 and 18C had no incidences and serotype 1 had only one incidence in the data set. Therefore, below we use only serotypes 6B (31 cases), 14 (19 cases) and 19F (31 cases).

The antibodies of serotypes 6B, 14 and 19F were measured from infants at six different ages. At the age of 6 weeks, none of the infants carried Pnc of serotype 6B or 14 and only two infants carried serotype 19F. Therefore, the first analyzed age was 10 weeks. From the other four age groups, 14 weeks and 9 months were selected for further study. The number of observations is too small for automatic model selection and therefore we have used a full additive model in all experiments.

Experiment 5.7 (Pnc at Age of 10 Weeks) *From the total of 150 infants at the age of 10 weeks, there were three cases of serotypes 6B, 14 and 19F each. We performed three runs using the RK spline model*

$$\begin{aligned} Pnc^{10w} \quad = \quad & Mpre \oplus \mathcal{H}_1^{Mpre} \oplus Mpost \oplus \mathcal{H}_1^{Mpost} \oplus \\ & Mcord \oplus \mathcal{H}_1^{Mcord} \oplus I6w \oplus \mathcal{H}_1^{I6w} \oplus \\ & I10w \oplus \mathcal{H}_1^{I10w}, \end{aligned} \tag{5.8}$$

where variables Mpre, Mpost, Mcord, I6w and I10w denote the antibody concentration and Pnc^{10w} *whether the subject is a carrier of the corresponding* *serotype (6B, 14 or 19F).*

Figures 5.11 (a)-(c) summarize the linear terms and Figures 5.12 (a)-(c) the nonlinear terms of the model. Figure 5.12 (c) shows the influence of mother's post-vaccination antibody to the likelihood of 19F. The nonlinearity is due to one abnormal training observation with high antibody concentration (76 mg/l).

Experiment 5.8 (Pnc at Age of 14 Weeks) *The total of 148 infants were serotyped at the age of 14 weeks. We used the RK spline model*

$$Pnc^{14w} = Pnc^{10w} \oplus I14w \oplus \mathcal{H}_1^{I14w},$$

where Pnc^{10w} *includes the terms of the model* (5.8).

Figures 5.11 (d)-(f) summarize the linear terms and Figures 5.12 (d)-(f) the nonlinear terms.

Experiment 5.9 (Pnc at Age of 9 Months) *Infants at the age of 9 months should not benefit directly from mother's vaccination since the half time of Pnc antibodies is approximately 32 days. The number of observation in this age group was 143. We used the RK spline model*

$$Pnc^{9m} = Pnc^{14w} \oplus I18w \oplus \mathcal{H}_1^{I18w} \oplus I9m \oplus \mathcal{H}_1^{I9m}.$$

The model has the infant's antibody level at the age of 6, 10, 14, 18 weeks and 9 months with mother's pre-vaccination, post-vaccination and cord blood antibodies. Figures 5.11 (g)-(i) show the linear terms and Figures 5.12 (g)-(i) the nonlinear terms. Figure 5.12 (i) shows the influence of infant's antibody at the age of 9 months to likelihood of 19F. The nonlinearity is due to one abnormal observation with high antibody concentration (9.5 mg/l).

5.4.3 Conclusion of Pnc Carriage Estimation

Medical: The infants whose mothers received Pnc vaccine during the pregnancy had less Pnc carriage in the first year of life than infants whose mothers were not vaccinated. The half time of Pnc antibodies is too short to explain the phenomenon alone.

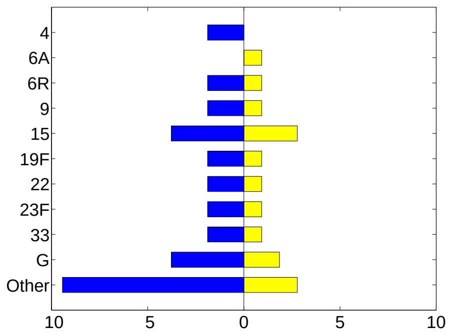
Pnc *serotype 6B* is the most common serotype among 9 and 10 month old infants. The significant findings with 95% CI were that the infant's antibody of 6B at the age of 6 weeks protects them from the corresponding carriage at the age of 10 weeks and the infant's antibody at the age of 14 weeks protects them from carriage at the age of 9 months.

In all age groups, the percentage of Pnc *serotype 14* carriers in the intervention group is greater than or equal to the control group. Therefore vaccination during the pregnancy seems not to protect children from Pnc carriage of serotype 14. It is important to notice that the number of observation does not justify the conclusion that vaccination would expose to serotype 14 carriage. High serotype 14 antibody concentration during infancy seems to protect infants from the corresponding carriage at the age of 9 months.

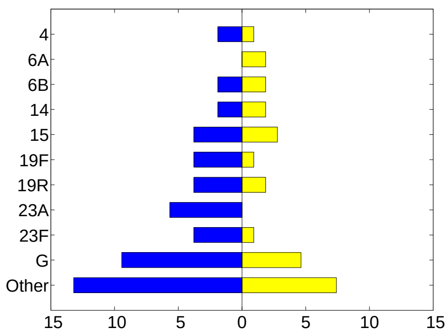
In 5 of 6 age groups of the study, the percentage of positive *serotype 19F* observations was greater in the control than intervention group. However, we could not find a clear pattern on how the antibodies protect from carriage of 19F.

In none of the 9 analyses performed, mother's post-vaccination nor cord blood antibody gave protection against serotype 6B, 14 or 19F carriage with the 95% CI. However in 7 of 9 analyses the mother's post-vaccination antibody, and in 8 of 9 analyses the cord blood antibody, gave some protection. This suggests that the 23-valent Pnc polysaccharide vaccine given to mother during the pregnancy protects the infant from carriage of Pnc serotypes 6B, 14 and/or 19F. The number of positive carriers in the data set was too small to determine the overall protective effect of maternal vaccination on the infant's Pnc carriage.

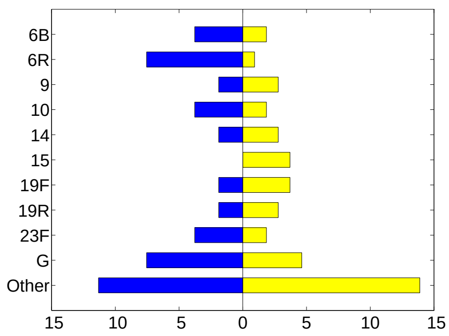
Methodological: Figures 5.12 (c) and 5.12 (i) show how one abnormal training observation can influence the RK spline. In both cases, the nonlinearity is due to one abnormal training observation. RK splines are sensitive to outliers in the data. This can be used to advantage if RK splines are used for preliminary data analysis. Most data analysis tools perform better without outliers and RK splines can be used to remove outliers from the training data.



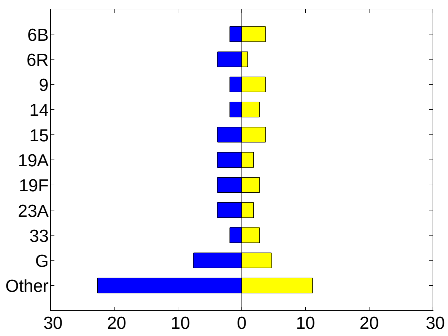
(a) 6 weeks (n=149).



