PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Pearce

First name: Nathan Other name/s: Douglas

Abbreviation for degree as given in the University calendar: PhD (Mathematics)

School: School of Mathematics and Statistics Faculty: Science

Title: Kernel Methods for Data Analysis

Abstract 350 words maximum: (PLEASE TYPE)

In this thesis, statistical inference is made using reproducing kernel methods. Kernel methods are a versatile tool, and allow a seamless transition from parametric to nonparametric methods. A flagship example of the use of kernel methods is the support vector machine.

An effective and elegant method for classification problems, the support vector machine is one of many applications of kernel methods. By embedding penalised splines within kernel methods, the support vector machine is given an interpretable, even additive structure. Addressed in detail are the large scale computational issues involved with support vector machines.

Kernel methods are further used to make explicit links to longitudinal data analysis. In doing so, the broad kernel machine methodology can incorporate the repeated measurements of longitudinal data analysis. Additionally, the links are made explicit between the degrees of freedom and kernel methods.

Bayes methodology is addressed with kernel methods. A variational Bayes approach is used for linear mixed models and generalised linear mixed models. The approach is shown to be computationally efficient. Moreover, classical methods such as restricted maximum likelihood and penalised quasi-likelihood are shown to be special cases of variational Bayes.

The final chapter of this thesis addresses the issue of model selection with only minimal assumptions. The robustness of such an approach is verified through extensive testing.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

.....
Signature f Witness' Date

The University recognises that there-may-be exceptional-circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional

FOR OFFICE USE ONLY Date of Completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

COPYRIGHT STATEMENT

'I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.
I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International (this is applicable to doctoral theses only).
I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.'

Signed

Date

AUTHENTICITY STATEMENT

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

Signed ..

Date ..

Kernel Methods for Data Analysis

28th February, 2010

A thesis presented to

The School of Mathematics and Statistics
The University of New South Wales

in fulfilment of the thesis requirement
for the degree of

Doctor of Philosophy

by NATHAN DOUGLAS PEARCE

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ..

Date ..

Abstract

In this thesis, statistical inference is made using reproducing kernel methods. Kernel methods are a versatile tool, and allow a seamless transition from parametric to non-parametric methods. A flagship example of the use of kernel methods is the support vector machine.

An effective and elegant method for classification problems, the support vector machine is one of many many applications of kernel methods. By embedding penalised splines within kernel methods, the support vector machine is given an interpretable, even additive structure. Addressed in detail are the large scale computational issues involved with support vector machines.

Kernel methods are further used to make explicit links to longitudinal data analysis. In doing so, the broad kernel machine methodology can incorporate the repeated measurements of longitudinal data analysis. Additionally, the links are made explicit between the degrees of freedom and kernel methods.

Bayes methodology is addressed with kernel methods. A variational Bayes approach is used for linear mixed models and generalised linear mixed models. The approach is shown to be computationally efficient. Moreover, classical methods such as restricted maximum likelihood and penalised quasi-likelihood are shown to be special cases of variational Bayes.

The final chapter of this thesis addresses the issue of model selection with only minimal assumptions. The robustness of such an approach is verified through extensive testing.

Supervisor: Professor Matt P. Wand.

Acknowledgements

I take this opportunity to thank a few of the people who have played a role in the preparation of this thesis. By far the greatest thanks must go to Prof. Matt Wand, my supervisor, for his guidance and support over my candidature. His deep insight and close attention to detail have been invaluable.

I would also like to thank Helen Armstrong, Jan Baldeaux, Hugo Bowne-Anderson, Prof. Inge Koch, John Ormerod, Michael Roper and Ben Warhurst for the many helpful discussions. This work has benefitted considerably from their comments and ideas.

I would also like to acknowledge financial support and grants of both MASCOS and of the School of Mathematics and Statistics. This research was also supported financially by an Australian Postgraduate Award. Finally, my family provided the love and support that they have given over my time at university. Without this, I would not have made it this far.

Nathan Pearce, November 2009.

Contents

Statements of Authorship and Originality

This thesis represents the work of N. D. Pearce, with supervision by Prof. M. P. Wand. Chapters 2 and 3 of this thesis are based on the publications:

Pearce, N. D. and Wand, M. P. (2006) Penalized Splines and Reproducing Kernel Methods, *The American Statistician*, **60**, 233–240.

Pearce, N. D. and Wand, M. P. (2009) Explicit Connections Between Longitudinal Data Analysis and Kernel Machines, *Electronic Journal of Statistics*, **3**, 797–823.

"I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgment is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the projects design and conception or in style, presentation and linguistic expression is acknowledged."

Signed ...

Date

List of Figures

List of Tables

List of Algorithms

Chapter One

Introduction

This thesis is devoted to the use of kernel methods for the analysis of data. Many of the topics explored are on the interplay between traditional statistical approaches and those of machine learning. A particular focus is therefore on nonparametric regression and classification tasks. For classification tasks, the Machine Learning literature focuses on decision functions, while traditional statistics tends to focus more on the conditional probabilities. Such conditional probabilities may be obtained through either a frequentist or Bayesian framework.

There are some major drivers in the push towards effective, large-scale machine learning capabilities. Firstly, there is an ever increasing availability of computational power. Secondly, there has been a massive increase in the availability of large-scale, quality data to analyse. The data come from such areas as detecting credit card fraud, search engine optimisation, medical applications, handwritten document recognition and bioinformatics.

Kernel methods are an adaptive and versatile tool. Recent years have seen kernel methods used extensively in statistics, medicine, computer science and engineering. The support vector machine is one such example of the use of kernel methods to produce high quality inference. On large data sets, there are also computational advantages to using support vector machines. With Vapnik and Lerner (1963), Wahba (1969) and Boser, Guyon and Vapnik (1992), support vector machines were derived through geometric arguments. Nowadays, support vector machines are investigated primarily as special cases of kernel machines (Steinwart and Christmann, 2008). An important question of support vector machines, and of kernel machines generally, is the optimal choice of parameters.

An attractive aspect of the support vector machine is the lack of assumptions on the underlying probability distribution. Traditionally, statistical models impose normality and homoscedasticity, however real-life data will rarely hold to such high standards. We need to be able to model the data without imposing unnecessary assumptions – it is the data that is to guide the model.

1.1 Thesis Overview

The thesis is structured around topics on support vector machines, penalised splines, longitudinal data analysis, regression, classification, quantile regression, robust regression, variational Bayes, degrees of freedom, optimisation techniques and kernel methods. The rest of the thesis is divided into six chapters.

Chapter 2: Penalised Splines and Reproducing Kernel Methods

This chapter is largely based on Pearce and Wand (2006). We show how penalised splines are embedded in the class of reproducing kernel methods. Penalised splines have a simple structure that may be used in conjunction with the support vector machine and other reproducing kernel methods. Key computational benefits are achieved without significant losses in accuracy.

Chapter 3: Explicit Links Between Longitudinal Data Analysis and Kernel Machines

Much of the material in Chapter 3 originally appeared in Pearce and Wand (2009). Longitudinal data is characterised by repeated measurements of individuals over time. The chapter gives explicit links between longitudinal data analysis and kernel machines. Indeed, it is shown that many longitudinal data analysis techniques are special types of kernel machines. The links shown in this chapter allow kernel machine methodology to incorporate repeated measurements.

Chapter 4: Semiparametric Regression via Variational Bayes

In here we present a variational Bayes approach to parameter selection. The Bayesian approach will often lead to intractable integrals, but with variational Bayes, the objective becomes obtainable. We show that there exists a close relationship between variational Bayes, and classical approaches such as restricted maximum likelihood and penalised quasi-likelihood.

Chapter 5: Impact of Kernel Parameters on Degrees of Freedom

The degrees of freedom of a model is an established concept in the Statistical literature. The degrees of freedom give a intuitive and scale free assessment of the amount of fitting applied. In this chapter, we investigate the relationship between the degrees of

freedom and the model parameters. The degrees of freedom is extended to encompass such instances as quantile regression and support vector machines.

Chapter 6: Active Set Optimisation of Support Vector Machines

The fast and reliable training of support vector machines remains a topic of much interest in the Machine Learning community. With Chapter 6, a large scale optimisation algorithm is detailed for support vector machine training. The algorithm, active set support vector machine (AS-SVM), allows for the fast training of an support vector machines despite the large scale nature of the problem. Experimental evidence of time comparisons with existing methods show significant improvements in the time it takes to train a support vector machine.

Chapter 7: On Model Validation and Selection

Chapter 7 presents a new method for both model validation and selection. Model validation involves the testing of a parametric null against a nonparametric alternative. Naturally, the question is then raised, if not parametric, how do we decide what the nonparametric fit should be? This chapter puts forward a novel criterion, called the parameter information criterion. The parameter information criterion gives a regression fit to data, and does so with minimal assumptions.

Chapter Two

Penalised Splines and Reproducing Kernel Methods

2.1 Introduction

The mid-1990s saw the parallel emergence of two important areas of data analysis research.[1] Although built on ideas that had accumulated over the previous decades, they were both ignited by several key papers and results. One area of data analysis research, as found in the Statistics literature, is a nonparametric regression technique known as *penalised splines*. The other area, *reproducing kernel methods*, are founded primarily in the Machine Learning literature, and have been used in a broad range of applications. This chapter builds a bridge between these two sets of literature.

The main stimulus for the emergence of penalised spline research was Eilers and Marx (1996), while another key reference is Hastie (1996). The essential underlying ideas have been around for much longer, such as those given in Schoenberg (1969); Parker and Rice (1985) and Wahba (1990, Chapter 7). The focus of this penalised spline research is the generalisation of ordinary smoothing splines to knot sequences different from, and usually much smaller than, the observed predictor variables. Hastie (1996) and Marx and Eilers (1998) illustrated the benefits for additive models. Brumback, Ruppert and Wand (1999) identified simple mixed model representations which allowed, for example, straightforward incorporation of longitudinal data into nonparametric regression. Other developments include simpler incorporation of measurement error (Berry, Carroll and Ruppert, 2002) and geostatistical data (Kammann and Wand, 2003). Much of the work on penalised splines up until about 2002 is summarised in the book by Ruppert, Wand and Carroll (2003).

A major stimulus in the emergence of both support vector machines and *reproducing kernel methods* was Boser, Guyon and Vapnik (1992), with Cortes and Vapnik (1995) being

[1]This chapter is based on the publication: Pearce, N. D. and Wand M. P. (2006). Penalized Splines and Reproducing Kernel Methods. *The American Statistician*, **60**, 233–240.

another key reference. An essential idea behind the early development of the support vector machine, margin maximisation, is much older, with Vapnik and Lerner (1963) and Wahba (1969). Around the same time, with Aizerman, Braverman and Rozonoer (1964), reproducing kernel methods were researched. The support vector machine has since blossomed into a huge literature, and has been the main catalyst for what have become known as *reproducing kernel methods*, or simply *kernel methods*, in machine learning. These titles should not be confused with kernel smoothing methods in the nonparametric regression literature (e.g., Wand and Jones, 1995).

A comprehensive overview of reproducing kernel methods is provided by Burges (1998); Evgeniou, Pontil and Poggio (2000); Cristianini and Shawe-Taylor (2000); Schölkopf and Smola (2002); Berlinet and Thomas-Agnan (2004) and Steinwart and Christmann (2008). Before the emergence of support vector machines, reproducing kernel methods were prominent in the nonparametric regression literature as a framework for smoothing spline methodology, as summarised in Wahba (1990). However, the adoption of these ideas by the machine learning community has widened the scope of reproducing kernel methods considerably.

This chapter shows how penalised splines are embedded in the class of reproducing kernel methods and thus builds a bridge between these two bodies of research. Reproducing kernel representation of penalised splines is relatively simple compared with smoothing splines representation. It is envisaged that support vector machine research has the most to gain from this connection. The reduced knot aspect of penalised splines allows for big savings in computational complexity, as we explain in Section 2.7. This last feature is particularly relevant since sample sizes in classification applications are subject to continual increase. In addition, much of the support vector machine research is done within the machine learning discipline, and largely oblivious to many statistical principles such as interpretation, model building, diagnosis, low-dimensional structure and proper accounting for data dependencies. Kernels based on penalised splines offer the opportunity to incorporate some of these principles more straightforwardly than commonly used kernels. Similar recent work has been done using the ideas of smoothing spline analysis of variance; see Lin and Zhang (2006) and Lee, Kim, Lee and Koo (2006).

An illustration of a support vector machine classifier which utilises low-dimensional structure and is immediately interpretable is given in Figure 2.1. It arises from use of additive model penalised spline kernels (Sections 2.5.3 and 2.7) to build a classifier for

the "spam" data, described in Hastie, Tibshirani and Freidman (2001), with spam e-mail messages coded as $+1$ and ordinary messages coded as -1. Each panel shows the slice of the classification surface for the labelled predictor, with all other predictors set to their medians. It is seen, for example, that frequency of the word "free" has a monotonic effect on classification while frequency of exclamation marks (`ch!`) has a non-monotonic effect.

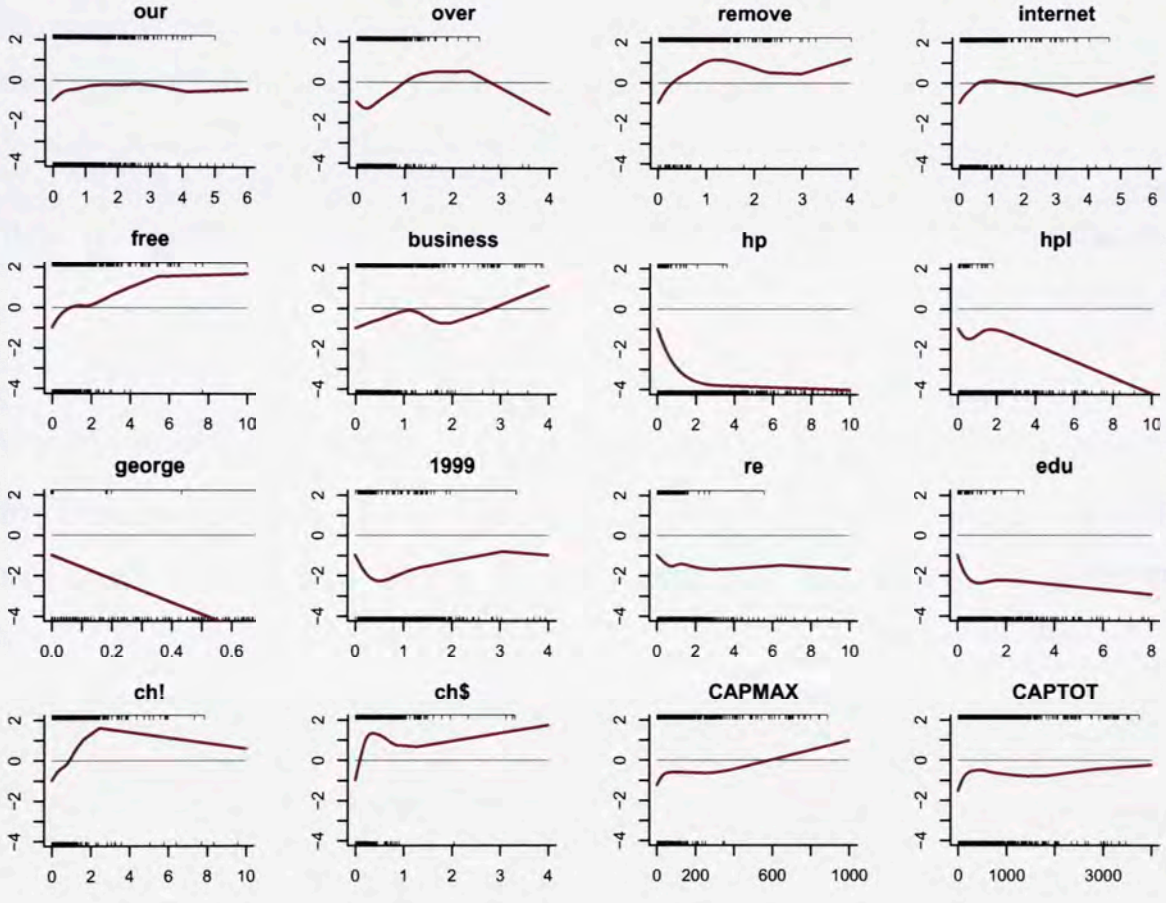


Figure 2.1. *Visualisation of a penalised spline support vector classifier for the "spam" data. Each panel shows the slice of the classifier with all other predictors set to their medians. The tick-marks show the predictor values: spam e-mail messages along the top, ordinary e-mail messages along the bottom.*

The next section provides a brief description of the simplest version of penalised splines. Section 2.3 describes the basics of reproducing kernel methods. The link between these two concepts is laid out in Section 2.4. Various extensions are treated in Section 2.5. Alternatives to reproducing kernel methods are given in Section 2.6. Section 2.7 is devoted to the special case of support vector machines and advantages of the penalised spline approach are explained. We close with some summary remarks for this

chapter in Section 2.8.

2.2 Penalised Splines

For the moment we will consider only the regression situation where the observed data are $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $1 \leq i \leq n$, and both variables are continuous. The simplest penalised spline model is

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+ + \varepsilon_i. \tag{2.1}$$

Here $x_+ = \max(0, x)$, $\kappa_1, \ldots, \kappa_K$ are a set of knots over the range of the x_i's and the ε_i are independent mean zero random variables with common variance σ_ε^2. Fitting is typically performed via penalised least squares:

$$\min_{\beta, u} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+ \right)^2 + \lambda \sum_{k=1}^{K} u_k^2 \right\} \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\mathsf{T}$, $\boldsymbol{u} = (u_1, \ldots, u_K)^\mathsf{T}$, and where $\lambda > 0$ is a smoothing parameter. The smoothing parameter controls the trade-off between bias and overfitting.

A matrix formulation of (2.1) is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon} \tag{2.3}$$

where

$$\boldsymbol{X} = [1 \;\; x_i]_{1 \leq i \leq n}, \quad \boldsymbol{Z} = \underset{1 \leq k \leq K}{[(x_i - \kappa_k)_+]_{1 \leq i \leq n}}$$

and $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ contain the respective subscripted variables. Thus (2.2) becomes

$$\min_{\beta, u} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\|^2 + \lambda \|\boldsymbol{u}\|^2 \right) \tag{2.4}$$

where $\|v\| = \sqrt{v^\mathsf{T} v}$ denotes the norm of the vector v. The solution is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}, \quad \widehat{\boldsymbol{u}} = \boldsymbol{Z}^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \tag{2.5}$$

with $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \lambda \boldsymbol{I}$. The notation of (2.3) suggests a linear mixed model and (2.5) corresponds exactly to best linear unbiased prediction if $\boldsymbol{u}$ is treated as a random effects vector with covariance matrix $(\sigma_\varepsilon^2 / \lambda)\boldsymbol{I}$ (Brumback *et al.*, 1999).

While we use the term "penalised splines", it should be pointed out that there are several alternative names for what is essentially the same general approach. These include low-rank splines, P-splines, pseudosplines and reduced knot splines.

2.3 Kernel Machines

In this section, we provide definitions and some fundamental theorems of reproducing kernel methods. This facilitates the kernel representation of penalised splines in the next section, and representations called upon in later chapters. Reproducing kernel methods are performed within the functional analytic structure known as a reproducing kernel Hilbert space (RKHS). The theory of RKHSs was developed by Kolmogorov (1941) and Aronszajn (1950). Contemporary summaries include Wahba (1990); Evgeniou, Pontil and Poggio (2000); Schölkopf and Smola (2002) and Steinwart and Christmann (2008). Of particular relevance to penalised splines are penalisations over subspaces that are based on projection operators. Relevant background material on Hilbert space projection theory may be found in Simmons (1963) and Berg, Christensen and Ressel (1984).

2.3.1 Review of Reproducing Kernel Hilbert Spaces

We start with some fundamental definitions and results, beginning with the following definition.

Definition 2.1. *Let $\mathcal{X}$ be a non-empty set. A function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** on $\mathcal{X}$ if there exists a Hilbert space, $\mathcal{H}$, and a map $\Phi\colon \mathcal{X} \to \mathcal{H}$ such that for all $s, t \in \mathcal{X}$,*

$$k(s,t) = \langle \Phi(s), \Phi(t) \rangle.$$

*The map Φ is the **feature map** and $\mathcal{H}$ is the **feature space** of k.*

It follows that a kernel k must be symmetric, that is $k(s,t) = k(t,s)$ for all $s, t \in \mathcal{X}$. Moreover, the function $k(s,\cdot)\colon \mathcal{X} \to \mathbb{R}$ has $k(s,\cdot) \in \mathcal{H}$ for all $s \in \mathcal{X}$. For a given kernel, neither the feature map nor feature space are unique. We wish to determine whether a function k is a kernel. It may not be straightforward to find a feature space and feature map for the kernel. The following definition of a positive definite function is often helpful in determining whether a function is a kernel.

Definition 2.2. *A function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called **positive definite** if it is symmetric, and for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$, we have*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0. \tag{2.6}$$

The $n \times n$ matrix, $\boldsymbol{K}$, with entries $k(x_i, x_j)$, $1 \leq i, j \leq n$, is called the *Gram* matrix. For a function k being positive definite, the inequality in (2.6) is equivalent to the Gram

matrix being positive semidefinite. The following theorem, given as stated by Steinwart and Christmann (2008, Theorem 4.16), shows the equivalence between positive definite functions and kernels.

Theorem 2.3. *A function is a kernel if and only if it is positive definite.*

We now have necessary and sufficient conditions for a function to be a kernel. In particular, a kernel may be expressed as either as a positive definite function, or as an inner product, with feature space and feature map. For a kernel, however, it remains that the feature space and feature map are not unique. For an element $x \in \mathcal{X}$, a Dirac functional $\delta_x \colon \mathcal{H} \to \mathbb{R}$ is such that $\delta_x(f) \equiv f(x)$. That is, δ_x maps $f \in \mathcal{H}$ to the value f has at x. As a linear functional, δ_x is bounded if and only if it is continuous (e.g., Dudley, 2002, page 190). Making use of the Dirac functional, we have the following fundamental definition.

Definition 2.4. *A Hilbert space $\mathcal{H}$ is called a* ***reproducing kernel Hilbert space*** *over $\mathcal{X}$ if for all $x \in \mathcal{X}$ the Dirac functional δ_x is a bounded linear functional.*

A well known result is that not only does every RKHS have a unique kernel, but every kernel has a unique RKHS. This result is expressed in the following theorem, proven by Aronszajn (1950) and attributed to E. H. Moore.

Theorem 2.5. *Assume $\mathcal{H}$ is an RKHS over $\mathcal{X}$, and $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ has the property*

$$k(s,t) = \langle \delta_s, \delta_t \rangle_{\mathcal{H}}, \quad \text{for all } s,t \in \mathcal{X}.$$

Then $\mathcal{H}$ uniquely determines k, and k uniquely determines $\mathcal{H}$.

Due to this uniqueness property, we can denote the RKHS by $\mathcal{H}_k$. Of all Hilbert spaces, only for an RKHS does $\delta_x(f) = 0$ for all $x \in \mathcal{X}$ imply $\|f\|_{\mathcal{H}} = 0$. The adjective "reproducing" arises from the important result

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = \delta_x(f) = f(x), \quad \text{for all } f \in \mathcal{H}_k. \tag{2.7}$$

In particular,

$$\langle k(s,\cdot), k(t,\cdot) \rangle_{\mathcal{H}_k} = \langle \delta_s, \delta_t \rangle_{\mathcal{H}_k} = k(s,t), \quad \text{for all } s,t \in \mathcal{X}.$$

A topological space is called *separable* if it contains a countable dense subset. If $\mathcal{X}$ is separable, and k is a continuous kernel on $\mathcal{X}$, then $\mathcal{H}_k$ is also separable (Steinwart, 2001).

The steps for a separable RKHS construction from k are:

Name	$k(s,t)$
Linear	$s^T t$
Polynomial	$(1+s^T t)^l$
Gaussian	$\exp(-\gamma\|s-t\|^2)$
Laplacian	$\exp(-\gamma\|s-t\|)$

Table 2.1. *Examples of commonly used kernels with input space* $\mathcal{X} = \mathbb{R}^d$. *We have* $\gamma > 0$ *and* $l \in \mathbb{N}$.

i) Determine the eigen-decomposition of the kernel, $k(s,t) = \sum_{j=0}^{q} \lambda_j \phi_j(s)\phi_j(t)$, with $\{q \in \mathbb{N} \cup \infty\}$. This series is assumed to be well-defined (e.g., uniformly convergent).

ii) Define the pre-Hilbert space (i.e., an inner product space), $\mathcal{H}_{pre}$, of real-valued functions on $\mathcal{X}$:

$$\mathcal{H}_{pre} = \left\{ f : f = \sum_{j=0}^{q} a_j \phi_j, \quad \text{such that} \quad \sum_{j=0}^{q} a_j^2 / \lambda_j < \infty \right\}.$$

iii) Endow $\mathcal{H}_{pre}$ with the inner product

$$\left\langle \sum_{j=0}^{q} a_j \phi_j, \sum_{j=0}^{q} a'_j \phi_j \right\rangle_{\mathcal{H}_{pre}} = \sum_{j=0}^{q} a_j a'_j / \lambda_j.$$

iv) Complete the pre-Hilbert space.

A more general construction, allowing for non-separable RKHS, is given by Steinwart and Christmann (2008). Trivially, the RKHS norm of $f = \sum_{j=0}^{q} a_j \phi_j$ in $\mathcal{H}_{pre}$ is

$$\|f\|_{\mathcal{H}_k} \equiv \sqrt{\langle f, f \rangle_{\mathcal{H}_{pre}}} = \left(\sum_{j=0}^{q} a_j^2 / \lambda_j \right)^{1/2}.$$

We will see that penalised splines give rise to separable, finite dimensional RKHSs. Examples of kernels on $\mathbb{R}^d$ are given in Table 2.1. Linear and polynomial kernels have finite dimensional RKHSs. Gaussian and Laplacian kernels have separable RKHSs; explicit descriptions of their RKHSs are given by Bach and Jordan (2002) and Steinwart, Hush and Scovel (2006).

Name	$\mathcal{Y}$	$\mathcal{L}(a,b)$
Squared error	$\mathbb{R}$	$(a-b)^2$
Absolute value	$\mathbb{R}$	$\lvert a-b\rvert$
ϵ-insensitive	$\mathbb{R}$	$(\lvert a-b\rvert-\epsilon)_+$
Heaviside	$\{-1,1\}$	$I_{(ab<0)}+\frac{1}{2}I_{(ab=0)}$
Bernoulli log-likelihood	$\{-1,1\}$	$\log(1+e^{-ab})$
Hinge loss	$\{-1,1\}$	$(1-ab)_+$

Table 2.2. *Examples of commonly used loss functions, together with appropriate closed domains. We have* $\epsilon > 0$ *and* $(b)_+ \equiv \max(0,b)$.

2.3.2 Loss Functions and Objectives

For regression and binary classification tasks, we wish to find a function[2] $f\colon \mathcal{X} \to \mathbb{R}$, so that $f(x)$ is a prediction of y at x. We measure how good such a predictor is at x through the *loss* function.

Definition 2.6. *Let* $\mathcal{Y} \subset \mathbb{R}$ *be closed. A* ***loss*** *function on* $\mathcal{Y}$ *is a function* $\mathcal{L}\colon \mathcal{Y}\times\mathbb{R} \to [0,\infty)$. *The loss function is* ***(strictly) convex*** *if* $\mathcal{L}(y,\cdot)\colon \mathbb{R}\to[0,\infty)$ *is (strictly) convex for all* $y \in \mathcal{Y}$.

We would interpret $\mathcal{L}(y,f(x))$ as being the cost, or loss, of predicting y by $f(x)$. A small $\mathcal{L}(y,f(x))$ is preferred, indicating that a good prediction of y at x has been made. Some examples of common loss functions are given in Table 2.2. The most commonly used loss function is the squared error loss. The choice of loss can be made in consideration of the model application.

Empirical risk minimisation (ERM) over $\mathcal{H}_k$ involves the directly minimising the average loss over the observed data,

$$\min_{f\in\mathcal{H}_k}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i,f(x_i))\right\}. \tag{2.8}$$

For many loss functions, ERM is an *ill-posed* problem (Tarantola, 2005). It is not clear if even such a minimiser exists. For strictly monotonic loss, such as the Bernoulli log-likelihood loss, it is well known that there are data sets for which there is no empirical risk minimiser. A characterisation of such data sets is given by Silvapulle (1981) and

[2]Broader functional forms would encompass instances such as multiclass classification (e.g., Lee, Lin and Wahba, 2004; Zhu and Hastie, 2005) and unsupervised learning (e.g., Steinwart, Hush and Scovel, 2005).

Albert and Anderson (1984). We now provide a more general characterisation of the existence of an empirical risk minimiser.

Theorem 2.7. *Let $\mathcal{L}$ be convex. Then there exist some $f \in \mathcal{H}_k$ such that:*

i) $f(x_i) \geq 0$ *for all* $\mathcal{L}(y_i, \cdot)$ *not monotonically decreasing,*

ii) $f(x_i) \leq 0$ *for all* $\mathcal{L}(y_i, \cdot)$ *not monotonically increasing, and*

iii) $f(x_i) \neq 0$ *for some* $\mathcal{L}(y_i, \cdot)$ *strictly monotonic,*

if and only if there does not exist an empirical risk minimiser over $\mathcal{H}_k$.

The proof of Theorem 2.7 is given in Appendix 2.A.1. For convex loss functions such as squared error, ϵ-insensitive and hinge loss, $\mathcal{L}(y, \cdot)$ is not strictly monotonic for all $y \in \mathcal{Y}$. As a consequence of Theorem 2.7, for each of these losses, a fit to the ERM exists.

For non-convex loss, the resulting ERM optimisation in (2.8) may be NP-hard. In particular, for Heaviside loss, minimising the ERM is NP-hard (Minsky and Papert, 1988). The hinge loss serves as a convex upper bound for the Heaviside loss. ERM with hinge loss then gives an upper bound to ERM with Heaviside loss, and ensures that the minimisation problem is tractable (Lin, Lee and Wahba, 2002).

2.3.3 Regularisation and Representation

In practice, ERM can lead to overfitting. The fit may follow the data too closely, and extrapolate poorly to new observations. It is a standard procedure to minimise the empirical loss and squared RKHS norm of the fit. The trade-off is controlled by a *smoothing* parameter, $\lambda > 0$. A fit over $\mathcal{H}_k$, with respect to (x_i, y_i), $1 \leq i \leq n$, $\mathcal{L}$ and λ, is any solution to

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}_k}^2 \right\}. \tag{2.9}$$

Such a combination of loss and RKHS norm penalty is known as a *kernel machine*. Discovered by Cortes and Vapnik (1995), the use of hinge loss in (2.9) results in what is known as support vector classification, while using ϵ-insensitive loss results in support vector regression (e.g., Drucker *et al.*, 1997). Collectively, support vector classification and regression are known as support vector machines. Kernel machines with squared error loss (e.g., Suykens *et al.*, 2002) include popular statistical methods such as kriging (e.g., Cressie, 1993; Stein, 1999), smoothing splines (e.g., Wahba, 1990; Green and Silverman, 1994) and additive models (e.g., Hastie and Tibshirani, 1990). Recently Zhu

and Hastie (2005) explored the use of binomial log-likelihood loss in the kernel machine framework and coined the term *kernel logistic regression*.

The following theorem is often useful when fitting a kernel machine, and was first shown for least squares loss by Kimeldorf and Wahba (1971).

Theorem 2.8 (Representer Theorem I). *Let f be a fit over $\mathcal{H}_k$. Then f admits a representation of the form*

$$f(x) = \sum_{i=1}^{n} c_i k(x, x_i). \tag{2.10}$$

for some $c_i \in \mathbb{R}$, $1 \leq i \leq n$.

The representation of the fit in (2.10) is known as the *dual form* of the solution. The c_i are dependent on the data and the choice of $\mathcal{L}$ and λ. A corollary for the uniqueness and existence is given by Steinwart and Christmann (2008, pages 168 and 201).

Corollary 2.9. *If the loss is continuous, then a fit exists. If the loss is convex, then any fit is unique.*

Even though the fit may be unique, the choice of $c_1, \ldots, c_n$ in the dual form need not be. The "kernel trick" is that we do not need to calculate the eigenfunctions of k in order to find the dual form. The kernel trick is a popular theme in kernel methods, and has allowed many linear algorithms to be easily converted into non-linear algorithms (Schölkopf and Smola, 2002). The corresponding primal form of the solution is a linear combination of the eigenfunctions. When the eigenfunctions are easily calculated, as with the linear kernel, the primal form may offer a simpler and more intuitive form of the solution.

It is often desirable that certain functions in $\mathcal{H}_k$ are unpenalised. Let $\mathcal{H}_0$ be such a subspace of $\mathcal{H}_k$ for which penalisation is not desired. Mathematically, this means that fits over $\mathcal{H}_0$ are found by empirical risk minimisation. Let $\mathcal{H}_1 = \mathcal{H}_0^{\perp}$ be the orthogonal complement of $\mathcal{H}_0$ in $\mathcal{H}_k$, so that $\mathcal{H}_k = \mathcal{H}_0 \oplus \mathcal{H}_1$. The projection operator $P_1 \colon \mathcal{H}_k \to \mathcal{H}_k$ denotes the orthogonal projection onto $\mathcal{H}_1$. It can be shown (e.g., Aronszajn, 1950) that $\mathcal{H}_0$ and $\mathcal{H}_1$ are reproducing kernel Hilbert spaces in their own right, with kernels k_0 and k_1 such that $k_0 + k_1 = k$. With respect to the null space $\mathcal{H}_0$, loss function $\mathcal{L}$, and smoothing parameter λ, we define fits according to

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda \left\| P_1 f \right\|_{\mathcal{H}_k}^2 \right\}. \tag{2.11}$$

The following theorem and corollary show the representation and uniqueness of fits (Kimeldorf and Wahba, 1971; Schölkopf, Herbrich, Smola and Williamson, 2001).

Theorem 2.10 (Representer Theorem II). *Let f be a fit over $\mathcal{H}_k$ with respect to null space $\mathcal{H}_0 = \text{span}\{\psi_0, \ldots, \psi_p\}$, with $\mathcal{H}_1 = \mathcal{H}_0^{\perp}$. Then f admits a dual representation of the form*

$$f(x) = \sum_{i=0}^{p} \beta_i \psi_i(x) + \sum_{i=1}^{n} c_i k_1(x, x_i)$$

for some $\beta_i \in \mathbb{R}$, $0 \leq i \leq p$ and $c_i \in \mathbb{R}$, $1 \leq i \leq n$.

Corollary 2.11. *If $\mathcal{L}$ is convex then $P_1 f$ is unique. If $\mathcal{L}$ is strictly convex, then $f(x_i)$ is unique for any $1 \leq i \leq n$. Furthermore, if the $n \times (p+1)$ matrix $[\psi_{j-1}(x_i)]_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p+1}}$ has rank $p+1$, then f is unique.*

Corollary 2.11 shows sufficient conditions for the uniqueness of a solution. The corollary does not imply the existence of a solution. From Kimeldorf and Wahba (1971) we know that a solution to (2.11) exists for squared error loss. For convex loss, we now present necessary and sufficient conditions for the existence of a fit.

Theorem 2.12 (Existence of a solution). *Let $\mathcal{L}$ be a convex loss. Then there exists a fit with respect to null space $\mathcal{H}_0$ if and only if there exists an empirical risk minimiser over $\mathcal{H}_0$.*

A proof of Theorem 2.12 is given in Appendix 2.A.2. Theorem 2.12 shows the existence of a solution to the projected RKHS minimisation (2.11) is equivalent to existence of the empirical risk minimiser over the null space (2.8). For squared error, ϵ-insensitive and hinge loss, amongst others, the existence of a solution is guaranteed.

Following Wahba (1990, Chapter 10), we consider changing of the norm of the Hilbert space. For RKHS $\mathcal{H}_k = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_r$, let $\mathcal{H}_{k'}$ be an RKHS, and $L \colon \mathcal{H}_k \to \mathcal{H}_{k'}$ a linear operator, such that for all $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$:

$$f(x) = Lf(x) \quad \text{and} \quad \|Lf\|_{\mathcal{H}_{k'}}^2 = \lambda_1 \|P_1 f\|_{\mathcal{H}_k}^2 + \cdots + \lambda_r \|P_r f\|_{\mathcal{H}_k}^2 .$$

Then the kernel of $\mathcal{H}_{k'}$ is given by $k'(s,t) = \sum_{j=1}^{r} k_j(s,t)/\lambda_j$. This fact leads to the following theorem of Wahba (1990, page 128).

Theorem 2.13 (Representer Theorem III). *Let $\mathcal{H}_0 = \text{span}\{\psi_0, \ldots, \psi_p\}$, and $\mathcal{H}_0, \ldots, \mathcal{H}_r$ be mutually orthogonal with $\mathcal{H}_k = \mathcal{H}_0 \oplus \cdots \oplus \mathcal{H}_r$. Furthermore, let $P_j \colon \mathcal{H}_k \to \mathcal{H}_k$ be the projection onto $\mathcal{H}_j$, with smoothing parameter λ_j, for all $1 \leq j \leq q$. Then, any minimiser of*

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda_1 \|P_1 f\|_{\mathcal{H}_k}^2 + \cdots + \lambda_r \|P_r f\|_{\mathcal{H}_k}^2 \right\},$$

admits a representation of the form

$$f(x) = \sum_{i=0}^{p} \beta_i \psi_i(x) + \sum_{i=1}^{n} a_i \left\{ \sum_{j=1}^{r} k_j(x, x_i)/\lambda_j \right\},$$

for some $\beta_i \in \mathbb{R}$, $0 \leq i \leq p$ and $a_i \in \mathbb{R}$, $1 \leq i \leq n$.

We require the set $\{\mathcal{H}_0, \ldots, \mathcal{H}_r\}$ to be mutually orthogonal. If the set were not orthogonal, $\mathcal{H} = \mathcal{H}_0 \oplus \cdots \oplus \mathcal{H}_r$ would be a Hilbert space, though would not be an RKHS. By the combination of Corollary 2.11, Theorem 2.12 and Theorem 2.13, we have a characterisation of the representation, uniqueness, and existence of a fit to the kernel machine.

2.4 Reproducing Kernel Representation of Penalised Splines

We now show how penalised splines are a special case of reproducing kernel methods. In particular, penalised splines correspond to a finite dimensional RKHS. Here we explicitly lay out the reproducing kernel representation of penalised splines with its terminology and notation. This is very worthwhile as, for example, it allows exponents of penalised splines to see how their various principles (e.g., additive modelling) can be extended to other settings such as support vector classification.

Consider the setting of Section 2.2 with pre-specified knots $\kappa_1, \ldots, \kappa_K$. The kernel that allows penalised splines to be set within an RKHS framework is

$$k(s,t) = 1 + st + \sum_{k=1}^{K} (s-\kappa_k)_+ (t-\kappa_k)_+.$$

The eigenfunctions are, trivially,

$$\phi_0(x) = 1,\ \phi_1(x) = x,\ \phi_{k+1}(x) = (x-\kappa_k)_+,\ 1 \le k \le K$$

with eigenvalues $\gamma_0 = \gamma_1 = \cdots = \gamma_{K+1} = 1$. As such, the eigenfunctions also form an orthonormal basis. The RKHS is

$$\mathcal{H}_k = \left\{ f : f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x-\kappa_k)_+ \right\}$$

with inner product

$$\left\langle \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x-\kappa_k)_+,\ \beta_0' + \beta_1' x + \sum_{k=1}^{K} u_k' (x-\kappa_k)_+ \right\rangle_{\mathcal{H}_k} = \beta_0\beta_0' + \beta_1\beta_1' + \sum_{k=1}^{K} u_k u_k'.$$

In particular,

$$\|f\|_{\mathcal{H}_k}^2 = \|\boldsymbol{\beta}\|^2 + \|\boldsymbol{u}\|^2.$$

The penalised spline RKHS is a particularly simple Hilbert space in that it is finite-dimensional and isomorphic to $\mathbb{R}^{K+2}$. This means that projections in $\mathcal{H}_k$ correspond to familiar Euclidean projections of the coefficients, as illustrated in the next paragraph.

For penalised splines the subspace of unpenalised functions is the linear component

$$\mathcal{H}_0 = \{f : f(x) = \beta_0 + \beta_1 x\}$$

and the orthogonal complement

$$\mathcal{H}_1 = \mathcal{H}_0^{\perp} = \left\{ f : f(x) = \sum_{k=1}^{K} u_k (x - \kappa_k)_+ \right\}$$

is the spline basis function component. The projection of $f \in \mathcal{H}_k$ onto $\mathcal{H}_1$ is given by

$$P_1 \left(\beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x - \kappa_k)_+ \right) = \sum_{k=1}^{K} u_k (x - \kappa_k)_+$$

and, hence, $\|P_1 f\|_{\mathcal{H}_k}^2 = \|\boldsymbol{u}\|^2$. Therefore (2.11) is equivalent to (2.4) for squared error loss. For more general loss, (2.11) reduces to

$$\min_{\boldsymbol{\beta}, \boldsymbol{u}} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_i) + \lambda \|\boldsymbol{u}\|^2 \right\}.$$

Define $\boldsymbol{X}_x = [1 \; x]$ and $\boldsymbol{Z}_x = [(x - \kappa_1)_+ \cdots (x - \kappa_K)_+]$. Then the primal form of the solution is $\widehat{f}(x) = \boldsymbol{X}_x \widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_x \widehat{\boldsymbol{u}}$ while the dual form is $\widehat{f}(x) = \sum_{i=1}^{n} \widehat{c}_i (\boldsymbol{X}_x \boldsymbol{X}_{x_i}^{\mathsf{T}} + \boldsymbol{Z}_x \boldsymbol{Z}_{x_i}^{\mathsf{T}})$ for suitable $\widehat{c}_i$, $1 \leq i \leq n$.

2.5 Extensions

Sections 2.2 and 2.4 only considered penalised splines for scalar predictors and truncated line basis functions. However, as shown in this section, the reproducing kernel representations apply for general penalised spline models such as those involving other spline basis functions, higher dimensional smoothing and additive structure.