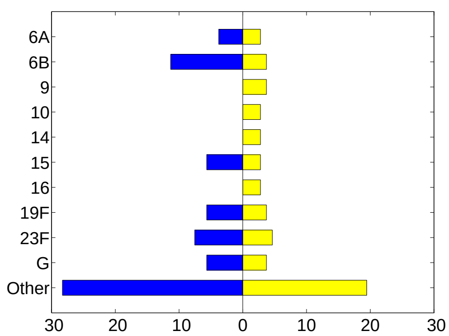
(b) 10 weeks (n=150).



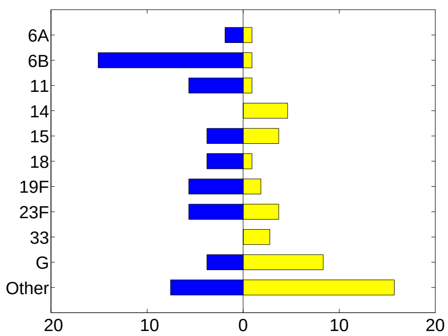
(c) 14 weeks (n=148).



(d) 18 weeks (n=145).

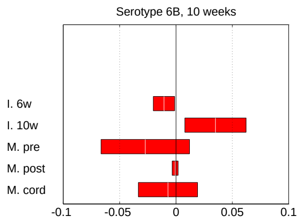


(e) 9 months (n=143).

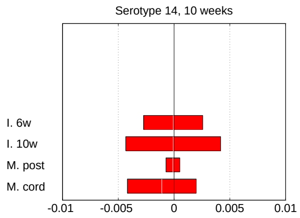


(f) 10 months (n=141).

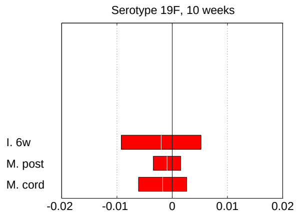
Figure 5.10: Percentage of Pnc carriage for the ten most common serotypes in each age group. Black bars represent control and gray bars intervention children. Label *G* denotes serotypes 29, 34, 35, 42* and 47*.



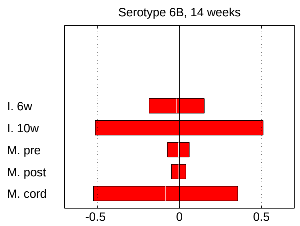
(a) Serotype 6B at age of 10 weeks.



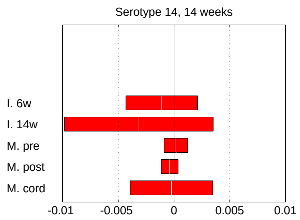
(b) Serotype 14 at age of 10 weeks.



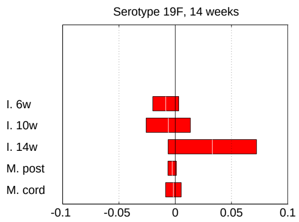
(c) Serotype 19F at age of 10 weeks.



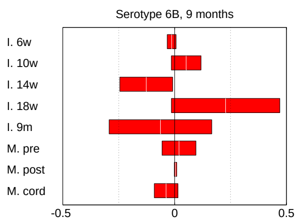
(d) Serotype 6B at age of 14 weeks.



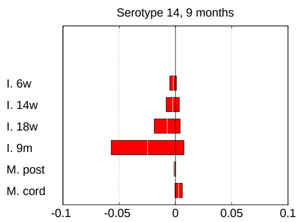
(e) Serotype 14 at age of 14 weeks.



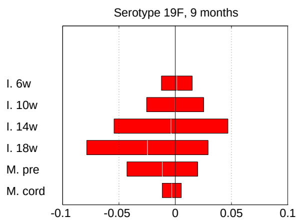
(f) Serotype 19F at age of 14 weeks.



(g) Serotype 6B at age of 9 months.

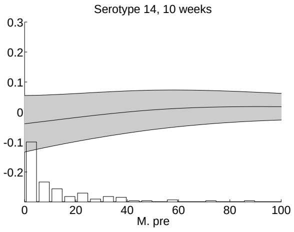


(h) Serotype 14 at age of 9 months.

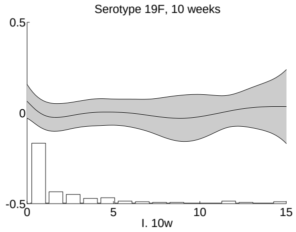


(i) Serotype 19F at age of 9 months.

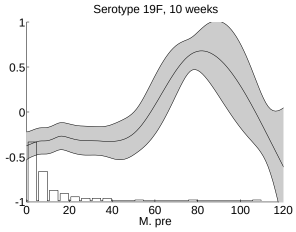
Figure 5.11: Linear terms with 95% CI. A negative coefficient means that the probability of the serotype decreases as a function of the antibody and a positive coefficient the opposite. A low antibody concentration makes infant prone to carriage of Pnc. A bar on the negative side of the mean denotes that the corresponding variable protects from Pnc carriage. A bar on the positive side means that the variable enhances carriage. The coefficients of the mother were multiplied by 10 to obtain approximately the same scale with the coefficients of infant.



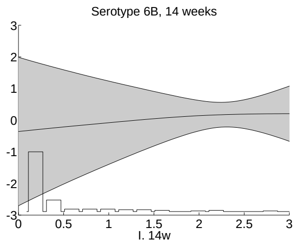
(a) Mother's pre-vaccination and serotype 14 at age of 10 weeks.



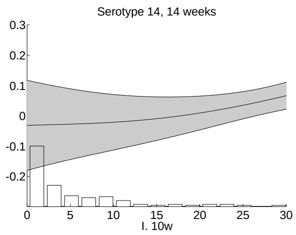
(b) Infant's 10 week antibody and serotype 19F at age of 10 weeks.



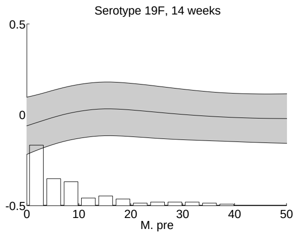
(c) Mother's pre-vaccination and serotype 19F at age of 10 weeks.



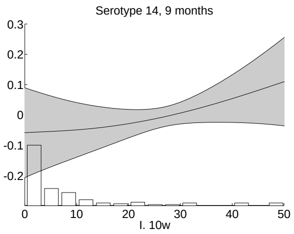
(d) Infant's 14 week antibody and serotype 6B at age of 14 weeks.



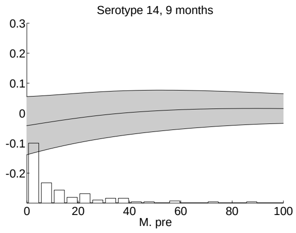
(e) Infant's 10 week antibody and serotype 14 at age of 14 weeks.



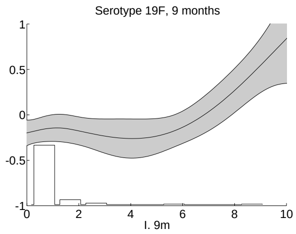
(f) Mother's pre-vaccination and serotype 19F at age of 14 weeks.



(g) Infant's 10 week antibody and serotype 14 at age of 9 months.



(h) Mother's pre-vaccination and serotype 14 at age of 9 months.



(i) Infant's 9 month antibody and serotype 19F at age of 9 months.

Figure 5.12: Nonlinear terms with 95% CI.