2.5.1 Other Spline Basis Functions

For $x \in \mathbb{R}$, general penalised spline models can be written as

$$f(x) = \boldsymbol{X}_x \boldsymbol{\beta} + \boldsymbol{Z}_x \boldsymbol{u} \tag{2.12}$$

where $\boldsymbol{X}_x = [1 \; x \; \cdots \; x^p]$ for some $p \geq 0$ and $\boldsymbol{Z}_x$ is a set of spline basis functions. Without loss of generality, the penalty on $\boldsymbol{u}$ can be taken to be $\|\boldsymbol{u}\|^2$ by appropriate transformation of the functions in $\boldsymbol{Z}_x$. Beyond the truncated line model (2.2) the simplest basis is

$$\boldsymbol{Z}_x = \underset{1 \leq k \leq K}{\left[(x - \kappa_k)_+^p \right]},$$

corresponding to truncated polynomials of degree p. For numerical stability reasons, it is usually advantageous to linearly transform the truncated polynomial basis functions to, say, B-spline basis functions (e.g., Eilers and Marx, 1996). A suitable adjustment needs to be made to the penalisation component.

Another family of bases is that corresponding to thin plate splines (French, Kammann and Wand, 2001) and takes the form $\boldsymbol{X}_x = [1 \; x \; \cdots \; x^{m-1}]$ and

$$\boldsymbol{Z}_x = \left[|x - \kappa_k|^{2m-1}\right]_{1\le k\le K} \boldsymbol{\Omega}^{-1/2}, \quad \boldsymbol{\Omega} = \left[|\kappa_k - \kappa_{k'}|^{2m-1}\right]_{1\le k,k'\le K}.$$

These have an advantage of simple extension to higher dimensional x (Section 2.5.2). At this level of generality, the appropriate kernel is

$$k(s,t) = \boldsymbol{X}_s\boldsymbol{X}_t^{\mathsf{T}} + \boldsymbol{Z}_s\boldsymbol{Z}_t^{\mathsf{T}},$$

and the RKHS representation of (2.12) ensues.

2.5.2 Higher Dimensional Predictors

There are a number of ways by which spline basis functions can be extended to accommodate higher dimensional predictors. For example, the extension of the thin plate spline bases for $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ is

$$f(\boldsymbol{x}) = \boldsymbol{X}_x\boldsymbol{\beta} + \boldsymbol{Z}_x\boldsymbol{u}$$

where the columns of $\boldsymbol{X}_x$ consist of all d-dimensional polynomials in $x_1, \ldots, x_d$ with degree less than m and

$$\boldsymbol{Z}_x = \left[r_{md}\left(\|\boldsymbol{x} - \kappa_k\|\right)\right]_{1\le k\le K} \boldsymbol{\Omega}^{-1/2}, \quad \boldsymbol{\Omega} = \left[r_{md}\left(\|\kappa_k - \kappa_{k'}\|\right)\right]_{1\le k,k'\le K}$$

with

$$r_{md}(x) = \begin{cases} x^{2m-d}, & d \text{ odd}, \\ x^{2m-d}\log(x), & d \text{ even}, \end{cases}$$

(e.g., Green and Silverman, 1994). For $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^d$ the appropriate kernel is

$$k(\boldsymbol{s},\boldsymbol{t}) = \boldsymbol{X}_s\boldsymbol{X}_t^{\mathsf{T}} + \boldsymbol{Z}_s\boldsymbol{Z}_t^{\mathsf{T}}.$$

2.5.3 Additive Models

For two predictors x_1 and x_2 the linear penalised spline model is of the form

$$y_i = f(x_{1i}, x_{2i}) + \varepsilon_i$$

where

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \sum_{k=1}^{K_1} u_{1k}(x_1 - \kappa_{1k})_+ + \beta_2 x_2 + \sum_{k=1}^{K_2} u_{2k}(x_2 - \kappa_{2k})_+. \tag{2.13}$$

The fitting criterion is

$$\min_{\beta_j, u_{1k}, u_{2k}} \left\{ \sum_{i=1}^{n} (y_i - f(x_{1i}, x_{2i}))^2 + \lambda_1 \sum_{k=1}^{K_1} u_{1k}^2 + \lambda_2 \sum_{k=1}^{K_2} u_{2k}^2 \right\}, \tag{2.14}$$

where λ_1 and λ_2 are, respectively, smoothing parameters for variables x_1 and x_2.

Let $\mathcal{H}_0$, $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively denote the reproducing kernel Hilbert spaces generated by the kernels

$$k_0(\boldsymbol{s}, \boldsymbol{t}) = 1 + \boldsymbol{s}^\mathsf{T}\boldsymbol{t}, \quad k_1(\boldsymbol{s}, \boldsymbol{t}) = \sum_{k=1}^{K_1} (s_1 - \kappa_{1k})_+ (t_1 - \kappa_{1k})_+,$$

$$\text{and} \quad k_2(\boldsymbol{s}, \boldsymbol{t}) = \sum_{k=1}^{K_2} (s_2 - \kappa_{2k})_+ (t_2 - \kappa_{2k})_+,$$

where $\boldsymbol{s} = [s_1\ s_2]^\mathsf{T}$ and $\boldsymbol{t} = [t_1\ t_2]^\mathsf{T}$. Then

$$\mathcal{H}_k = \mathcal{H}_0 \oplus \mathcal{H}_1 \oplus \mathcal{H}_2$$

is the RKHS generated by $k = k_0 + k_1 + k_2$, with $\mathcal{H}_0$, $\mathcal{H}_1$ and $\mathcal{H}_2$ mutually orthogonal subspaces of $\mathcal{H}_k$. For $f \in \mathcal{H}_k$ let $P_1 f$ denote the projection of f onto $\mathcal{H}_1$. Then, using the notation of (2.13),

$$P_1 f(x_1, x_2) = \sum_{k=1}^{K_1} u_{1k}(x_1 - \kappa_{1k})_+ \quad \text{and} \quad \|P_1 f\|_{\mathcal{H}_k}^2 = \sum_{k=1}^{K_1} u_{1k}^2.$$

The projection operator P_2 is defined analogously and (2.14) may be written as

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} (y_i - f(x_{1i}, x_{2i}))^2 + \lambda_1 \|P_1 f\|_{\mathcal{H}_k}^2 + \lambda_2 \|P_2 f\|_{\mathcal{H}_k}^2 \right\}.$$

For general loss functions the criterion is

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_{1i}, x_{2i})) + \lambda_1 \|P_1 f\|_{\mathcal{H}_k}^2 + \lambda_2 \|P_2 f\|_{\mathcal{H}_k}^2 \right\}.$$

The extension to other basis functions and several predictors is straightforward. The same applies to additive models with higher dimensional components (e.g., Kammann and Wand, 2003).

2.5.4 Semiparametric Regression Models

General semiparametric regression models contain both smooth functional (nonparametric) and ordinary linear (parametric) components. The simplest is

$$y_i = \beta_0 + \beta_z z_i + f(x_i) + \varepsilon_i$$

which is often referred to as a partially linear model. If f has representation (2.12) then the appropriate kernel is $k = k_0 + k_1$ where

$$k_0(s,t) = \mathbf{X}_{s_1}\mathbf{X}_{t_1}^{\mathsf{T}} + s_2 t_2 \quad \text{and} \quad k_1(s,t) = \mathbf{Z}_{s_1}\mathbf{Z}_{t_1}^{\mathsf{T}},$$

$\boldsymbol{s} = [s_1\ s_2]^{\mathsf{T}}$ and $\boldsymbol{t} = [t_1\ t_2]^{\mathsf{T}}$. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be the reproducing kernel Hilbert spaces generated by k_0 and k_1, respectively. Then $\mathcal{H}_k = \mathcal{H}_0 \oplus \mathcal{H}_1$ and the problem takes the same form as (2.11) with null space $\mathcal{H}_0$.

2.5.5 Varying Coefficient Models

With varying coefficient models (Hastie and Tibshirani, 1993), we have both predictor variable, $x \in \mathbb{R}$, and modifying predictor variable $s \in \mathbb{R}$. With varying coefficient models, we assume that for fixed predictor variable, the model is linear in terms of the modifying predictor variable,

$$y_i = \alpha(s_i) + \beta(s_i)x_i + \varepsilon_i.$$

The intercept coefficients, $\alpha(.)$ and slope coefficients, $\beta(.)$, are functions of the modifying predictor variable, s. With a penalised spline form for both intercept and slope, the penalised spline version of the model is

$$y_i = f(s_i, x_i) + \varepsilon_i$$

where

$$f(s,x) = \alpha_0 + \alpha_1 x + \sum_{k=1}^{K_1} u_k^{\alpha}(s-\kappa_k)_+ + \left(\beta_0 + \beta_1 s + \sum_{k=1}^{K_2} u_k^{\beta}(s-\kappa_k)_+\right) x.$$

Penalising the spline terms, the representation $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}$ is obtained by setting

$$\boldsymbol{X} = [1\ s_i\ x_i\ s_i x_i]_{1\le i\le n}, \quad \boldsymbol{Z} = \left[\begin{array}{cc} (s_i-\kappa_k)_+ & x_i\,(s_i-\kappa_k)_+ \\ {\scriptstyle 1\le k\le K} & {\scriptstyle 1\le k\le K} \end{array}\right]_{1\le i\le n}$$

we find the appropriate kernel is $k = k_0 + k_1$ where

$$k_0(s,t) = \mathbf{X}_{s_1}\mathbf{X}_{t_1}^{\mathsf{T}} + s_2 t_2 \quad \text{and} \quad k_1(s,t) = \mathbf{Z}_{s_1}\mathbf{Z}_{t_1}^{\mathsf{T}}.$$

2.6 Alternative Penalties

Aside from squared RKHS regularisation, we have also considered empirical risk minimisation (2.8). Examples of empirical risk minimisation include ordinary least squares and linear quantile regression (e.g., Koenker and Park, 1996). Typically q is small and

fixed. Empirical risk minimisation is a special case of the projected squared RKHS norm penalty (2.11). Empirical risk minimisation is obtained either through the projection operator projecting everything to 0, or more simply, allowing $\lambda = 0$. There has been a considerable amount of research on regularisation penalties. This section considers several alternatives to the squared RKHS norm.

2.6.1 A Generalisation of the RKHS Norm

More general than the squared RKHS norm penalty is the penalty term $\Omega(\|\cdot\|_{\mathcal{H}_k})$, where $\Omega\colon [0,\infty) \to \mathbb{R}$ is a strictly monotonicly increasing function (Schölkopf and Smola, 2002). This more general form can be seen as reparameterisation of λ. Suppose that $\hat{f}$ is a solution to the optimisation problem with penalty term $\Omega(\|\cdot\|_{\mathcal{H}_k})$. It is seen that $\hat{f}$ is also a solution to the representation in (2.9), for some $\lambda \in [0,\infty]$ (Schölkopf and Smola, 2002, page 90).

2.6.2 l^1-norm Penalty

An alternative to the RKHS norm is the l^1-norm penalty (Tibshirani, 1996; Cristianini and Shawe-Taylor, 2000; Antoniadis and Fan, 2001; Koenker, 2005). Here we restrict ourselves to the finite dimensional case. Where $f = \sum_{j=0}^{q} a_j \phi_j$, with $p < q$, the l^1-norm penalty is

$$\min_{f=\sum_{j=0}^{q} a_j\phi_j} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda \sum_{j=p+1}^{q} |a_j| \right\}. \tag{2.15}$$

The l^1-norm penalty appears to be particularly useful when there are a large number of irrelevant variables (Tibshirani, 1996; Candes and Tao, 2007). An attractive computational aspect of the l^1-norm penalty is that if $\mathcal{L}$ is piecewise linear, then the minimisation in (2.15) results in a linear program (Zhu, Kosset, Hastie and Tibshirani, 2004).

2.6.3 l^p-norm Penalty

Consider the following optimisation problem

$$\min_{f=\sum_{j=0}^{q} a_j\phi_j} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda \left(|a_0|^p + \cdots + |a_q|^p \right)^{1/p} \right\},$$

where $p \geq 1$. With $p = 1$, we obtain the l^1-norm penalty as given directly above. If we take $p = 2$, we obtain the RKHS norm, that is, $\Omega(a) = a$, for $a \geq 0$.

2.6.4 Banach Space Penalty

We have seen in Section 2.3 that an inner product space can be developed into a Hilbert space. The squared norm of an element in the pre-Hilbert space is then given by the inner product, $\langle f, f \rangle = \|f\|^2_{\mathcal{H}_k}$. Although we have shown that many penalised spline formulations may be expressed making use of the Hilbert space norm, the use of the Hilbert space norm is restrictive in that the squared norm must be an inner product.

Banach spaces are defined as complete normed vector spaces. This requires a set of functions, $\{\phi_0, \phi_1, \ldots\}$, together with some norm, $\|\cdot\|_{\mathcal{B}}$. For suitable choices of $\|\cdot\|_{\mathcal{B}}$, we have the optimisation

$$\min_{f \in \mathcal{B}} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{B}} \right\}.$$

The Banach space norm penalty subsumes the l^p-norm penalty, amongst others. The completion of the Banach space does not in itself ensure that the minimum exists for infinite dimensional case. In practice the squared RKHS norm penalty is, however, the more common choice. The squared RKHS norm penalty also subsumes many alternatives, for example Wahba (1990) and the splines of Section 2.5.

2.7 Support Vector Classifiers

Squared error and likelihood-based losses (e.g., logistic, Poisson) for penalised splines have received a great deal of attention in the literature (e.g., Eilers and Marx, 1996; Ruppert *et al.*, 2003). In this section we focus on the case of hinge loss, $\mathcal{L}(a, b) = (1 - ab)_+$, corresponding to support vector classifiers. In addition, we will focus on the situation where the sample size n is much larger than the dimension of the predictors d. The reverse situation, sometimes called high dimension/low sample size, has been the subject of a great deal of attention in the recent literature; especially due to the advent of microarray gene expression data (e.g., Dudoit, Fridlyand and Speed, 2002; Liu, Lin and Ghosh, 2007). Penalised splines seem to be more advantageous for the classical $n \gg d$ situation.

We consider the generalisation of the two-component additive model described in Section 2.5.3 corresponding to $\boldsymbol{x}_i \in \mathbb{R}^d$:

$$f(\boldsymbol{x}_i) = (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_i = \left(\boldsymbol{X}\boldsymbol{\beta} + \sum_{\ell=1}^{L} \boldsymbol{Z}_\ell \boldsymbol{u}_\ell \right)_i \tag{2.16}$$

for design matrices $\boldsymbol{X}$, $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_L]$, where each subvector $\boldsymbol{u}_\ell$ has its own smoothing

parameter. The criterion to minimise is then

$$\sum_{i=1}^{n}(1-y_i f(\boldsymbol{x}_i))_+ + \sum_{\ell=1}^{L}\lambda_\ell\|\boldsymbol{u}_\ell\|^2, \tag{2.17}$$

where $y_i \in \{-1,1\}$. Note that (2.13) and (2.14) correspond to the situation where $d = L = 2$,

$$\boldsymbol{X} = [1\ x_{1i}\ x_{2i}]_{1\le i\le n} \quad \text{and} \quad \boldsymbol{Z} = [\boldsymbol{Z}_1\ \boldsymbol{Z}_2] = \Big[\underset{1\le k\le K_1}{(x_{1i}-\kappa_{1k})_+}\ \underset{1\le k\le K_2}{(x_{2i}-\kappa_{2k})_+}\Big]_{1\le i\le n}.$$

While this example involves two univariate smooths, it should be noted that higher-dimensional smooths can also be accommodated by (2.16) and (2.17) (e.g., Kammann and Wand, 2003).

Unlike least squares loss and Bernoulli log-likelihood loss, hinge loss is usually handled via Lagrangian optimisation methods. A summary is provided by Chapter 5 of Cristianini and Shawe-Taylor (2000). See also Section 12.2 and 12.3 of Hastie (1996). Minimisation of (2.17) is equivalent to the constrained optimisation problem

$$\begin{aligned}&\min_{\boldsymbol{\beta},\mathbf{u}}\left(\textstyle\sum_{\ell=1}^{L}\lambda_\ell\|\boldsymbol{u}_\ell\|^2 + \sum_{i=1}^{n}\xi_i\right)\\ &\text{subject to } \xi_i \ge 0,\ y_i(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_i \ge 1-\xi_i, \text{ for all } 1\le i\le n.\end{aligned}$$

The Lagrangian primal function is

$$L_P = \sum_{\ell=1}^{L}\lambda_\ell\|\boldsymbol{u}_\ell\|^2 + \sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_i - (1-\xi_i)\} - \sum_{i=1}^{n}\tau_i\xi_i \tag{2.18}$$

where $\alpha_i, \tau_i \ge 0$ for all $1 \le i \le n$. Setting the derivatives of L_P with respect to $\boldsymbol{\beta}$, $\boldsymbol{u}_\ell$ and ξ_i to zero results in the equalities

$$\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{\alpha}\odot\boldsymbol{y}) = \mathbf{0};\ \ \boldsymbol{u}_\ell = (2\lambda_\ell)^{-1}\boldsymbol{Z}_\ell^{\mathsf{T}}(\boldsymbol{\alpha}\odot\boldsymbol{y}),\ 1\le\ell\le L;\ \text{ and }\ \tau_i = 1-\alpha_i,\ 1\le i\le n,$$

where here, and subsequently, $\boldsymbol{A}\odot\boldsymbol{B}$ denotes the element-wise product of equal-sized matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Substitution into (2.18) leads to the Lagrangian dual function

$$L_D = \mathbf{1}^{\mathsf{T}}\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{\alpha} \quad \text{where} \quad \boldsymbol{D} = \tfrac{1}{2}(\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}})\odot(\boldsymbol{Z}\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}^{\mathsf{T}}) \tag{2.19}$$

and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1\mathbf{1},\cdots,\lambda_L\mathbf{1})$. The fitted $\widehat{\alpha}_i$ values are then found by solving the quadratic programming problem

$$\begin{aligned}&\min_{\boldsymbol{\alpha}}\left(\tfrac{1}{2}\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{\alpha} - \mathbf{1}^{\mathsf{T}}\boldsymbol{\alpha}\right)\\ &\text{subject to } 0\le\alpha_i\le 1, \text{ for all } 1\le i\le n, \text{ and } \boldsymbol{X}^{\mathsf{T}}(\boldsymbol{\alpha}\odot\boldsymbol{y}) = \mathbf{0}.\end{aligned} \tag{2.20}$$

The Karush-Kuhn-Tucker constraints include

$$\alpha_i[y_i(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_i - (1-\xi_i)] = 0,\ \tau_i\xi_i = 0 \quad \text{and} \quad y_i(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_i - (1-\xi_i) \ge 0$$

for all $1 \leq i \leq n$.

Let $\widehat{\boldsymbol{\alpha}}$ be a solution to (2.20). The fitted $\boldsymbol{u}$ is then

$$\widehat{\boldsymbol{u}} = \tfrac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}^{\mathsf{T}}(\widehat{\boldsymbol{\alpha}} \odot \boldsymbol{y}).$$

A fitted value for $\boldsymbol{\beta}$ is often determined by the non–bounded support vectors. These are given by $\{\boldsymbol{x}_i : 0 < \widehat{\alpha}_i < 1\}$. Let $\mathcal{M}$ be the set of $1 \leq i \leq n$ such that $\boldsymbol{x}_i$ is a non–bounded support vector. For each $i \in \mathcal{M}$, $\widehat{\xi}_i = 0$ and from the first Karush-Kuhn-Tucker constraint we obtain the set of equations:

$$(\boldsymbol{X}\beta)_i = (1/y_i) - (\boldsymbol{Z}\boldsymbol{u})_i = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{u})_i, \quad i \in \mathcal{M}. \tag{2.21}$$

(the last equality follows from $y_i \in \{-1, 1\}$).

If p is the length of $\boldsymbol{\beta}$ and m is the cardinality of $\mathcal{M}$ then (2.21) represents a system of p unknowns with m linear equations. Most of the support vector machine literature only treats the case $p = 1$, corresponding to an unpenalised intercept. For the case $p = 1$, Cristianini and Shawe-Taylor (2000) solve for $\boldsymbol{\beta} = \beta_0$ using an arbitrary margin point, Hastie (1996) recommends averaging over all m margin points, while Hush, Kelly, Scovel and Steinwart (2006) recommends choosing β_0 in order to minimise the primal (2.17). The system of linear equations can be both over-specified and under-specified. Hastie *et al.* (2004) gave extensive treatment to the under-specified case with $p = 1$. For general p our current recommendation for obtaining $\widehat{\boldsymbol{\beta}}$ is to minimise the primal, and then to minimise $\|P_0 f\|_{\mathcal{H}_k}$. That is,

$$\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\beta}, \quad \text{subject to } \sum_{i=1}^{n} \mathcal{L}(y_i, (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\widehat{\boldsymbol{u}})_i) = \operatorname*{argmin}_{\boldsymbol{\beta}'} \sum_{i=1}^{n} \mathcal{L}(y_i, (\boldsymbol{X}\boldsymbol{\beta}' + \boldsymbol{Z}\widehat{\boldsymbol{u}})_i).$$

We note that many quadratic programming algorithms will implicitly or explicitly find a fitted value $\widehat{\boldsymbol{\beta}}$. These include, for example, Platt (1999); Fine and Scheinberg (2001); Scheinberg (2006) and Ormerod, Wand and Koch (2008).

The bulk of the computation is concerned with the solution of (2.20). For penalised splines kernels (2.19) shows that the Gram matrix $\boldsymbol{K} = [k_1(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{1 \leq i,j \leq n}$ admits the factorisation

$$\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}^{\mathsf{T}} = \left\{\boldsymbol{Z}\boldsymbol{\Lambda}^{-1/2}\right\}\left\{\boldsymbol{Z}\boldsymbol{\Lambda}^{-1/2}\right\}^{\mathsf{T}}$$

and thus has rank bound above by the number of columns in $\boldsymbol{Z}$. Fine and Scheinberg (2001) describe interior point algorithms that take advantage of such low-rank kernels. The algorithms involve $O(nK^2)$ operations per iteration, where K is the rank of the Gram matrix and corresponds to the number of columns in $\boldsymbol{Z}$ for penalised splines. For fixed

K, the number of iterations required for termination is typically sub-linear in n. Since full-rank kernels are $O(n^3)$ this can result in large computational savings when applying interior point methods in the $n \gg K$ situation. Figure 3 of Fine and Scheinberg (2001) illustrates a more than 20-fold improvement in computation time for a particular example. There are also big reductions in storage when compared with full-rank interior point algorithms. Under looser convergence criteria than those typically used for interior point methods, there are algorithms that have faster rates of convergence. Joachims (2006) gives a cutting-plane algorithm that involves only $O(nK)$ operations in total for convergence. Although the convergence criteria is not as strict as that used by Fine and Scheinberg (2001), it is the first solver to achieve $O(nK)$ convergence. For full-rank kernels, $O(n^2)$ convergence to an approximate primal solution was achieved by Hush, Kelly, Scovel and Steinwart (2006). Large computational savings may be made using Joachims (2006) cutting-plane algorithm in the $n \gg K$ situation. Low-rank kernels for support vector machines have also been studied by Smola and Schölkopf (2000); Williams and Seeger (2001) and Ormerod, Wand and Koch (2008).

The significance of low-rank kernels in machine learning and related fields such as data mining and bioinformatics cannot be overstated. Sample sizes tend to be constantly on the increase in applications, and algorithms with $O(nK)$ operations will become a necessity. Penalised splines are inherently of this order without significant losses in accuracy.

2.7.1 "Skin of the Orange" Example

We tested additive penalised spline support vector classifiers on the "skin of the orange" simulation settings described in Section 12.3.4 of Hastie *et al.* (2001). Table 2.3 is mostly a reproduction of their Table 12.2 but with addition of classifier 7 — and lists the mean misclassification rates from the simulation study (along with standard errors). Classifier 1 is a support vector machine with linear kernel. Classifiers 2–4 are support vector machines with polynomial kernels of dimensions 2, 5 and 10 respectively. Classifier 5 is BRUTO algorithm of Hastie and Tibshirani (1990) and classifier 6 the MARS algorithm of Hastie *et al.* (2001). Based on ideas in the current chapter, classifier 7 is described in the next paragraph. At the time of writing, data from the Hastie *et al.* (2001) simulation study are available on the internet[3] and classifier 7 was applied to those data, making the results directly comparable. Note that the Bayes error for each setting is 0.029 and

[3]available at `http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/orange/`

represents a lower bound on the expected misclassification rate.

Classifier 7 involved the 4- and 10-dimensional extension of the truncated line additive model (2.13) with 20 knots in each direction. The matter of having a good choice of smoothing parameters in the additive penalised spline classifier is non-trivial. For now, we have made a relatively simplistic rule. We roughly mimicked the "4 degrees of freedom per smooth function" default used in the S-PLUS function `gam()` (Chambers and Hastie, 1991). For hinge loss the usual degrees of freedom definitions for penalised spline additive models (e.g., Ruppert, Wand and Carroll, 2003, Section 11.4) are not immediate due to its non-differentiability. We got around this by using the Bernoulli log-likelihood loss as a rough approximation.

Table 2.3 shows that this "rough-and-ready" additive penalised spline support vector classifier performs quite well compared with the classifiers from the original study. Classifier 5 (BRUTO) performs better than classifier 7 in both settings, but uses much more sophisticated smoothing parameter and variable selection strategies. Classifier 2 performs better than classifier 7 when there are no additional noise features, but the 2-degree polynomial kernel is ideal for the spherical Bayes classification boundary of this setting. It should also be mentioned that classifiers 1–4 had their smoothing parameters chosen for optimal performance using the test data; while classifiers 5–7 used data-driven rules for smoothing parameter selection, and possibly variable selection, using only the training data.

2.8 Discussion

The connection between penalised splines and reproducing kernel methods has the potential to be very fruitful. As is made clear in Section 2.7, support vector machines, which are not seriously hindered by large sample sizes, are a major payoff from this connection. It is also anticipated that many features of semiparametric regression including variable selection, smoothing parameter selection, interpretability, robustness, low-dimensional structure will prove to be beneficial in data mining and machine learning applications. The simple structure of penalised splines will aid research in this direction.

Some of the properties of RKHSs were presented in Sections 2.3.1 and 2.3.3. We have shown penalised splines to be related to the use of a special class of RKHSs. These RKHSs are of finite dimension, and isomorphic to Euclidean space. Since Pearce and Wand (2006), penalised spline support vector classifiers have been applied to a variety of

	classifier	no noise features (4 dimensions)	six noise features (10 dimensions)
1	SVC/orig.	0.450 (0.003)	0.472 (0.003)
2	SVC/poly. 2	0.078 (0.003)	0.152 (0.004)
3	SVC/poly. 5	0.180 (0.004)	0.370 (0.004)
4	SVC/poly. 10	0.230 (0.003)	0.434 (0.002)
5	BRUTO	0.084 (0.003)	0.090 (0.003)
6	MARS	0.156 (0.004)	0.173 (0.005)
7	**SVC/add. pen. spline**	**0.095 (0.004)**	**0.123 (0.003)**
	Bayes error	0.029	0.029

Table 2.3. *Mean (standard error of the mean) misclassification rates over 50 simulations for the "skin of the orange" example. Classifiers 1–6 are described in Section 12.3.4 of Hastie (1996). Classifier 7 is a support vector classifier with additive penalised spline kernel as described in Section 2.7.*

machine learning problems. Ormerod, Wand and Koch (2008) showed positive results when compared with the use of Gaussian kernels.

2.A Appendix

There are proofs for two theorems in this chapter, Theorems 2.7 and 2.12. These two theorems are related, and give necessary and sufficient conditions for the existence of a minimiser.

2.A.1 Existence of an Empirical Risk Minimiser

Let us assume that $\mathcal{H}_k$ has a finite dimensional eigen-decomposition, with eigenvectors $\{\phi_0, \ldots, \phi_q\}$. For convex loss, $\mathcal{L}$, let $\Lambda\colon \mathbb{R}^{q+1} \to [0, \infty)$ be defined as

$$\Lambda(v) \equiv \sum_{i=1}^{n} \mathcal{L}(y_i, \sum_{j=0}^{q} v_{j+1}\phi_j(x_i)). \tag{2.22}$$

Clearly, Λ is a finite sum of convex functions, and therefore a convex function itself. For each $f \in \operatorname{span}\{\phi_0, \phi_1, \ldots, \phi_q\}$, we have $f(\cdot) = \sum_{j=0}^{q} \beta_j \phi_j(\cdot)$, for some $\beta_0, \ldots, \beta_q \in \mathbb{R}$. Therefore,

$$\Lambda((\beta_0, \ldots, \beta_q)^\mathsf{T}) = \sum_{i=1}^{n} \mathcal{L}(y_i, \sum_{j=0}^{q} \beta_j \phi_j(x_i)) = \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)).$$

Let us make another definition (e.g., Rockafellar, 1970).

Definition 2.14. *A vector, $v \in \mathbb{R}^{q+1}$ is called a **direction of (strict) recession** of Λ if for all $u \in \mathbb{R}^{q+1}$, $\Lambda(u + \cdot\, v)\colon \mathbb{R} \to [0, \infty)$ is (strictly) monotonically decreasing, i.e., $h(\cdot)\colon \mathbb{R} \to [0, \infty)$ is (strictly) monotonically decreasing, where $h(a) = \Lambda(u + av)$.*

We now characterise the directions of strict recession of Λ.

Lemma 2.15. *A vector, $v \in \mathbb{R}^{q+1}$ is a direction of strict recession of Λ if and only if $f(\cdot) = \sum_{j=0}^{q} v_j \phi_j(\cdot)$ has the properties*

i) $f(x_i) \geq 0$ *for all* $\mathcal{L}(y_i, \cdot)$ *not monotonically decreasing,*

ii) $f(x_i) \leq 0$ *for all* $\mathcal{L}(y_i, \cdot)$ *not monotonically increasing, and*

iii) $f(x_i) \neq 0$ *for some* $\mathcal{L}(y_i, \cdot)$ *strictly monotonic.*

Proof. First, we assume that f satisfies the properties *i)-iii)*. Let $w \in \mathbb{R}^n$ be an arbitrary vector. Then, from properties *i)* and *ii)*, $\mathcal{L}(y_i, w_i + \cdot\, f(x_i))\colon \mathbb{R} \to [0, \infty)$ is a monotonically decreasing function for all $1 \leq i \leq n$. From property *iii)*, we then find $\mathcal{L}(y_i, w_i + \cdot\, f(x_i))$ is a strictly monotonically decreasing function, for some $1 \leq i \leq n$. On combining these, we find $\sum_{i=1}^{n} \mathcal{L}(y_i, w_i + \cdot\, f(x_i))$ to be strictly monotonically decreasing. Hence, v is a direction of strict recession of Λ.

We now show the converse. Since loss functions have range within $[0, \infty)$, we have $\sum_{i=1}^{n} \mathcal{L}(y_i, 0) < \infty$. If f does not satisfy property *i)* or *ii)*, then for some $1 \leq i \leq n$, $\mathcal{L}(y_i, \cdot\, f(x_i))$ is not a monotonically decreasing function. As $\mathcal{L}$ is convex, $\mathcal{L}(y_i, tf(x_i)) \to \infty$ as $t \to \infty$, and v is not a direction of recession. Instead, if f satisfies properties *i)* and *ii)* but not *iii)*, we find $\sum_{i=1}^{n} \mathcal{L}(y_i, \cdot\, f(x_i))$ to not be strictly monotonically decreasing, and hence v is not a direction of strict recession. □

Clearly, if such an f exists, there cannot be a minimiser; if $g \in \operatorname{span}\{\phi_0, \ldots, \phi_p\}$, then $\sum_{i=1}^{n} \mathcal{L}(y_i, g(x_i)) > \sum_{i=1}^{n} \mathcal{L}(y_i, g(x_i) + f(x_i))$. The converse is less clear; that if no such f exists, that a minimiser does exist. The function $\Lambda\colon \mathbb{R}^d \to \mathbb{R} \cup \infty$ is called *proper* as $\Lambda(z) < \infty$ for at least one $z \in \mathbb{R}^d$. We have the following theorem of Rockafellar (1970, Theorem 27.1 b, page 264).

Theorem 2.16. *Let Λ be any proper convex function. Then a minimum of Λ exists if and only if Λ has no directions of strict recession.*

By Theorem 2.16 and Lemma 2.15, it is clear that an empirical risk minimiser exists, for finite dimensional $\mathcal{H}_k$, if and only if there does not exist f that simultaneously satisfies properties *i)-iii)*. We have however, relied on the assumption that $\mathcal{H}_k$ has a finite

dimensional eigen-decomposition, as is usually the case for ERM. Consider the subspace of $\mathcal{H}_k$ given by $\mathcal{H}_{\|} = \text{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$. Then $\mathcal{H}_{\|}$ has a finite dimensional eigen-decomposition. Moreover, if $P_{\|} : \mathcal{H}_k \to \mathcal{H}_k$ is the projection operator onto $\mathcal{H}_{\|}$, then

$$f(x_i) = P_{\|} f(x_i) \quad \text{for all } 1 \leq i \leq n.$$

The ERM optimisation problem over $\mathcal{H}_k$ is then equivalent to

$$\min_{f \in \mathcal{H}_{\|}} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)).$$

It is also clear that properties *i)-iii)* in Theorem 2.7 are equivalent to properties *i)-iii)* in Lemma 2.15 with $q = n - 1$ and $\phi_j(\cdot) = k(\cdot, x_{j+1})$ for all $0 \leq j \leq q$. As such, for finite or infinite dimensional RKHS, the conditions for the existence of an ERM are given by properties *i)-iii)*.

2.A.2 Equivalence of Existence

Let us now prove Theorem 2.12. As given in Theorem 2.10, we know that if a fit exists, that it has a representation of the form

$$f(x) = \sum_{j=0}^{p} \beta_j \psi_j(x) + \sum_{i=1}^{n} c_i k_1(x, x_i), \tag{2.23}$$

so that $f(x_i) = \sum_{j=0}^{p} \beta_j \psi_j(x_i) + (\boldsymbol{K}_1 \boldsymbol{c})_i$, where $\boldsymbol{K}_1$ is the Gram matrix of k_1. For $v \in \mathbb{R}^{q+1}$ and $\boldsymbol{c} \in \mathbb{R}^n$, consider the function $\Lambda_\lambda : \mathbb{R}^{q+1+n} \to [0, \infty)$,

$$\Lambda_\lambda \begin{pmatrix} v \\ \boldsymbol{c} \end{pmatrix} \equiv \sum_{i=1}^{n} \mathcal{L}(y_i, \sum_{j=0}^{q} v_{j+1} \psi_j(x_i) + (\boldsymbol{K}_1 \boldsymbol{c})_i) + \lambda \boldsymbol{c}^{\mathsf{T}} \boldsymbol{K}_1 \boldsymbol{c}.$$

We then have

$$\Lambda_\lambda((\beta_0, \ldots, \beta_q, c_1, \ldots, c_n)^{\mathsf{T}}) = \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda \|P_1 f\|_{\mathcal{H}_k}^2,$$

where f is given in (2.23). Hence, any direction of recession of Λ_λ requires $\boldsymbol{c}^{\mathsf{T}} \boldsymbol{K}_1 \boldsymbol{c} = 0$, equivalently $(\boldsymbol{K}_1 \boldsymbol{c})_i = 0$ for all $1 \leq i \leq n$. The existence of a direction of strict recession for Λ_λ is the same as existence of a direction of strict recession for Λ in (2.22). By application of Theorem 2.16, the theorem is proved.

Chapter Three

Explicit Links Between Longitudinal Data Analysis and Kernel Machines

3.1 Introduction

Longitudinal data is characterised by there being repeated measurements of individuals over time.[1] Such data sets abound in medical literature, where longitudinal studies are a dominant fixture. Since the seminal work of Harville (1977) and of Laird and Ware (1982), linear mixed models have been the mainstay of longitudinal data analyses. The predominant distinguishing feature of linear mixed models, when compared with linear models, is the dichotomisation of effects into fixed and random types. The fitting of fixed and random effects differ in that the latter is subject to a degree of shrinkage, or penalisation, dependent on the values of covariance parameters in the model. The concept of best linear unbiased prediction appealingly accommodates the handling of both types of effects (e.g., Robinson, 1991). Expositions on longitudinal data analysis, including the role of linear mixed models, can be found in Diggle, Heagerty, Liang and Zeger (2002); Fitzmaurice, Laird and Ware (2004); McCulloch, Searle and Neuhaus (2008) and Verbeke and Molenberghs (2000).

The main goal of this chapter is to expose the commonalities shared by longitudinal data analysis and kernel machines. We show, explicitly, that many popular longitudinal fitting procedures are in fact special types of kernel machine. Their representation as kernel machines offers some key benefits to the practitioner of longitudinal data analysis as well to the practitioner of kernel machines. There are at least two potential payoffs from such links:

i) The enrichment of longitudinal models to cope with non-linear predictor effects.

ii) The adaptation of kernel machine classifiers to account for within-subject correlation when applied to longitudinal data.

[1]This chapter is based the publication: Pearce, N. D. and Wand, M. P. (2009). Explicit Links Between Longitudinal Data Analysis and Kernel Machines. *Electronic Journal of Statistics*, **3**, 797–823.

Sections 3.3.2–3.3.5 gives some details on *i)*. Sections 3.3.10–3.3.12 contains some illustrations of *ii)*.

Some recent related work is Gianola, Fernando and Stella (2006) and Liu, Lin and Ghosh (2007), each of whom combine linear mixed models with kernel machines to analyse very high-dimensional genetic data-sets. However, neither of these papers deal with regular longitudinal data analysis models. James and Hastie (2001) and Müller (2005) are examples of articles concerned with classification when the data are longitudinal.

The connections between longitudinal data analysis and kernel machines are not as strong in the case of classification tasks. The next section gives a concise overview of continuous response longitudinal data analysis. Section 3.3 forms the main body of the chapter and gives an explicit case-by-case description of kernel machine representations of popular longitudinal data analytic models, as well as explaining some non-linear (kernel-based) extensions. Generalised response models and kernel machines are treated in Section 3.4. Concluding discussion is given in Section 3.5.

3.2 Gaussian Linear Mixed Model

In this section, and the following section, we suppose that the response variables are Gaussian. In this case, the main vehicle for longitudinal data analysis is the linear mixed model

$$y = X\beta + Zu + \varepsilon, \quad \begin{bmatrix} u \\ \varepsilon \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \right). \tag{3.1}$$

The use of (3.1) for longitudinal data analysis dates back to Laird and Ware (1982). Good summaries of estimation and prediction within this linear mixed model structure may be found in, for example, McCulloch, Searle and Neuhaus (2008); Robinson (1991); Ruppert, Wand and Carroll (2003, Chapter 4) and Verbeke and Molenberghs (2000). We will just present the main results here.

For given covariance matrices G and R the the theory of best linear unbiased prediction (BLUP) can be used to guide choice of β and u, and results in the criterion:

$$(y - X\beta - Zu)^{\mathsf{T}} R^{-1} (y - X\beta - Zu) + u^{\mathsf{T}} G^{-1} u.$$

This is minimised by

$$\begin{aligned} \beta_{\text{BLUP}} &= (X^{\mathsf{T}} V^{-1} X)^{-1} X^{\mathsf{T}} V^{-1} y, \\ u_{\text{BLUP}} &= GZ^{\mathsf{T}} V^{-1} (y - X\beta_{\text{BLUP}}) \end{aligned} \tag{3.2}$$

where $\boldsymbol{V} = \text{Cov}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathsf{T}} + \boldsymbol{R}$. Expressions in (3.2) are known as the BLUPs of $\boldsymbol{\beta}$ and $\boldsymbol{u}$.

In practice, longitudinal data are fitted via the steps:

i) Estimation of $\boldsymbol{G}$ and $\boldsymbol{R}$. Usually, these matrices are restricted to a parametrised class of covariances matrices. Most commonly this is achieved though maximum likelihood, or restricted maximum likelihood (REML), under the normality assumption (3.1).

ii) Substitution of the estimated covariance matrices into (3.2). The resulting estimators, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{u}}$, are commonly known as estimated BLUPs, or EBLUPs for short.

The EBLUP phrase can be transferred to any linear function of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{u}}$. Thus, $\boldsymbol{A}\widehat{\boldsymbol{\beta}} + \boldsymbol{B}\widehat{\boldsymbol{u}}$ is the EBLUP of $\boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{u}$ for any pair of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ for which $\boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{u}$ is defined. These two steps show a division into two types of estimation targets that arise in longitudinal data analysis: the covariance parameters in the $\boldsymbol{G}$ and $\boldsymbol{R}$ matrices, and the effects $\boldsymbol{\beta}$ and $\boldsymbol{u}$. The strong connections between longitudinal data analysis and kernel machines occur at the EBLUP step for estimation of the fixed and random effects. For this reason, we will not dwell on the estimation of the covariance parameters, and instead refer the reader to Pinheiro and Bates (2000). In further sections with Gaussian response variables, the covariance parameters will be taken as given.

3.3 Explicit Links for Gaussian Longitudinal Analysis

In this section we show, explicitly, how longitudinal data analysis is connected to kernel machine methodology. General kernel machines can be formulated in a number of ways. Among the most common are: optimisation and projection within reproducing Hilbert spaces (e.g., Kimeldorf and Wahba, 1971), maximum a posteriori estimation in Gaussian processes (e.g., Rasmussen and Williams, 2005) and Tikhonov regularisation of ill-posed problems (Tarantola, 2005). Due to its prominence in the Statistics literature (e.g., Wahba, 1990; Berlinet and Thomas-Agnan, 2004) we will use the first of these formulations.

We show that all longitudinal data analyses that use EBLUPs are actually just fitting a special type of kernel machine. To make these connections clear, we first treat some special cases of (3.1). We build up to complete generality in the later subsections.

3.3.1 Random Intercept Model

The simple linear random intercept model is

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + U_i + \varepsilon_{ij}, \quad 1 \le j \le n_i, \quad 1 \le i \le m, \tag{3.3}$$

where $(x_{ij}, y_{ij}) \in (\mathbb{R} \times \mathbb{R})$ is the jth predictor/response pair for subject i, and the ε_{ij} are independent $\mathrm{N}(0, \sigma_\varepsilon^2)$ within-subject errors. The regression coefficients β_0 and β_1 are fixed effects, while the subject-specific intercepts

$$U_i \overset{\text{ind.}}{\sim} \mathrm{N}(0, \sigma_u^2),$$

are random effects.