Chapter 6

Summary

The aim of pattern recognition is to find rules or functions to refine data. Usually, and especially in clinical medicine, the rules or functions are generated from a sample called training data. The goal is not to learn exact representation of the training data itself, but to build a statistical model of the process which generated the data. A function that is closely fitted to the training data will tend to have a large expected error when used to estimate unseen data. If we smoothen the function, the expected error on unseen data will first decrease. However, if we continue smoothing too far, the expected error will raise again. The aim of generalization is to find the best smoothing level. We showed a theoretical background of generalization in function approximation and derived an error measure (MSE) minimizing the expected error.

We have used RK splines to fit a function of several variables to scattered, noisy data. The fitted function is presented as a sum of functions of one variable plus a sum of functions of two variables, and so forth. The different variables may have different value domains, and therefore also different RK functions. In the present study, we have shown how to generate a RKHS for a specific data. Matrix decompositions can be utilized in the calculation of RK spline fits while adaptively choosing multiple smoothing parameters with the GCV. To alleviate the computational time requirement of GCV, we have presented a parallel version of GCV. Parallel GCV utilizes block cyclic distribution of matrices.

Variable selection is an important aspect in nonparametric approximation when the number of variables is large. In medical data, the independent variables usually correlate with each other and therefore there is no optimal variable combination. In this study, we have presented a heuristic term selection algorithm. It attempts to minimize the number of variables needed in approximation. We tested the term selection using both artificial and medical data sets. The experiments suggest that heuristic model selection should be favored for the standard breadth-first search.

To evaluate RK splines in the analysis of clinical data, we demonstrated the use of the method on several medical data sets. Serum LDL cholesterol is often estimated using a linear model. Our model selection algorithm discovered that the product of total cholesterol and triglycerides is statitically very significant ($p < 0.000001$) and it should be included in the model. In another experiment, we studied nutrient intake in a low-cholesterol diet. One of the input variables was a ratio of two variables and we were able to detect its nonlinearity using RK splines.

We performed two separate experiments to study Down syndrome screening. In the first study, the aim was to compare variables influencing the reference values of hCG and AFP hormones between four Finnish cities. Several variables proved to have a nonlinear effect on reference values. GCV oversmoothed the insignificant nonlinear terms and the terms with outliers. This was the only data set that required preprocessing before applying RK splines. The second Down syndrome experiment was similar to the first one except that the comparison was performed between Taiwanese and British women. RK splines detected a nonlinearity in gestational age. In the UK gestational age was measured in days, and in Taiwan it was measured in weeks. RK splines detected the difference and the confidence intervals of gestational age were found to be different between the countries.

In the last experiment we studied pneumonia. It kills over 3 million children every year in developing countries and currently there are no effective vaccines that would protect infants from serious Pnc infections. We studied whether Pnc vaccine given to mother during the pregnancy would protect infants from infections. In 15 of the 18 experiments performed with RK splines vaccination provided some protection from Pnc infections.

Chapter 7

Conclusion

The use of the nonparametric methods in the analysis of medical data has received increasing attention lately. The nonparametric methods provide the same flexibility as the neural networks and other semiparametric methods, but require no parameter setting. RK splines are a computationally intensive method, but in medicine where the number of observations is usually low, that is a few hundred or a few thousand, the computational complexity does not set serious limitations to the data analysis. If the number of observations is large, there are heuristics that can be attached to RK splines

In medicine, the confidence intervals are the conventional manner to express the reliability of the results. RK splines provide a model with Bayesian confidence intervals. Therefore, the results can be reported using the same format as in the traditional parametric regression. Also the influence of the individual terms can be extracted from the model and displayed visually.

Further study on RK splines is required before the method can be used safely as an alternative to traditional parametric regression. The current version of automatic model space selection requires a separate data set that is used to evaluate and compare the alternative model spaces. In medical data, the number of observations is usually low and therefore it would be useful to perform model space selection using the training data itself. The automatic elimination of outliers would also increase reliability of the method.

One interesting topic for further study would be to develop a GCV algorithm that includes an automatic model space selection. The current version of automatic model space selection improves the performance of the model space by adding new elementary spaces from the set of untested elementary spaces. The influence of the individual smooth terms is determined separately using GCV. Instead of running GCV in an inner loop and model space selection in an outer loop, a novel GCV could perform both in the same loop. The coefficients of the untested smooth elementary spaces and the smooth elementary spaces that have been tested but discarded from the model space can be set to zero. The nonsmooth elementary spaces can be given a binary coefficient, i.e., one or zero, depending on whether the elementary space is in the model space or not. The novel GCV would probably speed up the training of RK splines.

We have shown that RK splines are a potential data mining tool in the analysis of medical data. As the computational barriers of nonparametric estimation are collapsing, RK splines are likely to find their position among other nonparametric approximation methods. Our case studies support this, because we were able to find with RK splines several phenomena that were not discovered with conventional methods.

Appendix A

Description of Data Sets

The appendix describes briefly the data sets used in chapters 4 and 5. We use the following names for the data sets:

Name of data set	Chapter
Pima Indian diabetes data	4
Heart disease data	4
Friedewald data	5.1
STRIP data	5.2
Down syndrome data sets	
Helsinki data	5.3
Tampere data	5.3
Turku data	5.3
Oulu data	5.3
Taiwan data	5.3
UK data	5.3
Pneumoniae data sets	
Serotype 6B data	5.4
Serotype 14 data	5.4
Serotype 19F data	5.4

Table A.1: The names of the data sets.

For continuous variables, we show the minimum, maximum, mean, standard deviation and percentage of missing values, and for discrete variables the number of observations in each category and the percentage of missing values. The number of observation is presented in the form $c = n$, where c indicates the category and n the number of observations. The category $> c$ indicates the total number of observation in categories $c + 1, \ldots, \max(c)$.

A.1 Pima Indian Diabetes Data

The Pima Indian diabetes data set was originally donated by Vincent Sigillito, Johns Hopkins University, Laurel, MD 20707 [MST94]. Several constraints were placed on the selection of these instances from a larger database held by the National Institute of Diabetes and Digestive and Kidney Diseases. In particular, all patients here are females at least 21 years old of Pima Indian heritage living near Phoenix, Arizona, USA.

The problem associated with the data set is to predict whether a patient would test positive for diabetes given a number of physiological measurements and medical test results. Michie et al. have used this data set to compare the performance of machine learning algorithms, neural networks and statistical methods in [MST94].

The data set comprises 768 observations. Table A.2 summarizes the statistics of the variables. According to World Health Organization criteria a patient has diabetes if 2-hour post-load plasma glucose is at least 200 mg/dl. In Table A.2, *plasma glucose* denotes the plasma glucose concentration in an oral glucose tolerance test, *blood pressure* the diastolic blood pressure, *skin fold* the triceps skin fold thickness, *serum insulin* the 2-hour serum insulin, *BMI* the body mass index and *diabetes pedigree* the diabetes pedigree function.

The data set contains few missing values that are denoted by zero in the original data set. We have not converted these zeros to missing values, but we have kept them in order to alleviate comparison between our test results and the results of previous research projects that have utilized the same data set.

Variable	Min	Max	Mean	SD	Missing
Diabetes	500 positive, 268 negative				0 %
Times pregnant	0	17	3.85	3.37	0 %
Plasma glucose	0	199	120.89	31.97	0 %
Blood pressure (mm Hg)	0	122	69.11	19.36	0 %
Skin fold (mm)	0	99	20.54	15.95	0 %
Serum insulin (μU/ml)	0	846	79.80	115.24	0 %
BMI (kg/m^2)	0	67.1	31.99	7.88	0 %
Diabetes pedigree	0.078	2.42	0.47	0.33	0 %
Age (years)	21	81	33.24	11.76	0 %

Table A.2: Summary of the Pima Indian diabetes data.