Given estimates $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\varepsilon^2$ of the variance components, the fitted line for subject i is

$$\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{U}_i, \quad 1 \le i \le m, \tag{3.4}$$

where $\widehat{\beta}_0$, $\widehat{\beta}_1$ and the $\widehat{U}_i$ are EBLUPs, as given by (3.2) with

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n_1} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{m1} \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_{m1} \\ \vdots & \vdots \\ 1 & x_{mn_m} \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{3.5}$$

Figure 3.1 shows the EBLUPs for data on longitudinally recorded weights of 48 pigs (source: Diggle, Heagerty, Liang and Zeger, 2002), with σ_u^2 and σ_ε^2 estimated via REML. We now explain how (3.4) and the fitted lines in Figure 3.1 can be obtained as a solution to an RKHS optimisation problem – thereby making them a special case of kernel machines. In the following discussion, we assume that the estimates of σ_u^2 and σ_ε^2 have been obtained (either via REML, or some other means) and are equal to $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\varepsilon^2$, respectively.

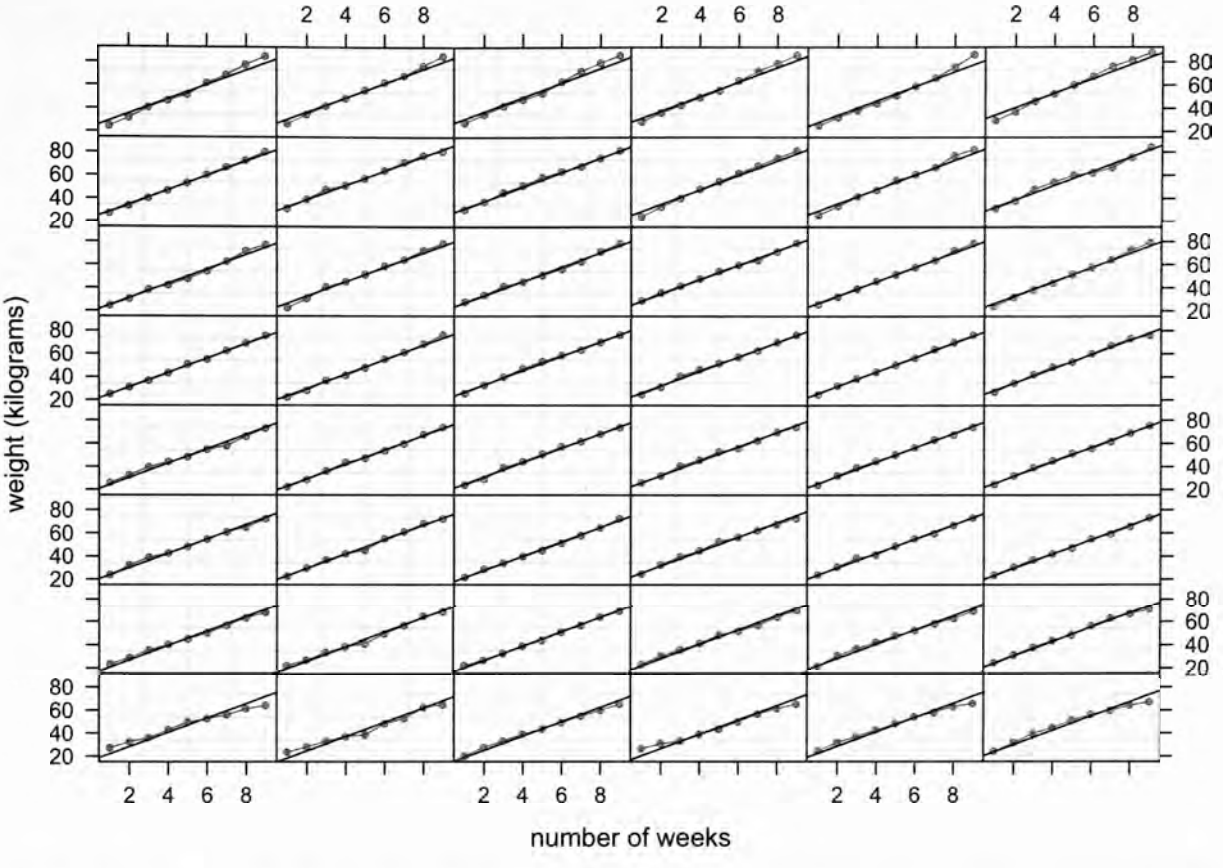


Figure 3.1. *The EBLUP-fitted lines to the pig-weights data for the simple linear random intercept model. The panels are ordered according to the size of the 48 pigs.*

Let $n = \sum_{i=1}^{m} n_i$ and re-subscript the (x_{ij}, y_{ij}) and ε_{ij} sequentially; i.e., according to the map:

$$\begin{array}{ccccccccccc} (1,1), & \ldots, & (1,n_1), & (2,1), & \ldots, & (2,n_2), & \ldots, & (m,1), & \ldots, & (m,n_m) \\ \downarrow & \cdots & \downarrow & \downarrow & \cdots & \downarrow & \cdots & \downarrow & \cdots & \downarrow \\ 1, & \ldots, & n_1, & n_1+1, & \ldots, & n_1+n_2, & \ldots, & \sum_{j=1}^{m-1} n_j + 1, & \ldots, & n. \end{array} \tag{3.6}$$

This leads to the representation

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{m} U_j Z_{ij} + \varepsilon_i, \quad 1 \le i \le n,$$

where Z_{ij} is (i,j) entry of $\boldsymbol{Z}$ as given in (3.5) and is an indicator of (x_i, y_i) being measurements for subject j ($1 \le i \le n$, $1 \le j \le m$). Next, form the RKHS of real-valued functions on $\mathbb{R}^{m+1}$:

$$\mathcal{H}_k = \left\{ f \colon f(x, z_1, \ldots, z_m) = \beta_0 + \beta_1 x + \sum_{j=1}^{m} U_j z_j, \right\}, \tag{3.7}$$

with kernel

$$k(\boldsymbol{s}, \boldsymbol{t}) = k((s_1, \ldots, s_{m+1}), (t_1, \ldots, t_{m+1})) = 1 + s_1 t_1 + \sum_{j=1}^{m} s_{1+j} t_{1+j}. \tag{3.8}$$

Note that, while $\mathcal{H}_k$ is defined on the whole of $\mathbb{R}^{m+1}$, its members of interest in longitudinal data analysis are actually on:

$$\mathbb{R} \times (1,0,0,\ldots,0) \times (0,1,0,\ldots,0) \times (0,0,0,\ldots,1) \subset \mathbb{R}^{m+1}.$$

Let

$$\mathcal{H}_\beta = \{f : f(x,z_1,\ldots,z_m) = \beta_0 + \beta_1 x\} \tag{3.9}$$

be a subspace of $\mathcal{H}_k$.

Theorem 3.1. *Let $(x_i, y_i, Z_{i1},\ldots,Z_{im})$, $1 \le i \le n$, be a sequentially subscripted longitudinal data set. Consider the RKHS $\mathcal{H}_k$ given by* (3.7) *and* (3.8)*, and subspace $\mathcal{H}_\beta$ given by* (3.9)*. Let P_u be the projection operator onto $\mathcal{H}_u = \mathcal{H}_\beta^\perp$. Then the solution to the RKHS optimisation problem*

$$\min_{f\in\mathcal{H}_k}\left[\sum_{i=1}^{n}\{y_i - f(x_i,Z_{i1},\ldots,Z_{im})\}^2 + \lambda \|P_u f\|^2_{\mathcal{H}_k}\right] \tag{3.10}$$

with $\lambda = \widehat{\sigma}^2_\varepsilon/\widehat{\sigma}^2_u$ corresponds to the EBLUPs of (3.4)*. Explicitly, the solution to* (3.10) *is*

$$\begin{aligned}
\widehat{f}(x,1,0,\ldots,0) &= \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{U}_1,\\
\widehat{f}(x,0,1,\ldots,0) &= \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{U}_2,\\
&\vdots\\
\text{and}\quad \widehat{f}(x,0,0,\ldots,1) &= \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{U}_m,
\end{aligned}$$

where $x \in \mathbb{R}$, $\widehat{\beta}_0$, $\widehat{\beta}_1$ and the $\widehat{U}_i$ are given by (3.2) *with $G = \widehat{\sigma}^2_u \boldsymbol{I}$, $\boldsymbol{R} = \widehat{\sigma}^2_\varepsilon \boldsymbol{I}$, and both $\boldsymbol{X}$ and $\boldsymbol{Z}$ given by* (3.5)*.*

Proof. As $\mathcal{H}_k$ is finite dimensional, any $f \in \mathcal{H}_k$ may be expressed as

$$f(x,z_1,\ldots,z_m) = \beta_0 + \beta_1 x + \sum_{j=1}^{m} U_j z_j,$$

so that

$$P_u f(x,z_1,\ldots,z_m) = \sum_{j=1}^{m} U_j z_j, \quad \text{and} \quad \|P_u f(x,z_1,\ldots,z_m)\|^2_{\mathcal{H}_k} = \|\boldsymbol{u}\|^2.$$

Also,

$$y_i - f(x_i,z_{i1},\ldots,z_{im}) = (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})_i,$$

so that

$$\sum_{i=1}^{n}\{y_i - f(x_i,z_{i1},\ldots,z_{im}\}^2 = \|\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}\|^2.$$

The RKHS problem in (3.10) is then equivalent to

$$\min_{\boldsymbol{\beta},\boldsymbol{u}} (1/\widehat{\sigma}^2_\varepsilon)\|\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}\|^2 + (1/\widehat{\sigma}^2_u)\|\boldsymbol{u}\|^2$$

which corresponds to EBLUP for the random intercept model. □

3.3.2 Kernel based Extension to General Mean Curves

Note that the kernel for the simple linear random intercept model can be written as

$$k((s_1,\ldots,s_{m+1}),(t_1,\ldots,t_{m+1})) = k_\beta(\boldsymbol{s},\boldsymbol{t}) + k_u(\boldsymbol{s},\boldsymbol{t})$$

where $k_u(\boldsymbol{s},\boldsymbol{t}) \equiv \sum_{j=1}^{m} s_{1+j}t_{1+j}$ corresponds to the random intercept structure in the model, and $k_\beta(\boldsymbol{s},\boldsymbol{t}) \equiv 1 + s_1t_1$ corresponds to the population mean structure. More general population mean structures can be obtained by

$$k_\beta(\boldsymbol{s},\boldsymbol{t}) = k_0(\boldsymbol{s},\boldsymbol{t}) + k_c(\boldsymbol{s},\boldsymbol{t}) = k_{0,1}(s_1,t_1) + k_{c,1}(s_1,t_1)$$

for kernels $k_{0,1}\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, and $k_{c,1}\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$. The kernel $k_{0,1}$ corresponds to unpenalised functions, and typically $k_{0,1}(s_1,t_1) = 1$. We take $k_{c,1}$ to be a kernel on $\mathbb{R}\times\mathbb{R}$. Examples include:

$$k_{c,1}(s_1,t_1) = \begin{cases} \exp\{-\gamma\,(s_1-t_1)^2\} \\ (1+\gamma\,|s_1-t_1|)\exp(-\gamma\,|s_1-t_1|) \end{cases}$$

where $\gamma > 0$ is a kernel parameter. The later kernel is known as the Matérn kernel (Matérn, 1960; Seeger, Kakade and Foster, 2008). Each of these kernels have infinite-length eigen-decompositions and result in an infinite dimensional, separable RKHS. The kernel trick ensures that fitting and representation do not require an eigen-decomposition.

Let $\mathcal{H}_0$, $\mathcal{H}_c$, and $\mathcal{H}_u$, be the RKHS generated by k_0, k_c and k_u respectively. Then

$$\mathcal{H}_k = \mathcal{H}_0 \oplus (\mathcal{H}_0 \oplus \mathcal{H}_u)^\perp \oplus \mathcal{H}_u \tag{3.11}$$

is an RKHS. Moreover, if $\mathcal{H}_0$ and $\mathcal{H}_c$ are orthogonal, then $\mathcal{H}_c = (\mathcal{H}_0 \oplus \mathcal{H}_u)^\perp$, and $\mathcal{H}_k$ has kernel $k = k_0 + k_c + k_u$. Let $P_c\colon \mathcal{H}_k \to \mathcal{H}_k$ be the projection operator corresponding to projection onto $(\mathcal{H}_0 \oplus \mathcal{H}_u)^\perp$, and let P_u be the projection operator onto $\mathcal{H}_u$. Then a mean curve, with random intercept shifts, can be fitted via the RKHS minimisation problem

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}\{y_i - f(x_i,Z_{i1},\ldots,Z_{im})\}^2 + \lambda_c\,\|P_cf\|_{\mathcal{H}_k}^2 + \lambda_u\,\|P_uf\|_{\mathcal{H}_k}^2\right\}, \tag{3.12}$$

where $\lambda_c > 0$ and $\lambda_u = \sigma_\varepsilon^2/\sigma_u^2$ are smoothing parameters. With multiple penalisations, we consider the solution to a generalised RKHS minimisation problem. By the Representer Theorem, a solution to (3.12) admits representation of the form

$$\begin{aligned} f(x,z_1,\ldots,z_m) = \beta_0 + \sum_{i=1}^{n} a_i\{&k_c((x,z_1,\ldots,z_m),(x_i,Z_{i1},\ldots,Z_{im}))/\lambda_c \\ &+ k_u((x,z_1,\ldots,z_m),(x_i,Z_{i1},\ldots,Z_{im}))/\lambda_u\}, \end{aligned} \tag{3.13}$$

where $a_1, \ldots, a_n \in \mathbb{R}$. Substituting the representation of (3.13) into (3.12) then gives the matrix criterion:

$$\min_{\beta_0, \boldsymbol{a}} \|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{K}_\lambda \boldsymbol{a}\| + \boldsymbol{a}^\mathsf{T} \boldsymbol{K}_\lambda \boldsymbol{a}, \tag{3.14}$$

where $\boldsymbol{K}_\lambda$ is the $n \times n$ matrix with (i,j) entries

$$k_c((x_i, Z_{i1}, \ldots, Z_{im}), (x_j, Z_{j1}, \ldots, Z_{jm}))/\lambda_c + k_u((x_i, Z_{i1}, \ldots, Z_{im}), (x_j, Z_{j1}, \ldots, Z_{jm}))/\lambda_u.$$

The matrix criterion is minimised by

$$\widehat{\beta}_0 = (\mathbf{1}^\mathsf{T} \boldsymbol{V}^{-1} \mathbf{1})^{-1} \mathbf{1}^\mathsf{T} \boldsymbol{V}^{-1} \boldsymbol{y},$$
$$\widehat{\boldsymbol{a}} = \boldsymbol{V}^{-1}(\boldsymbol{y} - \mathbf{1}\widehat{\beta}_0),$$

where $\boldsymbol{V} \equiv \boldsymbol{K}_\lambda + \boldsymbol{I}$.

We may express the representation in (3.13) in a more intuitive form, as

$$f(x) = \beta_0 + \sum_{i=1}^{n} c_i k_c(x, x_i) + \sum_{i=1}^{m} U_i Z_{ij},$$

where $\boldsymbol{c} = \boldsymbol{a}/\lambda_c$, and $\boldsymbol{U} = \boldsymbol{Z}\boldsymbol{a}/\lambda_u$. Such an expression delineates the within-subject effects. The fitted values are for $\widehat{\boldsymbol{c}}$ and $\widehat{\boldsymbol{u}}$ are given by $\widehat{\boldsymbol{c}} = \widehat{\boldsymbol{a}}/\lambda_c$ and $\widehat{\boldsymbol{u}} = \boldsymbol{Z}\widehat{\boldsymbol{a}}/\lambda_u$. Explicitly, for $x \in \mathbb{R}$ we have,

$$\widehat{f}(x, 1, 0, \ldots, 0) = \widehat{\beta}_0 + \sum_{i=1}^{n} \widehat{c}_i k(x, x_i) + \widehat{U}_1,$$
$$\widehat{f}(x, 0, 1, \ldots, 0) = \widehat{\beta}_0 + \sum_{i=1}^{n} \widehat{c}_i k(x, x_i) + \widehat{U}_2,$$
$$\vdots$$
$$\text{and} \quad \widehat{f}(x, 0, 0, \ldots, 1) = \widehat{\beta}_0 + \sum_{i=1}^{n} \widehat{c}_i k(x, x_i) + \widehat{U}_m.$$

It still remains to choose the kernel, which we now briefly address.

3.3.3 On the Selection of Kernel

For the longitudinal data analysis of the Section 3.3.2, the user is required to select a kernel. There are a wide variety of choices that can be made for $k_{c,1}$. A popular choice is the Gaussian kernel, typically with some data-dependent parameter γ. We briefly look at the issue of selecting the kernel. To the analyse the properties of the kernels, and classes of kernels, we have the following definition.

Definition 3.2. *A subset of a vector space is called a* **cone** *if it is closed under multiplication by positive scalars. The cone of a set, $\mathcal{A}$, is the smallest cone containing $\mathcal{A}$.*

We have the following lemma as an immediate consequence of the positive definiteness of a kernel (i.e., Definition 2.2).

Lemma 3.3. *Let k_1 and k_2 be kernels, and $\lambda_1, \lambda_2 \geq 0$. Then $\lambda_1 k_1 + \lambda_2 k_2$ is a kernel.*

Lemma 3.3 shows the set of kernels to be a cone. We now restrict these cones to particular varieties of kernels. For example, the cone of the Gaussian kernels, C_{Gauss}, comprises all kernels, k, that can be expressed as

$$\begin{aligned} k(x, x') &= \int_0^\infty k_{t\text{-Gauss}}(x, x') d\mu(t) \\ &= \int_0^\infty \exp(-t \left|x - x'\right|^2) d\mu(t), \end{aligned}$$

for some measure μ, where $k_{t\text{-Gauss}}$ denotes the Gaussian kernel with parameter t. The following example shows that the cone of the Laplacian kernels lies within cone of the Gaussian kernels.

Theorem 3.4. *Denote by $C_{Laplace}$ and C_{Gauss} the cone of Laplacian and Gaussian kernels respectively. Then $C_{Laplace} \subset C_{Gauss}$.*

Proof. For each $\gamma \in [0, \infty)$, we search for some function, $g_\lambda : [0, \infty) \to [0, \infty]$, such that

$$\int_0^\infty k_{t,-Gauss}(x, x') g_\lambda(t) dt = k_{\gamma\text{-Laplace}}(x, x'), \quad \text{for all } x, x' \in \mathbb{R},$$

where $k_{\gamma\text{-Laplace}}$ denotes the Laplacian kernel, $k_{\gamma\text{-Laplace}}(x, x') = \exp(-\gamma \left|x - x'\right|)$. Let $s = \left|x - x'\right|^2$. Then we have

$$\int_0^\infty e^{-st} g_\gamma(t) dt = e^{-\gamma\sqrt{s}}, \quad \text{for all } s \geq 0. \tag{3.15}$$

We recognise (3.15) as a Laplace transform. Inverting the transform (e.g., Korn and Korn, 2000, Appendix D),

$$g_\lambda(t) = \frac{\gamma e^{-\frac{\gamma^2}{4t}}}{2\sqrt{\pi} t^{3/2}}.$$

As g_λ is non-negative, we conclude that $C_{Laplace} \subset C_{Gauss}$. □

Theorem 3.4 provides a helpful interpretation of the Laplacian kernel. We may suspect that a Gaussian kernel is appropriate, though need to choose a value for γ. The Laplacian kernel may be expressed as an integral over Gaussian kernels.

3.3.4 Extension to Additional Linear Predictors

Our final extension of the random intercept model involves the possible inclusion of additional predictors, assumed to have a linear effect on the mean of the response variable.

Corresponding to each y_i, $1 \leq i \leq n$, let $\boldsymbol{x}_i^\ell$ a $p \times 1$ vector of such predictors. Then we should replace (3.11) by

$$\mathcal{H}_k = \mathcal{H}_0 \oplus \mathcal{H}_c \oplus \mathcal{H}_u$$

where each of these RKHSs are now on $\mathbb{R}^{m+p+1}$ and

$$\mathcal{H}_0 = \{f\colon f(\boldsymbol{x}^\ell, x, z_1, \ldots, z_m) = [1\ (\boldsymbol{x}^\ell)^\mathsf{T}]\boldsymbol{\beta}\}$$

corresponds to the fixed effects. The RKHS minimisation problem is now of the form

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}\left(y_i - f(\boldsymbol{x}_i^\ell, x_i, Z_{i1}, \ldots, Z_{im})\right)^2 + \lambda_c \|P_c f\|_{\mathcal{H}_k}^2 + \lambda_u \|P_u f\|_{\mathcal{H}_k}^2\right\}.$$

By the Representer Theorem, the minimisation reduces to

$$\min_{\boldsymbol{\beta},\boldsymbol{a}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}_\lambda \boldsymbol{a}\|^2 + \boldsymbol{a}^\mathsf{T}\boldsymbol{K}_\lambda \boldsymbol{a}, \tag{3.16}$$

where $\boldsymbol{X} = [1\ (\boldsymbol{x}_i^\ell)^\mathsf{T}]_{1\leq i\leq n}$ and $\boldsymbol{K}_\lambda$ has terms

$$[\boldsymbol{K}_\lambda]_{ij} = k_{c,1}(x_i, x_j)/\lambda_c + \boldsymbol{Z}_i^\mathsf{T}\boldsymbol{Z}_j/\lambda_u.$$

The minimisation of (3.16) leads to the solutions:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{V}^{-1}\boldsymbol{y}, \quad \text{and} \quad \widehat{\boldsymbol{a}} = \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}),$$

where $\boldsymbol{V} = \boldsymbol{K}_\lambda + \boldsymbol{I}$. Let $\widehat{\boldsymbol{c}} = \widehat{\boldsymbol{a}}/\lambda_c$, and $\widehat{\boldsymbol{u}} = \boldsymbol{Z}^\mathsf{T}\widehat{\boldsymbol{a}}/\lambda_u$. The fit can then be expressed as

$$\begin{aligned}
\widehat{f}(\boldsymbol{x}^\ell, x, 1, 0, \ldots, 0) &= [1\ (\boldsymbol{x}^\ell)^\mathsf{T}]\boldsymbol{\beta} + \sum_{i=1}^{n}\widehat{c}_i k(x, x_i) + +\widehat{U}_1,\\
\widehat{f}(\boldsymbol{x}^\ell, x, 0, 1, \ldots, 0) &= [1\ (\boldsymbol{x}^\ell)^\mathsf{T}]\boldsymbol{\beta} + \sum_{i=1}^{n}\widehat{c}_i k(x, x_i) + \widehat{U}_2,\\
&\vdots\\
\text{and}\quad \widehat{f}(\boldsymbol{x}^\ell, x, 0, 0, \ldots, 1) &= [1\ (\boldsymbol{x}^\ell)^\mathsf{T}]\boldsymbol{\beta} + \sum_{i=1}^{n}\widehat{c}_i k(x, x_i) + \widehat{U}_m.
\end{aligned}$$

We now provide illustration of fits for this most general random intercept model. The longitudinal data set on spinal bone mineral density was originally analysed by Bachrach *et al.* (1999). The study comprised some 230 girls and young women. Many of the individuals in the study had repeated measurements, with a total of some 405 measurements across the study. The subjects are categorised as belonging to one of four ethnicity groups: Asian, Black, Hispanic and White. With double subscript notation, the model is

$$y_{ij} = [1\ (\boldsymbol{x}_{ij}^\ell)^\mathsf{T}]\boldsymbol{\beta} + c(x_{ij}) + U_i + \varepsilon_{ij}$$

where the y_{ij} are spinal bone mineral measurements (g/cm^2), the x_{ij}^{ℓ} contain indicators for ethnicity and the x_{ij} are age measurements. The function $c\colon \mathbb{R} \to \mathbb{R}$ indicates a curve corresponding to the kernel k_c.

We used the Gaussian kernel with $\gamma = 0.05$, $\lambda_c = 1$ and $\lambda_u = 1$. The fitted curves in the upper part of Figure 3.2 show an increase in spinal bone mineral density up to the age of 22, and a higher spinal bone mineral density for Blacks. The mean age effect is clearly non-linear and is estimated well by the Gaussian kernel. The discussion of Section 3.3.3, suggests a Laplacian kernel-based fit. The lower part of Figure 3.2 also shows a Laplacian kernel-based fit, with parameters $\gamma = 0.0005$, $\lambda_c = 1$ and $\lambda_u = 1$. For this example, the Laplacian curve appears to be less smooth that the Gaussian curve. Both curves model the observed data well.

3.3.5 Extension to Multivariate Kernels

We briefly mention one last extension: the replacement of $c(x_i)$ by $c(\boldsymbol{x}_i)$ where the $\boldsymbol{x}_i \in \mathbb{R}^d$. This can be achieved by making k_c a d-variate kernel as opposed to the univariate kernels treated so far in this section. The relevant RKHS is now on $\mathbb{R}^{m+p+d}$, and the kernel k_c is on $\mathbb{R}^d \times \mathbb{R}^d$. Models of a similar type were recently considered by Liu *et al.* (2007). Kernels methods allow the input domain to be very broad, many examples of the possibilities are given in Shawe-Taylor and Cristianini (2004).

3.3.6 The Linear Mixed Model as a Kernel Machine

We now review the relationship between the linear mixed model and kernel machines. This helps us facilitate the longitudinal analysis of later sections. For inputs $x \in \mathbb{R}$ and $z \in \mathbb{R}^q$, we seek a function, f, so that $f(x, z)$ predicts y. For a set of mutually orthogonal functions $\psi_j\colon \mathbb{R} \to \mathbb{R}$, $0 \leq j \leq p$, the functional form for f is

$$f(x, z_1, \ldots, z_q) = \sum_{j=0}^{p} \beta_j \psi_j(x) + \sum_{j=1}^{q} u_j z_j.$$

For a strictly positive definite $q \times q$ matrix, G, consider

$$k((x, z_1, \ldots, z_q), (x', z_1', \ldots, z_q')) = \sum_{j=0}^{p} \psi_j(x)\psi_j(x') + \boldsymbol{z}^{\mathsf{T}} G \boldsymbol{z}', \tag{3.17}$$

$$= \sum_{j=0}^{q} \psi_j(x)\psi_j(x') + \sum_{j=1}^{q} (\boldsymbol{z}^{\mathsf{T}} G^{1/2})_j (\boldsymbol{z}'^{\mathsf{T}} G^{1/2})_j. \tag{3.18}$$

It is then clear from the expression in (3.18) that

i) k is a kernel, and

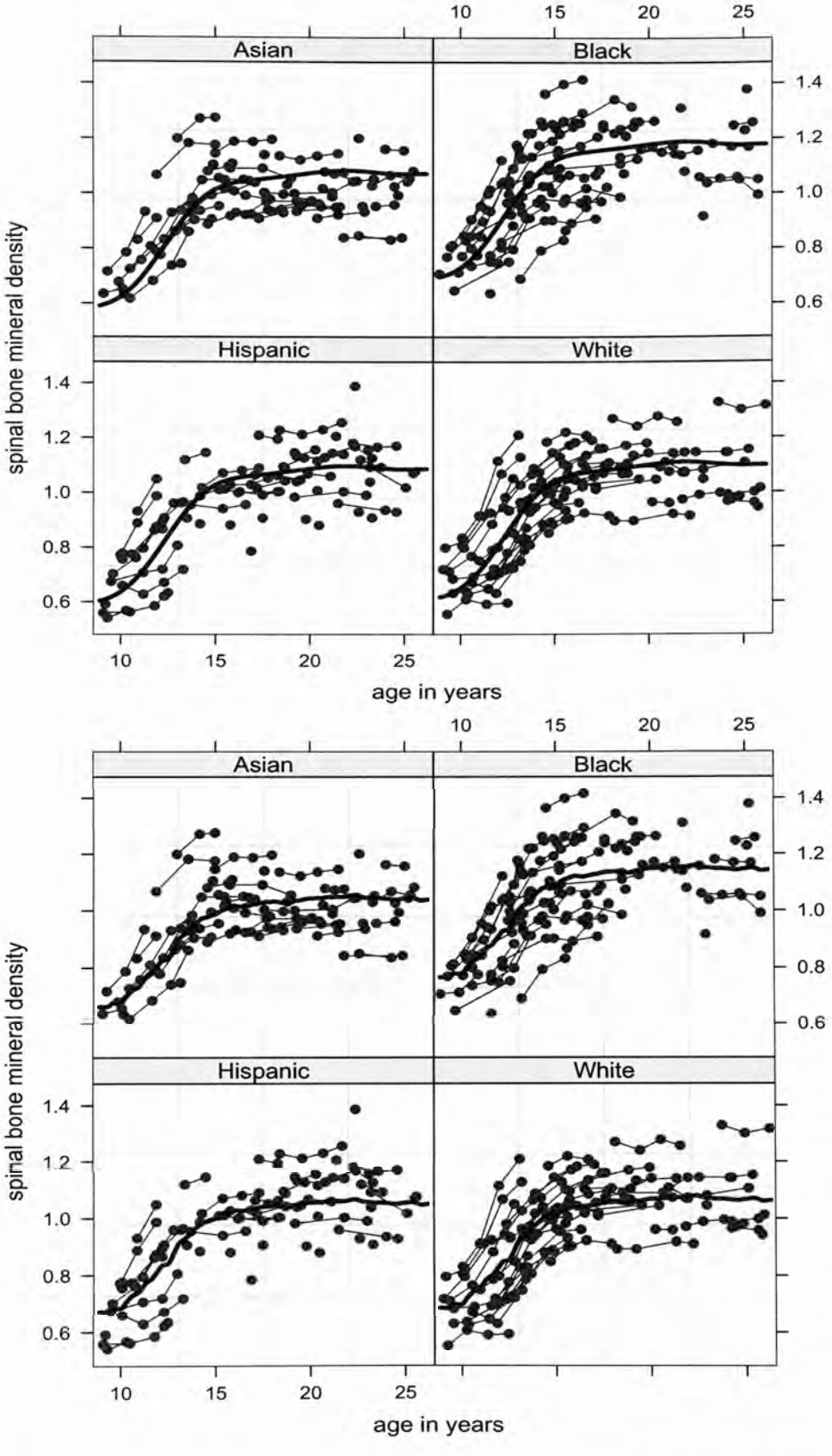


Figure 3.2. *Upper part is the Gaussian kernel-based fit to spinal bone mineral,* $\gamma = 0.05$, $\lambda_c = 1$ *and* $\lambda_u = 1$. *Lower part is the Laplacian kernel-based fit to spinal bone mineral,* $\gamma = 0.0005$, $\lambda_c = 1$ *and* $\lambda_u = 1$.

ii) $\mathcal{H}_k$ has an orthonormal basis in

$$\{\psi_0(x), \ldots, \psi_p(x), (\boldsymbol{z}^\top \boldsymbol{G}^{1/2})_1, \ldots, (\boldsymbol{z}^\top \boldsymbol{G}^{1/2})_q\}.$$

With the RKHS corresponding to the null space denoted by $\mathcal{H}_\beta$, let

$$\mathcal{H}_\beta = \left\{ f : f(x, z_1, \ldots, z_q) = \sum_{j=0}^{p} \beta_j \psi_j(x) \right\}. \tag{3.19}$$

In the following theorem, we consider $\boldsymbol{R} = \sigma_\varepsilon^2 \boldsymbol{I}$. The more general case will be considered in Section 3.3.9.

Theorem 3.5. *Let $\mathcal{H}_k$ be an RKHS with kernel given by* (3.17)*, and subspace $\mathcal{H}_\beta$ given by* (3.19)*. Furthermore, let $P_u : \mathcal{H}_k \to \mathcal{H}_k$ be the projection operator onto the orthogonal complement of $\mathcal{H}_\beta$. Then the solution to the kernel machine*

$$\min_{f \in \mathcal{H}_k} \sum_{i=1}^{q} \mathcal{L}(y_i, f(x_i, z_{i1}, \ldots, z_{iq})) + \lambda \left\| P_u f \right\|_{\mathcal{H}_k}^2 \tag{3.20}$$

with $\lambda = \widehat{\sigma}_\varepsilon^2$ corresponds to that of the observed BLUP. Explicitly, the solution to (3.20) *is*

$$\widehat{f}(x, z_1, \ldots, z_q) = \sum_{j=0}^{p} \widehat{\beta}_j \psi_j(x) + \sum_{j=1}^{q} \widehat{U}_j z_j,$$

where $x \in \mathbb{R}$, $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$ and the $\widehat{U}_1, \ldots, \widehat{U}_q$ are given by (3.2) *with*

$$\boldsymbol{X} = \begin{bmatrix} \psi_0(x_1) & \cdots & \psi_q(x_1) \\ \vdots & \ddots & \vdots \\ \psi_0(x_n) & \cdots & \psi_q(x_n) \end{bmatrix}, \tag{3.21}$$

$\boldsymbol{Z}$ a matrix with terms z_{ij}, $\boldsymbol{R} = \sigma_\varepsilon^2 \boldsymbol{I}$ and $\boldsymbol{G}$ the matrix in (3.17).

Proof. By the representer theorem, the solution to (3.20) admits a dual form represention

$$f(x, z_1, \ldots, z_q) = \sum_{j=0}^{p} \psi_j(x) \beta_i + \sum_{i=1}^{n} c_i k_u((x, z_1, \ldots, z_q), (x_i, Z_{i1}, \ldots, Z_{iq})),$$

where it is clear that

$$k_u((x, z_1, \ldots, z_q), (x', z_1', \ldots, z_q')) = \boldsymbol{z}^\top \boldsymbol{G} \boldsymbol{z}'$$

is the kernel with RKHS $\mathcal{H}_u$. We then have

$$f(x, z_1, \ldots, z_q)) = \sum_{j=0}^{p} \psi_j(x) \beta_i + \sum_{i=1}^{q} c_i \boldsymbol{Z}_i^\top \boldsymbol{G} \boldsymbol{z}.$$

In particular, for $1 \leq \ell \leq n$,

$$\begin{aligned} f(x_\ell, z_{\ell 1}, \ldots, z_{\ell q})) &= \sum_{j=0}^{p} \psi_j(x_\ell)\beta_i + \sum_{i=1}^{q} c_i \boldsymbol{Z}_i^\mathsf{T} \boldsymbol{G} \boldsymbol{Z}_\ell \\ &= (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}_u \boldsymbol{c})_\ell, \end{aligned} \tag{3.22}$$

where $\boldsymbol{K}_u$ has terms $k_u((x_i, z_{i1}, \ldots, z_{iq}), (x'_j, z'_{j1}, \ldots, z'_{jq}))$. On substituting (3.22) into (3.20), we find

$$\min_{\boldsymbol{\beta}, \boldsymbol{c}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}_u \boldsymbol{c})^\mathsf{T} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}_u \boldsymbol{c}) + \widehat{\sigma}_\varepsilon^2 \boldsymbol{c}^\mathsf{T} \boldsymbol{K}_u \boldsymbol{c}. \tag{3.23}$$

Some algebra then gives the minimiser of (3.23) as

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\mathsf{T} \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\mathsf{T} \boldsymbol{V}^{-1} \boldsymbol{y}, \\ \widehat{\boldsymbol{c}} &= \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}), \end{aligned} \tag{3.24}$$

where $\boldsymbol{V} = \boldsymbol{K}_u + \sigma_\varepsilon^2 \boldsymbol{I}$.

Let $\widehat{\boldsymbol{u}} = \boldsymbol{G}\boldsymbol{Z}^\mathsf{T}\widehat{\boldsymbol{c}}$. Then since $\boldsymbol{K}_u = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\mathsf{T}$,

$$\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{K}_1 \widehat{\boldsymbol{c}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{u}},$$

and (3.24) becomes

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\mathsf{T} \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\mathsf{T} \boldsymbol{V}^{-1} \boldsymbol{y}, \\ \widehat{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}^\mathsf{T} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \end{aligned}$$

which is the same as for the BLUP. □

We have shown a connection between kernel machines and the BLUP of the simple linear mixed model. In particular, the covariance of $\boldsymbol{u}$ in (3.1) is seen as a feature of the expression for the kernel in (3.17). This is used in the following sections, as we return specifically to longitudinal data analysis.

3.3.7 Random Intercept and Slope Model

Close inspection of Figure 3.1 shows that the parallel lines restriction imposed by the random intercept model is questionable. A more realistic model is one that allows each pig to have his/her own slope. This is achieved through the random intercept and slope model

$$y_{ij} = \beta_0 + V_i + (\beta_1 + W_i) x_{ij} + \varepsilon_{ij}, \quad 1 \leq j \leq n_i, \quad 1 \leq i \leq m, \tag{3.25}$$

where, as with (3.3), $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$, while

$$\begin{bmatrix} V_i \\ W_i \end{bmatrix} \overset{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_w \\ \rho\sigma_v\sigma_w & \sigma_w^2 \end{bmatrix} \right) \tag{3.26}$$

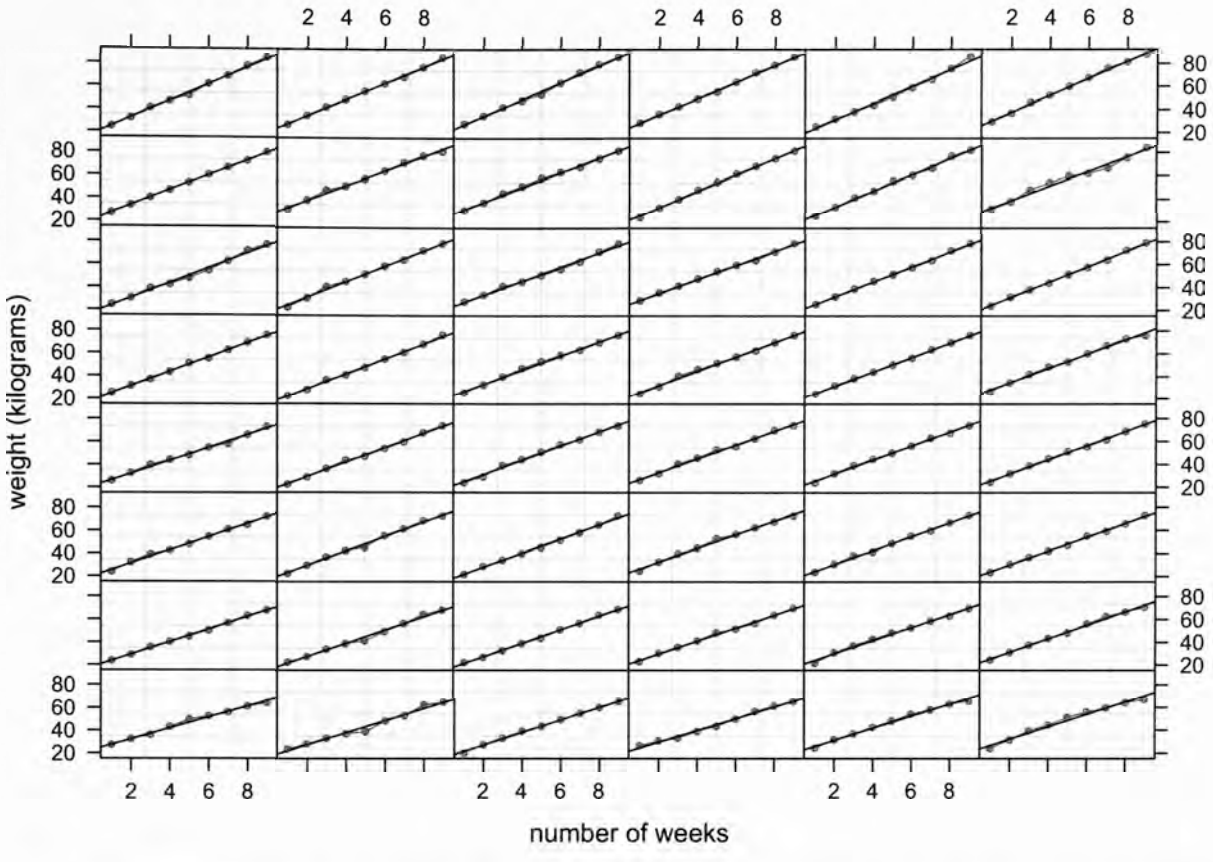


Figure 3.3. *The EBLUP-fitted lines to the pig-weights data for the simple linear random intercept and slope model. The panels are ordered according to the final weights of each of the 48 pigs in the sample.*

allow for subject specific deviations in both intercept and slope from the mean line $\beta_0 + \beta_1 x$. Figure 3.3 shows an EBLUPs fit of this model to the pig weights data, with the covariance matrix parameters estimated via REML. The resulting fits compare favourably with the random intercept model of the same data shown in Figure 3.1. It appears that the pigs do have different growth rates. A first step is to switch from the double subscripting of longitudinal data analysis to single subscripting notation via the map (3.6). The single subscript version of the random intercept and slope model (3.25) is

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{m} (V_j + x_i W_j) Z_{ij} + \varepsilon_i, \quad 1 \leq i \leq n \tag{3.27}$$

where, as before, Z_{ij} is (i, j) entry of $\mathbf{Z}$ as given in (3.5).

The extension of (3.27) may be made to obtain a canonical form. To achieve canonical form, let

$$\boldsymbol{u} = [V_1 \; W_1 \; \cdots \; V_m \; W_m]^{\mathsf{T}}, \tag{3.28}$$

and the replacement of $\boldsymbol{Z}$ and $\boldsymbol{G}$ by

$$\boldsymbol{Z} = \begin{bmatrix} 1 & x_{11} & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{1m} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & x_{21} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2m} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & x_{n1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & x_{nm} \end{bmatrix} \quad \text{and} \quad \boldsymbol{G} = \underset{1 \le i \le m}{\text{blockdiag}} \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_w \\ \rho\sigma_v\sigma_w & \sigma_w^2 \end{bmatrix} \tag{3.29}$$

in the BLUP equations (3.2). From these replacements we now describe RKHS representation of these EBLUPs.