A.2 Heart Disease Data

The heart disease data comes from Cleveland Foundation and was supplied by Robert Detrano, V. A. Medical Center, Long Beach, CA. It is part of the collection of databases at the University of California, Irvine collated by David Aha. The data set has previously been studied in [MST94].

The problem associated with the data set is to predict whether a patient has heart disease given a number of physiological measurements and medical test results. The data set comprises 270 observations. Table A.3 summarizes the statistics of the variables. Label *electrocardiographic* denotes the resting electrocardiographic results (0,1 or 2), *old peak* the ST depression induced by exercise relative to rest, *slope of peak* the slope of the peak exercise ST segment, and *number of major vessels* the number of major vessels (0-3) colored by fluoroscopy. All variables excluding *chest pain type*, *electrocardiographic* and *thal* were considered either continuous or binary.

Variable	Min	Max	Mean	SD	Missing
Heart disease	120 positive, 150 negative				0 %
Age	29	77	54.43	9.11	0 %
Resting Blood pressure	94	200	131.34	17.86	0 %
Serum cholesterol (mg/dl)	126	564	249.66	51.69	0 %
Maximum heart rate	71	202	149.68	23.17	0 %
Old peak	0	6.2	1.05	1.15	0 %
Slope of peak	1	3	1.59	0.61	0 %
Number of major vessels	0	3	0.67	0.94	0 %
Sex	183 males, 87 females				0 %
Chest pain type	1 = 20, 2 = 42, 3 = 79, 4 = 129				0 %
Blood sugar > 120 mg/dl	40 positive, 230 negative				0 %
Electrocardiographic	0 = 131, 2 = 2, 3 = 137				0 %
Exercise induced angina	89 positive, 181 negative				0 %
Thal	normal = 152, fixed defect = 14 reversible defect = 104				0 %

Table A.3: Summary of the heart disease data.

A.3 Friedewald Data

The Friedewald data contains cholesterol values of Finnish normal population. The data set was donated by Jukka Marniemi and Olli Impivaara, Research and Development Centre, Social Insurance Institution, Turku. Friedewald data comprises 809 observations. Table A.4 summarizes the statistics of the variables.

Variable	Min	Max	Mean	SD	Missing
Total cholesterol (mmol/l)	2.96	10.93	6.43	1.18	0 %
HDL (mmol/l)	0.58	3.44	1.34	0.32	0 %
Triglycerides (mmol/l)	0.33	6.85	1.50	0.79	0 %
LDL (mmol/l)	1.88	8.95	4.58	1.07	0 %

Table A.4: Summary of the Friedewald data.

A.4 STRIP Data

The purpose of the STRIP baby project (Special Turku coronary Risk factor Intervention Project for babies) was to find out whether it is possible safely modify exposure to coronary risk factors by beginning the intervention in infancy. The project was carried out at the Department of Pediatrics, University Central Hospital, Turku.

The STRIP data comprises 630 observations. Each subject was measured for serum lipids, average daily intake of 62 nutrients and average daily dietary divided into 98 food groups. Table A.5 shows the key statistics of lipids and nutrients.

Variable	Min	Max	Mean	SD	Missing
Group	307 interventions, 323 controls				0 %
Total cholesterol (mmol/l)	2.50	8.50	4.44	0.72	11.6 %
HDL (mmol/l)	0.68	2.41	1.29	0.23	11.4 %
Triglycerides (mmol/l)	0.20	2.10	0.67	0.27	11.6 %
Fat (g)	19.20	117.90	54.61	12.98	0 %
SAFA (g)	5.34	54.88	22.08	6.08	0 %
MUFA (g)	6.19	41.84	19.20	5.03	0 %
PUFA (g)	2.35	19.03	9.41	2.87	0 %
P/S	0.11	1.00	0.44	0.14	0 %
Carbohydrate (g)	92.10	354.10	204.53	36.75	0 %
Fiber (g)	4.10	29.00	13.16	3.58	0 %

Table A.5: Summary of the STRIP data.

A.5 Down Syndrome Data Sets

There are six Down syndrome data sets used in this study. The British and Taiwanese data sets were donated by Kevin Spencer. The Finnish data sets were collected by Finnish university hospitals as a part of the Nesumed project funded by the National Technology Agency.

The Helsinki data contains 487 observations. Table A.6 shows the key statistics of the data set. Smoking has three categories: 0 = nonsmoking, 1 = smoked during the first trimester, 2 = smoking. Tables A.7, A.8, A.9, A.10 and A.11 show correspondingly the key figures of the Tampere, Turku, Oulu, Taiwan and UK data sets. The Tampere data comprises 314, Turku 1105, Oulu 782, Taiwan 14 003, and UK 20232 observations.

Variable	Min	Max	Mean	SD	Missing
hCG (kU/l)	NA	NA	NA	NA	100 %
AFP (kU/l)	6	88	28.9	10.9	0 %
Gestational age (days)	105	137	111.0	5.3	0 %
Maternal weight (kg)	44	103	65.1	9.6	9.4 %
Previous births	0 = 201, 1 = 185, 2= 76,				0 %
	3 = 20, 4 = 5				
Smoking	0 = 417, 1 = 4, 2= 55,				2.3 %

Table A.6: Summary of the Helsinki data.

Variable	Min	Max	Mean	SD	Missing
hCG (kU/l)	3.8	94.9	17.4	12.3	0 %
AFP (kU/l)	9.4	76.7	26.2	10.3	0 %
Gestational age (days)	99	127	109.8	3.8	0 %
Maternal weight (kg)	43	112	65.5	10.6	8.9 %
Previous births	0 = 106, 1 = 143, 2= 52,				0 %
	3 = 11, 4 = 2				
Smoking	0 = 277, 1 = 9, 2= 28,				0 %

Table A.7: Summary of the Tampere data.

Variable	Min	Max	Mean	SD	Missing
hCG (kU/l)	2.5	252.1	19.6	16.3	0 %
AFP (kU/l)	9.6	116.9	27.9	10.6	0 %
Gestational age (days)	94	134	109.0	3.6	0 %
Maternal weight (kg)	45	141	65.6	10.6	7.8 %
Previous births	0 = 402, 1 = 477, 2= 180,				0 %
	3 = 35, 4 = 9, $> 4 = 2$				
Smoking	0 = 956, 1 = 7, 2= 126,				1.4 %

Table A.8: Summary of the Turku data.

A.6 Pneumoniae Data Sets

The purpose of the ARIVAC (Acute Respiratory Infections VACcine) project was to study the influence of conventional Pnc polysaccharide vaccine given

Variable	Min	Max	Mean	SD	Missing
hCG (kU/l)	1.7	265	19.7	18.5	0 %
AFP (kU/l)	8.3	158	28.1	11.2	0 %
Gestational age (days)	99	132	111.1	4.6	0 %
Maternal weight (kg)	39	118	64.7	10.3	0.1 %
Previous births	0 = 302, 1 = 286, 2= 133,				0 %
	3 = 40, 4 = 10, > 4 = 11				
Smoking	0 = 643, 1 = 11, 2= 109,				2.4 %

Table A.9: Summary of the Oulu data.