Let us form the RKHS of real-valued functions on $\mathbb{R}^{2m+1}$:

$$\mathcal{H}_k = \left\{ f \colon f(x, z_1, \ldots, z_m) = \beta_0 + \beta_1 x + \sum_{j=1}^{2m} U_j z_j \right\} \tag{3.30}$$

with kernel

$$k(\boldsymbol{s}, \boldsymbol{t}) = k((x, z_1 \ldots, z_{2m}), (x', z_1', \ldots, z_{2m}')) = 1 + xx' + \boldsymbol{z}^{\mathsf{T}} \boldsymbol{G} \boldsymbol{z}'. \tag{3.31}$$

Note that

$$k(\boldsymbol{s}, \boldsymbol{t}) = k_0(\boldsymbol{s}, \boldsymbol{t}) + k_u(\boldsymbol{s}, \boldsymbol{t}),$$

where

$$\begin{aligned} k_\beta((x, z_1 \ldots, z_{2m}), (x', z_1', \ldots, z_{2m}')) &= 1 + xx', \\ \text{and} \quad k_u((x, z_1 \ldots, z_{2m}), (x', z_1', \ldots, z_{2m}')) &= \boldsymbol{z}^{\mathsf{T}} \boldsymbol{G} \boldsymbol{z}'. \end{aligned} \tag{3.32}$$

Let $\mathcal{H}_\beta$ and $\mathcal{H}_u$ be the RKHSs generated by k_β and k_u respectively. Since $\mathcal{H}_\beta$ and $\mathcal{H}_u$ are mutually orthogonal, we have RKHS

$$\mathcal{H}_k = \mathcal{H}_\beta \oplus \mathcal{H}_u.$$

The following example is a straightforward application of Theorem 3.5.

Example 3.6. Let $(x_i, y_i, Z_{i1}, \ldots, Z_{im})$, $1 \le i \le n$, be a sequentially subscripted longitudinal data set. Consider the RKHS $\mathcal{H}_k$ defined by (3.30) and (3.31) and subspaces $\mathcal{H}_\beta$ and

$\mathcal{H}_u$ be generated by (3.32). Let P_u be the projection operator onto $\mathcal{H}_u$. Then the solution to the RKHS optimisation problem

$$\min_{f\in\mathcal{H}_k}\left[\sum_{i=1}^{n}\{y_i - f(x_i, Z_{i1},\ldots,Z_{im})\}^2 + \lambda_u \|P_u f\|^2_{\mathcal{H}_k}\right] \tag{3.33}$$

with $\lambda_u = \widehat{\sigma}^2_\varepsilon$ corresponds to the EBLUPs. Explicitly, the solution to (3.33) is

$$\begin{aligned}
\widehat{f}(x,1,x,0,0,\ldots 0,0) &= \widehat{\beta}_0 + \widehat{V}_1 + (\widehat{\beta}_1 + \widehat{W}_1)x,\\
\widehat{f}(x,0,0,1,x,\ldots 0,0) &= \widehat{\beta}_0 + \widehat{V}_2 + (\widehat{\beta}_1 + \widehat{W}_2)x,\\
&\vdots\\
\widehat{f}(x,0,0,0,0,\ldots 1,x) &= \widehat{\beta}_0 + \widehat{V}_m + (\widehat{\beta}_1 + \widehat{W}_m)x,
\end{aligned}$$

where $x \in \mathbb{R}$, $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{V}_i = \widehat{U}_{2i-1}$ and $\widehat{W}_i = \widehat{U}_{2i}$ for $1 \le i \le m$, are given by (3.2) with $\boldsymbol{X}$ given by (3.5), $\boldsymbol{Z}$ given by (3.29), $\boldsymbol{R} = \sigma^2_\varepsilon \boldsymbol{I}$ and $\boldsymbol{G}$ given by (3.29). ◁

The above example shows the random intercept and slope model (3.25)-(3.26) to be a special case of the kernel machine. We can make the connection more explicit by considering the orthonormal basis of $\mathcal{H}_k$. The singular value decomposition (or spectral decomposition) of the random effects matrix is

$$\begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_w \\ \rho\sigma_v\sigma_w & \sigma_w^2 \end{bmatrix} = \begin{bmatrix} \alpha & \sqrt{1-\alpha^2} \\ \sqrt{1-\alpha^2} & -\alpha \end{bmatrix} \begin{bmatrix} d_u & 0 \\ 0 & d_w \end{bmatrix} \begin{bmatrix} \alpha & \sqrt{1-\alpha^2} \\ \sqrt{1-\alpha^2} & -\alpha \end{bmatrix}$$

where the eigenvalues d_v and d_w are given by

$$\begin{aligned}
d_v = d_v(\sigma_v,\sigma_w,\rho) &\equiv (\sigma_v^2+\sigma_w^2)/2 + \sqrt{(\sigma_v^2-\sigma_w^2)^2/4 + (\sigma_v\sigma_w\rho)^2},\\
\text{and } d_w = d_w(\sigma_v,\sigma_w,\rho) &\equiv (\sigma_v^2+\sigma_w^2)/2 - \sqrt{(\sigma_v^2-\sigma_w^2)^2/4 + (\sigma_v\sigma_w\rho)^2}.
\end{aligned}$$

The first normalised eigenvector component α takes the form

$$\alpha = \alpha(\sigma_v,\sigma_w,\rho) \equiv \begin{cases} \sigma_v\sigma_w\rho/\sqrt{(\sigma_v\sigma_w\rho)^2 + (\sigma_v^2 - d_v)^2}, & \text{if } \rho \neq 0 \text{ or } \sigma_v \neq \sigma_w,\\ 0, & \text{otherwise.}\end{cases}$$

The matrix

$$\boldsymbol{U} \equiv \begin{bmatrix} \alpha & \sqrt{1-\alpha^2} \\ \sqrt{1-\alpha^2} & -\alpha \end{bmatrix}$$

is orthonormal: $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$. From (3.18), an orthonormal basis for $\mathcal{H}_k$ is then

$$\left\{1, x, \sqrt{d_v}\alpha z_1 + \sqrt{d_w}\sqrt{1-\alpha^2}z_2, \sqrt{d_v}\sqrt{1-\alpha^2}z_1 - \sqrt{d_w}\alpha z_2, \right.$$
$$\left. \ldots, \sqrt{d_v}\alpha z_{2m-1} + \sqrt{d_w}\sqrt{1-\alpha^2}z_{2m}, \sqrt{d_v}\sqrt{1-\alpha^2}z_{2m-1} - \sqrt{d_w}\alpha z_{2m}\right\}.$$

3.3.8 Kernel Extension to Random Intercept and Slope

As in Section 3.3.2 we can extend the random intercept and slope model to allow for non-linear mean structure. The representation of an RKHS optimisation problem given by (3.33) allows a kernel-based extension for nonlinearities. As well as maintaining the random intercept and slope, a nonlinear component is included.

The extension to (3.25) considered here is of the form

$$y_{ij} = \beta_0 + V_i + (\beta_1 + W_i)x_{ij} + c(x_{ij}) + \varepsilon_{ij}, \quad 1 \le j \le n_i, \quad 1 \le i \le m,$$

where $c\colon \mathbb{R} \to \mathbb{R}$. In this model $\beta_0 + \beta_1 \cdot + c(\cdot)$ is the smooth overall function. Changing to single subscript notation, we have the canonical form

$$y_i = \beta_0 + \beta_1 x_i + c(x_i) + \sum_{j=1}^{m} Z_{ij} U_i + \varepsilon_i, \quad 1 \le i \le n,$$

where $\boldsymbol{u}$ and $\boldsymbol{Z}$ are given by (3.28) and (3.29).

Subject specific deviations in both intercept and slope are allowed. The relevant RKHS over $\mathbb{R}^{2m+1}$ is

$$\mathcal{H}_k = \mathcal{H}_0 \oplus \mathcal{H}_c \oplus \mathcal{H}_u$$

where $\mathcal{H}_0 = \{f\colon f(x) = \beta_0 + \beta_1 x\}$, $\mathcal{H}_c$ is any RKHS over $\mathbb{R}$ and orthogonal to $\mathcal{H}_0$, and $\mathcal{H}_u$ is the RKHS corresponding to k_u in (3.32). The kernels are of the form

$$k_0((x,z),(x',z')) = 1 + xx', \quad k_c((x,z),(x',z')) = k_{c,1}(x,x'),$$
$$\text{and} \quad k_u((x,z),(x',z')) = \boldsymbol{z}^{\mathsf{T}} \boldsymbol{G} \boldsymbol{z}'.$$

With projections P_c and P_u, the RKHS optimisation problem is then

$$\min_{f \in \mathcal{H}_k} \left[\sum_{i=1}^{n} \{y_i - f(x_i, Z_{i1}, \ldots, Z_{i2m}\}^2 + \lambda_c \|P_c f\|_{\mathcal{H}_k}^2 + \lambda_u \|P_u f\|_{\mathcal{H}_k}^2 \right],$$

where $\lambda_c \in \mathbb{R}$ and $\lambda_u = \sigma_\varepsilon^2$. Applying Theorem 3.5, the fit takes the form

$$\widehat{f}(x,1,x,0,0,\ldots 0,0) = \widehat{\beta}_0 + \widehat{V}_1 + (\widehat{\beta}_1 + \widehat{W}_1)x + \sum_{i=1}^{n} \widehat{c}_i k_{c,1}(x,x_i),$$
$$\widehat{f}(x,0,0,1,x,\ldots 0,0) = \widehat{\beta}_0 + \widehat{V}_2 + (\widehat{\beta}_1 + \widehat{W}_2)x + \sum_{i=1}^{n} \widehat{c}_i k_{c,1}(x,x_i),$$
$$\vdots$$
$$\widehat{f}(x,0,0,0,0,\ldots 1,x) = \widehat{\beta}_0 + \widehat{V}_m + (\widehat{\beta}_1 + \widehat{W}_m)x + \sum_{i=1}^{n} \widehat{c}_i k_{c,1}(x,x_i),$$

where $\widehat{\boldsymbol{c}} = \widehat{\boldsymbol{a}}/\lambda_c$ and

$$[\widehat{V}_1\ \widehat{W}_1\ \cdots \widehat{V}_m\ \widehat{W}_m]^{\mathsf{T}} = \boldsymbol{G}\boldsymbol{Z}^{\mathsf{T}}\widehat{\boldsymbol{a}}.$$

The coefficients $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{a}}$ are the solution to

$$\min_{\boldsymbol{\beta},\boldsymbol{a}}\{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{K}_\lambda\boldsymbol{a}\|^2+\boldsymbol{a}^\mathsf{T}\boldsymbol{K}_\lambda\boldsymbol{a}\},$$

where $\boldsymbol{K}_\lambda$ has terms $k_{c,1}(x_i,x_j)/\lambda_c+\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_j^\mathsf{T}$.

We illustrate this method with the rats data set. The rats data set is from Gelfand, Hills, Racine-Poon and Smith (1990). The data consists of the weight measurements of 30 rats. The rats were weighted weekly, for a total of 5 measurements each. Each rat portrays an almost linear increase in weight over the time of the study. A quadratic kernel was found to fit well. The parameterisations, $\boldsymbol{G}$ and λ_c, were estimated via REML. The random intercept–and–slope model does not give a good fit to the data. There is a noticeable curvature that is adequately modelled under the kernel extension, as shown in Figure 3.4.

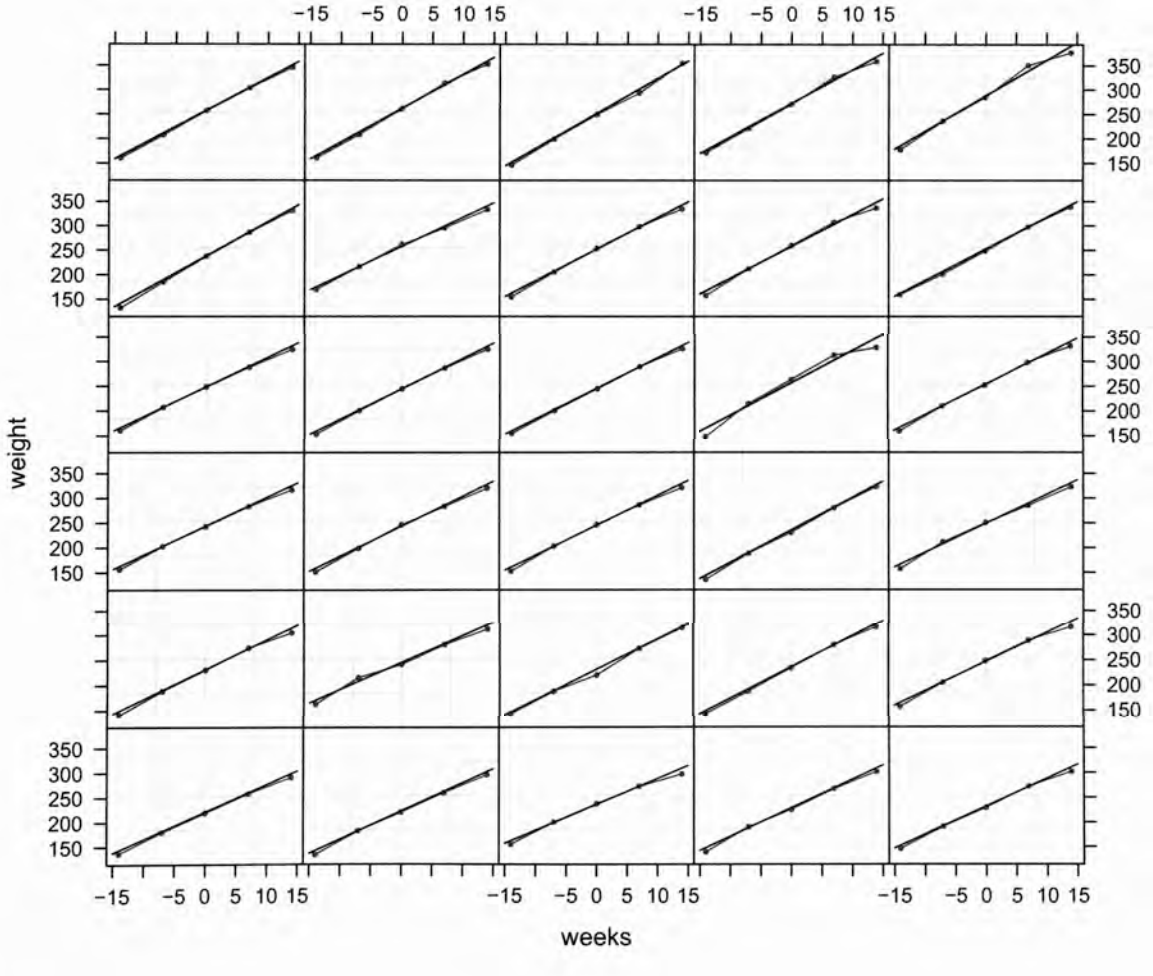


Figure 3.4. *Kernel-based fit to rats data. Each of the 30 rats shows an increase in weight. The is a noticeable curvature, and this is adequately modelled by the kernel extension.*

3.3.9 Extension to General Random Effects Structure

The general form of the $X\beta + Zu$, $u \sim (0, G)$, structure for parametric longitudinal data analysis has

$$X = \begin{bmatrix} 1 & X_1^F \\ \vdots & \vdots \\ 1 & X_m^F \end{bmatrix}, \quad Z = \underset{1\le i\le m}{\text{blockdiag}}(X_i^R), \quad \text{and} \quad u = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}$$

with

$$G = \text{Cov}(u) = \underset{1\le i\le m}{\text{blockdiag}}(\Sigma).$$

Here X_i^F is an $m \times p$ matrix corresponding to the ith subject's fixed effects contribution $(X_i^F\beta)$, X_i^R is an $m \times q$ matrix and u_i is a $q \times 1$ random effects vector corresponding the ith subject's contribution $(X_i^R u_i)$ and Σ is an unstructured $q \times q$ covariance matrix satisfying $\text{Cov}(u_i) = \Sigma$, $1 \le i \le m$. The BLUPs for β and u minimise

$$(1/\sigma_\varepsilon^2)\,\|y - X\beta - Zu\|^2 + u^T G u. \tag{3.34}$$

Theorem 3.1 and Example 3.6 can be generalised to the situation where BLUP corresponds to the solution of an RKHS optimisation problem. The relevant RKHS, $\mathcal{H}_k$, consists of real-valued functions on $\mathbb{R}^{p+mq}$ with kernel

$$k(s,t) \equiv k((s_1,\ldots,s_{p+mq}),(t_1,\ldots,t_{p+mq})) = 1 + \sum_{j=1}^{p} s_j t_j + \sum_{i,j=1}^{mq} s_{p+i}[G]_{ij} t_{p+j}.$$

Subspaces of interest are those generated by

$$k_F(s,t) \equiv 1 + \sum_{j=1}^{p} s_j t_j \quad \text{and} \quad k_R(s,t) \equiv \sum_{i,j=1}^{m} s_{p+i}[\Sigma]_{ij} t_{p+j}.$$

We denote these by $\mathcal{H}_F$ and $\mathcal{H}_R$ respectively. We have

$$\mathcal{H}_k = \mathcal{H}_F \oplus \mathcal{H}_R.$$

Let Z_i be the ith row of Z. Then the BLUPs given by (3.34) correspond to the RKHS optimisation problem

$$\sum_{i=1}^{n} (y_i - f(x_i, z_{i1},\ldots,z_{im}))^2 + \lambda\,\|P_R f\|_{\mathcal{H}_k}^2,$$

where $\lambda = \sigma_\varepsilon^2$, and P_R is the projection operator corresponding to projection onto $\mathcal{H}_R$.

3.3.10 Correlated Errors

Each of the longitudinal models considered so far have

$$\boldsymbol{R} = \mathrm{Cov}(\varepsilon) = \sigma_\varepsilon^2 \boldsymbol{I}.$$

However, in longitudinal data analysis it is common to allow more general structure in the $\boldsymbol{R}$ matrix. Longitudinal data models such as these do not fit as comfortably into the RKHS framework. The RKHS corresponding to general positive definite $\boldsymbol{R}$ is given by the following theorem.

Theorem 3.7. *Let $\mathcal{H}_k$ be an RKHS with kernel given by* (3.17)*, and subspace $\mathcal{H}_\beta$ given by* (3.19)*. Furthermore, let $P_u \colon \mathcal{H}_k \to \mathcal{H}_k$ be the projection operator onto $\mathcal{H}_\beta^\perp$, and $\boldsymbol{R}$ be a strictly positive definite $n \times n$ matrix. Then the solution to the kernel machine*

$$\min_{f \in \mathcal{H}_k} \sum_{i,j=1}^{n} (y_i - f(x_i, z_{i1}, \ldots, z_{im}))[\boldsymbol{R}^{-1}]_{ij}(y_j - f(x_j, z_{j1}, \ldots, z_{jm})) + \|P_u f\|_{\mathcal{H}_k}^2 \tag{3.35}$$

is the same as that for the BLUP, with $\boldsymbol{X}$ given by (3.21)*, $\boldsymbol{Z}$ a matrix with terms z_{ij}, $\boldsymbol{G}$ is given by* (3.1)*.*

Proof. By the representer theorem, the solution to (3.35) admits a represention of the form

$$f(x, z_1, \ldots, z_m)) = \sum_{j=0}^{p} \Phi_j(x)\beta_i + \sum_{i=1}^{n} c_i k_u((x, z_1, \ldots, z_m), (x_i, z_{i1}, \ldots, z_{im})).$$

In particular, for $1 \leq i \leq n$,

$$f(x_i, z_{i1}, \ldots, z_{im}) = \sum_{j=0}^{p} \Phi_j(x_i)\beta_i + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{c}.$$

On substituting into (3.35), we find

$$\min_{\boldsymbol{\beta}, \boldsymbol{c}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{c})^\top \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{c}) + \boldsymbol{c}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{c}. \tag{3.36}$$

Some algebra then gives the minimiser of (3.36) as

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{y}, \\ \widehat{\boldsymbol{c}} &= \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}), \end{aligned} \tag{3.37}$$

where $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top + \boldsymbol{R}$. Let $\widehat{\boldsymbol{u}} = \boldsymbol{G}\boldsymbol{Z}^\top \widehat{\boldsymbol{a}}$. Then (3.37) becomes

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{y}, \\ \widehat{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}), \end{aligned}$$

which is the same as for the EBLUP. □

Example 3.8. The random intercept model with first-order autoregressive (AR(1)) errors has

$$y_{ij} = \beta_0 + U_i + \beta_1 x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} = \rho\varepsilon_{i,j-1} + \xi_{ij},$$

for $1 \leq i \leq m$, $1 \leq j \leq n_i$, where $|\rho| < 1$, $U_i \sim \mathrm{N}(0, \sigma_u^2)$, and the $\xi_{ij} \sim \mathrm{N}(0, \sigma_\xi^2)$ are independent. The $\boldsymbol{R}$ matrix in this case is

$$\boldsymbol{R} = \sigma_\epsilon^2 \operatorname*{blockdiag}_{1\leq i\leq m} \begin{bmatrix} 1 & \rho & \cdots & \rho^{n_i-1} \\ \rho & 1 & \cdots & \rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \cdots & 1 \end{bmatrix}.$$

The kernel machine is then

$$\min_{f\in\mathcal{H}_k} \sum_{i,j=1}^{n} (y_i - f(x_i, z_{i1}, \ldots, z_{im}))[\boldsymbol{R}^{-1}]_{ij}(y_j - f(x_j, z_{j1}, \ldots, z_{jm})) + \|P_u f\|_{\mathcal{H}_k}^2,$$

and is equivalent to

$$\min_{\boldsymbol{\beta},\boldsymbol{u}} (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^\mathsf{T} \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \frac{1}{\sigma_u^2} \|\boldsymbol{u}\|^2.$$

The fit is given by

$$\begin{aligned} \widehat{f}(x, 1, 0, 0, \ldots 0) &= \widehat{\beta}_0 + \widehat{U}_1 + \widehat{\beta}_1 x, \\ \widehat{f}(x, 0, 1, 0, \ldots 0) &= \widehat{\beta}_0 + \widehat{U}_2 + \widehat{\beta}_1 x, \\ &\vdots \\ \widehat{f}(x, 0, 0, 0, \ldots 1) &= \widehat{\beta}_0 + \widehat{U}_m + \widehat{\beta}_1 x, \end{aligned}$$

where $(\widehat{\beta}_0, \widehat{\beta}_1)^\mathsf{T} = \boldsymbol{\beta}_{\text{BLUP}}$, $(\widehat{U}_1, \ldots, \widehat{U}_m)^\mathsf{T} = \boldsymbol{u}_{\text{BLUP}}$, and $\boldsymbol{G} = \sigma_u^2 \boldsymbol{I}$. ◁

3.3.11 Alternative Regression Loss Functions

So far in this section we have only considered squared error loss $\mathcal{L}_{\text{LS}}(a, b) \equiv (a - b)^2$. A range of alternatives for regression are available to the practitioner. We call a loss function a *regression loss* if it admits the representation $\mathcal{L}(a, b) = h(a - b)$, for some function $h\colon \mathbb{R} \to [0, \infty)$. The Statistics literature identifies various reasons why we would choose a regression loss other than least squares. These include:

i) The distribution of the errors may be non-Gaussian.

ii) To improve the robustness of the model.

iii) We may be interested in a quantity other that the conditional mean, such as the conditional median.

iv) It may be computationally cheaper to use an alternative loss function.

Recall that the Gaussian linear mixed model (3.1) assumes that the errors are from the normal distribution. More generally, for some distribution f_ε, we have

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}), \quad \text{and } \varepsilon_i \overset{\text{ind.}}{\sim} f_\varepsilon. \tag{3.38}$$

The log-likelihood of (3.38) is then

$$\log(p) = \sum_{i=1}^{n} \log f_\varepsilon((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})_i) - \tfrac{1}{2}\boldsymbol{u}^\mathsf{T}\boldsymbol{G}^{-1}\boldsymbol{u} - \tfrac{1}{2}\log|\boldsymbol{G}| - \tfrac{n}{2}\log(2\pi).$$

We then choose $(\boldsymbol{\beta}, \boldsymbol{u})$ by the maximising the log-likelihood,

$$\max_{\boldsymbol{\beta},\boldsymbol{u}} \left\{ \sum_{i=1}^{n} \log f_\varepsilon((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})_i) - \tfrac{1}{2}\boldsymbol{u}^\mathsf{T}\boldsymbol{G}^{-1}\boldsymbol{u} \right\}. \tag{3.39}$$

The connection between the maximum likelihood and the kernel machine is well established in the literature (e.g., Poggio and Girosi, 1990; Green and Silverman, 1994). It is made clear by the following theorem.

Theorem 3.9. *Let* $\mathcal{L}(a, b) = -\log f_\varepsilon(a - b)$. *Then the solution to the kernel machine*

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda \left\| P_u f \right\|_{\mathcal{H}_k}^2 \right\}$$

corresponds to the maximum likelihood estimator in (3.39).

Proof. By the representer theorem, any fit admits the representation

$$f(x) = \sum_{j=0}^{q} \phi_j(x)\beta_j + \sum_{i=1}^{n} c_i k_u(x, x_i)$$

for some β_j, $1 \leq j \leq q$ and c_i, $1 \leq i \leq n$. The result then follows by substituting for $1 \leq i \leq n$. □

It is well known (e.g., Huber and Wiley, 1981) that the least squares loss is non-robust against outliers. The motivation with robust statistics is to produce estimators that are not unduly affected by small departures from model assumptions. In particular, we are concerned with departures in normality in the error component.

The use of the t-distribution for modelling the errors has attracted some interest in robust modelling, for example Lange, Little and Taylor (1989); Peel and McLachlan (2000) and Staudenmayer, Lake and Wand (2009). An attractive aspect is that is that we may maintain an elegant mixed model framework. For a t-distribution with degrees of freedom, ν, and scale parameter, σ, the probability density function is given by:

$$f_{\nu,\sigma}(x) \equiv \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu\sigma^2}\right)^{-\left(\frac{\nu+1}{2}\right)},$$

where $\Gamma(\cdot)$ is the Gamma function, $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$. As an alternative to the Gaussian linear mixed model (3.1), we now have

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}), \quad \text{and } \varepsilon_i \overset{\text{ind.}}{\sim} t_{\nu,\sigma}. \tag{3.40}$$

The maximum likelihood estimator of $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ in (3.40) is then given as the solution to

$$\min_{\boldsymbol{\beta},\boldsymbol{u}} \left\{ \sum_{i=1}^{n} \mathcal{L}_{\nu,\sigma\text{-tdist}}(y_i, (\boldsymbol{X\beta} + \boldsymbol{Zu})_i) + \sigma \boldsymbol{u}^{\mathsf{T}} \boldsymbol{G}^{-1} \boldsymbol{u} \right\},$$

where,

$$\mathcal{L}_{\nu,\sigma\text{-tdist}}(a,b) = (\nu+1)\log\left(1 + \nu^{-1}\left(\frac{a-b}{\sigma}\right)^2\right).$$

The relationship between $\mathcal{L}_{\nu,\sigma\text{-tdist}}$ and $f_{\nu,\sigma}(x)$ is given by

$$\mathcal{L}_{\nu,\sigma\text{-tdist}}(a,b) = -2\log f_{\nu,\sigma}(a-b) + c$$

where c is a constant, independent of a and b.

The relationship of the loss to twice the negative log-likelihood of the error distribution has been well established in the literature (Green and Silverman, 1994). Consider the double exponential distribution, again with scale parameter σ,

$$f_\sigma(x) = \frac{\exp(-|x|/(2\sigma))}{4\sigma}.$$

Minimising the double exponential loss is then equivalent to minimising over the absolute value loss,

$$\begin{aligned} \mathcal{L}_{\text{av}}(a,b) &\equiv \frac{|a-b|}{\sigma} \\ &= -2\log(f_\sigma(x)) + c, \end{aligned} \tag{3.41}$$

for constant c. The absolute value loss has been use to find a median regression curve (Barrodale, 1968). A similar approach can be made using a nonsymmetric double exponential distribution. For some $0 < \tau < 1$, we have the quantile regression loss,

$$\mathcal{L}_{\tau\text{-qr}}(a,b) \equiv \rho_\tau(a-b) \equiv \begin{cases} -(1-\tau)(a-b), & (a-b) > 0, \\ \tau(a-b), & (a-b) < 0. \end{cases}$$

The quantile regression loss has attracted some recent attention in the machine learning literature (Takeuchi, Le, Sears and Smola, 2006; Christmann and Steinwart, 2008). For some $\delta > 0$, Huber's loss, $\mathcal{L}_{\delta\text{-Huber}}$, is given by

$$\mathcal{L}_{\delta\text{-Huber}}(a,b) \equiv \begin{cases} (a-b)^2, & |a-b| \le \delta, \\ 2\delta|a-b| - \delta^2, & |a-b| > \delta. \end{cases}$$

Name	Prior distribution on ε	$\mathcal{L}(a-b)$
BLUP	Normal	$(a-b)^2$
Robust regression	t-distribution	$\log\left(1+t(a-b)^2\right)$
Median regression	Double exponential	$\lvert a-b\rvert$
Quantile regression	Weighted double exponential	$\rho_\tau(a-b)$
Support vector regression	$\frac{1}{2\sigma+\epsilon}\exp\left(-\frac{(\lvert x\rvert-\epsilon)_+}{2\sigma}\right)$	$(\lvert a-b\rvert-\epsilon)_+$

Table 3.1. *Regression formulations with corresponding regression loss functions and priors. We require the parameterisations $t>0$, $0<\tau<1$, and $\epsilon>0$.*

For small values of $|a-b|$, Huber's loss is equivalent to the least squares loss, while for large values the penalty is linear. An alternative to the Huber's loss is the ϵ-insensitive loss $\mathcal{L}_{\epsilon\text{-insens}}(a,b)$, first given by Drucker, Burges, Kaufman, Smola and Vapnik (1997),

$$\mathcal{L}_{\epsilon\text{-insens}}(a,b) \equiv (|a-b|-\epsilon)_+ .$$

Note that the ϵ-insensitive loss ignores deviances smaller than ϵ, and has a linear penalty for larger values of $|a-b|$. Both Huber's loss and the ϵ-insensitive loss may be expressed as a constant (not dependent on x), plus twice the negative log-likelihood of a distribution. Girosi (1998) showed that the density

$$f_{\epsilon,\sigma}(x) = \frac{1}{2\sigma+\epsilon}\exp\left(-\frac{(|x|-\epsilon)_+}{2\sigma}\right)$$

leads to the use of the support vector regression loss. The median regression loss, Huber's loss and the support vector regression loss are all known to be robust. It is of no suprise that the densities that generate them are also fat tailed. The linear mixed model has

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}), \quad \text{and } \varepsilon_i \overset{\text{ind.}}{\sim} f_\varepsilon, \tag{3.42}$$

with solution given by

$$\min_{\boldsymbol{\beta},\boldsymbol{u}} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, (\boldsymbol{X\beta}+\boldsymbol{Zu})_i) + \boldsymbol{u}^{\mathsf{T}}\boldsymbol{G}^{-1}\boldsymbol{u} \right\},$$

where $\mathcal{L}(a,b) = -2\log f_\epsilon(a-b) + c$.

3.3.12 Example: Median Longitudinal Regression

The traditional approach to regression estimation is concerned with finding the conditional mean. For many problems, we may instead be more interested with finding the conditional median, or by extension, the conditional quantiles.

Along with the book Koenker (2005) there has been a increase in interest in median regression as a helpful data analysis tool. Median and quantile regression has had a substantial interest in the ecology, economics and statistics literature with Albrecht, Bjrklund and Vroman (2003); Buchinsky (1994); Cade and Noon (2003); Chaudhuri, Doksum and Samarov (1997); Engle and Manganelli (2004); Knight and Ackerly (2002) and Yu and Jones (1998) to name but a few. Recent literature, such as Takeuchi, Le, Sears and Smola (2006),Li, Liu and Zhu (2007) and Christmann and Steinwart (2008) have shown the appropriateness of reproducing kernel methods for the task, though do not consider longitudinal data.

The spinal bone mineral data set was previously analysed in Section 3.3.4, whereby the conditional mean spinal bone densities were modelled. For measurements of spinal bone mineral density, it may be of more interest to model the conditional medians. Authors, such as Koenker (2005), have argued that for many practical problems, it is the conditional median that is of interest.

We now detail the use of the absolute value loss (3.41) for median longitudinal regression. The RKHS problem is of the form:

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}\left|y_i - f(\boldsymbol{x}_i^{\ell}, x_i, Z_{i1},\ldots,Z_{im})\right| + \lambda_c \left\|P_c f\right\|_{\mathcal{H}_k}^2 + \lambda_u \left\|P_u f\right\|_{\mathcal{H}_k}^2\right\}, \tag{3.43}$$

where the parameter σ in (3.41) has been absorbed into λ_c and λ_u. (As REML is for least squares loss, we do not have ready estimates for σ.) By Theorem 2.13, any minimiser of 3.43 may be expressed as

$$f(\boldsymbol{x}^{\ell}, x, z) = [1\ (\boldsymbol{x}^{\ell})^{\mathsf{T}}]\boldsymbol{\beta} + \sum_{i=1}^{n}(\lambda_c^{-1}k_{c,1}(x, x_i) + \lambda_u^{-1}k_u(z, z_i))c_i.$$

Evaluating (3.3.12) at each observation,

$$f(\boldsymbol{x}_i^{\ell}, x_i, z_i) = (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}_{\lambda}\boldsymbol{c})_i.$$

Substituting (3.3.12) into (3.3.12) gives the optimisation problem

$$\min_{\boldsymbol{\beta},\boldsymbol{c}}\left\{\sum_{i=1}^{n}|(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}_{\lambda}\boldsymbol{c})_i| + \boldsymbol{c}^{\mathsf{T}}\boldsymbol{K}_{\lambda}\boldsymbol{c}\right\}.$$

Via the use of Lagrangian multipliers, a dual form of (3.43) arises as the QP

$$\min_{\boldsymbol{a}}\left(\tfrac{1}{2}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}_{\lambda}\boldsymbol{a} - \boldsymbol{y}^{\mathsf{T}}\boldsymbol{a}\right)$$

$$\text{subject to } -1 \le a_i \le 1, \text{ for all } 1 \le i \le n, \text{ and } \boldsymbol{X}^{\mathsf{T}}\boldsymbol{a} = 0.$$

The illustration given in Figure 3.5 shows a fitted median curve to the males cohort of the spinal bone mineral data set. Like the support vector machine, median regression

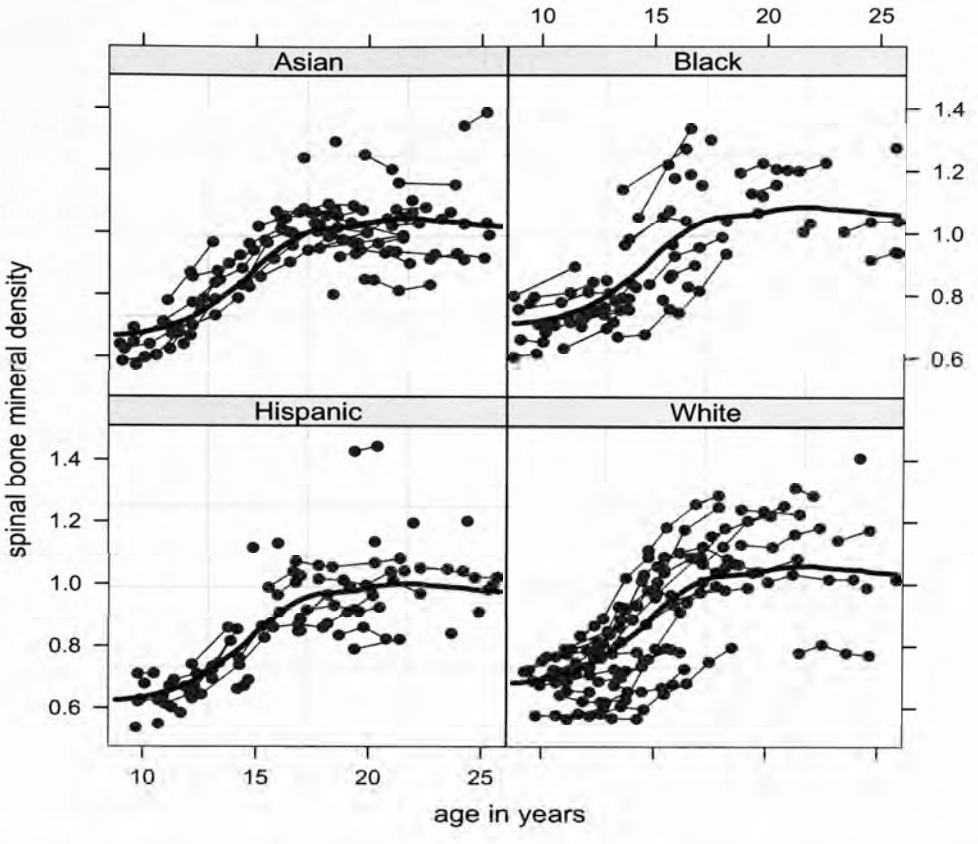


Figure 3.5. *Median regression applied to the males in the spinal bone mineral data set.*

results in a quadratic program. The kernel $k_{c,1}$ was chosen to be Gaussian with $\gamma = 0.05$. The parameterisations $\lambda_c = 0.1$ and $\lambda_u = 0.1$ were chosen by hand. The curves show an increase in median spinal bone mineral density up to about the age of 22. The curves also show a higher median spinal bone mineral density for Black males, and lower medians for the Hispanic cohort.

3.4 Generalised Response Extension

We have relied on the assumption of homoscedasticity of the errors, more specifically, that the variation of the errors is independent of the conditional mean. Many longitudinal studies have a non-continuous response, such as count or binary variable. With a binary variable, the conditional variance is dependent on the conditional mean. In such circumstances the linear mixed model in (3.42) is not appropriate and alternative approaches are required. The most common involve generalised linear mixed models (GLMM) and generalised estimating equations (GEE). In this section we describe explicit connections between kernel machines and the popular penalised quasi-likelihood (PQL) methodology for fitting GLMMs to generalised response longitudinal data.

To keep the notation simple, we will work with GLMMs confined to the canonical

one-parameter exponential family framework:

$$f(\boldsymbol{y}|\boldsymbol{u}) = \exp\{\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}^{\mathsf{T}}b(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) + \mathbf{1}^{\mathsf{T}}c(\boldsymbol{y})\}, \quad \boldsymbol{u} \sim (\mathbf{0}, \boldsymbol{G}) \tag{3.44}$$

where $f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u})$ denotes the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ and $\boldsymbol{u}$, and b and c depend upon the family member. The most common examples are Bernoulli, with $b(s) = \log(1+e^s)$, $c(s) = 0$, and Poisson with $b(s) = e^s$, $c(s) = -\log(s!)$. The one-parameter exponential family makes the assumption that there is a functional relationship from the conditional mean to the conditional variance (e.g., Rabe-Hesketh and Skrondal, 2008). The matrices in the linear predictor $\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}$, as well as $\boldsymbol{G}$, have definition and structure identical to those in the continuous response situation described in Sections 3.2 and 3.3. The simplest example is the generalised response random intercept model

$$f(y_{ij}|U_1,\ldots,U_m) = \exp\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{y_{ij}(\beta_0+\beta_1 x_{ij}+U_i) - b(\beta_0+\beta_1 x_{ij}+U_i)\}\right] \tag{3.45}$$

with $U_i \overset{\text{ind.}}{\sim} \mathrm{N}(0,\sigma_u^2)$, $1 \le i \le m$, which corresponds to (3.44) with $\boldsymbol{X}$ and $\boldsymbol{Z}$ as in (3.5) and $\boldsymbol{G} = \sigma_u^2\boldsymbol{I}$.

A common approach to fitting GLMMs is maximum likelihood for $(\boldsymbol{\beta},\boldsymbol{G})$ and best prediction for $\boldsymbol{u}$ under the normality assumption $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0},\boldsymbol{G})$. However this requires numerical integration techniques and, especially if the integrals are multi-dimensional, approximations are used instead. The most common of these is PQL (e.g., Breslow and Clayton, 1993). However, we will not treat quasi-likelihoods here, so the label penalised likelihood (PL) is appropriate. For (3.44) with $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0},\boldsymbol{G})$ and $\boldsymbol{G}$ known this involves maximisation of the penalised likelihood,

$$\exp\{\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}^{\mathsf{T}}b(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) - \tfrac{1}{2}\boldsymbol{u}^{\mathsf{T}}\boldsymbol{G}^{-1}\boldsymbol{u}\} \tag{3.46}$$

to obtain the estimates $\widehat{\boldsymbol{\beta}}_{\text{PL}}$ and $\widehat{\boldsymbol{u}}_{\text{PL}}$.

We now show that the penalised likelihood (3.46) can be treated as an RKHS optimisation problem. Hence, obtaining $\widehat{\boldsymbol{\beta}}_{\text{PL}}$ and $\widehat{\boldsymbol{u}}_{\text{PL}}$ for a given $\boldsymbol{G}$ involves a particular kernel machine. Again, with simplicity in mind, we give the full explanation for the random intercept model (3.45). The general case follows via the linear algebraic arguments and structures given in Sections 3.3.7 and 3.3.9.

Re-subscript the (x_{ij},y_{ij}) sequentially (as in Section 3.3) and, as before, let Z_{ij} be the (i,j) entry of the matrix $\boldsymbol{Z}$ defined at (3.5). Then (3.45) is

$$f(y_i|U_1,\ldots,U_m) = \exp\sum_{i=1}^{n}\left\{y_i\left(\beta_0+\beta_1 x_i+\sum_{j=1}^{m}Z_{ij}U_j\right) - b\left(\beta_0+\beta_1 x_i+\sum_{j=1}^{m}Z_{ij}U_j\right)\right\}.$$

Model	Distribution	Link
Linear regression	Normal	μ
Logistic regression	Binomial	$\Lambda^{-1}(\mu)$
Probit regression	Binomial	$\Phi^{-1}(\mu)$
Poisson regression	Poisson	$\log(\mu)$
Gamma regression	Gamma	μ^{-1}
Inverse Gaussian regression	Inverse Gaussian	μ^{-2}

Table 3.2. *Some examples of commonly used GLMMs.*

Let $\mathcal{H}_k$, k and $\mathcal{H}_\beta$ be defined by (3.7), (3.8) and (3.9) respectively. Then penalised likelihood estimation of β and u is equivalent to the RKHS optimisation problem

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}\mathcal{L}(y_i,f(x_i,Z_{i1},\ldots,Z_{im}))+\lambda\left\|P_u f\right\|^2_{\mathcal{H}_k}\right\} \tag{3.47}$$

where P_u is the projection operator onto $\mathcal{H}_u=\mathcal{H}_\beta^{\perp}$, $\lambda=1/\sigma_u^2$ and the loss function is given by $\mathcal{L}(s,t)=-2\{st-b(t)\}$. For example,

$$\mathcal{L}(s,t)=\begin{cases}-2\left\{st-\log(1+e^t)\right\}, & \text{in the Bernoulli case,}\\ -2(st-e^t), & \text{in the Poisson case.}\end{cases}$$

If $\widehat{f}$ is the solution to (3.47) then

$$\begin{aligned}\widehat{f}(x,1,0,\ldots,0)&=\widehat{\beta}_0+\widehat{\beta}_1 x+\widehat{U}_1,\\ \widehat{f}(x,0,1,\ldots,0)&=\widehat{\beta}_0+\widehat{\beta}_1 x+\widehat{U}_2,\\ &\vdots\\ \text{and}\quad \widehat{f}(x,0,0,\ldots,1)&=\widehat{\beta}_0+\widehat{\beta}_1 x+\widehat{U}_m,\end{aligned}$$

where $(\widehat{\beta}_0,\widehat{\beta}_1)^{\mathsf{T}}=\widehat{\beta}_{\mathrm{PL}}$ and $(\widehat{U}_1,\ldots,\widehat{U}_m)^{\mathsf{T}}=\widehat{u}_{\mathrm{PL}}$.

The relationship between the fits, f, and the conditional mean, μ, is given by a bijective link function, $g(\mu(\cdot))=f(\cdot)$. For example,

$$f(\cdot)=g(\mu(\cdot))=\begin{cases}\log\frac{\mu(\cdot)}{1-\mu(\cdot)}, & \text{in the Bernoulli case,}\\ \log\mu(\cdot), & \text{in the Poisson case.}\end{cases}$$

On inverting the link function we find

$$\mu(\cdot)=g^{-1}(f(\cdot))=\begin{cases}\frac{\exp f(\cdot)}{1+\exp f(\cdot)}, & \text{in the Bernoulli case,}\\ \exp f(\cdot), & \text{in the Poisson case.}\end{cases}$$

3.4.1 Kernel Extension

With GLMMs, the link function gives an explicit relationship between the mean, μ, and some linear function, f. At times, there may no exist a suitable linear relationship, and a kernelised approach may be desired. A regularised setting for classification tasks is

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}\mathcal{L}(y_i,f(x_i,Z_{i1},\ldots,Z_{im}))+\lambda\|Pf\|^2_{\mathcal{H}_k}\right\},$$

where $\mathcal{L}$ is some loss function.

In order to fit a linear model, we have $\mathcal{H}_k=\left\{f\colon f=\beta_0+\beta_1x+\sum_{j=1}^{m}U_jZ_{ij}\right\}$. Let $\mathcal{H}_0$ and $\mathcal{H}_c$ denote the reproducing kernel Hilbert spaces generated by the kernels $k_0(s,t)$ and $k_c(s,t)$, where $k_0(s,t)=1+s_1t_1$ and $k_c(s,t)=\sum_{j=1}^{m}s_{1+j}t_{1+j}$. As before, let P_c be the projection of f onto $\mathcal{H}_c$. Like the EBLUP, we would like to allow for correlation among repeated measures.

For a non-linear model, especially for those with high dimensional Hilbert space, a regularisation of the nonlinear component is required. For example, consider: $k_0=1$, $k_c(s,t)=k_c(s_1,t_1)$ and $k_u(s,t)=\sum_{j=1}^{m}s_{1+j}t_{1+j}$, where k_c is kernel, such as $k_c(s_1,t_1)=\exp(-\gamma|s_1-t_1|^2)$. With $\mathcal{H}_0$, $\mathcal{H}_c$ and $\mathcal{H}_u$ corresponding to k_0, k_c and k_u respectively, $\mathcal{H}_k=\mathcal{H}_0\oplus\mathcal{H}_c\oplus\mathcal{H}_u$. Regularising both the nonlinear component and the U_j's we have

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}\mathcal{L}\left(y_i,f\left(x_i,Z_{i1},\ldots,Z_{im}\right)\right)+\lambda_c\left\|P_cf\right\|^2_{\mathcal{H}_k}+\lambda_u\left\|P_uf\right\|^2_{\mathcal{H}_k}\right\},\tag{3.48}$$

where P_c is the projection onto $\mathcal{H}_c$, and P_u is the projection onto $\mathcal{H}_u$. The question also remains as to the suitable choice of parameterisations, $\mathcal{L}$, $\mathcal{H}_k$, λ_c and λ_u.

3.4.2 Bernoulli Loss for Classification

An example of a Bernoulli response data involves longitudinal measurements on 275 Indonesian children from Diggle *et al.* (1995). The response variable is an indicator of respiratory infection. The study was conducted to determine the effects of vitamin A nourishment on the respiratory health of children. The aim was to see if vitamin A supplementation would be of benefit. For our purposes, we look to see the effect of age on the presence of respiratory infection. The analysis also needs to account for correlation among repeated measures on the same child as well a possibly non-linear age effect. We have included, as a fixed effect, both the sex of the child and whether they are vitamin A deficient.

A non-linear fit to the age of the individuals is used, making use of the Gaussian kernel. For each individual in the study, we have between 1 and 6 measurements over time. Due to the small size of the data set possible interactions between the predictors are excluded from the model. However, analysis needs to account for correlation among repeated measures on the same child as well a possibly non-linear age effect. The Bernoulli log-likelihood is used in logistic regression and kernel logistic regression, for example Green and Yandell (1985) and Zhu and Hastie (2005). The RKHS problem is

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}\mathcal{L}\left(y_i, f\left(\boldsymbol{x}_i^{\ell}, x_i, Z_{i1},\ldots,Z_{im}\right)\right)+\lambda_c\left\|P_c f\right\|_{\mathcal{H}_k}^2+\lambda_u\left\|P_u f\right\|_{\mathcal{H}_k}^2\right\}, \tag{3.49}$$

where $\mathcal{L}(s,t)=-2\left\{st+\log(1+e^t)\right\}$. Using the Bernoulli loss, and applying the Representer Theorem to (3.48), we have the matrix optimisation problem

$$\min_{\boldsymbol{\beta},\boldsymbol{a}}\left\{\sum_{i=1}^{n}\mathcal{L}\left(y_i,(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{K}_\lambda\boldsymbol{a})_i\right)+\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}_\lambda\boldsymbol{\alpha}\right\}, \tag{3.50}$$

where $\boldsymbol{K}_\lambda$ is given by (3.14). The minimisation in (3.50) is convex, and can be solved through standard optimisation techniques such as quasi-Newton optimisation (Nocedal and Wright, 1999; Zhu and Hastie, 2005).

For an individual in the study, the fit is then given by

$$\widehat{f}(\boldsymbol{x}^{\ell},x,z_1,\ldots,z_m)=\widehat{\beta}_0+\widehat{\beta}_1x_1^{\ell}+\widehat{\beta}_2x_2^{\ell}+\sum_{i=1}^{n}\widehat{c}_ik(x,x_i)+\boldsymbol{Z}\widehat{\boldsymbol{u}}_i,$$

where $\widehat{\boldsymbol{c}}=\widehat{\boldsymbol{a}}/\lambda_c$, and $\widehat{\boldsymbol{U}}=\boldsymbol{Z}\widehat{\boldsymbol{a}}/\lambda_u$, with $\widehat{\boldsymbol{a}}$ a solution to (3.50). The *discriminant* is given simply by

$$\widehat{f}(\boldsymbol{x}^{\ell},x,z_1,\ldots,z_m)=\widehat{\beta}_0+\widehat{\beta}_1x_1^{\ell}+\widehat{\beta}_2x_2^{\ell}+\sum_{i=1}^{n}\widehat{c}_ik(x,x_i),$$

and provides a fit to the data, excluding the subject-specific random effects.

Figure 3.6 contains plots of the discriminants for Bernoulli log-likelihood loss. They were fitted with the Gaussian kernel with $\gamma=5$. In all the plots, the smoothing parameter for withon subject correlation was chosen as $\lambda_u=10$. For the upper part of Figure 3.6, we chose $\lambda_c=10$, and for the lower part we chose $\lambda_c=1$. With $\lambda_c=1$ we have a less smooth fit, the curve better follows the data. We find that the model shows that having a vitamin A deficiency indicates a higher probability of respiratory infection. A similar level of increase was noticed for males; the fits show that males showed a higher probability of respiratory infection.

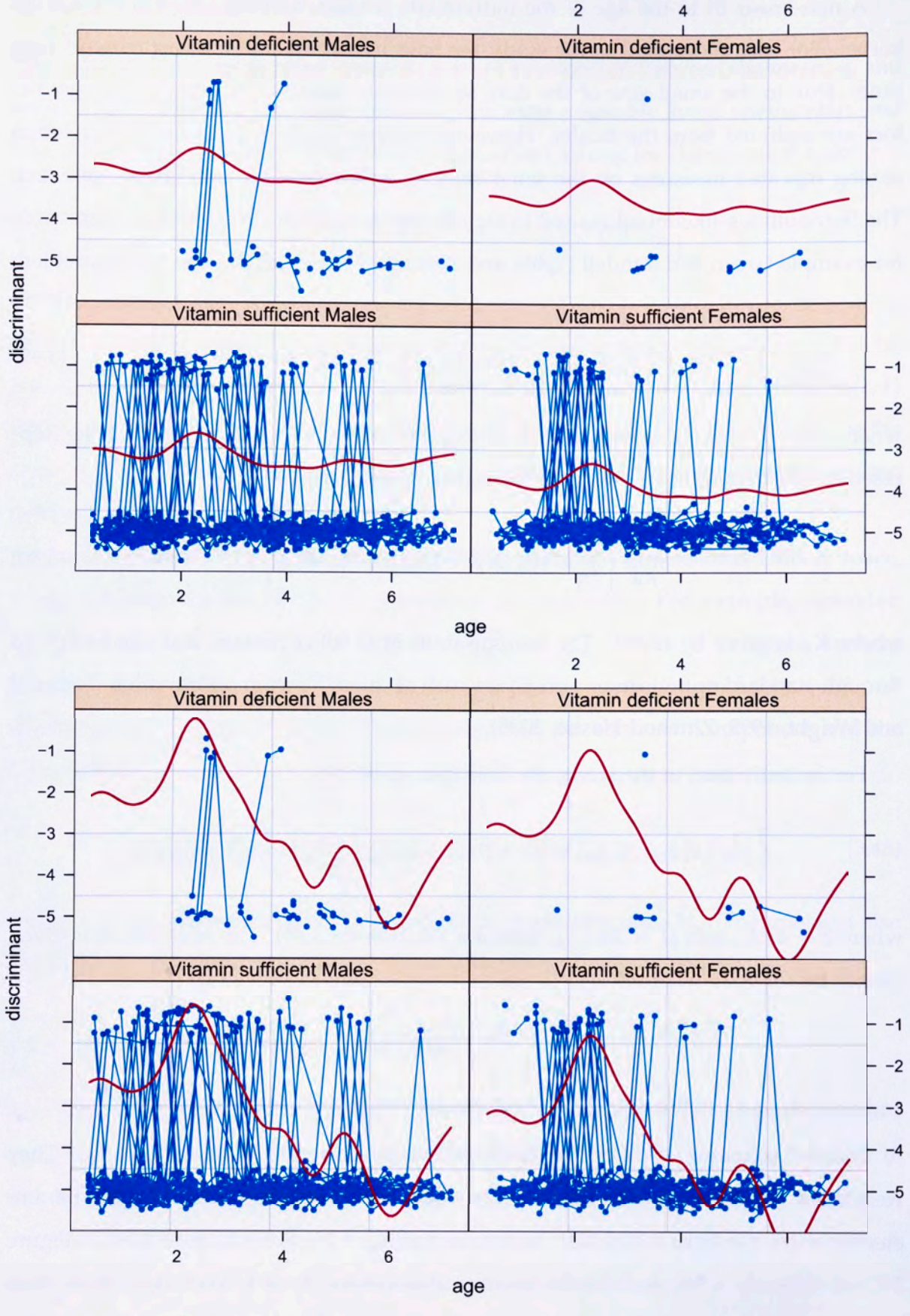


Figure 3.6. *Results of fitting Bernoulli log-likelihood loss, with 2 different values of the regularisation parameter λ_c. If viewed as a classification problem then the curves correspond to discriminants. The longitudinal data are jittered to enhance visualisation.*

3.4.3 Alternative Loss Functions for Classification

In the classification setting, with now $y \in \{-1, 1\}$, examples of loss functions include

$$\mathcal{L}(a,b) = \begin{cases} \log(1+e^{-ab}) & \text{(Bernoulli log-likelihood)} \\ (1-ab)_+ & \text{(hinge loss).} \end{cases}$$

We have seen the Bernoulli log-likelihood loss being used in the previous subsection. The Bernoulli log-likelihood falls within the scope of GLMMs, and the use of Bernoulli log-likelihood can be justified from a maximum likelihood standpoint. There are, however, popular loss functions that do not conform to the GLMM framework. The hinge loss, in particular, does not conform to the GLMM framework.

The use of the hinge loss comes at the cost of not having asymptotically consistent estimates of the conditional probabilities (Steinwart, 2001). What the hinge loss produces a classifier in the sense that it classifies new observations as being either in one class or the other. The support vector classifier avoids the somewhat intermediatory step of predicting the probabilities. In the Indonesian children's example, we are interested in whether the children have respiratory infection. In our sample, around 9.5% of the cases have a respiratory infection at a given time. Using hinge loss, we can produce a support vector machine that predicts whether they are at a high risk of respiratory infection.

A weighted hinge loss function is

$$\mathcal{L}(a,b) = \begin{cases} C_1(1-b)_+, & \text{if } a = 1, \\ C_{-1}(1+b)_+, & \text{if } a = -1, \end{cases}$$

for some positive constants, C_1 and C_{-1}. In one example, shown in in the upper part of Figure 3.7, we have chosen $C_1 = 1$ and $C_{-1} = 0.05$. These costs relate to the cost of misclassification, estimating the cost of falsely diagnosing the presence or respiratory disease to be twenty times less that not detecting the disease when it is present. In the second example in Figure 3.7, we have $C_1 = 1$ and $C_{-1} = 0.1$. This is an indication of the relative costs of misdiagnosis. We would consider the relative cost of falsely diagnosing a patient with respiratory infection as being around one tenth the cost of missing a diagnosis on a patient with the infection. The RKHS problem is given by (3.49), and the class predictions are made with $\text{sign}(f(\boldsymbol{x}^\ell, x, z_1, \ldots, z_m))$.

The upper part of Figure 3.7 shows the discriminant to be greater than zero in most cases. It is only with vitamin sufficient females over the age of five and a half years does the discriminant drops below zero. In the lower part of Figure 3.7, the ratio of C_1 to

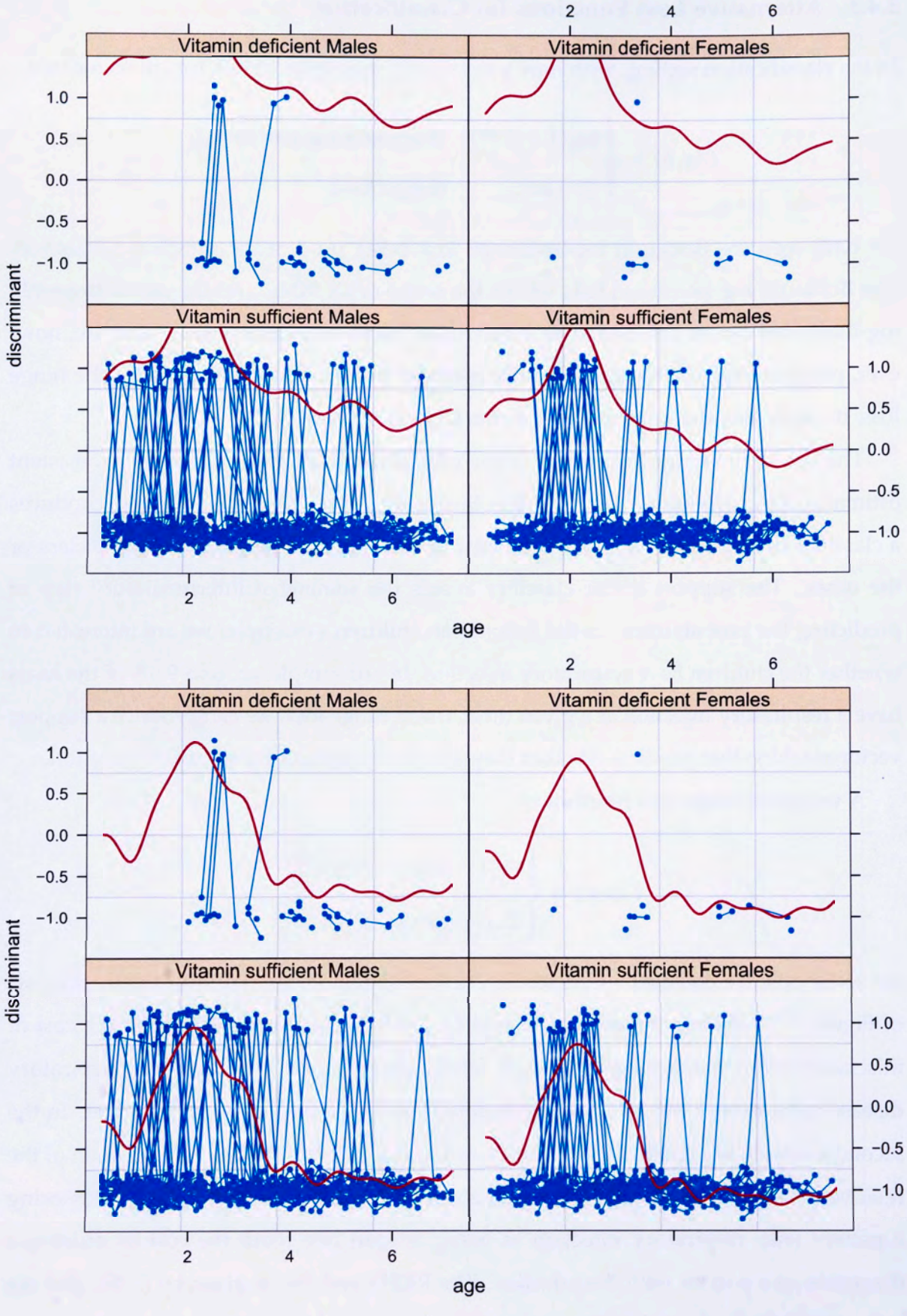


Figure 3.7. *Results of fitting hinge loss, with 2 different values of the cost parameter for the smaller class. If viewed as a classification problem then the curves correspond to discriminants. The longitudinal data are jittered to enhance visualisation.*

C_{-1} better reflects the observed number of observations per class. Like with Bernoulli log-likelihood, we find the discriminant to be higher with vitamin A deficiency.

3.5 Discussion

In this chapter we have shown that two ostensibly different areas of research – longitudinal data analysis and kernel machines – are, in fact, very similar in their underlying mathematics. It is anticipated that the explicit connections that have been established here will facilitate a more fluid exchange of ideas between the two fields. For longitudinal data analysis, there is the possibility of using kernel machines to better deal with non-linearity and to develop improved classification procedures. From the kernel machine perspective, we envisage kernel methodology that is tailored to longitudinal data models and accounts for complications such as within-unit correlation.

Chapter Four

Semiparametric Regression via Variational Bayes

4.1 Introduction

Bayesian inference is an effective and popular method for learning tasks. If θ is a vector of Bayesian model parameters, and y the observed data, Bayesian inference is based on the posterior density

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)}.$$

For many models of practical interest, the posterior distribution does not have a closed form. Moreover, it is often computationally demanding to calculate expectations with respect to the posterior (Bishop, 2006). For continuous distributions, taking expectations over the joint distribution, $p(\theta, y)$ may result in an analytically intractable integral.

Variational approximation techniques offer an approximate solution to Bayes learning. Widely applicable, variational approximation has its roots in the "calculus of variations" (Gelfand and Fomin, 2000), that is, in finding the optimum of a functional. The learning framework is popular in statistical physics (e.g., Feynman, 1972; Callen, 1985), under such names as *mean-field variational approximation* and *free-energy minimisation*. For Bayesian learning, this learning framework is known as *variational Bayes*. With Jordan, Ghahramani, Jaakkola and Saul (1999) and Jaakkola and Jordan (2000), variational Bayes has become a popular way to learn otherwise intractable models. Recent books on the topic include MacKay (2003) and Bishop (2006).

The essential idea is to introduce a set of approximating densities to $p(\cdot|y)$. These approximations are then optimised so as to minimise the discrepancy between them and the true posteriors, using some measure of the difference. The optimisation is carried out by varying the parameters of these approximations, thus giving the approximation its name.

An alternative to variational Bayes is the numerical evaluation of the posterior distribution, such as Markov chain Monte Carlo algorithms. A variety of numerical inte-

gration methods have been developed with Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953); Hastings (1970); Gelfand and Smith (1990) and Gilks and Spiegelhalter (1996). An overview of developments is given by Robert and Casella (2004). Monte Carlo algorithms have been developed and applied to Bayesian inference (Pearl, 1988; Gilks, Thomas and Spiegelhalter, 1994). Numerical methods such as Monte Carlo can offer convergence to the true posterior distribution. However, a complex, high–dimensional integral may make make numerical integration methods prohibitively expensive. The Monte Carlo methods can be slow to converge, with convergence hard to diagnose (Cowles and Carlin, 1996).

This chapter examines the use of mean-field variational Bayes for semiparametric regression. We show that a close relationship exists between mean-field variational Bayes and classical techniques such as maximum likelihood (ML) and resticted maximum likelihood (REML). In particular, we derive REML as a special case of mean-field variational Bayes when applied to semiparametric regression. The Bayesian framework allows elegant generalisations. Following the approach of Albert and Chib (1995) and Girolami and Rogers (2006), we apply the mean-field variational approximation to the binary response situation. The resulting estimators are closely related to those of penalised quasi-likelihood, as given by Breslow and Clayton (1993) and Wolfinger and O'Connell (1993).

In the next section, we review some properties of mean-field variational approximation. In Section 4.3, we apply variational Bayes to the Bayesian linear mixed model and show the relationship with REML. Section 4.4 delves into the Baysian generalised linear mixed model, in particular, the probit mixed model. A discussion of this chapter is given in Section 4.5. Pseudo-code for algorithms of this chapter are given in Sections 4.A.6 and 4.A.7.

4.2 Mean Field Variational Approximation

The most common type of variational approximation involves the notion of *Kullback-Leibler convergence* applied to a *Bayesian network*. Bayesian networks correspond to models with hierarchical dependence structure, such as mixed models and empirical Bayes models. With nodes corresponding to parameters and to observed data, a *directed acyclic graph* (DAG) describes the dependence structure of a Bayesian network. Variational inference approximation has a wide literature. Our focus here is on Bayesian inference for semiparametric regression.

4.2.1 Kullback-Leibler Divergence

For arbitrary density functions q and p over Θ, the Kullback-Leibler (KL) divergence from q to p is

$$\mathrm{KL}[q\|p] = \int_\Theta q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{4.1}$$

It was shown by Kullback and Leibler (1951) that for all densities q, the divergence satisfies the inequality

$$\mathrm{KL}[q\|p] \geq 0. \tag{4.2}$$

Furthermore, if q is absolutely continuous, then there is equality in (4.2) if and only if $q = p$.

By a standard manipulation,

$$\begin{aligned}
\log p(\boldsymbol{y}) &= \log p(\boldsymbol{y}) \int_\Theta q(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int_\Theta q(\boldsymbol{\theta}) \log p(\boldsymbol{y})\, d\boldsymbol{\theta} \\
&= \int_\Theta q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})/q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})/q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta} \\
&= \int_\Theta q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta} + \int_\Theta q(\boldsymbol{\theta}) \log \left\{ \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} \right\} d\boldsymbol{\theta} \\
&= L(q) + \mathrm{KL}[q\|p(\cdot|\boldsymbol{y})]
\end{aligned}$$

where we have defined

$$L(q) \equiv \int_\Theta q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta}. \tag{4.3}$$

The quantity $L(q)$ serves as a lower bound for $\log p(\boldsymbol{y})$. Their difference, $\mathrm{KL}[q\|p(\cdot|\boldsymbol{y})]$, is the Kullback-Leibler divergence from the density q to the true posterior $p(\cdot|\boldsymbol{y})$. Having a small Kullback-Leibler divergence indicates that the probability distribution q is, in the sense of (4.1), close to the true $p(\cdot|\boldsymbol{y})$. In maximising $L(q)$, we serve to minimise $\mathrm{KL}[q\|p(\cdot|\boldsymbol{y})]$.

4.2.2 Factorised Density Transforms

With variational Bayes, we restrict the space of functions approximating densities q to a smaller class. The aim in doing so is to ensure that the integrals in (4.3) are tractable. The restriction that we use is that of a factorised density. Let us partition the elements of $\boldsymbol{\theta} \in \mathbb{R}^m$ into groups denoted by $\boldsymbol{\theta}_i$ for $1 \leq i \leq M$. We make the restriction on q that these groups are statistically independent. That is, q admits the factorisation

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i). \tag{4.4}$$

The factorisation restriction (4.4) is known in physics as the mean-field approximation (e.g., Parisi, 1988). This factorisation is the only restriction being placed on q; it is non-parametric in its assumptions. Note that if $M = 1$, then no restrictions are being made. The literature contains several alternatives to mean-field approximation. These include the Laplace approximation (e.g., Tierney and Kadane, 1986), as well as parametric assumptions (e.g., Attias, 2000). The mean-field assumption will often implicitly subsume the alternatives.

Amongst all distributions $q(\boldsymbol{\theta})$ having the form (4.4), we now seek the distribution for which the lower bound $L(q)$ is the largest. By a standard manipulation,

$$\begin{aligned}
L(q) &= \int_{\Theta} \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{\prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i)} \right\} d\boldsymbol{\theta} \\
&= \int_{\Theta} \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i) \left\{ \log p(\boldsymbol{y}, \boldsymbol{\theta}) - \sum_{i=1}^{M} \log q_i(\boldsymbol{\theta}_i) \right\} d\boldsymbol{\theta} \\
&= \int_{\Theta} \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i) \log p(\boldsymbol{y}, \boldsymbol{\theta}) \, d\boldsymbol{\theta} - \sum_{i=1}^{M} \int_{\Theta} q_i(\boldsymbol{\theta}_i) \log q_i(\boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_i \\
&= \int_{\Theta_j} q_j(\boldsymbol{\theta}_j) \left\{ \int_{\Theta_{i \neq j}} \log p(\boldsymbol{y}, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_{i \neq j} \right\} d\boldsymbol{\theta}_j - \sum_{i=1}^{M} \int_{\Theta_i} q_i(\boldsymbol{\theta}_i) \log q_i(\boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_i \\
&= \int_{\Theta_j} q_j(\boldsymbol{\theta}_j) \left\{ \mathsf{E}_{-\boldsymbol{\theta}_i} \log p(\boldsymbol{\theta}, \boldsymbol{y}) \right\} d\boldsymbol{\theta}_j - \sum_{i=1}^{M} \int_{\Theta_i} q_i(\boldsymbol{\theta}_i) \log q_i(\boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_i \qquad (4.5) \\
&= - \int_{\Theta_j} q_j(\boldsymbol{\theta}_j) \log \left\{ \frac{q_j(\boldsymbol{\theta}_j)}{\exp(\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}, \boldsymbol{y}))} \right\} d\boldsymbol{\theta}_j - \sum_{i \neq j} \int_{\Theta_i} q_i(\boldsymbol{\theta}_i) \log q_i(\boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_i,
\end{aligned}$$

where (4.5) has $\mathsf{E}_{-\boldsymbol{\theta}_j}$ denoting the expectation with respect to the density $\prod_{i \neq j} q_i(\boldsymbol{\theta}_i)$, for $1 \leq j \leq M$. Let z_j be the normalisation factor for $\exp(\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}, \boldsymbol{y}))$, i.e.,

$$z_j \equiv \int_{\Theta_{i \neq j}} \exp(\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}, \boldsymbol{y})) d\boldsymbol{\theta}_{i \neq j}.$$

On optimising $L(q)$ over $q_j(\boldsymbol{\theta}_j)$, we have

$$L(q) = - \int_{\Theta_j} q_j(\boldsymbol{\theta}_j) \log \left\{ \frac{q_j(\boldsymbol{\theta}_j)}{\exp(\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}, \boldsymbol{y}))/z_j} \right\} d\boldsymbol{\theta}_j + C,$$

where C is a generic constant, not dependent on $q_j(\boldsymbol{\theta}_j)$. We then recognise

$$\int_{\Theta_j} q_j(\boldsymbol{\theta}_j) \log \left\{ \frac{q_j(\boldsymbol{\theta}_j)}{\exp(\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}, \boldsymbol{y}))/z_j} \right\} d\boldsymbol{\theta}_j$$

as being the Kullback-Leibler divergence between $q_j(\boldsymbol{\theta}_j)$ and $\exp(\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}, \boldsymbol{y}))/z_j$. On maximising $L(q)$ with respect to $q_j(\boldsymbol{\theta}_j)$, we then find

$$q_j^*(\boldsymbol{\theta}_j) \propto \exp\{\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}, \boldsymbol{y})\}. \qquad (4.6)$$

The optimal factor $q_j^*(\boldsymbol{\theta}_j)$ is dependent on the choice of the remaining $q_i^*(\boldsymbol{\theta}_i)$, $i \neq j$. As j is any $1 \leq j \leq M$, equation (4.6) represents a set of M consistency conditions. It is required that each condition be satisfied simultaneously.

It is typical for variational Bayes for a coordinate ascent procedure to be used (e.g., Blei and Jordan, 2005; Ormerod and Wand, 2009). The coordinate ascent procedure is as follows. After initialising the factors, we cycle through $q_1^*(\boldsymbol{\theta}_1), \ldots, q_M^*(\boldsymbol{\theta}_M)$, maximising $L(q)$ with respect to each of the M factors individually. This is a recursive procedure, repeated until the change in $L(q)$ becomes negligible. Upon convergence, we have our optimal parameters. It is well known that the Kullback-Leibler divergence is convex in its first parameter, $\mathrm{KL}(\cdot, q)\colon \mathsf{P} \to \mathbb{R}$. As such, $L(\cdot)$ is concave, moreover, it is concave under the factorisation restriction (4.4). Under some mild assumptions (e.g., Luenberger and Yinyu, 2003, page 253), the convergence of coordinate ascent is guaranteed. We consider an alternative to coordinate ascent in Section 4.3.4.

4.2.3 Markov Blankets

A directed acyclic graph is an important component in the representing of a Bayesian network. For a DAG, a probability distribution p is called a Bayesian network with respect to the DAG if p admits the representation

$$p(\boldsymbol{y}, \theta_1, \ldots, \theta_m) = p(\boldsymbol{y}|\text{parents of } \boldsymbol{y}) \prod_{i=1}^{m} p(\theta_i|\text{parents of } \theta_i),$$

where the parents of each of the θ_i are given by the DAG.

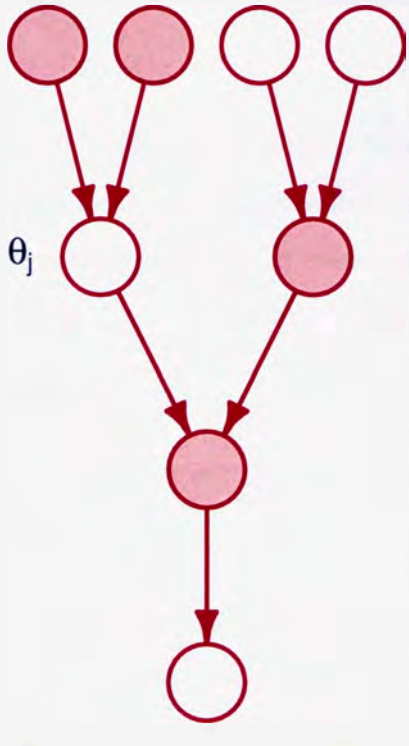


Figure 4.1. *An example of a directed acyclic graph. The four nodes comprising the Markov blanket of θ_j are shaded.*

The DAG representation of Bayesian models gives rise to a useful result arising from the notion of a *Markov blanket*. The term was coined by Pearl (1988). The Markov blanket of a node is the parents, children and other parents of the children. An example of a Markov blanket on a DAG is given in Figure 4.1. In the figure, the element labelled θ_j has two parents, one child, and one other parent of the child. The set of these four nodes comprise the Markov blanket of θ_j. The

key element of the Markov blanket is in the result

$$p(\theta_j|\text{rest}) = p(\theta_j|\text{Markov blanket of } \theta_j). \tag{4.7}$$

This means that determination of the required full conditionals involves only the Markov blanket. The Markov blanket is localised on the DAG, comprising only nearby nodes. It is hence only the nearby nodes that determine the conditional distribution. It follows from this fact and expression (4.6) that the factorised density approach involves only local calculations on the DAG. Mathematically, this is

$$\begin{aligned} q_j^*(\boldsymbol{\theta}_j) &\propto \exp\{\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}, \boldsymbol{y})\} \\ &\propto \exp\{\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}_j|\text{rest})\} \\ &= \exp\{\mathsf{E}_{-\boldsymbol{\theta}_j} \log p(\boldsymbol{\theta}_j|\text{Markov blanket of } \boldsymbol{\theta}_j)\}, \end{aligned}$$

where the last line follows from the Markov blanket result (4.7).

4.3 Gaussian Response Semiparametric Regression

The Bayesian version of the Gaussian linear mixed model takes the general form

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{G}, \boldsymbol{R} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}), \quad \boldsymbol{u}|\boldsymbol{G} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}) \tag{4.8}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{u}$ is vector of random effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are corresponding design matrices and $\boldsymbol{G}$ and $\boldsymbol{R}$ are covariance matrices. Several possibilities exist for $\boldsymbol{G}$ and $\boldsymbol{R}$ (e.g., McCulloch, Searle and Neuhaus, 2008). For now, we restrict attention to variance component models with

$$\boldsymbol{G} = \text{blockdiag}(\sigma_{u1}^2 \boldsymbol{I}_{K_1}, \ldots, \sigma_{ur}^2 \boldsymbol{I}_{K_r}) \quad \text{and} \quad \boldsymbol{R} = \sigma_\varepsilon^2 \boldsymbol{I}. \tag{4.9}$$

We also impose the conjugate priors:

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}), \quad \sigma_{ui}^2 \sim \mathrm{IG}(A_{ui}, B_{ui}),\ 1 \leq i \leq r, \quad \sigma_\varepsilon^2 \sim \mathrm{IG}(A_\varepsilon, B_\varepsilon) \tag{4.10}$$

for some $\sigma_\beta^2, A_{ui}, B_{ui}, A_\varepsilon, B_\varepsilon > 0$. The DAG representation of the Bayesian Gaussian linear mixed model (4.8)–(4.10) is displayed in Figures 4.2 and 4.3.

We find that a tractable solution arises with the two-component factorisation

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2) = q_{\boldsymbol{\beta},\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{u}) q_{\sigma^2}(\sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2). \tag{4.11}$$

It is shown in Appendix 4.A.1 that application of (4.6) leads to the optimal densities taking the form

$$q_{\boldsymbol{\beta},\boldsymbol{u}}^*(\boldsymbol{\beta}, \boldsymbol{u}) \sim \mathrm{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q), \quad \text{and} \quad q_{\sigma^2}^*(\sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2) = q_{\sigma_\varepsilon^2}(\sigma_\varepsilon^2) \prod_{i=1}^{r} q_{\sigma_{ui}^2}(\sigma_{ui}^2), \tag{4.12}$$

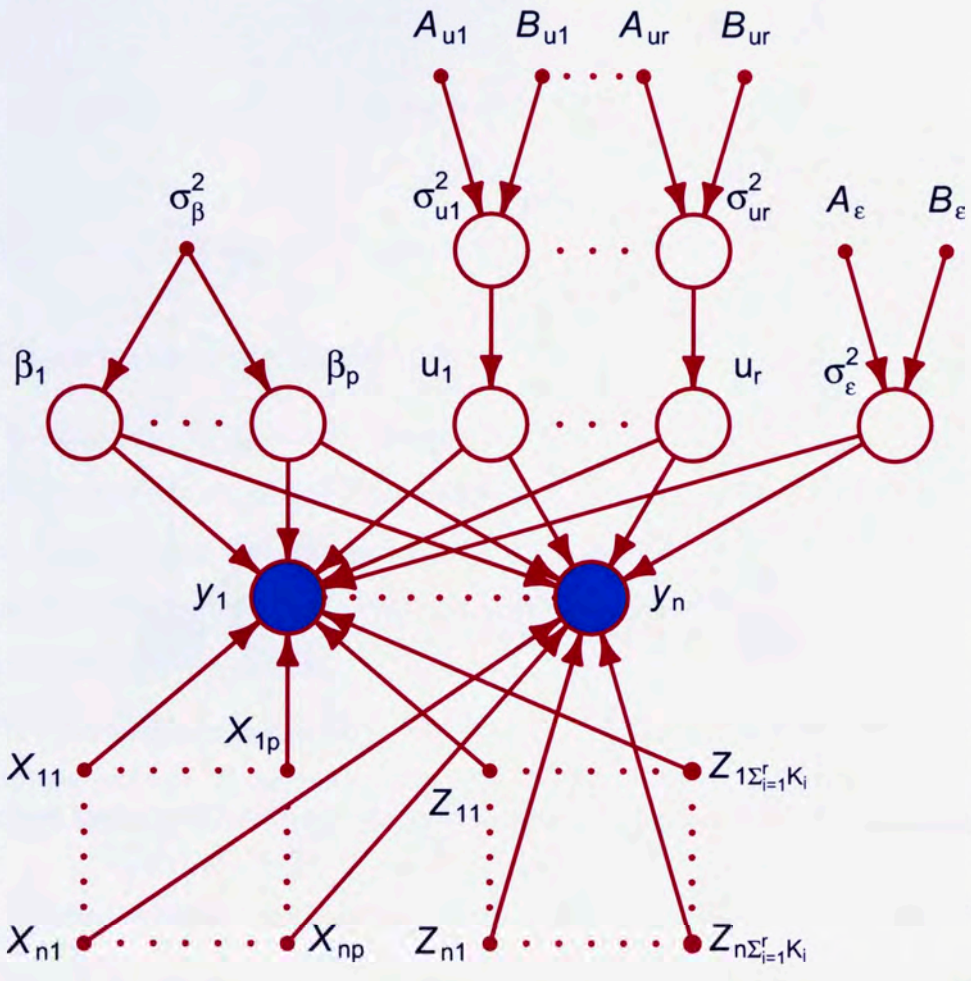


Figure 4.2. *Directed acyclic graph representing the Bayesian linear mixed model* (4.8)-(4.10). *Large nodes correspond to scalar random variables in the model, with the observed random variables shaded. The smaller nodes correspond to constants.*

with

$$q_{\sigma^2_\varepsilon} \sim \text{IG}(A_{q,\varepsilon}, B_{q,\varepsilon}), \quad \text{and} \quad q_{\sigma^2_{ui}} \sim \text{IG}(A_{q,ui}, B_{q,ui}). \tag{4.13}$$

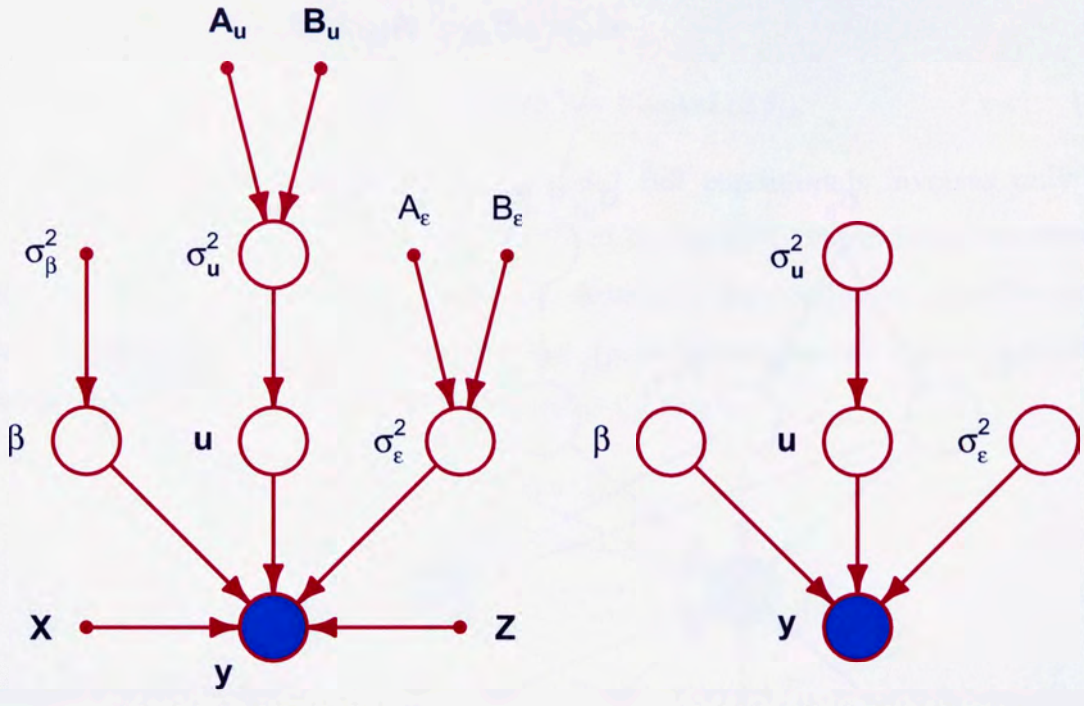
The parameters $A_{q,\varepsilon}$ and $A_{q,ui}$ are deterministic,

$$A_{q,\varepsilon} = A_\varepsilon + \tfrac{n}{2} \quad \text{and} \quad A_{q,ui} = A_{ui} + \tfrac{K_i}{2},\ 1 \le i \le r.$$

The parameters $\boldsymbol{\mu}_q$ and $\boldsymbol{\Sigma}_q$ in (4.12), and $B_{q,\varepsilon}$ and $B_{q,ui}$ in (4.13), however, are dependent on q. Let $(\boldsymbol{\mu}_q)_{\boldsymbol{\beta}}$ and $(\boldsymbol{\mu}_q)_{\boldsymbol{u}_i}$ denotes the components of $\boldsymbol{\mu}_q$ corresponding to $\boldsymbol{\beta}$ and $\boldsymbol{u}_i$ respectively;

$$(\boldsymbol{\mu}_q)_{\boldsymbol{\beta}} \equiv \begin{bmatrix} (\boldsymbol{\mu}_q)_1 \\ \vdots \\ (\boldsymbol{\mu}_q)_p \end{bmatrix}, \quad \text{and} \quad (\boldsymbol{\mu}_q)_{\boldsymbol{u}_i} \equiv \begin{bmatrix} (\boldsymbol{\mu}_q)_{p+1+\sum_{j=1}^{i} K_j} \\ \vdots \\ (\boldsymbol{\mu}_q)_{p+\sum_{j=1}^{i+1} K_j} \end{bmatrix}, \quad \text{for} \quad 1 \le i \le r.$$

Similarly, let $(\boldsymbol{\Sigma}_q)_{\boldsymbol{\beta}}$ denote the $p \times p$ matrix corresponding to the $\boldsymbol{\beta}$ components of $\boldsymbol{\Sigma}_q$, and $(\boldsymbol{\Sigma}_q)_{\boldsymbol{u}_i}$ to the $K_i \times K_i$ matrix corresponding to the $\boldsymbol{u}_i$ components of $\boldsymbol{\Sigma}_q$. Furthermore,



(a) *With fixed effects shown* **(b)** *Without fixed effects shown*

Figure 4.3. *Directed acyclic graph representing the Bayesian linear mixed model.*

letting $C \equiv [X\ Z]$, coordinate ascent gives the following four update equations:

$$\begin{aligned}
\boldsymbol{\Sigma}_q &\leftarrow \left\{ \frac{A_{q,\varepsilon}}{B_{q,\varepsilon}} \boldsymbol{C}^{\mathsf{T}}\boldsymbol{C} + \text{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p, \frac{A_{q,u1}}{B_{q,u1}}\boldsymbol{I}_{K_1}, \ldots, \frac{A_{q,ur}}{B_{q,ur}}\boldsymbol{I}_{K_r} \right) \right\}^{-1}, \\
\boldsymbol{\mu}_q &\leftarrow \left(\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}} \right) \boldsymbol{\Sigma}_q \boldsymbol{C}^{\mathsf{T}}\boldsymbol{y}, \\
B_{q,\varepsilon} &\leftarrow B_\varepsilon + \tfrac{1}{2}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_q\|^2 + \text{tr}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\Sigma}_q)\}, \quad \text{and} \\
B_{q,ui} &\leftarrow B_{ui} + \tfrac{1}{2}\{\|(\boldsymbol{\mu}_q)_{\boldsymbol{u}_i}\|^2 + \text{tr}((\boldsymbol{\Sigma}_q)_{\boldsymbol{u}_i})\}, \quad \text{for} \quad 1 \le i \le r,
\end{aligned}$$

These update equations arise from maximising $L(q)$ with respect to either $q_{\boldsymbol{\beta},\boldsymbol{u}}$ or q_{σ^2}, and are derived in Appendix 4.A.1. Furthermore, pseudo-code is given in Appendix 4.A.6, and allows for the $B_{q,\varepsilon}$ update to be performed in an efficient manner. After a complete cycle though the updates, the lower bound $L(q)$ takes the form:

$$\begin{aligned}
L(q) = \frac{p + \sum_{i=1}^r K_i}{2} - \frac{n}{2}\log(2\pi) - \frac{p}{2}\log(\sigma_\beta^2) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_q| - \frac{\|(\boldsymbol{\mu}_q)_{\boldsymbol{\beta}}\|^2 + \text{tr}((\boldsymbol{\Sigma}_q)_{\boldsymbol{\beta}})}{2\sigma_\beta^2} \\
+ A_\varepsilon \log(B_\varepsilon) - A_{q,\varepsilon}\log(B_{q,\varepsilon}) + \log\Gamma(A_{q,\varepsilon}) - \log\Gamma(A_\varepsilon) \\
+ \sum_{i=1}^r A_{ui}\log(B_{ui}) - A_{q,ui}\log(B_{q,ui}) + \log\Gamma(A_{q,ui}) - \log\Gamma(A_{ui}).
\end{aligned} \tag{4.14}$$

We refer the reader to Appendix 4.A.2 for the derivation of this expression.