Variable	Min	Max	Mean	SD	Missing
hCG (kU/l)	0.04	100	16.2	13.9	0 %
AFP (kU/l)	2.8	150	47.0	23.5	0 %
Maternal age (years)	13	44	28.8	4.0	0 %
Gestational age (days)	98	160	120.4	13.1	0 %
Maternal weight (kg)	24.8	156	55.4	7.8	0 %

Table A.10: Summary of the Taiwan data.

during pregnancy and potential of maternally transferred antibodies for protection of infants against Pnc diseases. The study was carried out by the National Health Institute, Finland and the Research Institute for Tropical Medicine, the Philippines. The Pneumoniae data sets contain 161 observations. Table A.12 summarizes the number of positive serotype 1, 5, 6B, 14, 18C and 19F findings in all six age groups.

Serotypes 6B, 14 and 19F were selected for further study. Tables A.13, A.14 and A.15 show the statistics of Pnc antibodies 6B, 14 and 19F. The antibodies were measured from mothers before vaccination (Mother pre vac.), about one month after vaccination (Mother post vac.) and at delivery (Mother cord). Infants were measured for antibodies at age of 6, 10, 14, 18 weeks, 9 and 10 months.

Variable	Min	Max	Mean	SD	Missing
hCG (kU/l)	0.10	100	16.7	12.4	0 %
AFP (kU/l)	0.10	150	37.0	17.1	0 %
Maternal age (years)	15	47	28.7	5.0	0 %
Gestational age (days)	89	160	112.3	7.8	0 %
Maternal weight (kg)	38	153	65.8	12.7	0 %

Table A.11: Summary of the UK data.

Age	1	5	6B	14	18C	19F
6 weeks	0	0	0	0	0	2
10 weeks	1	0	3	3	0	3
14 weeks	0	0	4	4	0	5
18 weeks	0	0	5	4	0	5
9 months	0	0	10	3	0	7
10 months	0	0	9	5	0	9

Table A.12: Number of the positive Pnc serotype findings.

Variable	Min	Max	Mean	SD	Missing
Mother pre vac. (mg/l)	0.09	44.07	5.13	6.09	1.24 %
Mother post vac. (mg/l)	0.21	360.72	21.44	43.80	36.65%
Mother cord (mg/l)	0.09	41.85	6.34	8.26	23.60%
Infant 6 weeks (mg/l)	0.09	20.58	1.93	3.26	9.32%
Infant 10 weeks (mg/l)	0.09	8.17	0.81	1.26	8.70%
Infant 14 weeks (mg/l)	0.09	4.32	0.39	0.63	10.56%
Infant 18 weeks (mg/l)	0.09	1.56	0.21	0.27	13.04%
Infant 9 months (mg/l)	0.09	0.96	0.15	0.18	14.29%
Infant 10 months (mg/l)	0.09	1.72	0.20	0.30	22.36%

Table A.13: Summary of the serotype 6B data.

Variable	Min	Max	Mean	SD	Missing
Mother pre vac. (mg/l)	0.14	210.42	15.82	27.75	2.48%
Mother post vac. (mg/l)	0.14	1652.93	112.67	217.48	37.27%
Mother cord (mg/l)	0.14	413.05	35.83	60.16	24.22%
Infant 6 weeks (mg/l)	0.14	95.85	13.67	18.25	10.56%
Infant 10 weeks (mg/l)	0.14	52.2	7.02	10.86	9.94%
Infant 14 weeks (mg/l)	0.14	45.36	3.57	5.79	11.80%
Infant 18 weeks (mg/l)	0.14	20.79	1.71	2.80	13.04%
Infant 9 months (mg/l)	0.14	4.88	0.50	0.60	16.15%
Infant 10 months (mg/l)	0.14	6.04	0.59	0.81	24.22%

Table A.14: Summary of the serotype 14 data.

Variable	Min	Max	Mean	SD	Missing
Mother pre vac. (mg/l)	0.17	114.12	9.33	13.34	0.62%
Mother post vac. (mg/l)	0.61	366.96	37.90	56.46	35.40%
Mother cord (mg/l)	0.17	223	13.24	23.78	22.98%
Infant 6 weeks (mg/l)	0.04	21.33	3.81	4.50	9.32%
Infant 10 weeks (mg/l)	0.17	15.8	1.97	2.86	8.07%
Infant 14 weeks (mg/l)	0.17	8.37	1.02	1.44	9.94%
Infant 18 weeks (mg/l)	0.17	5.22	0.57	0.72	12.42%
Infant 9 months (mg/l)	0.17	9.52	0.52	1.02	14.29%
Infant 10 months (mg/l)	0.17	7.4	0.72	1.22	21.12%

Table A.15: Summary of the serotype 19F data.

Appendix B

Mathematics

B.1 Hilbert Space

Most of the definitions shown here are presented in [SM94].

Let $\mathbb{R}$ be the field of real numbers and V a nonempty set.

Definition B.1 (Vector space) *The set V is a* vector space *if it has the operations* vector addition $V \times V \to V$ *and* scalar multiplication $\mathbb{R} \times V \to V$ *satisfying the following properties for all $a, b \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$:*

i. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$
ii. $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$
iii. $\mathbf{0} + \mathbf{x} = \mathbf{x}$
iv. $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$
v. $a(b\mathbf{x}) = (ab)\mathbf{x}$
vi. $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$
vii. $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$
viii. $1\mathbf{x} = \mathbf{x}$

Definition B.2 (Inner product space) *The inner product space is a vector space with an* inner product $V \times V \to \mathbb{R}$ *satisfying, for every $a, b \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, the properties:*

i. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$
ii. $\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle$
iii. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$,

where the notation $\langle \cdot\, , \cdot \rangle$ is used for inner product. The inner product space has a *norm* $\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2}$.

Definition B.3 *A sequence of elements $\mathbf{x}_n$ of an inner product space is called a* Cauchy sequence *if for every $\epsilon > 0$ there is an integer N such that*

$\|\mathbf{x}_n - \mathbf{x}_m\| < \epsilon$ *whenever* $n, m \geq N$.

Definition B.4 *An inner product space is called* complete *if for any Cauchy sequence* $\mathbf{x}_n$ *there is an element* $\mathbf{x}$ *for which* $\|\mathbf{x}_n - \mathbf{x}\| \to 0$ *as* $n \to \infty$.

Definition B.5 (Hilbert space) *A* Hilbert space *is a complete inner product space.*

Definition B.6 (Orthogonality) *The vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *are said to be* orthogonal *if* $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. *The spaces* H_1 *and* H_2 *are said to be orthogonal if every element of* H_1 *is orthogonal to every element of* H_2.

Definition B.7 (Projection) *Let* H *be a Hilbert space and* M *a closed subspace of* H. *The* orthogonal projection *of* $h \in H$ *onto* M *is the element* $m \in M$ *satisfying*

$$\|h - m\| \leq \|h - y\|$$

for all $y \in M$. *The element* m *is unique and* $h - m$ *is orthogonal to* M. *Projections are linear operators, that is, linear functions from* H *into itself.*

Definition B.8 (Direct sum) *Let* H_1 *and* H_2 *be subspaces of the Hilbert space* H. *If for every* $h \in H$ *there is a unique representation* $h = h_1 + h_2$, *where* $h_1 \in H_1$ *and* $h_2 \in H_2$, *we say that* H *is a direct sum of* H_1 *and* H_2, *denoted as* $H = H_1 \oplus H_2$.