Upon convergence to $\boldsymbol{\mu}_q^*, \boldsymbol{\Sigma}_q^*, B_{q,u1}^*, \ldots, B_{q,ur}^*$ and $B_{q,\varepsilon}^*$, the approximate posteriors are:

$$p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y}) \simeq \text{the } \text{N}(\boldsymbol{\mu}_q^*, \boldsymbol{\Sigma}_q^*) \text{ density in } (\boldsymbol{\beta}, \boldsymbol{u})$$

and

$$p(\sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2 | \boldsymbol{y}) \simeq \text{product of IG}(A_{q,ui}^*, B_{q,ui}^*),\ 1 \leq i \leq r,$$
$$\text{and IG}(A_{q,\varepsilon}^*, B_{q,\varepsilon}^*) \text{ densities,}$$

where $A_{q,ui}^* = A_{ui} + \frac{K_i}{2}$ for $1 \leq i \leq r$, and $A_{q,\varepsilon}^* = A_\varepsilon + \frac{n}{2}$.

4.3.1 Characterising the Optimality

We find that for large K_i, that the convergence of the coordinate ascent is slow. We have a more serious problem in infinite dimensions; if $K_i = \infty$ (as with Gaussian processes) the algorithm breaks down. To use a rich kernel, such as the Gaussian kernel, an alternative algorithm is to be found. Similar problems have been noted by Gibbs and MacKay (2000) and Opper and Winther (2000).

We now consider the optimality conditions, in order to ascertain the properties of the maximiser of $L(q)$. These optimality conditions include

$$B_{q,ui}^* = B_{ui} + \tfrac{1}{2}\{\|(\boldsymbol{\mu}_q^*)_{\boldsymbol{u}_i}\|^2 + \operatorname{tr}((\boldsymbol{\Sigma}_q^*)_{\boldsymbol{u}_i})\}, \quad \text{for} \quad 1 \leq i \leq r,$$

therefore,

$$\frac{B_{q,ui}^*}{A_{q,ui}^*} = \frac{B_{ui} + \frac{1}{2}\left\{\left\|\left(\frac{A_{q,\varepsilon}^*}{B_{q,\varepsilon}^*}\boldsymbol{\Sigma}_q^* \boldsymbol{C}^\mathsf{T}\boldsymbol{y}\right)_{\boldsymbol{u}_i}\right\|^2 + \operatorname{tr}((\boldsymbol{\Sigma}_q^*)_{\boldsymbol{u}_i})\right\}}{A_{ui} + \frac{K_i}{2}}. \tag{4.15}$$

Some algebra shows that $\operatorname{tr}((\boldsymbol{\Sigma}_q^*)_{\boldsymbol{u}_i})$ admits the expression

$$\operatorname{tr}((\boldsymbol{\Sigma}_q^*)_{\boldsymbol{u}_i}) = \frac{B_{q,ui}^*}{A_{q,ui}^*}\left\{K_i - \frac{A_{q,\varepsilon}^*}{B_{q,\varepsilon}^*}\operatorname{tr}((\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_q^*)_{\boldsymbol{u}_i})\right\}. \tag{4.16}$$

Substituting (4.16) into (4.15) and rearranging then gives

$$\left\{A_{ui} + \frac{A_{q,\varepsilon}^*}{2B_{q,\varepsilon}^*}\operatorname{tr}((\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_q^*)_{\boldsymbol{u}_i})\right\}\frac{B_{q,ui}^*}{A_{q,ui}^*} = B_{ui} + \tfrac{1}{2}\left\|\left(\frac{A_{q,\varepsilon}^*}{B_{q,\varepsilon}^*}\boldsymbol{\Sigma}_q^*\boldsymbol{C}^\mathsf{T}\boldsymbol{y}\right)_{\boldsymbol{u}_i}\right\|^2. \tag{4.17}$$

We now make a similar argument for $A_{q,\varepsilon}^*$ and $B_{q,\varepsilon}^*$. The optimality conditions give

$$\begin{aligned}\frac{B_{q,\varepsilon}^*}{A_{q,\varepsilon}^*} &= \frac{B_\varepsilon + \frac{1}{2}\left\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_q^*\|^2 + \operatorname{tr}(\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_q^*)\right\}}{A_\varepsilon + \frac{n}{2}} \\ &= \frac{B_\varepsilon + \frac{1}{2}\left\{\left\|\boldsymbol{y} - \frac{A_{q,\varepsilon}^*}{B_{q,\varepsilon}^*}\boldsymbol{C}\boldsymbol{\Sigma}_q^*\boldsymbol{C}^\mathsf{T}\boldsymbol{y}\right\|^2 + \operatorname{tr}(\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_q^*)\right\}}{A_\varepsilon + \frac{n}{2}}.\end{aligned} \tag{4.18}$$

Let $\boldsymbol{V}^* \equiv \text{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p, \frac{A_{q,u1}^*}{B_{q,u1}^*}\boldsymbol{I}_{K_1}, \ldots, \frac{A_{q,ur}^*}{B_{q,ur}^*}\boldsymbol{I}_{K_r}\right)$. Then $\operatorname{tr}(\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_q^*)$ admits the expression

$$\operatorname{tr}(\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_q^*) = \frac{B_{q,\varepsilon}^*}{A_{q,\varepsilon}^*}\left\{p + \left(\sum_{i=1}^r K_i\right) - \operatorname{tr}(\boldsymbol{V}^*\boldsymbol{\Sigma}_q^*)\right\}. \tag{4.19}$$

Substituting (4.19) into (4.18) and rearranging then gives

$$\left\{A_\varepsilon + \tfrac{1}{2}\left(n - p - \sum_{i=1}^{r} K_i\right) + \tfrac{1}{2}\mathrm{tr}(\boldsymbol{V}^*\boldsymbol{\Sigma}_q^*)\right\}\frac{B_{q,\varepsilon}^*}{A_{q,\varepsilon}^*} = B_\varepsilon + \tfrac{1}{2}\left\|\boldsymbol{y} - \frac{A_{q,\varepsilon}^*}{B_{q,\varepsilon}^*}\boldsymbol{C}\boldsymbol{\Sigma}_q^*\boldsymbol{C}^\mathsf{T}\boldsymbol{y}\right\|^2. \quad (4.20)$$

The solution $\left\{\frac{B_{q,u1}^*}{A_{q,u1}^*}, \ldots, \frac{B_{q,ur}^*}{A_{q,ur}^*}, \frac{B_{q,\varepsilon}^*}{A_{q,\varepsilon}^*}\right\}$ to equations (4.17) and (4.20) may be considered as comprising part of the solution to the variational Bayes optimisation. That is, we may solve in terms of the ratios $\frac{B_{q,ui}^*}{A_{q,ui}^*}$, without the requirement that $B_{q,ui}^* < \infty$. A similar approach to Bayesian problems have been used in Wahba (1985).

4.3.2 A Dual Space Formulation

Unlike the coordinate ascent, the optimality conditions of (4.17) and (4.20) allow for a dual space formulation. This ensures that the algorithm may be kernelised. Other authors, such as Harville (1974); Hastie and Tibshirani (2004) and Friston *et al.* (2006) have given versions of algorithms in both primal and dual forms. Let us begin by setting

$$\boldsymbol{\Pi}_q^* \equiv \left\{\sigma_\beta^2\boldsymbol{X}\boldsymbol{X}^\mathsf{T} + \sum_{i=1}^{r}\frac{B_{q,ui}^*}{A_{q,ui}^*}\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T} + \frac{B_{q,\varepsilon}^*}{A_{q,\varepsilon}^*}\boldsymbol{I}_n\right\}^{-1}. \quad (4.21)$$

We now express the equations (4.17) and (4.20) in a manner amenable to kernelisation. It is shown in Appendix 4.A.3 that the solution to the Bayesian Gaussian linear mixed model requires

$$\left\{A_{ui} + \frac{B_{q,ui}^*}{2A_{q,ui}^*}\mathrm{tr}(\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\Pi}_q^*)\right\}\frac{B_{q,ui}^*}{A_{q,ui}^*} = B_{ui} + \tfrac{1}{2}\left(\frac{B_{q,ui}^*}{A_{q,ui}^*}\right)^2\boldsymbol{y}^\mathsf{T}\boldsymbol{\Pi}_q^*\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\Pi}_q^*\boldsymbol{y}, \quad (4.22)$$

for all $1 \leq i \leq r$, and,

$$\left\{A_\varepsilon + \frac{B_{q,\varepsilon}^*}{2A_{q,\varepsilon}^*}\mathrm{tr}(\boldsymbol{\Pi}_q^*)\right\}\frac{B_{q,\varepsilon}^*}{A_{q,\varepsilon}^*} = B_\varepsilon + \tfrac{1}{2}\left(\frac{B_{q,\varepsilon}^*}{A_{q,\varepsilon}^*}\right)^2\|\boldsymbol{\Pi}_q^*\boldsymbol{y}\|^2. \quad (4.23)$$

The formulation of (4.22) and (4.23) allows the Bayesian Gaussian linear mixed model to be expressed in terms of inner products. Let $\boldsymbol{K}_i \equiv \boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}$ for $1 \leq i \leq r$ in (4.21). Then (4.22) simplifies to

$$\left\{A_{ui} + \frac{B_{q,ui}^*}{2A_{q,ui}^*}\mathrm{tr}(\boldsymbol{K}_i\boldsymbol{\Pi}_q)\right\}\frac{B_{q,ui}^*}{A_{q,ui}^*} = B_{ui} + \tfrac{1}{2}\left(\frac{B_{q,ui}^*}{A_{q,ui}^*}\right)^2\boldsymbol{y}^\mathsf{T}\boldsymbol{\Pi}_q^*\boldsymbol{K}_i\boldsymbol{\Pi}_q^*\boldsymbol{y}. \quad (4.24)$$

In the expression (4.24), the to reference to $\boldsymbol{Z}$ occurs only through the inner products, $\boldsymbol{K}_i = \boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}$, for $1 \leq i \leq r$. As such, the Gaussian linear mixed model (4.8)-(4.10) may be kernelised; with positive definite functions k_i, for each $1 \leq i \leq r$.

4.3.3 Relation to Restricted Maximum Likelihood

We now show that ML and REML are a special cases of the variational Bayes framework. This helps improve the interpretation of the mean-field approximation variational Bayes. Introduced by Patterson and Thompson (1971), REML is a mature and well–regarded method method for estimation of covariance matrices or variance parameters in semi-parametric regression problems. There are stable algorithms for performing either ML or REML. These can be adapted for the variational Bayes framework.

Theorem 4.1. *Let* $\left\{ \frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}, \frac{B^*_{q,u1}}{A^*_{q,u1}}, \ldots, \frac{B^*_{q,ur}}{A^*_{q,ur}} \right\}$ *be a stationary point to the likelihood function of the Gaussian linear mixed model* (4.8)-(4.10). *Let*

$$\psi_0 \equiv \frac{B_{q,\varepsilon}}{A_{q,\varepsilon}} \quad \text{and} \quad \psi_i \equiv \frac{B_{q,ui}}{A_{q,ui}}, \quad \text{for} \quad 1 \leq i \leq r.$$

Then $\{\psi_0^*, \psi_1^*, \ldots, \psi_r^*\} \equiv \left\{ \frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}, \frac{B^*_{q,u1}}{A^*_{q,u1}}, \ldots, \frac{B^*_{q,ur}}{A^*_{q,ur}} \right\}$ *is a stationary point of*

$$\max_{\psi_0,\psi_1,\ldots,\psi_r} \tfrac{1}{2} \log |\boldsymbol{\Pi}_q| - \tfrac{1}{2} \boldsymbol{y}^{\mathsf{T}} \boldsymbol{\Pi}_q \boldsymbol{y} - A_\varepsilon \log \psi_0 - B_\varepsilon \psi_0^{-1} - \sum_{i=1}^{r} \left(A_{ui} \log \psi_i + B_{ui} \psi_i^{-1} \right), \quad (4.25)$$

where

$$\boldsymbol{\Pi}_q = \left\{ \sigma_\beta^2 \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}} + \sum_{i=1}^{r} \psi_i \boldsymbol{Z}_i \boldsymbol{Z}_i^{\mathsf{T}} + \psi_0 \boldsymbol{I}_n \right\}^{-1}.$$

Proof of Theorem 4.1 is given in Appendix 4.A.4. The Gaussian linear mixed model (4.8)-(4.10) has the prior parameters A_{ui}, B_{ui}, A_ε, B_ε and σ_β. The prior distributions are improper in the limit as A_{ui}, B_{ui}, A_ε, $B_\varepsilon \to 0^+$ and $\sigma_\beta \to \infty$. It is noted that equation (4.25) bears some resemblance to the classical log-likelihood. The expression in (4.25) is recognised as a constant plus the log-likelihood of an $n + r + 1$-dimensional Gaussian distribution. We now clarify a simple connection, as a direct consequence of Theorem 4.1.

Theorem 4.2. *Let* A_{ui}, B_{ui}, A_ε, $B_\varepsilon \to 0^+$. *Then a maximal point to the optimisation in* (4.25) *gives a stationary point to*

$$\max_{\{\psi_0,\psi_1,\ldots,\psi_r\} \in \Omega} \tfrac{1}{2} \log |\boldsymbol{\Pi}_q| - \tfrac{1}{2} \boldsymbol{y}^{\mathsf{T}} \boldsymbol{\Pi}_q \boldsymbol{y}, \quad (4.26)$$

where the domain is

$$\Omega = \{\psi_0 \geq 0, \ldots, \psi_r \geq 0\}.$$

We recognise (4.26) as a constant plus the log-likelihood of an n-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Pi}_q^{-1}$. It is well known that such likelihoods may have local maxima. For a recent discussion on such bimodality, see

Welham and Thompson (2009). As $\sigma_\beta \to \infty$, the equivalent optimisation problem (e.g., Patterson and Thompson, 1971; Harville, 1977) is

$$\min_{\psi_0,\psi_1,\ldots,\psi_r} \tfrac{1}{2}\log|\boldsymbol{\Pi}_R| - \tfrac{1}{2}\log|\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{\Pi}_R\boldsymbol{X}^*| - \tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\mathsf{T}}\boldsymbol{\Pi}_R(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}), \tag{4.27}$$

where

$$\boldsymbol{\Pi}_R = \left\{\sum_{i=1}^{r}\psi_i\,\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathsf{T}} + \psi_0\boldsymbol{I}_n\right\}^{-1},$$

and $\boldsymbol{X}^*$ is a matrix made up of linearly independent columns of $\boldsymbol{X}$, with

$$\text{rank}(\boldsymbol{X}^*) = \text{rank}(\boldsymbol{X}),$$

and

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y},$$

where $(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-}$ denotes the Moore-Penrose generalised inverse of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$. The quantity in (4.27) is know as the restricted log-likelihood. The stationary points of (4.27) over Ω are REML estimates.

The relationship between REML and empirical Bayes is well known (Harville, 1974). It perhaps should not then be surprising that improper Bayesian conjugate priors lead to REML estimates. A close relationship between parametric empirical Bayes and REML is shown in Friston *et al.* (2002). The link between Laplacian approximations and REML is shown in Friston *et al.* (2006). Moreover, it is shown in Wipf *et al.* (2007) that, regardless of the choice of prior, there is always a relationship between Bayesian models and REML. The given technique, however, does not specify the functional form (4.25). With proper priors, variational Bayes gives a lower bound for $\log p(\boldsymbol{y})$ in (4.14). The existence of posterior probability estimates then distinguishes variational Bayes from either REML or from empirical Bayes. With a lower bound on $\log p(\boldsymbol{y})$, it is then straightforward to approximate the deviance information criterion for comparing models (Spiegelhalter *et al.*, 2002).

4.3.4 Optimising the Parameters of Variational Bayes

We now detail an alternative to the coordinate ascent algorithm. We have shown the maximiser of $L(q)$ under the factorisation restriction (4.11) is the same as the maximiser of REML-style optimisation in (4.25). It is now (4.25) to which we apply a standard optimisation technique. Let $\psi_0 \equiv \frac{B_{q,\varepsilon}}{A_{q,\varepsilon}}$ and $\psi_i \equiv \frac{B_{q,ui}}{A_{q,ui}}$, for $1 \leq i \leq r$. The dual space

formulation is

$$\begin{aligned}
\boldsymbol{\Pi}_q &= \left\{ \sigma_\beta^2 \boldsymbol{X}\boldsymbol{X}^\mathsf{T} + \sum_{i=1}^{r} \psi_i \boldsymbol{K}_i + \psi_0 \boldsymbol{I}_n \right\}^{-1}, \\
\psi_0 &= \frac{B_\varepsilon + \frac{1}{2}\psi_0^2 \|\boldsymbol{\Pi}_q \boldsymbol{y}\|^2}{A_\varepsilon + \frac{1}{2}\psi_0 \mathrm{tr}(\boldsymbol{\Pi}_q)}, \quad \text{and} \\
\psi_i &= \frac{B_{ui} + \frac{1}{2}\psi_i^2 \boldsymbol{y}^\mathsf{T} \boldsymbol{\Pi}_q \boldsymbol{K}_i \boldsymbol{\Pi}_q \boldsymbol{y}}{A_{ui} + \frac{1}{2}\psi_i \mathrm{tr}(\boldsymbol{K}_i \boldsymbol{\Pi}_q)}, \quad \text{for} \quad 1 \le i \le r.
\end{aligned}$$

A reliable method for finding ML estimates is in the method of successive approximations (e.g., Harville, 1977). We have the updates

$$\begin{aligned}
\boldsymbol{\Pi}_q &\leftarrow \left\{ \sigma_\beta^2 \boldsymbol{X}\boldsymbol{X}^\mathsf{T} + \sum_{i=1}^{r} \psi_i \boldsymbol{K}_i + \psi_0 \boldsymbol{I}_n \right\}^{-1}, \\
\psi_0 &\leftarrow \frac{B_\varepsilon + \frac{1}{2}\psi_0^2 \|\boldsymbol{\Pi}_q \boldsymbol{y}\|^2}{A_\varepsilon + \frac{1}{2}\psi_0 \mathrm{tr}(\boldsymbol{\Pi}_q)}, \quad \text{and} \\
\psi_i &\leftarrow \frac{B_{ui} + \frac{1}{2}\psi_i^2 \boldsymbol{y}^\mathsf{T} \boldsymbol{\Pi}_q \boldsymbol{K}_i \boldsymbol{\Pi}_q \boldsymbol{y}}{A_{ui} + \frac{1}{2}\psi_i \mathrm{tr}(\boldsymbol{K}_i \boldsymbol{\Pi}_q)}, \quad \text{for} \quad 1 \le i \le r.
\end{aligned}$$

Coordinate ascent can also be applied to the equivalent primal form. The optimality conditions, as update equations, are

$$\begin{aligned}
\boldsymbol{V} &\leftarrow \mathrm{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p, \psi_1^{-1}\boldsymbol{I}_{K_1}, \ldots, \psi_r^{-1}\boldsymbol{I}_{K_r}\right) \\
\boldsymbol{\Sigma}_q &\leftarrow \left\{ \psi_0^{-1}\boldsymbol{C}^\mathsf{T}\boldsymbol{C} + \boldsymbol{V} \right\}^{-1}, \\
\boldsymbol{\mu}_q &\leftarrow \psi_0^{-1}\boldsymbol{\Sigma}_q \boldsymbol{C}^\mathsf{T}\boldsymbol{y} \\
\psi_0 &\leftarrow \frac{B_\varepsilon + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_q\right\|^2}{A_\varepsilon + \frac{1}{2}\left(n - p - \sum_{i=1}^{r} K_i\right) + \frac{1}{2}\psi_0 \mathrm{tr}(\boldsymbol{V}\boldsymbol{\Sigma}_q)}, \quad \text{and} \\
\psi_i &\leftarrow \frac{B_{ui} + \frac{1}{2}\left\|\left(\boldsymbol{\mu}_q\right)_{\boldsymbol{u}_i}\right\|^2}{A_{ui} + \frac{1}{2}\left(K_i - \mathrm{tr}\left(\left(\boldsymbol{V}\boldsymbol{\Sigma}_q\right)_{\boldsymbol{u}_i}\right)\right)}, \quad \text{for} \quad 1 \le i \le r.
\end{aligned}$$

The sequence of updates $\{\psi_0, \ldots, \psi_r\}$ are equivalent whether made in the dual or primal form. Computationally, however, there are important differences. Naïvely, it takes $O(n^3)$ operations to calculate $\boldsymbol{\Pi}_q$, and $O((p + \sum_{i=1}^{r} K_i)^3)$ operations to calculate $\boldsymbol{\Sigma}_q$. Each ϕ_i update in the dual form requires $O(n^3)$ operations. As such, a complete iteration in the dual form takes $O((r+n)n^2)$. In the primal form, a $\boldsymbol{\mu}_q$ update requires some $O(n(p + \sum_{i=1}^{r} K_i))$, for an overall cost per iteration of $O((n + p + \sum_{i=1}^{r} K_i)(p + \sum_{i=1}^{r} K_i)^2)$ operations. The dual form is suitable for small n, and the primal form is suitable for small $p + \sum_{i=1}^{r} K_i$. The cost per iteration of the coordinate ascent is very similar to that of the primal form.

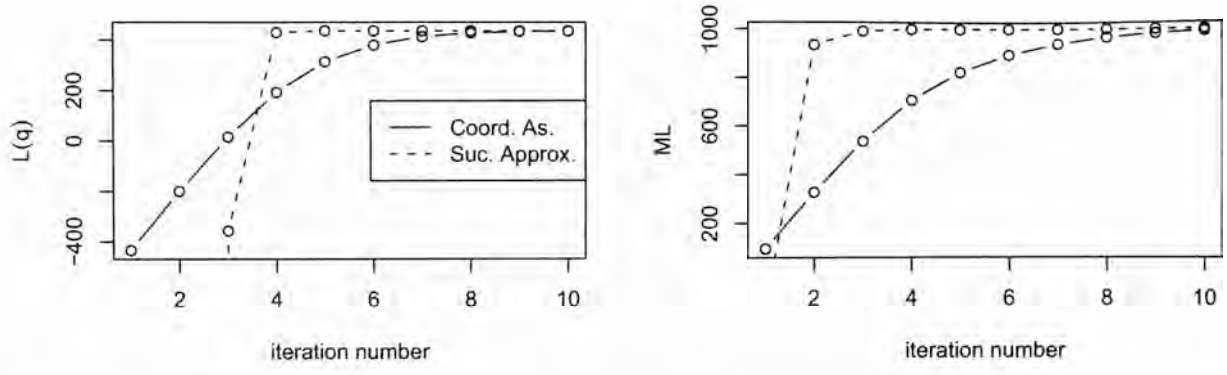


Figure 4.4. *Comparison of the convergence of coordinate ascent and successive approximation methods. The details of the penalised spline model used are given in Section 4.3.5. The comparison is made under two different optimisation criteria that share the same optimum. Under both criteria, successive approximation method exhibits faster convergence. Left: Convergence in* $L(q)$*, as given by equation* (4.34)*. For the coordinate ascent,* (4.34) *simplifies to* (4.14)*. Right: Convergence in the simpler optimisation given by* (4.25)*.*

A comparison with coordinate ascent is made. A simple spline model was fitted with twenty knots; the details are given in the next section. Figure 4.4 shows the convergence of the coordinate ascent and successive approximation methods. The left side plot shows the convergence in $L(q)$. The successive approximation method displays a better convergence than does that of coordinate ascent. On the right side of Figure 4.4, we make comparisons under the log-likelihood style optimisation criteria given by (4.25). There is a clear preference for the successive approximation method.

4.3.5 Spinal Bone Mineral Example

We now give an example of the spinal bone mineral density data set, as per Chapter 3 and Bachrach *et al.* (1999). The longitudinal data set includes a cohort of 193 young males, with subjects categorised as belonging to one of four ethnicity groups: Asian, Black, Hispanic and White. With single subscript notation, the model is

$$y_i = [1 \ (\boldsymbol{x}_i^{\ell})^{\mathsf{T}}]\boldsymbol{\beta} + c(x_i) + U_i + \varepsilon_i$$

where the y_i are spinal bone mineral measurements (g/cm^2), the $\boldsymbol{x}_i^{\ell}$ contain indicators for ethnicity and the x_i are age measurements. The function $c\colon \mathbb{R} \to \mathbb{R}$ indicates a curve.

Two different kernels were chosen to model the curvature. The first was a penalised spline kernel of Chapter 2, with 20 knots equally spaced over the observed domain of

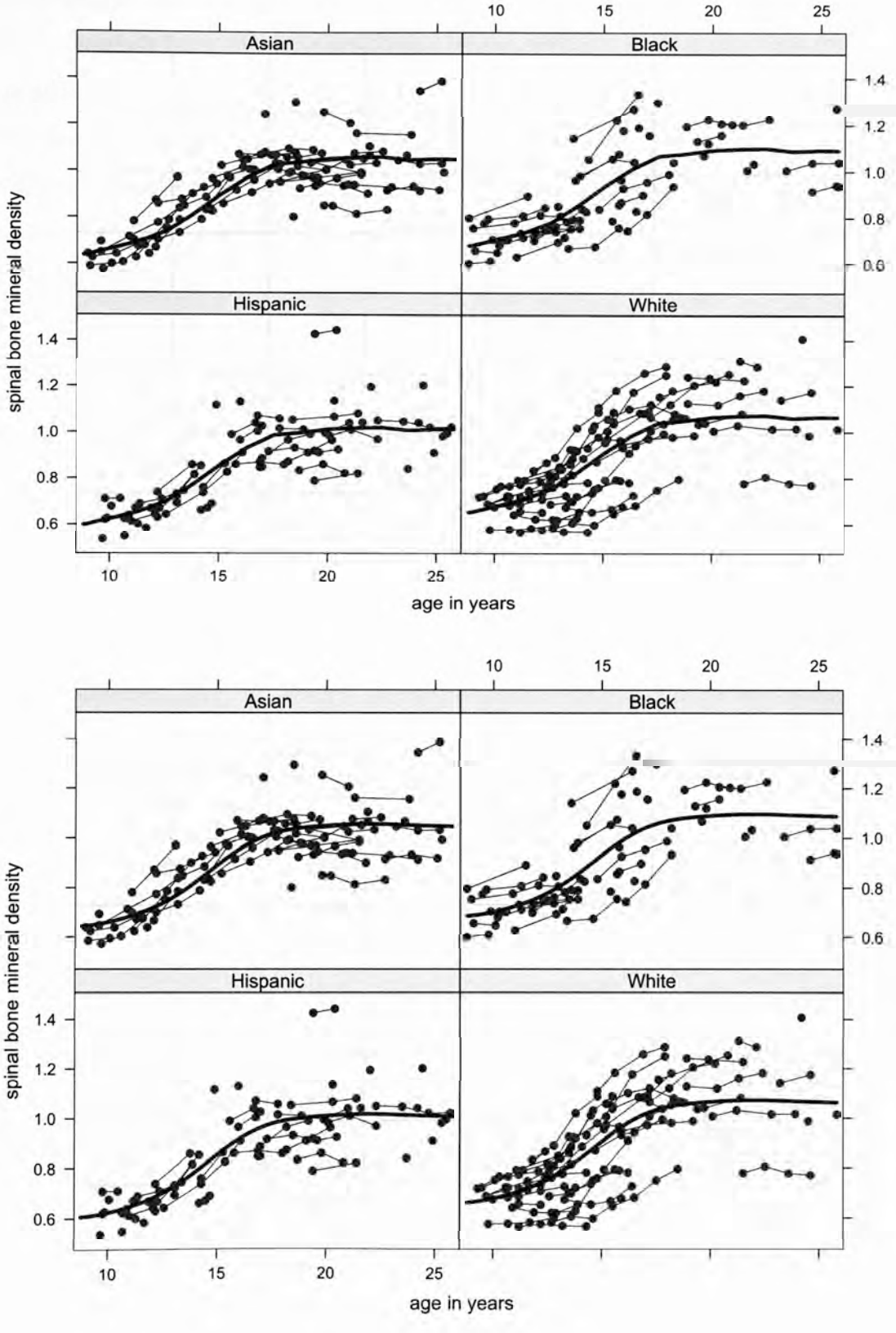


Figure 4.5. *Spinal bone mineral density measurements from the male cohort. Fits were made using Variational Bayes. Both curves fit the data well. Upper: Penalised spline fit, with twenty evenly spaced knots. Lower: Gaussian kernel based fit.*

ages. As a low-rank kernel, the primal updates were used. The second kernel was a Gaussian kernel with $\gamma = 0.05$. As the corresponding design matrix $\boldsymbol{Z}_1$ is infinite dimensional, the optimisation must be carried out in the dual. We used the uninformative prior parameters $A_\varepsilon = A_{u1} = A_{u2} = B_\varepsilon = B_{u1} = B_{u2} = 0.1$, and $\sigma_\beta = 10^8$. The spline fit gave $\widehat{\psi}_1 = 1.52 \times 10^{-2}$, $\widehat{\psi}_2 = 1.49 \times 10^{-2}$ and $\widehat{\psi}_0 = 1.98 \times 10^{-3}$. The Gaussian kernel gave $\widehat{\psi}_1 = 4.31 \times 10^{-2}$, $\widehat{\psi}_2 = 1.50 \times 10^{-2}$ and $\widehat{\psi}_0 = 1.94 \times 10^{-3}$. The fits are shown in Figure 4.5. Both fits appear appropriate, and would appear to neither oversmooth or overfit. It is of interest in comparison is the estimates of $\widehat{\psi}_0$. For spline fit we have $\widehat{\psi}_0 = 1.98 \times 10^{-3}$, and for Gaussian we have the slightly lower $\widehat{\psi}_0 = 1.94 \times 10^{-3}$. These serve as estimates for the errors, σ_ε^2, and would indicate that the Gaussian kernel is only slightly preferable.

4.4 Binary Response Semiparametric Regression

It is of interest to extend the Bayesian Gaussian regression model. In this section, it is the Bayesian probit regression model that is considered. We note that a similar approach may be made, for example, with Bayesian Poisson regression. The Bayesian version of the Gaussian probit mixed model takes the form

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{G} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\Phi(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})), \quad \boldsymbol{u}|\boldsymbol{G} \sim \text{N}(\boldsymbol{0}, \boldsymbol{G}) \tag{4.28}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of Bernoulli response variables, encoded as $\{0, 1\}$, and $\Phi(x) \equiv \int_{-\infty}^{\infty} \frac{\exp(-t^2/2)}{\sqrt{2\pi}} dt$. The components $\boldsymbol{\beta}$, $\boldsymbol{u}$, $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{G}$ are as in the previous section,

$$\boldsymbol{G} = \text{blockdiag}(\sigma_{u1}^2 \boldsymbol{I}_{K_1}, \ldots, \sigma_{ur}^2 \boldsymbol{I}_{K_r}). \tag{4.29}$$

We also impose the conjugate priors:

$$\boldsymbol{\beta} \sim \text{N}(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}), \quad \sigma_{ui}^2 \sim \text{IG}(A_{ui}, B_{ui}),\ 1 \leq i \leq r. \tag{4.30}$$

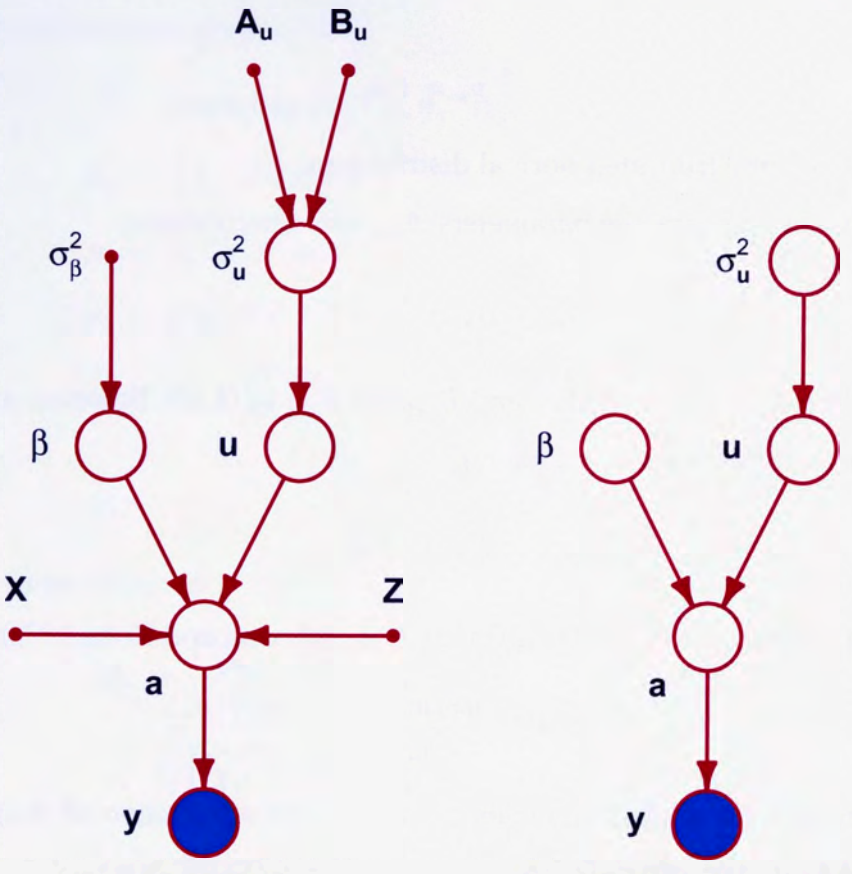
for some $\sigma_\beta^2 > 0$ and $A_{ui}, B_{ui} > 0$, for all $1 \leq i \leq r$.

Following Albert and Chib (1995) and Girolami and Rogers (2006), we introduce the vector of auxiliary variables $\boldsymbol{a} = (a_1, \ldots, a_n)$ where

$$a_i|\boldsymbol{\beta}, \boldsymbol{u} \overset{\text{ind.}}{\sim} \text{N}((\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_i, 1).$$

This allows us to write

$$p(y_i|a_i) = I(a_i \geq 0)^{y_i} I(a_i < 0)^{1-y_i}, \quad 1 \leq i \leq n.$$



(a) *With constants shown* **(b)** *Without constants shown*

Figure 4.6. *Directed acyclic graph representing the probit mixed model. The inclusion of the node* $\boldsymbol{a}$ *allows the problem to be tractable under the factorised density assumption* (4.31).

These associations are represented in Figure 4.6 as a DAG. In particular, the introduction of the auxiliary variables allows us to make the three-component factorisation,

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \sigma^2_{\boldsymbol{\beta}}) = q_{\boldsymbol{\beta},\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{u}) q_{\boldsymbol{a}}(\boldsymbol{a}) q_{\sigma^2_{\boldsymbol{u}}}(\sigma^2_{\boldsymbol{u}}). \tag{4.31}$$

It is shown in Appendix 4.A.1 that application of (4.6) leads to the optimal densities taking the form

$$q^*_{\boldsymbol{\beta},\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{u}) \sim \mathrm{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q),$$

and

$$q^*_{\sigma^2_{\boldsymbol{u}}}(\sigma^2_{u1}, \ldots, \sigma^2_{ur}) = \prod_{i=1}^{r} q_{\sigma^2_{ui}}(\sigma^2_{ui})$$

with

$$q_{\sigma^2_{ui}} \sim \mathrm{IG}(A_{q,ui}, B_{q,ui}).$$

We also have

$$q^*_{\boldsymbol{a}}(\boldsymbol{a}) \sim \prod_{i=1}^{n} [\mathrm{TN}((\boldsymbol{\mu}_{\boldsymbol{a}})_i, 1, 0, \infty)]^{y_i} [\mathrm{TN}((\boldsymbol{\mu}_{\boldsymbol{a}})_i, 1, -\infty, 0)]^{1-y_i},$$

with

$$\boldsymbol{\mu}_{\boldsymbol{a}} = \boldsymbol{C}\boldsymbol{\mu}_q,$$

and $\mathrm{TN}(\cdot,\cdot,\cdot,\cdot)$ is the truncated normal distribution.

Like the Gaussian case, the parameters $A_{q,ui}$ are deterministic,

$$A_{q,ui} = A_{ui} + \tfrac{K_i}{2},\ 1 \le i \le r.$$

The parameters $\boldsymbol{\mu}_q$ and $\boldsymbol{\Sigma}_q$ in (4.12), and $B_{q,\varepsilon}$ and $B_{q,ui}$ in (4.13), however, are dependent on q. Denote by $\boldsymbol{\eta}$ the mean of $\boldsymbol{a}$ under q,

$$\boldsymbol{\eta} \equiv \mathsf{E}_{q(\boldsymbol{a})}\boldsymbol{a}.$$

The mean of a truncated normal is given by (e.g., Johnson and Kotz, 1970):

$$\mathsf{E}_{q(\boldsymbol{a})}\boldsymbol{a} = \boldsymbol{\mu}_{\boldsymbol{a}} + \frac{\boldsymbol{y}\boldsymbol{\phi}(\boldsymbol{\mu}_{\boldsymbol{a}})}{\boldsymbol{\Phi}(\boldsymbol{\mu}_{\boldsymbol{a}})} - \frac{(\mathbf{1}-\boldsymbol{y})\boldsymbol{\phi}(\boldsymbol{\mu}_{\boldsymbol{a}})}{\mathbf{1}-\boldsymbol{\Phi}(\boldsymbol{\mu}_{\boldsymbol{a}})}.$$

We now minimise the Kullback-Leibler divergence, by application of (4.6). It is shown in Appendix 4.A.5 that coordinate ascent gives

$$\begin{aligned}
\boldsymbol{\Sigma}_q &\leftarrow \left\{\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C} + \mathrm{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p, \frac{A_{q,u1}}{B_{q,u1}}\boldsymbol{I}_{K_1}, \ldots, \frac{A_{q,ur}}{B_{q,ur}}\boldsymbol{I}_{K_r}\right)\right\}^{-1},\\
\boldsymbol{\mu}_q &\leftarrow \boldsymbol{\Sigma}_q\boldsymbol{C}^{\mathsf{T}}\boldsymbol{\eta},\\
\boldsymbol{\mu}_{\boldsymbol{a}} &\leftarrow \boldsymbol{C}\boldsymbol{\mu}_q,\\
\boldsymbol{\eta} &\leftarrow \boldsymbol{\mu}_{\boldsymbol{a}} + \frac{\boldsymbol{y}\boldsymbol{\phi}(\boldsymbol{\mu}_{\boldsymbol{a}})}{\boldsymbol{\Phi}(\boldsymbol{\mu}_{\boldsymbol{a}})} - \frac{(\mathbf{1}-\boldsymbol{y})\boldsymbol{\phi}(\boldsymbol{\mu}_{\boldsymbol{a}})}{\mathbf{1}-\boldsymbol{\Phi}(\boldsymbol{\mu}_{\boldsymbol{a}})}, \quad \text{and}\\
B_{q,ui} &\leftarrow B_{ui} + \tfrac{1}{2}\{\|(\boldsymbol{\mu}_q)_{u_i}\|^2 + \mathrm{tr}((\boldsymbol{\Sigma}_q)_{u_i})\}, \quad \text{for} \quad 1 \le i \le r,
\end{aligned}$$

until convergence.

Upon convergence to $\boldsymbol{\mu}^*_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}^*_q, B^*_{q,u1}, \ldots, B^*_{q,ur}$, the approximate posteriors are:

$$p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{y}) \simeq \text{the } \mathrm{N}(\boldsymbol{\mu}^*_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}^*_q) \text{ density in } (\boldsymbol{\beta},\boldsymbol{u}),$$

$$p(\sigma^2_{u1}, \ldots, \sigma^2_{ur}|\boldsymbol{y}) \simeq \text{product of } \mathrm{IG}(A^*_{q,ui}, B^*_{q,ui}) \text{ densities,}$$

and

$$\begin{aligned}
p^*_{\boldsymbol{a}}(\boldsymbol{a}|\boldsymbol{y}) \simeq\ &\text{product of } \mathrm{TN}((\boldsymbol{\mu}^*_{\boldsymbol{a}})_i, 1, -\infty, 0)\\
&\text{and } \mathrm{TN}((\boldsymbol{\mu}^*_{\boldsymbol{a}})_i, 1, 0, \infty) \text{ densities.}
\end{aligned}$$

As with the Bayesian Gaussian mixed model, the optimality criteria given by coordinate ascent may be expressed in different forms. In primal form, the method of

successive approximations gives

$$
\begin{aligned}
V &\leftarrow \text{blockdiag}\left(\sigma_\beta^{-2} I_p, \psi_1^{-1} I_{K_1}, \ldots, \psi_r^{-1} I_{K_r}\right) \\
\Sigma_q &\leftarrow \left\{\psi_0^{-1} C^\mathsf{T} C + V\right\}^{-1}, \\
\mu_q &\leftarrow \psi_0^{-1} \Sigma_q C^\mathsf{T} \eta, \\
\mu_a &\leftarrow C\mu_q, \\
\eta &\leftarrow \mu_a + \frac{y\phi(\mu_a)}{\Phi(\mu_a)} - \frac{(1-y)\phi(\mu_a)}{1-\Phi(\mu_a)}, \quad \text{and} \\
\psi_i &\leftarrow \frac{B_{ui} + \frac{1}{2}\left\|(\mu_q)_{u_i}\right\|^2}{A_{ui} + \frac{1}{2}(K_i - \text{tr}((V\Sigma_q)_{u_i})}, \quad \text{for} \quad 1 \le i \le r.
\end{aligned}
$$

A dual space formulation is

$$
\begin{aligned}
\Pi_q &\leftarrow \left\{XX^\mathsf{T} + \sum_{i=1}^{r} \psi_i K_i + I_n\right\}^{-1}, \\
\mu_a &\leftarrow \left(XX^\mathsf{T} + \sum_{i=1}^{r} \psi_i K_i\right) \Pi_q \eta, \\
\eta &\leftarrow \mu_a + \frac{y\phi(\mu_a)}{\Phi(\mu_a)} - \frac{(1-y)\phi(\mu_a)}{1-\Phi(\mu_a)}, \quad \text{and} \\
\psi_i &\leftarrow \frac{B_{ui} + \frac{1}{2}\psi_i^2 \mu_a^\mathsf{T} \Pi_q K_i \Pi_q \mu_a}{A_{ui} + \frac{1}{2}\psi_i \text{tr}(K_i \Pi_q)}, \quad \text{for} \quad 1 \le i \le r.
\end{aligned}
$$

The resulting estimators may be seen as a generalisation of the ML and REML estimators to the probit model. In particular, with improper priors, the resultant estimators appear to match those of the penalised quasi-likelihood approach of Breslow and Clayton (1993) and Wolfinger and O'Connell (1993). The mean, η, is known as the pseudo-data (e.g. Molenberghs and Verbeke, 2005). The literature contains several methods that can be seen as generalisations of ML or REML to non-Gaussian response. For binary response, it is common for such generalisations to display a bias in the mean estimates of the parameters (e.g., Lee and Nelder, 2003). That is, such generalisations are known to oversmooth.

There are alternative restrictions that lead to tractable solutions. Most notably, it would appear as if a Laplace or Taylor series approximation (e.g., Wolfinger and O'Connell, 1993) may both lead to a tractable solution, and to a generalisation of penalised quasi-likelihood.

4.4.1 Spam Data Example

We illustrate the Bayesian Probit mixed model with an example. The "spam" data is described in Hastie, Tibshirani and Freidman (2001), with spam e-mail messages coded

as 1 and ordinary messages coded as 0. For ease of presentation, some sixteen predictor variables out of a total of fifty-seven were selected. An additive model penalised spline kernel (Section 2.5.3) was used to build the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$. The prior parameters were set with

$$A_{u1} = \cdots = A_{u16} = B_{u1} = \cdots = B_{u16} = 0.01, \quad \text{and} \quad \sigma_\beta^2 = 10^8.$$

An illustration of the fit is given in Figure 4.7. Each panel shows the slice of the fitted surface for each labelled predictor, with all other predictors set to their medians. The resulting fitted surfaces are similar to those of the support vector classifier displayed of Section 2.1.

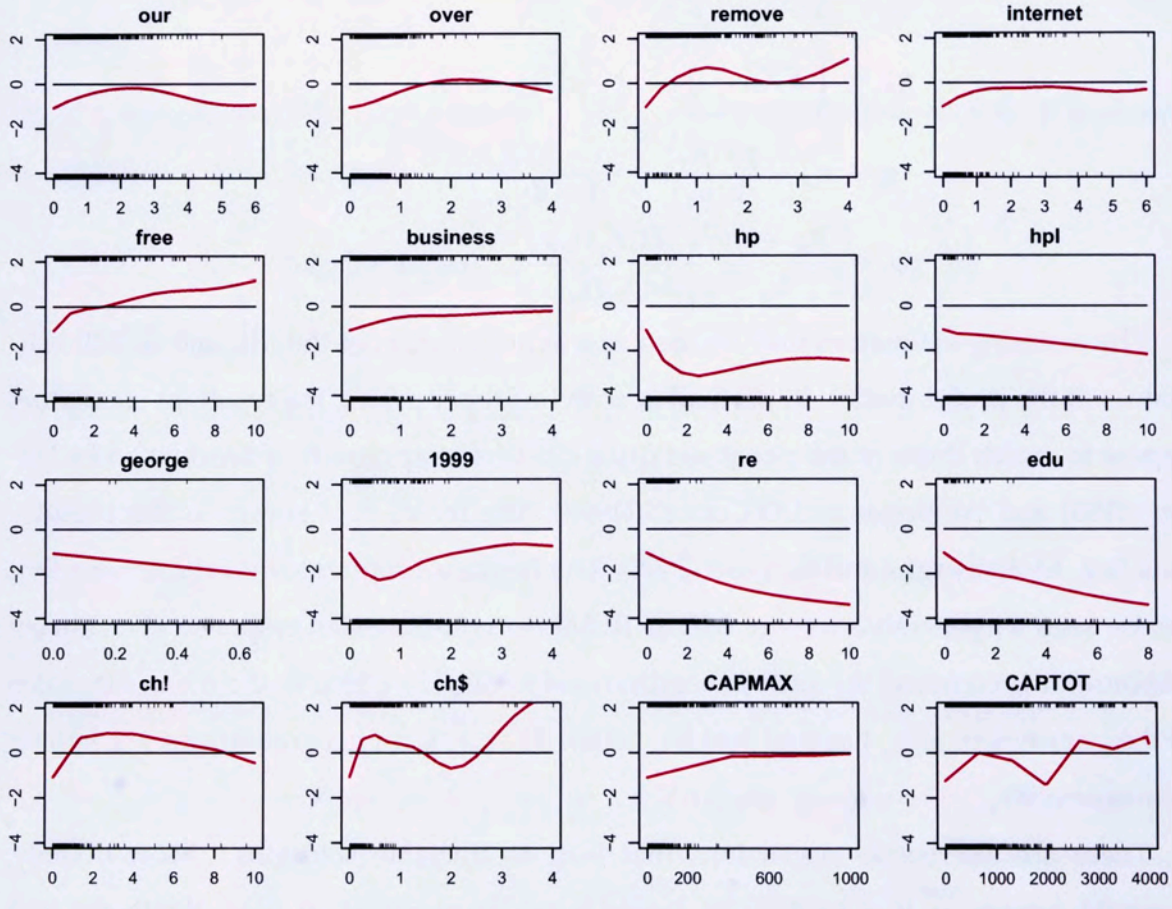


Figure 4.7. *Visualisation of a penalised spline Bayesian probit model fir for the "spam" data. Each panel shows the slice of the discriminant with all other predictors set to their medians. The tick-marks show the predictor values: spam e-mail messages along the top, ordinary e-mail messages along the bottom.*

4.5 Conclusion

Factorised density assumptions have lead to tractable solutions to the Bayesian Gaussian linear and probit mixed models. The optimality may be expressed in both dual and primal forms. The resultant optimisation criteria shows that the Bayesian Gaussian linear mixed model to be a generalisation of many existing techniques. We have shown that with improper priors, variational approximation of Bayesian GLMMs lead to the REML estimator. This is a benefit, not only to the interpretation of variational Bayes, but for the computation of the fits. The similarity of the resulting optimisation of variational Bayes to those of REML estimates ensures the extensive literature on REML may guide understanding of variational Bayes.

The extension to probit Bayesian mixed models shows a tractable problem, with penalised quasi-likelihood as a special case. A natural hypothesis is that such a relationship holds over the wide variety of generalised linear mixed models. Such links allow the elegant Bayesian framework to extend in a tractable manner to generalised linear mixed models, and by extension, to many types of kernel machine.

4.A Appendix

We begin with a helpful lemma for the Inverse Gamma distribution.

Lemma 4.3. *Let x and y be random variables with $x \sim \mathrm{IG}(A, B)$ and $y \sim \mathrm{IG}(A', B')$. Then*

i) $\log p_{\mathsf{x}}(x) = \log\left(\frac{B^A}{\Gamma(A)}\right) - (A+1)\log(x) - B/x.$

ii) $\mathsf{E}_{\mathsf{x}} \log x = \log B - \mathrm{digamma}(A), \quad \mathsf{E}_{\mathsf{x}} x^{-1} = \frac{A}{B}.$

iii) $\mathrm{KL}(p_{\mathsf{x}}, p_{\mathsf{y}}) = -\log\left(\frac{\Gamma(A')B^A}{\Gamma(A)B'^{A'}}\right) - (A - A')\left\{\log(B) - \mathrm{digamma}(A)\right\} + (B - B')\frac{A}{B}.$

Proof. i)-ii) See Johnson and Kotz (1970).

iii) Expanding the Kullback-Leibler divergence, and using *i)*,

$$\begin{aligned}
\mathrm{KL}(p_{\mathsf{x}}, p_{\mathsf{y}}) &= -\mathsf{E}_{\mathsf{x}}\left\{\log p_{\mathsf{y}}(x) - \log p_{\mathsf{x}}(x)\right\} \\
&= -\mathsf{E}_{\mathsf{x}}\left\{\log\left(\frac{B'^{A'}}{\Gamma(A')}\right) - (A'+1)\log(x) - B'/x \right. \\
&\qquad \left. - \log\left(\frac{B^A}{\Gamma(A)}\right) + (A+1)\log(x) + B/x\right\} \\
&= -\log\left(\frac{\Gamma(A')B^A}{\Gamma(A)B'^{A'}}\right) - \mathsf{E}_{\mathsf{x}}\left\{(A - A')\log(x) + \frac{B - B'}{x}\right\} \\
&= -\log\left(\frac{\Gamma(A')B^A}{\Gamma(A)B'^{A'}}\right) - (A - A')\left\{\log(B) - \mathrm{digamma}(A)\right\} + (B - B')\frac{A}{B}
\end{aligned}$$

where the last line follows from *ii)*. □

4.A.1 Gaussian Case

We derive the update equations for the Bayesian linear mixed model. There are two lemmas involved. The first, below, shows the optimal $q^*_{\sigma^2}$. The second, Lemma 4.5, shows the optimal $q^*_{\beta,u}$.

Lemma 4.4. *Let* $q_{\beta,u}$ *have mean* μ_q *and covariance* Σ_q, *that is,* $\mu_q \equiv \mathsf{E}_{q(\beta,u)} \begin{bmatrix} \beta \\ u \end{bmatrix}$ *and* $\Sigma_q \equiv \mathrm{Cov}_{q(\beta,u)} \begin{bmatrix} \beta \\ u \end{bmatrix}$. *Then the optimal* $q^*_{\sigma^2}$ *in the factorised density Bayesian linear mixed model* (4.8)-(4.11) *is*

$$q^*_{\sigma^2}(\sigma^2_{u1}, \ldots, \sigma^2_{ur}, \sigma^2_\varepsilon) = q^*_{\sigma^2_\varepsilon}(\sigma^2_\varepsilon) \prod_{i=1}^{r} q^*_{\sigma^2_{ui}}(\sigma^2_{ui})$$

where

$$q^*_{\sigma^2_\varepsilon} \sim \mathrm{IG}(A_{q,\varepsilon}, B_{q,\varepsilon}), \quad \text{and} \quad q^*_{\sigma^2_{ui}} \sim \mathrm{IG}(A_{q,ui}, B_{q,ui}).$$

The parameters $A_{q,\varepsilon}$ *and* $A_{q,ui}$ *are deterministic,*

$$A_{q,\varepsilon} = A_\varepsilon + \frac{n}{2} \quad \text{and} \quad A_{q,u} = A_{ui} + \frac{K_i}{2}, \quad \text{for} \quad 1 \le i \le r. \tag{4.32}$$

In contrast, the parameters $B_{q,\varepsilon}$ *and* $B_{q,ui}$ *are dependent on* μ_q *and* Σ_q,

$$\begin{aligned} B_{q,\varepsilon} &= B_\varepsilon + \tfrac{1}{2}\{\|y - C\mu_q\|^2 + \mathrm{tr}(C^\mathsf{T} C \Sigma_q)\}, \quad \text{and} \\ B_{q,ui} &= B_{ui} + \tfrac{1}{2}\{\|(\mu_q)_{u_i}\|^2 + \mathrm{tr}((\Sigma_q)_{u_i})\} \quad \text{for} \quad 1 \le i \le r, \end{aligned} \tag{4.33}$$

where $C \equiv [X\ Z]$.

Proof. Application of (4.6) leads to the optimal densities taking the form

$$q^*_{\sigma^2}(\sigma^2_{u1}, \ldots, \sigma^2_{ur}, \sigma^2_\varepsilon) \propto \exp\left\{\mathsf{E}_{q(\beta,u)} \log p(y, \beta, u, \sigma^2)\right\},$$

where $\sigma^2 = (\sigma^2_{u1}, \ldots, \sigma^2_{ur}, \sigma^2_\varepsilon)^\mathsf{T}$. By applying the Markov blanket to the DAG,

$$\begin{aligned} p(y, \beta, u, \sigma^2) &= p(y|\beta, u, \sigma^2) p(\beta|u, \sigma^2) p(u|\sigma^2) p(\sigma^2) \\ &= p(y|\beta, u, \sigma^2_\varepsilon) p(\beta) p(u|\sigma^2_u) p(\sigma^2). \end{aligned}$$

Therefore,

$$
\begin{aligned}
q^*_{\sigma^2}(\sigma^2_{u1},\ldots,\sigma^2_{ur},\sigma^2_\varepsilon) &\propto \exp\left[\mathsf{E}_{q(\boldsymbol{\beta},\boldsymbol{u})}\log\left\{p(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\right\}\right]\\
&\propto \exp\left[\mathsf{E}_{q(\boldsymbol{\beta},\boldsymbol{u})}\log\left\{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2_\varepsilon)p(\boldsymbol{u}|\sigma^2_{\boldsymbol{u}})p(\sigma^2)\right\}\right]\\
&\propto \exp\left[\mathsf{E}_{q(\boldsymbol{\beta},\boldsymbol{u})}\log\left\{(\sigma^2_\varepsilon)^{-n/2}\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}\|^2}{2\sigma^2_\varepsilon}\right)\right.\right.\\
&\qquad\times\left(\prod_{i=1}^r(\sigma^2_{ui})^{-K_i/2}\right)\exp\left(-\frac{1}{2}\boldsymbol{u}^{\mathsf{T}}(\boldsymbol{G})^{-1}\boldsymbol{u}\right)\\
&\qquad\times(\sigma^2_\varepsilon)^{-A_\varepsilon-1}\exp\left(\frac{-B_\varepsilon}{\sigma^2_\varepsilon}\right)\\
&\qquad\left.\left.\times\prod_{i=1}^r(\sigma^2_{ui})^{-A_{ui}-1}\exp\left(\frac{-B_{ui}}{\sigma^2_{ui}}\right)\right\}\right]\\
&= (\sigma^2_\varepsilon)^{-A_\varepsilon-1-n/2}(\sigma^2_{ui})^{-A_{ui}-1-K_i/2}\\
&\qquad\times\exp\left[-\tfrac{1}{\sigma^2_\varepsilon}\left\{B_\varepsilon+\tfrac{1}{2}\mathsf{E}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}\|^2\right\}\right.\\
&\qquad\left.-\sum_{i=1}^r\frac{1}{\sigma^2_{ui}}\left\{B_{ui}+\tfrac{1}{2}\mathsf{E}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|u_i\|^2\right\}\right]
\end{aligned}
$$

However,

$$\mathsf{E}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}\|^2 = \|\boldsymbol{y}-\boldsymbol{C\mu}_q\|^2+\mathrm{tr}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C\Sigma}_q),$$

and

$$\mathsf{E}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|u_i\|^2 = \|(\boldsymbol{\mu}_q)_{u_i}\|^2+\mathrm{tr}((\boldsymbol{\Sigma}_q)_{u_i}).$$

We then factorise $q^*_{\sigma^2}(\sigma^2_{u1},\ldots,\sigma^2_{ur},\sigma^2_\varepsilon)$ as

$$
\begin{aligned}
q^*_{\sigma^2}(\sigma^2_{u1},\ldots,\sigma^2_{ur},\sigma^2_\varepsilon) &\propto (\sigma^2_\varepsilon)^{-A_\varepsilon-n/2-1}(\sigma^2_{ui})^{-A_{ui}-q_1/2-1}\\
&\qquad\times\exp\left[-\tfrac{1}{\sigma^2_\varepsilon}\left[B_\varepsilon+\tfrac{1}{2}\left\{\|\boldsymbol{y}-\boldsymbol{C\mu}_q\|^2+\mathrm{tr}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C\Sigma}_q)\right\}\right]\right.\\
&\qquad\left.-\sum_{i=1}^r\frac{1}{\sigma^2_{ui}}\left[B_{ui}+\tfrac{1}{2}\left\{\|(\boldsymbol{\mu}_q)_{u_i}\|^2+\mathrm{tr}((\boldsymbol{\Sigma}_q)_{u_i})\right\}\right]\right],
\end{aligned}
$$

which we recognise as the product of inverse gamma densities,

$$q^*_{\sigma^2}(\sigma^2_{u1},\ldots,\sigma^2_{ur},\sigma^2_\varepsilon) = q^*_{\sigma^2_\varepsilon}(\sigma^2_\varepsilon)\prod_{i=1}^r q^*_{\sigma^2_{ui}}(\sigma^2_{ui}),$$

where

$$q^*_{\sigma^2_\varepsilon}\sim\mathrm{IG}(A_\varepsilon+\tfrac{1}{2}n, B_\varepsilon+\tfrac{1}{2}\{\|\boldsymbol{y}-\boldsymbol{C\mu}_q\|^2+\mathrm{tr}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C\Sigma}_q)\})$$

and,

$$q^*_{\sigma^2_{ui}}\sim\mathrm{IG}(A_{\boldsymbol{u}1}+\tfrac{K_i}{2}, B_{ui}+\tfrac{1}{2}\{\|(\boldsymbol{\mu}_q)_{u_i}\|^2+\mathrm{tr}((\boldsymbol{\Sigma}_q)_{u_i})\}) \quad\text{for all}\quad 1\le i\le r. \qquad \square$$

The next lemma considers the optimal $q^*_{\boldsymbol{\beta},\boldsymbol{u}}$ density.

Lemma 4.5. *Let $q^*_{\sigma^2}$ be a product of inverse gamma densities, with parameters,*

$$q^*_{\sigma^2}(\sigma^2_{u1},\ldots,\sigma^2_{ur},\sigma^2_\varepsilon) = q^*_{\sigma^2_\varepsilon}(\sigma^2_\varepsilon)\prod_{i=1}^{r} q^*_{\sigma^2_{ui}}(\sigma^2_{ui})$$

$$q^*_{\sigma^2_\varepsilon} \sim \mathrm{IG}(A_{q,\varepsilon}, B_{q,\varepsilon}), \quad \text{and} \quad q^*_{\sigma^2_{ui}} \sim \mathrm{IG}(A_{q,ui}, B_{q,ui}).$$

Then the optimal $q^*_{\beta,u}$ is $\mathrm{N}(\mu_q, \Sigma_q)$, where

$$\mu_q = \left(\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}}\right)\Sigma_q C^\mathsf{T} y, \quad \text{and}$$

$$\Sigma_q = \left\{\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}}C^\mathsf{T}C + \mathrm{blockdiag}\left(\sigma_\beta^{-2}I_p, \frac{A_{q,u1}}{B_{q,u1}}I_{K_1},\ldots,\frac{A_{q,ur}}{B_{q,ur}}I_{K_r}\right)\right\}^{-1}.$$

Proof. Application of (4.6) leads to the optimal densities taking the form

$$\begin{aligned} q^*_{\beta,u}(\beta,u) &\propto \exp\left[\mathsf{E}_{q(\sigma^2)}\log\left\{p(y|\beta,u,\sigma^2_\varepsilon)p(\beta)p(u|\sigma^2_u)p(\sigma^2)\right\}\right] \\ &\propto \exp\left[\mathsf{E}_{q(\sigma^2)}\log\left\{p(y|\beta,u,\sigma^2_\varepsilon)p(\beta)p(u|\sigma^2_u)\right\}\right] \\ &\propto \exp\left\{-\mathsf{E}_{q(\sigma^2)}\frac{\|y-X\beta-Zu\|^2}{2\sigma^2_\varepsilon} + \frac{\|\beta\|^2}{2\sigma^2_\beta} + \frac{1}{2}u^\mathsf{T}(\sigma^2_u I)^{-1}u\right\}. \end{aligned}$$

Now applying Lemma 4.3 *ii)*,

$$\begin{aligned} q^*_{\beta,u}(\beta,u) &\propto \exp\left[-\frac{A_{q,\varepsilon}\|y-X\beta-Zu\|^2}{2B^2_{q,\varepsilon}} - \frac{\|\beta\|^2}{2\sigma^2_\beta} - \sum_{i=1}^{r}\frac{A_{q,ui}u_i^\mathsf{T}u_i}{2B_{q,ui}}\right] \\ &\propto \exp\left[-\frac{A_{q,\varepsilon}}{2B_{q,\varepsilon}}\begin{bmatrix}\beta\\u\end{bmatrix}^\mathsf{T} C^\mathsf{T}C\begin{bmatrix}\beta\\u\end{bmatrix} - \frac{1}{2\sigma^2_\beta}\beta^\mathsf{T}\beta - \sum_{i=1}^{r}\frac{A_{q,ui}u_i^\mathsf{T}u_i}{2B_{q,ui}} + \frac{A_{q,\varepsilon}}{B_{q,\varepsilon}}\begin{bmatrix}\beta\\u\end{bmatrix}Cy\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\begin{bmatrix}\beta\\u\end{bmatrix}-\mu_q\right)^\mathsf{T}\Sigma_q^{-1}\left(\begin{bmatrix}\beta\\u\end{bmatrix}-\mu_q\right)\right], \end{aligned}$$

where $\mu_q = \left(\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}}\right)\Sigma_q C^\mathsf{T} y$, and

$$\Sigma_q = \left\{\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}}C^\mathsf{T}C + \mathrm{blockdiag}\left(\sigma_\beta^{-2}I_p, \frac{A_{q,u1}}{B_{q,u1}}I_{K_1},\ldots,\frac{A_{q,ur}}{B_{q,ur}}I_{K_r}\right)\right\}^{-1}.$$

We recognise $q^*_{\beta,u}$ as Gaussian. □

4.A.2 An Expression for the Lower Bound

We now derive $L(q)$, as given by (4.3), and a lower bound for $\log p(y)$.

$$\begin{aligned} L(q) =& \mathsf{E}_q\left\{\log p(y,\beta,u,\sigma^2) - \log q(\beta,u,\sigma^2)\right\} \\ =& \mathsf{E}_q\{\log p(y|\beta,u,\sigma^2_\varepsilon) + \log p(\beta) + \log p(u|\sigma^2_u) + \log p(\sigma^2) \\ & - \log q(\beta,u) - \log q(\sigma^2)\}. \end{aligned}$$

But

$$-\mathsf{E}_q \log q(\boldsymbol{\beta}, \boldsymbol{u}) = \tfrac{p+\sum_{i=1}^{r} K_i}{2}\{1 + \log(2\pi)\} + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_q|,$$

due to the entropy of a normal distribution. Also,

$$\mathsf{E}_q \log p(\boldsymbol{\beta}) = -\frac{p}{2}\log(2\pi\sigma_\beta^2) - \frac{\|(\boldsymbol{\mu}_q)_{\boldsymbol{\beta}}\|^2 + \operatorname{tr}(\boldsymbol{\Sigma}_{q,\boldsymbol{\beta}})}{2\sigma_\beta^2}.$$

We have

$$\begin{aligned}\mathsf{E}_q \log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2) &= -\frac{n}{2}\left\{\log(2\pi) + \mathsf{E}_q \log \sigma_\varepsilon^2\right\} - \mathsf{E}_q \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\|^2}{2\sigma_\varepsilon^2} \\ &= -\frac{n}{2}\left\{\log(2\pi) + \log(B_{q,\varepsilon}) - \operatorname{digamma}(A_{q,\varepsilon})\right\} \\ &\quad - \frac{A_{q,\varepsilon}}{2B_{q,\varepsilon}}\left(\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_q\|^2 + \operatorname{tr}(\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_q)\right),\end{aligned}$$

where we have used the properties of the Inverse Gamma distribution given by Lemma 4.3 *ii)*,

$$\mathsf{E}_q \log \sigma_\varepsilon^2 = \log(B_{q,\varepsilon}) - \operatorname{digamma}(A_{q,\varepsilon}), \quad \text{and} \quad \mathsf{E}_q\, \sigma_\varepsilon^{-2} = A_{q,\varepsilon}/B_{q,\varepsilon}.$$

Also, by Lemma 4.3 *iii)*,

$$\begin{aligned}&\mathsf{E}_q\left\{\log p(\sigma_\varepsilon^2) - \log q(\sigma_\varepsilon^2)\right\} \\ &\quad = \log\left(\frac{\Gamma(A_{q,\varepsilon})B_\varepsilon^{A_\varepsilon}}{\Gamma(A_\varepsilon)B_{q,\varepsilon}^{A_{q,\varepsilon}}}\right) + (A_{q,\varepsilon} - A_\varepsilon)(\log(B_{q,\varepsilon}) - \operatorname{digamma}(A_{q,\varepsilon})) + (B_{q,\varepsilon} - B_\varepsilon)\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}},\end{aligned}$$

so that,

$$\begin{aligned}&\mathsf{E}_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2) + \log p(\sigma_\varepsilon^2) - \log q(\sigma_\varepsilon^2)\right\} \\ &\quad = -\frac{n}{2}\log(2\pi) + \log\left(\frac{\Gamma(A_{q,\varepsilon})B_\varepsilon^{A_\varepsilon}}{\Gamma(A_\varepsilon)B_{q,\varepsilon}^{A_{q,\varepsilon}}}\right) \\ &\qquad + \left(A_{q,\varepsilon} - A_\varepsilon - \frac{n}{2}\right)\left\{\log(B_{q,\varepsilon}) - \operatorname{digamma}(A_{q,\varepsilon})\right\} \\ &\qquad + \left[B_{q,\varepsilon} - B_\varepsilon - \tfrac{1}{2}\left\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_q\|^2 + \operatorname{tr}(\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_q)\right\}\right]\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}}.\end{aligned}$$

We find

$$\begin{aligned}\mathsf{E}_q \log p(u_i|\sigma_{ui}^2) &= -\frac{K_i}{2}\left\{\log(2\pi) + \log(B_{q,ui}) - \operatorname{digamma}(A_{q,ui})\right\} \\ &\quad - \frac{A_{q,ui}}{2B_{q,ui}}\left\{\|(\boldsymbol{\mu}_q)_{ui}\|^2 + \operatorname{tr}(\boldsymbol{\Sigma}_{q,ui})\right\}.\end{aligned}$$

Again by Lemma 4.3 *iii)*,

$$\begin{aligned}&\mathsf{E}_q\left\{\log p(\sigma_{ui}^2) - \log q(\sigma_{ui}^2)\right\} \\ &\quad = \log\left(\frac{\Gamma(A_{q,ui})B_{ui}^{A_{ui}}}{\Gamma(A_{ui})B_{q,ui}^{A_{q,ui}}}\right) + (A_{q,ui} - A_{ui})(\log(B_{q,ui}) - \operatorname{digamma}(A_{q,ui})) \\ &\qquad + (B_{q,ui} - B_{ui})\frac{A_{q,ui}}{B_{q,ui}},\end{aligned}$$

so that

$$\begin{aligned}
&\mathsf{E}_q\left\{\log p(u_i|\sigma_{ui}^2)+\log p(\sigma_{ui}^2)-\log q(\sigma_{ui}^2)\right\}\\
&\quad=-\frac{K_i}{2}\log(2\pi)+\log\left(\frac{\Gamma(A_{q,ui})B_{ui}^{A_{ui}}}{\Gamma(A_{ui})B_{q,ui}^{A_{q,ui}}}\right)\\
&\qquad+(A_{q,\varepsilon}-A_\varepsilon-\tfrac{K_i}{2})\left\{\log(B_{q,ui})-\text{digamma}(A_{q,ui})\right\}\\
&\qquad+\left[B_{q,ui}-B_{ui}-\tfrac{1}{2}\left\{\|(\boldsymbol{\mu}_q)_{ui}\|^2+\text{tr}(\boldsymbol{\Sigma}_{q,ui})\right\}\right]\frac{A_{q,ui}}{B_{q,ui}}.
\end{aligned}$$

We then have $L(q)$;

$$\begin{aligned}
L(q)&=\mathsf{E}_q\{\log p(\boldsymbol{\beta})-\log q(\boldsymbol{\beta},\boldsymbol{u})\\
&\qquad+\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)+\log p(\sigma_\varepsilon^2)-\log q(\sigma_\varepsilon^2)\\
&\qquad+\sum_{i=1}^{r}\log p(u_i|\sigma_{ui}^2)+\log p(\sigma_{ui}^2)-\log q(\sigma_{ui}^2)\}\\
&=\frac{p+\sum_{i=1}^r K_i}{2}-\frac{n}{2}\log(2\pi)-\frac{p}{2}\log(\sigma_\beta^2)+\tfrac{1}{2}\log|\boldsymbol{\Sigma}_q|-\frac{\|(\boldsymbol{\mu}_q)_{\boldsymbol{\beta}}\|^2+\text{tr}(\boldsymbol{\Sigma}_{q,\boldsymbol{\beta}})}{2\sigma_\beta^2}\\
&\qquad+\log\left(\frac{\Gamma(A_{q,\varepsilon})B_\varepsilon^{A_\varepsilon}}{\Gamma(A_\varepsilon)B_{q,\varepsilon}^{A_{q,\varepsilon}}}\right)+(A_{q,\varepsilon}-A_\varepsilon-\tfrac{n}{2})\left\{\log(B_{q,\varepsilon})-\text{digamma}(A_{q,\varepsilon})\right\}\\
&\qquad+\left[B_{q,\varepsilon}-B_\varepsilon-\tfrac{1}{2}\left\{\|\boldsymbol{y}-\boldsymbol{C}\boldsymbol{\mu}_q\|^2+\text{tr}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\Sigma}_q)\right\}\right]\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}} \qquad (4.34)\\
&\qquad+\sum_{i=1}^{r}\left\{\log\left(\frac{\Gamma(A_{q,ui})B_{ui}^{A_{ui}}}{\Gamma(A_{ui})B_{q,ui}^{A_{q,ui}}}\right)+(A_{q,ui}-A_{ui}-\tfrac{K_i}{2})\left\{\log(B_{q,ui})-\text{digamma}(A_{q,ui})\right\}\right.\\
&\qquad\left.+\left[B_{q,ui}-B_{ui}-\tfrac{1}{2}\left\{\|(\boldsymbol{\mu}_q)_{ui}\|^2+\text{tr}(\boldsymbol{\Sigma}_{q,ui})\right\}\right]\frac{A_{q,ui}}{B_{q,ui}}\right\}.
\end{aligned}$$

Explicitly, for $q(\boldsymbol{\beta},\boldsymbol{u},\sigma_{u1}^2,\ldots,\sigma_{ur}^2,\sigma_\varepsilon^2)=q_{\boldsymbol{\beta},\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{u})q_{\sigma^2}(\sigma_{u1}^2,\ldots,\sigma_{ur}^2,\sigma_\varepsilon^2)$, where

$$\begin{aligned}
&q_{\boldsymbol{\beta},\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{u})\sim\text{N}(\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q),\quad q_{\sigma^2}(\sigma_{u1}^2,\ldots,\sigma_{ur}^2,\sigma_\varepsilon^2)\sim\text{IG}(A_\varepsilon,B_\varepsilon)\\
&\text{and}\quad q_{\sigma^2}(\sigma_{u1}^2,\ldots,\sigma_{ur}^2,\sigma_\varepsilon^2)\sim\text{IG}(A_{q,ui},B_{q,ui}),\quad\text{for all } 1\le i\le r,
\end{aligned}$$

then (4.34) gives the precise value for the likelihood. Various bounds may be achieved through the development of (4.34), depending on which update equations hold true. In particular, if $A_{q,\varepsilon}$, $B_{q,\varepsilon}$, $A_{q,ui}$, and $B_{q,ui}$, $1\le i\le r$ are at given by their updates (4.32) and (4.33). Then (4.34) simplifies to,

$$\begin{aligned}
L(q)&=\frac{p+\sum_{i=1}^r K_i}{2}-\frac{n}{2}\log(2\pi)-\frac{p}{2}\log(\sigma_\beta^2)+\tfrac{1}{2}\log|\boldsymbol{\Sigma}_q|-\frac{\|(\boldsymbol{\mu}_q)_{\boldsymbol{\beta}}\|^2+\text{tr}(\boldsymbol{\Sigma}_{q,\boldsymbol{\beta}})}{2\sigma_\beta^2}\\
&\qquad+\log\left(\frac{\Gamma(A_{q,\varepsilon})B_\varepsilon^{A_\varepsilon}}{\Gamma(A_\varepsilon)B_{q,\varepsilon}^{A_{q,\varepsilon}}}\right)+\sum_{i=1}^{r}\log\left(\frac{\Gamma(A_{q,ui})B_{ui}^{A_{ui}}}{\Gamma(A_{ui})B_{q,ui}^{A_{q,ui}}}\right).
\end{aligned}\qquad (4.35)$$

If all the update equations hold simultaneously, then (4.35) represents the lower bound over all distributions of the factorised density form given by (4.11). For numerical reasons, the bound in (4.35) is calculated as

$$\begin{aligned} L(q) = \frac{p + \sum_{i=1}^{r} K_i}{2} - \frac{n}{2}\log(2\pi) - \frac{p}{2}\log(\sigma_\beta^2) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_q| - \frac{\|(\boldsymbol{\mu}_q)_\beta\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q,\beta})}{2\sigma_\beta^2} \\ + A_\varepsilon \log(B_\varepsilon) - A_{q,\varepsilon}\log(B_{q,\varepsilon}) + \log\Gamma(A_{q,\varepsilon}) - \log\Gamma(A_\varepsilon) \\ + \sum_{i=1}^{r} \left\{ A_{ui}\log(B_{ui}) - A_{q,ui}\log(B_{q,ui}) + \log\Gamma(A_{q,ui}) - \log\Gamma(A_{ui}) \right\}, \end{aligned}$$

where $\log\Gamma(\cdot)$ directly calculates the log of the gamma function. The MATLAB function `gammaln`, and `R` function `lgamma` both calculate $\log\Gamma(\cdot)$ in a direct manner. A similar direct calculation can be made for $\log|\cdot|$.

4.A.3 Deriving the Dual Space Formulation

We wish to express (4.20) in a manner suitable for kernelisation. Let us begin by factorising $\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}} \boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T} + \boldsymbol{C}^\mathsf{T}$ as

$$\boldsymbol{C}^\mathsf{T}\left(\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T} + \boldsymbol{I}\right) = \left(\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{C}^\mathsf{T}\boldsymbol{C} + \boldsymbol{V}^*\right)\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T}.$$

Therefore,

$$\begin{aligned} \boldsymbol{C}\left(\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{C}^\mathsf{T}\boldsymbol{C} + \boldsymbol{V}^*\right)^{-1}\boldsymbol{C}^\mathsf{T} &= \boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T}\left(\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T} + \boldsymbol{I}_n\right)^{-1} \\ &= \frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\left\{\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T}\left(\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T} + \boldsymbol{I}_n\right)^{-1}\right\} \\ &= \frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\left\{\boldsymbol{I}_n - \frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\left(\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T} + \frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\boldsymbol{I}_n\right)^{-1}\right\} \end{aligned}$$

We then have

$$\boldsymbol{y}^\mathsf{T}\boldsymbol{C}\left(\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{C}^\mathsf{T}\boldsymbol{C} + \boldsymbol{V}^*\right)^{-1}\boldsymbol{C}^\mathsf{T}\boldsymbol{y} = \frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\boldsymbol{y}^\mathsf{T}\left\{\boldsymbol{I}_n - \frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\left(\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T} + \frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\boldsymbol{I}_n\right)^{-1}\right\}\boldsymbol{y}. \quad (4.36)$$

Differentiating both sides of (4.36) with respect to $\frac{A^*_{q,ui}}{B^*_{q,ui}}$ then gives

$$\frac{d}{d\frac{A^*_{q,ui}}{B^*_{q,ui}}}\left\{\boldsymbol{y}^\mathsf{T}\boldsymbol{C}\left(\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{C}^\mathsf{T}\boldsymbol{C} + \boldsymbol{V}^*\right)^{-1}\boldsymbol{C}^\mathsf{T}\boldsymbol{y}\right\} = \left\|\left(\boldsymbol{\Sigma}^*_q\boldsymbol{C}^\mathsf{T}\boldsymbol{y}\right)_{\boldsymbol{u}_i}\right\|^2,$$

and

$$\frac{d}{d\frac{A^*_{q,ui}}{B^*_{q,ui}}}\left[\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\boldsymbol{y}^\mathsf{T}\left\{\boldsymbol{I}_n-\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\left(\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T}+\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\boldsymbol{I}_n\right)^{-1}\right\}\boldsymbol{y}\right]$$
$$=\left(\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\right)^2\left(\frac{B^*_{q,ui}}{A^*_{q,ui}}\right)^2\boldsymbol{y}^\mathsf{T}\boldsymbol{\Pi}^*_q\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\Pi}^*_q\boldsymbol{y}.$$

We then have

$$\left\|\left(\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{\Sigma}^*_q\boldsymbol{C}^\mathsf{T}\boldsymbol{y}\right)_{\boldsymbol{u}_i}\right\|^2=\left(\frac{B^*_{q,ui}}{A^*_{q,ui}}\right)^2\boldsymbol{y}^\mathsf{T}\boldsymbol{\Pi}^*_q\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\Pi}^*_q\boldsymbol{y}. \tag{4.37}$$

The following lemma, known as Sylvester's identity, is proven in Bareiss (1968).