In Chapter 2 we are in a situation where the Hilbert space H is represented as the direct sum

$$H = H_0 \oplus H_1$$

where H_0 is finite dimensional and the subspaces H_0 and H_1 are orthogonal. Then both H_0 and H_1 are closed subspaces (since all finite dimensional subspaces are closed and orthogonal complements of arbitrary subsets are closed subspaces). Hence the orthogonal projections onto H_0 and H_1 are well defined in this situation.

Definition B.9 (Tensor product) *The Hilbert spaces we are interested in consist of real functions defined on some set. If* H_1 *is a Hilbert space of functions defined on* T_1 *with inner product* $< f, g >_1$ *and, similarly,* H_2 *is a Hilbert space of functions defined on* T_2 *with inner product* $< f, g >_2$, *then their* tensor product $H_1 \otimes H_2$ *is a Hilbert space of real function defined on* $T_1 \times T_2$ *which contains, among others, all functions of the form*

$$f(t_1, t_2) = f_1(t_1) f_2(t_2)$$

where $f_1 \in H_1$ *and* $f_2 \in H_2$*. The squared norm of such a function* f *is defined by*

$$< f, f >=< f_1, f_1 >_1 < f_2, f_2 >_2 .$$

Being a vector space, the tensor product $H_1 \otimes H_2$ *contains all finite linear combinations* $\sum_{i=1}^{k} f_{1i}(t_1) f_{2i}(t_2)$ *where every* $f_{1i} \in H_1$ *and every* $f_{i2} \in H_2$*. The Hilbert space* $H_1 \otimes H_2$ *is then obtained by a process of completion with respect to the inner product introduced above, see [Aro50] for details.*

Lemma B.1 *Let* H *be a Hilbert space of real-valued functions. The inner product of* $f \in H$ *and* $g \in H$ *can be calculated from the identity*

$$\langle f, g \rangle = \frac{1}{4}(\|f+g\|^2 - \|f-g\|^2).$$

B.2 Bernoulli Polynomials

Definition B.10 *Bernoulli polynomials are a sequence of polynomials defined inductively as*

$$\begin{aligned} B_0(x) &= 1 \\ B'_n(x) &= nB_{n-1}(x), \quad \text{and} \quad \int_0^1 B_n(x)dx = 0 \quad \text{if } n \geq 1 \end{aligned}$$

The first elements of the sequence are

$$\begin{aligned} B_0(x) &= 1 \\ B_1(x) &= x - 1/2 \\ B_2(x) &= x^2 - x + 1/6 \\ B_3(x) &= x^3 - 3/2x^2 + 1/2x \\ B_4(x) &= x^4 - 2x^3 + x^2 - 1/30. \end{aligned}$$

Bibliography

[All74] D. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

[Aro50] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[Att92] M. Atteia. *Hilbertian Kernels and Spline Functions.* North-Holland, Amsterdam, 1992.

[BF96] M. D. Buhmann and R. Fletcher. M. J. D. Powell's work in univariate and multivariate approximation theory and his contribution to optimization. Technical Report 96-16, Eidgenössische Technische Hochschule, Zürich, Oct 1996.

[BHSB+93] I. Bartels, B. Hoppe-Sievert, B. Bockel, S. Herold, and J. Caesar. Adjustment formulae for maternal serum alpha-fetoprotein, human chorionic gonadotropin, and unconjugated oestriol to maternal weight and smoking. *Prenatal Diagnosis*, 13(2):123–130, Feb 1993.

[BHT89] A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–555, Jun 1989.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[BN92] R. K. Beatson and G. N. Newsam. Fast evaluation of radial basis functions: I. *Computers Math. Applic.*, 24(12):7–19, 1992.

[BPL+89] L. Bernstein, M. C. Pike, R. A. Lobo, R. H. Depue, R. K. Ross, and B. E. Henderson. Cigarette smoking in pregnancy results in marked decrease in maternal hCG and oestradiol levels. *British Journal of Obstetrics and Gynaecology*, 96(1):92–96, Jan 1989.

[Buh90] M. D. Buhmann. Multivariate cardinal interpolation with radial-basis functions. *Constructive Approximation*, 6:225–255, 1990.

[BV93] A. Y. Bezhaev and V. A. Vasilenko. *Variational Spline Theory.* NCC Publisher, Novosibirsk, 1993.

[CDO+94] J. Choi, J. J. Dongarra, L. S. Ostrouchov, A. P. Petitet, D. W. Walker, and R. C. Whaley. The design and implementation of the ScaLAPACK LU, QR, and Cholesky factorization routines. Technical report, Engineering, Physics and Mathematics Division, U.S. Department of Energy, Sep 1994.

[CH91] J. J. Du Croz and N. J. Higham. Stability of methods for matrix inversion. Technical report, Department of Mathematics, University of Manchester, May 1991.

[Che86] E. W. Cheney. *Multivariate Approximation Theory: Selected Topics.* Society for Industrial and Applied Mathematics, Philadelphia, 1986.

[Che87] Z. Chen. A stepwise approach for the purely periodic interaction spline model. *Communications in Statistics, Theory and Methods*, 16(3):877–895, 1987.

[Chu88] C. K. Chui. *Multivariate Splines.* Society for Industrial and Applied Mathematics, Philadelphia, 1988.

[CJI+97] L. A. Cole, M. Jacobs, T. Isozaki, G. E. Pakomäki, R. O. Bahado-Singh, and M. J. Mahoney. Screening for Down syndrome using urine hCG free beta-subunit in the second trimester of pregnancy. *Prenatal Diagnosis*, 17(12):1107–1111, Dec 1997.

[CNS+95] M. R. Capeding, H. Nohynek, L. T. Sombrero, L. G. Pascual, E. S. Sunico, G. A. Esparar, E. Esko, M. Leinonen, and P. Ruutu. Evaluation of sampling sites for detection of upper respiratory tract carriage of Streptococcus pneumoniae and Haemophilus influenzae among healthy Filipino infants. *Journal of Clinical Microbiology*, 33(11):3077–3079, 1995.

[Com98] Committee on Nutrition. Cholesterol in childhood. *Pediatrics*, 101(1):141–147, Jan 1998.

[CW79] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

[Dac99] K. Dackland. *Design, Modeling, and Evaluation of High-Performance Algorithms for Matrix Factorizations and Matrix Pair Reductions.* PhD thesis, Department of Computing Science, Umeå University, May 1999.

[DD97] E. F. D'Azevedo and J. J. Dongarra. LAPACK working note 118: The design and implementation of the parallel out-of-core ScaLAPACK LU, QR and Cholesky factorization routines. Technical Report 347, Department of Computer Science, University of Tennessee, Jan 1997.

[Duc77] J. Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100, Berlin, 1977. Springer-Verlag.

[DW97] J. J. Dongarra and R. C. Whaley. LAPACK working note 94: A user's guide to the BLACS v1.1. Technical report, Department of Computer Science, University of Tennessee, May 1997.

[Efr81] B. Efron. Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics*, 9(2):139–172, 1981.

[Efr83] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

[EG83] B. Efron and G. Gong. A leisure look at the bootstrap, the Jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

[Fie86] M. Fiedler. *Special Matrices and Their Applications in Numerical Mathematics*. Martinus Nijhoff, Dordrecht, 1986.

[FLF72] W. T. Friedewald, R. I. Levy, and D. S. Fredrickson. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry*, 18(6):499–502, 1972.

[FMH+90] M. H. Frick, V. Manninen, J. K. Huttunen, O. P. Heinonen, L. Tenkanen, and M. Mänttäri. HDL-cholesterol as a risk factor in coronary heart disease. An update of the Helsinki Heart Study. *Drugs*, 40 Suppl(1):7–12, 1990.

[FMM+89] G. Farchi, S. Mariotti, A. Menotti, F. Seccareccia, S. Torsello, and F. Fidanza. Diet and 20-y mortality in two rural population groups of middle-aged men in Italy. *Journal of Clinical Nutrition*, 50:1095–1103, 1989.

[Fra82] R. Franke. Scattered data interpolation: tests of some methods. *Mathematics of Computation*, 38(157):181–200, 1982.

[Fri91] J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, Mar 1991.

[FS89] J. H. Friedman and B. W. Silverman. Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31:3–39, 1989.

[Gao93] F. Gao. On combining data from multiple sources with unknown relative weights. Technical Report 894, Department of Statistics, University of Wisconsin, Feb 1993.

[GBCW89] C. Gu, D. M. Bates, Z. Chen, and G. Wahba. The computation of generalized cross-validation functions through Householder tridiagonalization with applications to the fitting of interaction spline models. *SIAM Journal on Matrix Analysis and Applications*, 10(4):457–480, Oct 1989.

[Gir98] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

[GJP93] F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: from regularization to radial, tensor and additive splines. Technical Report A.I. Memo No.1430, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Jun 1993.

[GMG+89] M. Gratten, J. Montgomery, G. Gerega, H. Gratten, H. Siwi, A. Poli, and G. Koki. Multiple colonization of the upper respiratory tract of Papua New Guinea children with Haemophilus influenzae and Streptococcus pneumoniae. *Southeast Asian J. Trop. Med. Pub. Hlth.*, 20(4):501–509, 1989.

[GS94] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London, 1994.

[Gu89] C. Gu. Rkpack and its applications: Fitting smoothing spline models. Technical Report 857, Department of Statistics, University of Wisconsin, May 1989.

[Gu95] C. Gu. Smoothing spline density estimation: Conditional distribution. *Statistica Sinica*, 5:709–726, 1995.

[Gu96] C. Gu. Multidimensional smoothing with splines. Technical report, Purdue University, Nov 1996.

[Gu97] C. Gu. Structural multivariate function estimation: Some automatic density and hazard estimates. Technical report, Purdue University, 1997.

[GW78] W. J. Gordon and J. A. Wixom. Shepard's method of "metric interpolation" to bivariate and multivariate interpolation. *Mathematics of Computation*, 32(141):253–264, 1978.

[GW92] C. Gu and G. Wahba. Smoothing spline ANOVA with componentwise Bayesian "confidence intervals". Technical report, Purdue University, Nov 1992.

[HFK95] P. P. C. W. Huijbregts, E. J. M. Feskens, and D. Kromhout. Dietary patterns and cardiovascular risk factors in elderly men: The Zutphen elderly study. *International Journal of Epidemiology*, 24(2):313–320, 1995.

[HHSK98] J. J. Hsu, T. T. Hsieh, Y. K. Soong, and B. Kuo. The influence of maternal weight correction formulas in Asian Down syndrome screening using alpha-fetoprotein and free beta-human chorionic gonadotropin. *Journal of Maternal-Fetal Investigation*, 8(2):66–70, Jun 1998.

[HT86] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, Aug 1986.

[HT90] T. Hastie and R. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1990.

[Hut89] M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics, Simulation and Computation*, 18(3):1059–1076, 1989.

[KSKG92] J. E. Keil, S. E. Sutherland, R. G. Knapp, and P. C. Gazes. Serum cholesterol–risk factor for coronary disease mortality in younger and older blacks and whites. The Charleston Heart Study, 1960-1988. *Annals of Epidemiology*, 2(1–2):93–99, Jan–Mar 1992.

[KW70] G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[KW71] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

[Lap95] H. Lapinleimu. *Influence of Individualized Dietary Counseling and Apolipoprotein E Polymorphism on Serum Lipoproteins in Infancy.* PhD thesis, Social Insurance Institution, Finland, 1995.

[LBJ95] J. Loncar, V. M. Barnabei, and J. W. Larsen Jr. Advent of maternal serum markers for Down syndrome screening. *Obstetrical and Gynecological Survey*, 50(4):316–320, Apr 1995.

[LEOB+96] N. Lloyd-Evans, T. J. D. O'Dempsey, I. Baldeh, O. Secka, E. Demba, J. E. Todd, T. F. Mcardle, W. S. Banya, and B. M. Greenwood. Nasopharyngeal carriage of pneumococci in Gambian children and in their families. *Pediatric Infectious Disease Journal*, 15(10):866–871, 1996.

[LS94] D. Lee and J. J. H. Shiau. Thin plate splines with discontinuities and fast algorithms from their computation. *SIAM Journal of Scientific Computation*, 15(6):1311–1330, Nov 1994.

[Luo96a] Z. Luo. Backfitting in smoothing spline ANOVA. Technical Report 96-05, Department of Statistics, Pennsylvania State University, Oct 1996.

[Luo96b] Z. Luo. *Backfitting in Smoothing Spline ANOVA, with application to historical global temperature data.* PhD thesis, Department of Statistics, University of Wisconsin, Jul 1996.

[LVJ+95] H. Lapinleimu, J. Viikari, E. Jokinen, P. Salo, T. Routi, A. Leino, T. Rönnemaa, R. Seppänen, I. Välimäki, and O. Simell. Prospective randomized trial in 1062 infants of diet low in saturated fat and cholesterol. *Lancet*, 345(8948):471–476, 1995.

[Mac94a] D. J. C. MacKay. Bayesian methods for backpropagation networks. In E. Domany, J. L. van Hemmen, and K. Schulten, editors, *Models in Neural Networks III, Chapter 6*, New York, 1994. Springer-Verlag.

[Mac94b] S. MacMahon. Cholesterol reduction and death from non-coronary causes: evidence from randomized controlled trials. *Australian and New Zealand Journal of Medicine*, 24(1):120–123, 1994.

[Mac95a] D. J. C. MacKay. Bayesian non-linear modeling for the 1993 energy prediction competition. In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, Dordrecht, 1995. Kluwer.

[Mac95b] D. J. C. MacKay. Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6:469–505, 1995.

[Man97] H. Mannila. Methods and problems in data mining. Technical report, Department of Computer Science, University of Helsinki, 1997.

[MBP+94] F. Muller, L. Bussieres, M. C. Pelissier, J. F. Oury, C. Boue, S. Uzan, and A. Boue. Do racial differences exist in second-trimester maternal hCG levels? A study of 23,369 patients. *Prenatal Diagnosis*, 14(7):633–636, Jul 1994.

[MCFH92] M. C. Mogdasy, T. Camou, C. Fajardo, and M. Hortal. Colonizing and invasive strains of Streptococcus pneumoniae in Uruguayan children: type distribution and patterns of antibiotic resistance. *Pediatric Infectious Disease Journal*, 11(8):648–652, 1992.

[Mei79] J. Meinguet. Multivariate interpolation at arbitrary points made simple. *Journal of Applied Mathematics and Physics*, 30:292–304, 1979.

[Mit97] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[MST94] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York, 1994.

[NG94] P. Niyogi and F. Girosi. On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. Technical Report A.I. Memo No.1467, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Feb 1994.

[NW92] J. D. Neaton and D. Wentworth. Serum cholesterol, blood pressure, cigarette smoking, and death from coronary heart disease. Overall findings and differences by age for 316,099 white men. Multiple risk factor intervention trial research group. *Archives of Internal Medicine*, 152(1):56–64, Jan 1992.

[PG89] T. Poggio and F. Girosi. A theory of networks for approximation and learning. Technical Report A.I. Memo No.1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Jul 1989.

[PKH+93] G. E. Palomäki, G. J. Knight, J. E. Haddow, J. A. Canick, N. J. Wald, and A. Kennard. Cigarette smoking and levels of maternal serum alpha-fetoprotein, unconjugated estriol and hCG: impact on Down syndrome screening. *Obstetrics and Gynecology*, 81(5):675–678, May 1993.

[PKH94] G. Palomäki, G. J. Knight, and J. E. Haddow. Human chorionic gonadotropin and unconjugated oestriol measurements in insulin-dependent diabetic pregnant women being screened for fetal Down syndrome. *Prenatal Diagnosis*, 14(1):65–68, Jan 1994.

[Pow93] M. J. D. Powell. Truncated Laurent expansions for the fast evaluation of thin plate splines. *Numerical Algorithms*, 5:99–120, 1993.

[PTVF92] W. H. Press, S. A. Teukolsky, W. T. Vettering, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, New York, second edition, 1992.

[SH96] M. Sidani and B. Harrod. LAPACK working note 116: Parallel matrix distribution: Have we been doing it all right? Technical report, Cray Research, A Silicon Graphics Company, Nov 1996.

[Sha95] F. Shann. The management of pneumonia in children in developing countries. *Clinical Infectious Diseases*, 21(Suppl 3):S218–225, Dec 1995.

[She68] D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings–1968 ACM National Conference*, pages 517–524, 1968.

[SM94] C. G. Small and D. L. McLeish. *Hilbert Space Methods in Probability and Statistical Inference*. John Wiley & Sons, Inc., New York, 1994.

[Sør93] U. B. Sørensen. Typing of pneumococci by using 12 pooled antisera. *Journal of Clinical Microbiology*, 31(8):2097–2100, Aug 1993.

[TA96] C. Thomas-Agnan. Computing a family of reproducing kernels for statistical applications. *Numerical Algorithms*, 13:21–32, 1996.

[TDR92] K. L. Tucker, G. E. Dallal, and D. Rush. Dietary patterns of elderly Boston-area residents defined by cluster analysis. *Journal of the American Dietetic Association*, 92(12):1487–1491, 1992.

[Ver99] A. Verri. *Lecture notes: Networks for Learning: Regression and Classification*. MIT: Center for Biological and Computational Learning, 1999.

[VW83] M. A. Villalobos and G. Wahba. Multivariate thin plate spline estimates for the posterior probabilities in the classification problem. *Communications in Statistics: Theory and Methods*, 12(13):1449–1479, 1983.

[Wah78] G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of Royal Statistical Society, Series B*, 40:364–372, 1978.

[Wah83] G. Wahba. Bayesian “confidence intervals” for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B*, 45(1):133–150, 1983.

[Wah85] G. Wahba. A comparison of GCV and GLM for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 13(4):1378–1402, 1985.

[Wah90] G. Wahba. *Spline models for observational data*, volume 59 of *the CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

[Wah94] G. Wahba. Generalization and regularization in nonlinear learning systems. Technical Report 921, Department of Statistics, University of Wisconsin, May 1994.

[Wan94] Y. Wang. *Smoothing Spline Analysis of Variance of Data from Exponential Family*. PhD thesis, Department of Statistics, University of Wisconsin, Jul 1994.

[Wan95a] Y. Wang. GRKPACK: Fitting smoothing spline ANOVA models for exponential families. Technical Report 942, Department of Statistics, University of Wisconsin, Jan 1995.

[Wan95b] Y. Wang. Odds ratio estimation in Bernoulli smoothing spline ANOVA model. Technical Report 946, Department of Statistics, University of Wisconsin, Apr 1995.

[Wan96] Y. Wang. Smoothing spline models with correlated random errors. Technical Report 966, Department of Statistics, University of Wisconsin, Aug 1996.

[WDCC94] N. J. Wald, J. W. Densem, R. Cheng, and S. Collishaw. Maternal serum free alpha- and free beta-human chorionic gonadotrophin in pregnancies with insulin-dependent diabetes mellitus: implications for screening for Down's syndrome. *Prenatal Diagnosis*, 14(9):835–837, Sep 1994.

[WGB97] Y. Wang, W. Guo, and M. B. Brown. Spline smoothing for bivariate data with applications to associate between hormones. Technical report, Department of Biostatistics, University of Michigan, May 1997.

[WGWC93] G. Wahba, C. Gu, Y. Wang, and R. Chappell. Soft classification, a.k.a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance. Technical Report 899, Department of Statistics, University of Wisconsin, Jan 1993.

[WJ97] A. K. E. Wirfält and R. W. Jeffery. Using cluster analysis to examine dietary patterns: Nutrient intakes, gender, and weight status differ across pattern clusters. *Journal of the American Dietetic Association*, 97(3):272–279, 1997.

[WJGG94] G. Wahba, D. R. Johnson, F. Gao, and J. Gong. Adaptive tuning of numerical weather prediction models: Part I: randomized GCV and related methods in three and four dimensional data assimilation. Technical Report 920, Department of Statistics, University of Wisconsin, Apr 1994.

[WKHM97] N. J. Wald, A. Kennard, A. Hackshaw, and A. McGuire. Antenatal screening for Down's syndrome. *Journal of Medical Screening*, 4(4):181–246, 1997.

[WL95] G. Wahba and Z. Luo. Smoothing spline ANOVA fits for very large, nearly regular data sets, with application to historical global climate data. Technical report, Department of Statistics, University of Wisconsin, Dec 1995.

[WW93] G. Wahba and Y. Wang. Behavior near zero of the distribution of GCV smoothing parameter estimates. Technical Report 910, Department of Statistics, University of Wisconsin, Dec 1993.

[WW94] Y. Wang and G. Wahba. Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian "confidence intervals". Technical Report 913, Department of Statistics, University of Wisconsin, Jan 1994.

[WW96] N. J. Wald and H. C. Watt. Serum markers for Down's syndrome in relation to number of previous births and maternal age. *Prenatal Diagnosis*, 16(8):699–703, Aug 1996.

[WWG+93] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Structured machine learning for "soft" classification with smoothing spline ANOVA and stacked tuning, testing and evaluation. Technical Report 909, Department of Statistics, University of Wisconsin, Dec 1993.

[WWG+95a] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. Technical report, Department of Statistics, University of Wisconsin, May 1995.

[WWG+95b] Y. Wang, G. Wahba, C. Gu, R. Klein, and B. Klein. Using smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy. Technical Report 956, Department of Statistics, University of Wisconsin, Dec 1995.

[Xia96] D. Xiang. Model fitting and testing for non-Gaussian data with large data sets. Technical Report 957, Department of Statistics, University of Wisconsin, Jan 1996.

[XW94] D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. Technical Report 930, Department of Statistics, University of Wisconsin, Sep 1994.

[XW95] D. Xiang and G. Wahba. Testing the generalized linear model null hypothesis versus "smooth" alternatives. Technical Report 953, Department of Statistics, University of Wisconsin, Oct 1995.