Lemma 4.6 (Sylvester's identity). *For any $n \times m$ matrix $\boldsymbol{M}$,*

$$\left|\boldsymbol{M}\boldsymbol{M}^\mathsf{T}+\boldsymbol{I}_n\right|=\left|\boldsymbol{M}^\mathsf{T}\boldsymbol{M}+\boldsymbol{I}_m\right|.$$

Let $\boldsymbol{M}=\sqrt{\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}}\boldsymbol{C}\boldsymbol{V}^{*-1/2}$. Then, by Sylvester's identity,

$$\left|\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T}+\boldsymbol{I}_n\right|=\left|\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{V}^{*-1/2}\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{V}^{*-1/2}+\boldsymbol{I}_{p+\sum_{i=1}^M K_i}\right|. \tag{4.38}$$

Taking the natural logarithm of (4.38), and differentiating with respect to $\frac{A^*_{q,ui}}{B^*_{q,ui}}$ then gives

$$\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\mathrm{tr}((\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}^*_q)_{\boldsymbol{u}_i})=\frac{B^*_{q,ui}}{A^*_{q,ui}}\mathrm{tr}(\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\Pi}^*_q). \tag{4.39}$$

Now substitute (4.37) and (4.39) into (4.17). We thus have the optimality requirement

$$\left\{A_{ui}+\frac{B^*_{q,ui}}{2A^*_{q,ui}}\mathrm{tr}(\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\Pi}^*_q)\right\}\frac{B^*_{q,ui}}{A^*_{q,ui}}=B_{ui}+\tfrac{1}{2}\left(\frac{B^*_{q,ui}}{A^*_{q,ui}}\right)^2\boldsymbol{y}^\mathsf{T}\boldsymbol{\Pi}^*_q\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\Pi}^*_q\boldsymbol{y}.$$

On the right hand side of (4.20) we have

$$\left\|\boldsymbol{y}-\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\boldsymbol{C}\boldsymbol{\Sigma}^*_q\boldsymbol{C}^\mathsf{T}\boldsymbol{y}\right\|^2=\left\|\boldsymbol{y}-\left(\sigma^2_\beta\boldsymbol{X}\boldsymbol{X}^\mathsf{T}+\sum_{i=1}^r\frac{B^*_{q,ui}}{A^*_{q,ui}}\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\right)\boldsymbol{\Pi}^*_q\boldsymbol{y}\right\|^2$$
$$=\left(\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\right)^2\|\boldsymbol{\Pi}^*_q\boldsymbol{y}\|^2. \tag{4.40}$$

Multiplying both sides of (4.38) by $\left(\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\right)^n$ gives

$$\left|\boldsymbol{C}\boldsymbol{V}^{*-1}\boldsymbol{C}^\mathsf{T}+\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\boldsymbol{I}_n\right|=\left(\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\right)^{n-p-\sum_{i=1}^r K_i}\left|\frac{A^*_{q,\varepsilon}}{B^*_{q,\varepsilon}}\boldsymbol{V}^{*-1/2}\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{V}^{*-1/2}+\boldsymbol{I}_{p+\sum_{i=1}^M K_i}\right|. \tag{4.41}$$

Taking the logarithm of (4.41), differentiating with respect to $\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}$, and rearranging gives

$$n-p-\left(\sum_{i=1}^r K_i\right)+\mathrm{tr}(\boldsymbol{V}^*\boldsymbol{\Sigma}^*_q)=\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\mathrm{tr}(\boldsymbol{\Pi}^*_q). \tag{4.42}$$

On substituting (4.40) and (4.42) into (4.20), we have

$$\left\{A_\varepsilon + \frac{B^*_{q,\varepsilon}}{2A^*_{q,\varepsilon}}\mathrm{tr}(\boldsymbol{\Pi}^*_q)\right\}\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}} = B_\varepsilon + \tfrac{1}{2}\left(\frac{B^*_{q,\varepsilon}}{A^*_{q,\varepsilon}}\right)^2 \|\boldsymbol{\Pi}^*_q \boldsymbol{y}\|^2.$$

4.A.4 Proof of Theorem 4.1

We wish to find the stationary points of (4.25). Taking the derivative with respect to ψ_0,

$$\begin{aligned}\frac{d}{d\psi_0}&\left\{\tfrac{1}{2}\log|\boldsymbol{\Pi}_q| - \tfrac{1}{2}\boldsymbol{y}^\mathsf{T}\boldsymbol{\Pi}_q\boldsymbol{y} - A_\varepsilon\log\psi_0 - B_\varepsilon\psi_0^{-1} - \sum_{i=1}^{r}\left(A_{ui}\log\psi_i + B_{ui}\psi_i^{-1}\right)\right\}\\ &= -\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Pi}_q) + \tfrac{1}{2}\|\boldsymbol{\Pi}_q\boldsymbol{y}\|^2 - A_\varepsilon\psi_0^{-1} + B_\varepsilon\psi_0^{-2}. \qquad (4.43)\end{aligned}$$

Setting (4.43) equal to zero, and rearranging,

$$\left\{A_\varepsilon + \tfrac{1}{2}\psi_0^*\mathrm{tr}(\boldsymbol{\Pi}^*_q)\right\}\psi_0 = B_\varepsilon + \tfrac{1}{2}\psi_0^{*2}\|\boldsymbol{\Pi}^*_q\boldsymbol{y}\|^2.$$

Taking the derivative of (4.25) with respect to ψ_i, for $1 \le i \le r$, we have

$$\begin{aligned}\frac{d}{d\psi_i}&\left\{\tfrac{1}{2}\log|\boldsymbol{\Pi}_q| - \tfrac{1}{2}\boldsymbol{y}^\mathsf{T}\boldsymbol{\Pi}_q\boldsymbol{y} - A_\varepsilon\log\psi_0 - B_\varepsilon\psi_0^{-1} - \sum_{i=1}^{r}\left(A_{ui}\log\psi_i + B_{ui}\psi_i^{-1}\right)\right\}\\ &= -\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Pi}_q) + \tfrac{1}{2}\|\boldsymbol{\Pi}_q\boldsymbol{y}\|^2 - A_{ui}\psi_i^{-1} + B_{ui}\psi_i^{-2}. \qquad (4.44)\end{aligned}$$

Setting (4.44) equal to zero, and rearranging,

$$\left\{A_{ui} + \tfrac{1}{2}\psi_i^*\mathrm{tr}(\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\Pi}^*_q)\right\}\psi_i^* = B_{ui} + \tfrac{1}{2}\psi_i^{*2}\boldsymbol{y}^\mathsf{T}\boldsymbol{\Pi}^*_q\boldsymbol{Z}_i\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\Pi}^*_q\boldsymbol{y}, \qquad (4.45)$$

We now substitute $\psi_0^* = \frac{B^*_{q,\varepsilon}}{2A^*_{q,\varepsilon}}$ into (4.44), and $\psi_i^* = \frac{B^*_{q,ui}}{2A^*_{q,ui}}$ into (4.45). The resultant conditions are (4.22) and (4.23).

4.A.5 Optimal q Densities for the Bayesian Probit Mixed Model

We now derive the optimal densities for the Bayesian probit mixed model. There are three lemmas involved. The first, below, gives the optimal $q^*_{\sigma^2}$. The second, Lemma 4.9, gives the optimal $q^*_{\boldsymbol{\beta},\boldsymbol{u}}$. The third, Lemma 4.8, then gives the optimal $q^*_{\boldsymbol{a}}$.

Lemma 4.7. *Let $q_{\boldsymbol{\beta},\boldsymbol{u}}$ have mean $\boldsymbol{\mu}_q$ and covariance $\boldsymbol{\Sigma}_q$, that is, $\boldsymbol{\mu}_q \equiv \mathsf{E}_{q(\boldsymbol{\beta},\boldsymbol{u})}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}$ and $\boldsymbol{\Sigma}_q \equiv \mathrm{Cov}_{q(\boldsymbol{\beta},\boldsymbol{u})}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}$. Then the optimal $q^*_{\sigma^2}$ in the Bayesian probit mixed model (4.28)–(4.30) is*

$$q^*_{\sigma^2}(\sigma^2_{u1},\ldots,\sigma^2_{ur}) = \prod_{i=1}^{r} q^*_{\sigma^2_{ui}}(\sigma^2_{ui}),$$

where

$$q^*_{\sigma^2_{ui}} \sim \text{IG}(A_{q,ui}, B_{q,ui}), \quad \text{for} \quad 1 \leq i \leq r.$$

The parameters $A_{q,ui}$ are deterministic,

$$A_{q,ui} = A_{ui} + \tfrac{K_i}{2}, \quad \text{for} \quad 1 \leq i \leq r.$$

While the parameters $B_{q,ui}$ are dependent on $\boldsymbol{\mu}_q$ and $\boldsymbol{\Sigma}_q$,

$$B_{q,ui} = B_{ui} + \tfrac{1}{2}\{\|(\boldsymbol{\mu}_q)_{\boldsymbol{u}_i}\|^2 + \text{tr}((\boldsymbol{\Sigma}_q)_{\boldsymbol{u}_i})\} \quad \text{for} \quad 1 \leq i \leq r,$$

where $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$.

Proof. Application of (4.6) leads to the optimal densities taking the form

$$q^*_{\boldsymbol{\sigma}^2}(\sigma^2_{u1}, \ldots, \sigma^2_{ur}) \propto \exp\left\{\mathsf{E}_{-q(\boldsymbol{\sigma}^2)} \log p(\boldsymbol{\sigma}^2|\text{rest})\right\},$$

where $\boldsymbol{\sigma}^2 = (\sigma^2_{u1}, \ldots, \sigma^2_{ur})^\mathsf{T}$. However, by applying the Markov blanket result,

$$\begin{aligned} p(\boldsymbol{\sigma}^2|\text{rest}) &= p(\boldsymbol{\sigma}^2|\text{Markov blanket of } \boldsymbol{\sigma}^2) \\ &= p(\boldsymbol{\sigma}^2|\boldsymbol{u}) \end{aligned}$$

We then have:

$$q^*_{\boldsymbol{\sigma}^2}(\sigma^2_{u1}, \ldots, \sigma^2_{ur}) \propto \exp\left\{\mathsf{E}_{-q(\boldsymbol{\sigma}^2)} \log p(\boldsymbol{\sigma}^2|\boldsymbol{u})\right\}.$$

The stated result then follows from the proof of Lemma 4.4. □

Lemma 4.8. *Let $q^*_{\boldsymbol{\sigma}^2}$ be a product of inverse gamma densities, with parameters,*

$$q^*_{\boldsymbol{\sigma}^2}(\sigma^2_{u1}, \ldots, \sigma^2_{ur}, \sigma^2_\varepsilon) = q^*_{\sigma^2_\varepsilon}(\sigma^2_\varepsilon) \prod_{i=1}^{r} q^*_{\sigma^2_{ui}}(\sigma^2_{ui})$$

$$q^*_{\sigma^2_\varepsilon} \sim \text{IG}(A_{q,\varepsilon}, B_{q,\varepsilon}), \quad \text{and} \quad q^*_{\sigma^2_{ui}} \sim \text{IG}(A_{q,ui}, B_{q,ui}).$$

Furthermore, let $q^*_{\boldsymbol{\beta},\boldsymbol{u}} \sim \text{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$. Then the optimal $q^*_{\boldsymbol{a}}$ is given by

$$q^*_{\boldsymbol{a}}(\boldsymbol{a}) \sim \prod_{i=1}^{n} \left[\text{TN}((\boldsymbol{C}\boldsymbol{\mu}_q)_i, 1, 0, \infty)\right]^{y_i} \left[\text{TN}((\boldsymbol{C}\boldsymbol{\mu}_q)_i, 1, -\infty, 0)\right]^{1-y_i}$$

where $\text{TN}(\cdot,\cdot,\cdot,\cdot)$ is the truncated normal distribution.

Proof. The optimal $q^*_{\boldsymbol{a}}$ is given by

$$q^*_{\boldsymbol{a}}(\boldsymbol{a}) \propto \exp\{\mathsf{E}_{-\boldsymbol{a}} \log p(\boldsymbol{a}|\text{rest})\}$$

Also,

$$\begin{aligned}
p(\boldsymbol{a}|\text{rest}) &= p(\boldsymbol{a}|\text{Markov blanket of } \boldsymbol{a}) \\
&= p(\boldsymbol{a}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y}) \\
&\propto p(\boldsymbol{a}, \boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}) \\
&= p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\beta}, \boldsymbol{u}) p(\boldsymbol{a}|\boldsymbol{\beta}, \boldsymbol{u}) \\
&= p(\boldsymbol{y}|\boldsymbol{a}) p(\boldsymbol{a}|\boldsymbol{\beta}, \boldsymbol{u}) \\
&= \left\{ \prod_{i=1}^{n} I(a_i \geq 0)^{1-y_i} I(a_i < 0)^{y_i} \right\} (2\pi)^{-n/2} \exp\left(-\frac{1}{2} \left\| \boldsymbol{a} - \boldsymbol{C} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right\|^2 \right).
\end{aligned}$$

We have

$$\begin{aligned}
q_{\boldsymbol{a}}^*(\boldsymbol{a}) &\propto \exp\left[\mathsf{E}_{-\boldsymbol{a}} \left\{ \log p(\boldsymbol{y}|\boldsymbol{a}) + \log p(\boldsymbol{a}|\boldsymbol{\beta}, \boldsymbol{u}) \right\} \right] \\
&\propto \left\{ \prod_{i=1}^{n} I(a_i \geq 0)^{1-y_i} I(a_i < 0)^{y_i} \right\} \exp\left\{ \mathsf{E}_{\boldsymbol{\beta},\boldsymbol{u}} \left(-\frac{1}{2} \left\| \boldsymbol{a} - \boldsymbol{C} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right\|^2 \right) \right\} \\
&\propto \left\{ \prod_{i=1}^{n} I(a_i \geq 0)^{1-y_i} I(a_i < 0)^{y_i} \right\} \exp\{ -\tfrac{1}{2} \|\boldsymbol{a} - \boldsymbol{C}\boldsymbol{\mu}_q\|^2 \}.
\end{aligned}$$

Which we recognise as the truncated normal distribution. So the optimal $q_{\boldsymbol{a}}^*$ is

$$\begin{aligned}
q_{\boldsymbol{a}}^*(\boldsymbol{a}) =& (2\pi)^{-n/2} \exp\{ -\tfrac{1}{2} \|\boldsymbol{a} - \boldsymbol{C}\boldsymbol{\mu}_q\|^2 \} \\
& \times \prod_{i=1}^{n} \left[\left\{ \frac{I(a_i \geq 0)}{\Phi((\boldsymbol{C}\boldsymbol{\mu}_q)_i)} \right\}^{y_i} \left\{ \frac{I(a_i < 0)}{1 - \Phi((\boldsymbol{C}\boldsymbol{\mu}_q)_i)} \right\}^{1-y_i} \right].
\end{aligned}$$

□

Recall that he mean of the truncated normal, denoted by $\boldsymbol{\eta}$, is given by

$$\boldsymbol{\eta} \equiv \mathsf{E}_{q(\boldsymbol{a})}(\boldsymbol{a}) = \boldsymbol{\mu}_{\boldsymbol{a}} + \frac{\boldsymbol{y}\phi(\boldsymbol{\mu}_{\boldsymbol{a}})}{\Phi(\boldsymbol{\mu}_{\boldsymbol{a}})} - \frac{(\mathbf{1} - \boldsymbol{y})\phi(\boldsymbol{\mu}_{\boldsymbol{a}})}{\mathbf{1} - \Phi(\boldsymbol{\mu}_{\boldsymbol{a}})}.$$

The next lemma considers the optimal $q_{\boldsymbol{\beta},\boldsymbol{u}}^*$ density.

Lemma 4.9. *Let $q_{\sigma^2}^*$ be a product of inverse gamma densities, with parameters,*

$$q_{\sigma^2}^*(\sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2) = q_{\sigma_\varepsilon^2}^*(\sigma_\varepsilon^2) \prod_{i=1}^{r} q_{\sigma_{ui}^2}^*(\sigma_{ui}^2)$$

$$q_{\sigma_\varepsilon^2}^* \sim \text{IG}(A_{q,\varepsilon}, B_{q,\varepsilon}), \quad \text{and} \quad q_{\sigma_{ui}^2}^* \sim \text{IG}(A_{q,ui}, B_{q,ui}).$$

Furthermore, let

$$\boldsymbol{\eta} = \mathsf{E}_{q^*(\boldsymbol{a})}(\boldsymbol{a}).$$

Then the optimal $q_{\boldsymbol{\beta},\boldsymbol{u}}^*$ is $\mathrm{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$, where

$$\begin{aligned}
\boldsymbol{\mu}_q &= \left(\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}} \right) \boldsymbol{\Sigma}_q \boldsymbol{C}^{\mathsf{T}} \boldsymbol{\eta}, \quad \text{and} \\
\boldsymbol{\Sigma}_q &= \left\{ \frac{A_{q,\varepsilon}}{B_{q,\varepsilon}} \boldsymbol{C}^{\mathsf{T}}\boldsymbol{C} + \text{blockdiag}\left(\sigma_\beta^{-2} \boldsymbol{I}_p, \frac{A_{q,u1}}{B_{q,u1}} \boldsymbol{I}_{K_1}, \ldots, \frac{A_{q,ur}}{B_{q,ur}} \boldsymbol{I}_{K_r} \right) \right\}^{-1}.
\end{aligned}$$

Proof. We set

$$q^*(\sigma_{\boldsymbol{u}}^2) \propto \exp\{\mathsf{E}_{-\sigma_{\boldsymbol{u}}^2} \log p(\sigma_{\boldsymbol{u}}^2|\text{rest})\}.$$

However,

$$\begin{aligned} p(\sigma_{\boldsymbol{u}}^2|\text{rest}) &= p(\sigma_{\boldsymbol{u}}^2|\text{Markov blanket of } \sigma_{\boldsymbol{u}}^2) \\ &= p(\sigma_{\boldsymbol{u}}^2|\boldsymbol{u}) \\ &\propto p(\sigma_{\boldsymbol{u}}^2, \boldsymbol{u}) \\ &= p(\boldsymbol{u}|\sigma_{\boldsymbol{u}}^2)p(\sigma_{\boldsymbol{u}}^2) \\ &= \prod_{i=1}^{r} p(u_i|\sigma_{ui}^2)p(\sigma_{ui}^2), \end{aligned}$$

but,

$$\begin{aligned} p(u_i|\sigma_{ui}^2)p(\sigma_{ui}^2) &= (2\pi)^{-K_i/2}(\sigma_{ui}^2)^{-K_i/2} \exp\left(-\frac{\|u_i\|}{2\sigma_{ui}^2}\right)(\sigma_{ui}^2)^{-A_{ui}-1} \exp\left(-\frac{B_{ui}}{\sigma_{ui}^2}\right) \\ &\sim \text{IG}(A_{ui} + \tfrac{1}{2}K_i, B_{ui} + \tfrac{1}{2}\|\boldsymbol{u}\|^2). \end{aligned}$$

So

$$\begin{aligned} q^*(\sigma_{\boldsymbol{u}}^2) &\propto \exp\left\{\mathsf{E}_{-\sigma_{\boldsymbol{u}}^2} \sum_{i=1}^{r} \left(-A_{ui} - \tfrac{1}{2}K_i - 1\right)\log(\sigma_{ui}^2) - \left(B_{ui} + \tfrac{1}{2}\|\boldsymbol{u}\|^2\right)\frac{1}{\sigma_{ui}^2}\right\} \\ &= \prod_{i=1}^{r} \exp\left\{\sum_{i=1}^{r} \left(-A_{ui} - \tfrac{1}{2}K_i - 1\right)\log(\sigma_{ui}^2) - \left(B_{ui} + \mathsf{E}_{-\sigma_{\boldsymbol{u}}^2}\tfrac{1}{2}\|\boldsymbol{u}\|^2\right)\frac{1}{\sigma_{ui}^2}\right\} \\ &= \prod_{i=1}^{r} (\sigma_{ui}^2)^{-\frac{K_i}{2}-A_{ui}-1} \exp\left(-\frac{\mathsf{E}_q\|\boldsymbol{u}\|^2}{2}\right). \end{aligned}$$

We then have

$$q^*(\sigma_{\boldsymbol{u}}^2) = \prod_{i=1}^{r} q^*(\sigma_{ui}^2),$$

and

$$q^*(\sigma_{ui}^2) \sim \text{IG}(A_{q,ui}, B_{q,ui})$$

where

$$A_{q,ui} = A_{ui} + \frac{K_{ui}}{2}$$

and

$$B_{q,ui} = B_{ui} + \tfrac{1}{2}\left\{\|(\boldsymbol{\mu}_q)_{ui}\|^2 + \text{tr}((\boldsymbol{\Sigma}_q)_{ui})\right\}.$$

Application of (4.6) leads to the optimal densities taking the form

$$
\begin{aligned}
q^*_{\beta,u}(\beta,u) &\propto \exp\left[\mathrm{E}_{q(\sigma^2)q(a)} \log\left\{p(a|\beta,u)p(\beta)p(u|\sigma_u^2)p(\sigma^2)\right\}\right] \\
&\propto \exp\left[\mathrm{E}_{q(\sigma^2)q(a)} \log\left\{p(a|\beta,u)p(\beta)p(u|\sigma_u^2)\right\}\right] \\
&\propto \exp\left[-\mathrm{E}_{q(\sigma^2)q(a)}\left\{\frac{\|a-X\beta-Zu\|^2}{2}+\frac{\|\beta\|^2}{2\sigma_\beta^2}+\frac{1}{2}u^\mathsf{T}(\sigma_u^2 I)^{-1}u\right\}\right] \\
&= \exp\left[-\mathrm{E}_{q(\sigma^2)}\left\{\frac{\|\eta-X\beta-Zu\|^2+\mathrm{tr}(\mathrm{Cov}_{q(a)}a)}{2}+\frac{\|\beta\|^2}{2\sigma_\beta^2}+\frac{1}{2}u^\mathsf{T}(\sigma_u^2 I)^{-1}u\right\}\right] \\
&\propto \exp\left[-\mathrm{E}_{q(\sigma^2)}\left\{\frac{\|\eta-X\beta-Zu\|^2}{2}+\frac{\|\beta\|^2}{2\sigma_\beta^2}+\frac{1}{2}u^\mathsf{T}(\sigma_u^2 I)^{-1}u\right\}\right]
\end{aligned}
$$

where $\mathrm{Cov}_{q(a)}a$ is the covariance matrix corresponding to $q(a)$. Now applying Lemma 4.3 *ii)*,

$$
\begin{aligned}
q^*_{\beta,u}(\beta,u) &\propto \exp\left[-\frac{\|\eta-X\beta-Zu\|^2}{2}-\frac{\|\beta\|^2}{2\sigma_\beta^2}-\sum_{i=1}^{r}\frac{A_{q,ui}u_i^\mathsf{T}u_i}{2B_{q,ui}}\right] \\
&\propto \exp\left[-\frac{1}{2}\begin{bmatrix}\beta\\ u\end{bmatrix}^\mathsf{T} C^\mathsf{T}C\begin{bmatrix}\beta\\ u\end{bmatrix}-\frac{1}{2\sigma_\beta^2}\beta^\mathsf{T}\beta-\sum_{i=1}^{r}\frac{A_{q,ui}u_i^\mathsf{T}u_i}{2B_{q,ui}}+\begin{bmatrix}\beta\\ u\end{bmatrix}C\eta\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{\begin{bmatrix}\beta\\ u\end{bmatrix}-\mu_q\right\}^\mathsf{T}\Sigma_q^{-1}\left\{\begin{bmatrix}\beta\\ u\end{bmatrix}-\mu_q\right\}\right],
\end{aligned}
$$

where

$$
\Sigma_{q(\beta,u)} = \left\{C^\mathsf{T}C+\mathrm{blockdiag}\left(\sigma_\beta^{-2}I_p, \frac{A_{q,u1}}{B_{q,u1}}I_{K_1},\ldots,\frac{A_{q,ur}}{B_{q,ur}}I_{K_r}\right)\right\}^{-1},
$$

and

$$
\mu_q = \Sigma_{q(\beta,u)}C^\mathsf{T}\eta.
$$

We recognise $q^*_{\beta,u}$ as Gaussian. We therefore have

$$
q^*_{(\beta,u)} \sim \mathrm{N}(\mu_q, \Sigma_{q(\beta,u)}).
$$

□

4.A.6 Pseudo-code for the Gaussian Linear Mixed Model

Algorithm 4.1 Pseudo-code for the Gaussian coordinate ascent.

Require: $\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_r, \boldsymbol{y}, \sigma_\beta^2, A_\varepsilon, A_{u1}, \ldots, A_{ur}, B_\varepsilon, B_{u1}, \ldots, B_{ur}$

$\boldsymbol{C} \leftarrow [\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_r]$

$A_{q,\varepsilon} \leftarrow A_\varepsilon + \frac{n}{2}$

$B_{q,\varepsilon} \leftarrow A_{q,\varepsilon}$

for $i = 1$ to r **do**

  $A_{q,ui} \leftarrow A_{ui} + \frac{K_i}{2}$

  $B_{q,ui} \leftarrow A_{q,ui}$

end for

repeat

  $\boldsymbol{V} \leftarrow \text{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p, \frac{A_{q,u1}}{B_{q,u1}}\boldsymbol{I}_{K_1}, \ldots, \frac{A_{q,ur}}{B_{q,ur}}\boldsymbol{I}_{K_r}\right)$

  $\boldsymbol{\Sigma}_q \leftarrow \left\{\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}}\boldsymbol{C}^\mathsf{T}\boldsymbol{C} + \boldsymbol{V}\right\}^{-1}$

  $\boldsymbol{\mu}_q \leftarrow \left(\frac{A_{q,\varepsilon}}{B_{q,\varepsilon}}\right)\boldsymbol{\Sigma}_q\boldsymbol{C}^\mathsf{T}\boldsymbol{y}$

  $B_{q,\varepsilon} \leftarrow B_\varepsilon + \frac{1}{2}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_q\|^2 + \frac{B_{q,\varepsilon}}{A_{q,\varepsilon}}\left(p + (\sum_{i=1}^r K_i) - \text{tr}(\boldsymbol{V}\boldsymbol{\Sigma}_q)\right)\}$

  for $i = 1$ to r **do**

    $B_{q,ui} \leftarrow B_{ui} + \frac{1}{2}\{\|(\boldsymbol{\mu}_q)_{u_i}\|^2 + \text{tr}((\boldsymbol{\Sigma}_q)_{u_i})\}$

  end for

until convergence

return $A_{q,\varepsilon}, A_{q,u1}, \ldots, A_{q,ur}, B_{q,\varepsilon}, B_{q,u1}, \ldots, B_{q,ur}, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q$

Algorithm 4.2 Pseudo-code for the Gaussian primal optimisation.

Require: $\boldsymbol{X}$, $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_r$, $\boldsymbol{y}$, σ_β^2, A_ε, $A_{u1}, \ldots, A_{ur}$, B_ε, $B_{u1}, \ldots, B_{ur}$

$\boldsymbol{C} \leftarrow [\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_r]$

$A_{q,\varepsilon} \leftarrow A_\varepsilon + \frac{n}{2}$

$\psi_0 \leftarrow 1$

for $i = 1$ to r **do**

  $A_{q,ui} \leftarrow A_{ui} + \frac{K_i}{2}$

  $\psi_i \leftarrow 1$

end for

repeat

  $\boldsymbol{V} \leftarrow \text{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p, \psi_1^{-1}\boldsymbol{I}_{K_1} \ldots, \psi_r^{-1}\boldsymbol{I}_{K_r}\right)$

  $\boldsymbol{\Sigma}_q \leftarrow \left\{\psi_0^{-1}\boldsymbol{C}^\top\boldsymbol{C} + \boldsymbol{V}\right\}^{-1}$

  $\boldsymbol{\mu}_q \leftarrow \psi_0^{-1}\boldsymbol{\Sigma}_q\boldsymbol{C}^\top\boldsymbol{y}$

  $\psi_0 \leftarrow \dfrac{B_\varepsilon + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_q\right\|^2}{A_\varepsilon + \frac{1}{2}\left(n - p - \sum_{i=1}^r K_i\right) + \frac{1}{2}\psi_0 \text{tr}(\boldsymbol{V}\boldsymbol{\Sigma}_q)}$

  for $i = 1$ to r **do**

    $\psi_i \leftarrow \dfrac{B_{ui} + \frac{1}{2}\left\|\left(\boldsymbol{\mu}_q\right)_{\boldsymbol{u}_i}\right\|^2}{A_{ui} + \frac{1}{2}(K_i - \text{tr}((\boldsymbol{V}\boldsymbol{\Sigma}_q)_{\boldsymbol{u}_i}))}$

  end for

until convergence

$B_{q,\varepsilon} \leftarrow A_{q,\varepsilon}\psi_0$

for $i = 1$ to r **do**

  $B_{q,ui} \leftarrow A_{q,ui}\psi_i$

end for

return $A_{q,\varepsilon}$, $A_{q,u1}, \ldots, A_{q,ur}$, $B_{q,\varepsilon}$, $B_{q,u1}, \ldots, B_{q,ur}$, $\boldsymbol{\mu}_q$, $\boldsymbol{\Sigma}_q$

Algorithm 4.3 Pseudo-code for the Gaussian dual optimisation.

Require: $\boldsymbol{X}, \boldsymbol{K}_1, \ldots, \boldsymbol{K}_r, \boldsymbol{y}, \sigma_\beta^2, A_\varepsilon, A_{u1}, \ldots, A_{ur}, B_\varepsilon, B_{u1}, \ldots, B_{ur}$

for $i = 0$ to r **do**
  $\psi_i \leftarrow 1$
end for
repeat
  $\boldsymbol{\Pi}_q \leftarrow \left\{\sigma_\beta^2 \boldsymbol{X}\boldsymbol{X}^\mathsf{T} + \sum_{i=1}^r \psi_i \boldsymbol{K}_i + \psi_0 \boldsymbol{I}_n\right\}^{-1}$
  $\psi_0 \leftarrow \dfrac{B_\varepsilon + \frac{1}{2}\psi_0^2 \|\boldsymbol{\Pi}_q \boldsymbol{y}\|^2}{A_\varepsilon + \frac{1}{2}\psi_0 \mathrm{tr}(\boldsymbol{\Pi}_q)}$
  for $i = 1$ to r **do**
    $\psi_i \leftarrow \dfrac{B_{ui} + \frac{1}{2}\psi_i^2 \boldsymbol{y}^\mathsf{T} \boldsymbol{\Pi}_q \boldsymbol{K}_i \boldsymbol{\Pi}_q \boldsymbol{y}}{A_{ui} + \frac{1}{2}\psi_i \mathrm{tr}(\boldsymbol{K}_i \boldsymbol{\Pi}_q)}$
  end for
until convergence
return $\psi_0, \ldots, \psi_r, \boldsymbol{\Pi}_q$

4.A.7 Pseudo-code for the Bayesian Probit Mixed Model

Algorithm 4.4 Pseudo-code for the probit coordinate ascent.

Require: $\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_r, \boldsymbol{y}, \sigma_\beta^2, A_{u1}, \ldots, A_{ur}, B_{u1}, \ldots, B_{ur}$

$\boldsymbol{C} \leftarrow [\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_r]$

for $i = 1$ to r **do**

  $A_{q,ui} \leftarrow A_{ui} + \frac{K_i}{2}$

  $B_{q,ui} \leftarrow A_{q,ui}$

end for

$\boldsymbol{\eta} \leftarrow \boldsymbol{y} - \frac{1}{2}\mathbf{1}$

repeat

  $\boldsymbol{\Sigma}_q \leftarrow \left\{ \boldsymbol{C}^\top \boldsymbol{C} + \text{blockdiag}\left(\sigma_\beta^{-2} \boldsymbol{I}_p, \frac{A_{q,u1}}{B_{q,u1}} \boldsymbol{I}_{K_1}, \ldots, \frac{A_{q,ur}}{B_{q,ur}} \boldsymbol{I}_{K_r} \right) \right\}^{-1}$

  $\boldsymbol{\mu}_q \leftarrow \boldsymbol{\Sigma}_q \boldsymbol{C}^\top \boldsymbol{\eta}$

  $\boldsymbol{\mu}_a \leftarrow \boldsymbol{C}\boldsymbol{\mu}_q$

  $\boldsymbol{\eta} \leftarrow \boldsymbol{\mu}_a + \dfrac{\boldsymbol{y}\boldsymbol{\phi}(\boldsymbol{\mu}_a)}{\boldsymbol{\Phi}(\boldsymbol{\mu}_a)} - \dfrac{(\mathbf{1} - \boldsymbol{y})\boldsymbol{\phi}(\boldsymbol{\mu}_a)}{\mathbf{1} - \boldsymbol{\Phi}(\boldsymbol{\mu}_a)}$

  for $i = 1$ to r **do**

    $B_{q,ui} \leftarrow B_{ui} + \frac{1}{2}\{ \| (\boldsymbol{\mu}_q)_{\boldsymbol{u}_i} \|^2 + \text{tr}((\boldsymbol{\Sigma}_q)_{\boldsymbol{u}_i}) \}$

  end for

until convergence

return $A_{q,u1}, \ldots, A_{q,ur}, B_{q,u1}, \ldots, B_{q,ur}, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q, \boldsymbol{\mu}_a$

Algorithm 4.5 Pseudo-code for the probit primal optimisation.

Require: $\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_r, \boldsymbol{y}, \sigma_\beta^2, A_{u1}, \ldots, A_{ur}, B_{u1}, \ldots, B_{ur}$

$\boldsymbol{C} \leftarrow [\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_r]$

for $i = 1$ to r **do**

  $A_{q,ui} \leftarrow A_{ui} + \frac{K_i}{2}$

  $\psi_i \leftarrow 1$

end for

$\boldsymbol{\eta} \leftarrow \boldsymbol{y} - \frac{1}{2}\mathbf{1}$

repeat

  $\boldsymbol{V} \leftarrow \text{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p, \psi_1^{-1}\boldsymbol{I}_{K_1} \ldots, \psi_r^{-1}\boldsymbol{I}_{K_r}\right)$

  $\boldsymbol{\Sigma}_q \leftarrow \left\{\boldsymbol{C}^\mathsf{T}\boldsymbol{C} + \boldsymbol{V}\right\}^{-1}$

  $\boldsymbol{\mu}_q \leftarrow \boldsymbol{\Sigma}_q\boldsymbol{C}^\mathsf{T}\boldsymbol{\eta}$

  $\boldsymbol{\mu}_a \leftarrow \boldsymbol{C}\boldsymbol{\mu}_q$

  $\boldsymbol{\eta} \leftarrow \boldsymbol{\mu}_a + \dfrac{\boldsymbol{y}\phi(\boldsymbol{\mu}_a)}{\Phi(\boldsymbol{\mu}_a)} - \dfrac{(\mathbf{1} - \boldsymbol{y})\phi(\boldsymbol{\mu}_a)}{\mathbf{1} - \Phi(\boldsymbol{\mu}_a)}$

  for $i = 1$ to r **do**

    $\psi_i \leftarrow \dfrac{B_{ui} + \frac{1}{2}\left\|\left(\boldsymbol{\mu}_q\right)_{\boldsymbol{u}_i}\right\|^2}{A_{ui} + \frac{1}{2}(K_i - \text{tr}((\boldsymbol{V}\boldsymbol{\Sigma}_q)_{\boldsymbol{u}_i}))}$

  end for

until convergence

for $i = 1$ to r **do**

  $B_{q,ui} \leftarrow A_{q,ui}\psi_i$

end for

return $A_{q,\varepsilon}, A_{q,u1}, \ldots, A_{q,ur}, B_{q,\varepsilon}, B_{q,u1}, \ldots, B_{q,ur}, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q, \boldsymbol{\mu}_a$

Algorithm 4.6 Pseudo-code for the probit dual optimisation.

Require: $\boldsymbol{X}, \boldsymbol{K}_1, \ldots, \boldsymbol{K}_r, \boldsymbol{y}, \sigma_\beta^2, A_{u1}, \ldots, A_{ur}, B_{u1}, \ldots, B_{ur}$

for $i = 1$ to r **do**
  $\psi_i \leftarrow 1$
end for
$\boldsymbol{\eta} \leftarrow \boldsymbol{y} - \frac{1}{2}\mathbf{1}$
repeat
  $\boldsymbol{K}_q \leftarrow \sigma_\beta^2 \boldsymbol{X}\boldsymbol{X}^\top + \sum_{i=1}^{r} \psi_i \boldsymbol{K}_i$
  $\boldsymbol{\Pi}_q \leftarrow \left\{\boldsymbol{K}_q + \boldsymbol{I}_n\right\}^{-1}$
  $\boldsymbol{\mu}_a \leftarrow \boldsymbol{K}_q \boldsymbol{\Pi}_q \boldsymbol{\eta}$
  $\boldsymbol{\eta} \leftarrow \boldsymbol{\mu}_a + \dfrac{\boldsymbol{y}\phi(\boldsymbol{\mu}_a)}{\Phi(\boldsymbol{\mu}_a)} - \dfrac{(\mathbf{1}-\boldsymbol{y})\phi(\boldsymbol{\mu}_a)}{\mathbf{1}-\Phi(\boldsymbol{\mu}_a)}$
  for $i = 1$ to r **do**
    $\psi_i \leftarrow \dfrac{B_{ui} + \frac{1}{2}\psi_i^2 \boldsymbol{\mu}_a^\top \boldsymbol{\Pi}_q \boldsymbol{K}_i \boldsymbol{\Pi}_q \boldsymbol{\mu}_a}{A_{ui} + \frac{1}{2}\psi_i \mathrm{tr}(\boldsymbol{K}_i \boldsymbol{\Pi}_q)}$
  end for
until convergence
return $\psi_1, \ldots, \psi_r, \boldsymbol{\Pi}_q, \boldsymbol{\mu}_a$

Chapter Five

Impact of Kernel Parameters on Degrees of Freedom

5.1 Introduction

The degrees of freedom is a well-established concept in the Statistical literature. As a measure of the complexity of a model, the degrees of freedom gives an intuitive insight into the amount of fitting being performed. This chapter is an investigation into the degrees of freedom, both in the impact of kernel parameters, and in the extension of the degrees of freedom to a broad range of models.

Every application of a kernel machine requires choice of (a) the general form of the kernel, (b) the parameters inherent within that form and (c) the amount of regularisation – often controlled by a single "smoothing" parameter, which we denote by $\lambda > 0$. In machine learning contexts the most common choice for (a) is the Gaussian kernel. For a d-dimensional predictor, or feature, space they take the form

$$k_1(\boldsymbol{s}, \boldsymbol{t}) = \exp\{-\gamma \|\boldsymbol{s} - \boldsymbol{t}\|^2\}, \quad \boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^d \tag{5.1}$$

where $\gamma > 0$ is a scale parameter. Hence, if the Gaussian kernel is adopted then the user is left with specifying the value of the pair $(\lambda, \gamma) \in \mathbb{R}^+ \times \mathbb{R}^+$. The choice of such parameters is an old problem in nonparametric regression and other smoothing contexts such as kernel density estimation where it can be viewed in bias-variance trade-off terms. A large literature exists on automatic selection of such smoothing parameters (e.g., Breiman and Peters, 1992; Jones, Marron and Sheather, 1996).

Most nonparametric regression estimators are parameterised in such as way that there is direct relationship between the smoothing parameter and degree of regularisation. For example, smoothing splines almost invariably have $\lambda > 0$, the multiple of the penalty functional controlling the amount of regularisation. However, the link between a particular value of λ, e.g., setting $\lambda = 3$, and the resulting functional fit is less clear and at least depends on the scale of the data. This problem can be overcome through

the concept of *effective degrees of freedom* (Buja, Hastie and Tibshirani, 1989). It maps the smoothing parameter into a new parameter, often called *degrees of freedom* or *df* for short, that is much more interpretable and free of scale issues.

In this chapter we investigate the impact of kernel machine parameters on degrees of freedom. For example, how does the choice of (λ, γ) in Gaussian kernel machines effect *df*? Conversely, if the user wants a kernel machine with $df = 10$, say, then which values of (λ, γ) should be used? With questions such as these in mind we obtain expressions and results on the properties of *df* as a function of kernel parameters. Monotonicity relationships between *df* and kernel parameters are established and recommendations for their choice, given the *df* value, are given.

The degrees of freedom may be extended to the multiple kernel setting, whereby the degrees of freedom may be attributed to the various predictors. This allows us to express the contribution that an individual predictor plays toward the fit. The degrees of freedom may also be extended to such areas as quantile regression and support vector machines, which normally do not have an associated degree of freedom.

In the next section we consider least squares kernel machines, and show that these are part of the class of *linear smoothers*. Section 5.3 considers generalisations of the degrees of freedom to wider class of models. In Section 5.4 we specialise to classification tasks, and conclude with a discussion in Section 5.5. The proofs for all theorems in this chapter are given in Appendix 5.A.

5.2 Least Squares Kernel Machines

Let $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \leq i \leq n$, be a set of predictor/response pairs. Consider the general nonparametric regression model

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \tag{5.2}$$

where the ε_i are zero mean random variables. For now, fitting involves modelling f to be of the form

$$f \in \mathcal{H}_k, \quad \text{where} \quad \mathcal{H}_k = \mathcal{H}_0 \oplus \mathcal{H}_1.$$

The RKHS $\mathcal{H}_0$ corresponds to some kernel, k_0, and $\mathcal{H}_1$ corresponds to $\mathcal{H}_0^{\perp}$. For $\lambda > 0$, the least squares model is

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \left\| P_1 f \right\|_{\mathcal{H}_k}^2 \right\}.$$

The fit, $f \in \mathcal{H}_k$ may not be unique. However, by Corollary 2.11, for each $1 \leq i \leq n$, the fitted values $f(\boldsymbol{x}_i)$ are uniquely determined. The coefficients of the fit are obtained according to

$$\min_{\boldsymbol{\beta},\boldsymbol{c}} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{c}\|^2 + \lambda \boldsymbol{c}^\mathsf{T}\boldsymbol{K}\boldsymbol{c} \right), \tag{5.3}$$

so that

$$f(\boldsymbol{x}_i) = (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}\boldsymbol{c})_i, \quad \text{for all } 1 \leq i \leq n.$$

Here $\boldsymbol{y}$ and $\boldsymbol{c}$ are the vectors containing the y_i and c_i, $\boldsymbol{K} = [k_1(\boldsymbol{x}_i, \boldsymbol{x}_j)]$ is the Gram matrix, and $\boldsymbol{X}$ is an $n \times p$ matrix of rank p, corresponding to the null space. The solution to (5.3) may be expressed as

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y} \quad \text{and} \quad \widehat{\boldsymbol{c}} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

Alternatively, setting $\boldsymbol{H} \equiv \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-1}\boldsymbol{X}^\mathsf{T}$, and optimising (5.3) with respect to $\boldsymbol{\beta}$ and then $\boldsymbol{c}$ yields

$$\widehat{\boldsymbol{c}} = ((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{c}}).$$

The fitted values, $\widehat{\boldsymbol{y}}$, are then

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{K}\widehat{\boldsymbol{c}} = \boldsymbol{S}\boldsymbol{y}, \tag{5.4}$$

where

$$\boldsymbol{S} = \boldsymbol{H} + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}\{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K} + \lambda \boldsymbol{I}\}^{-1}(\boldsymbol{I} - \boldsymbol{H}).$$

Methods that admit the form (5.4) are known as *linear smoothers* (Buja *et al.*, 1989). For linear smoothers, the degrees of freedom, df, is defined as the trace of $\boldsymbol{S}$. The following lemma gives provides a helpful expression for $\boldsymbol{S}$.

Lemma 5.1. *For the kernel machine given by* (5.3)*, the smoother matrix,* $\boldsymbol{S}$*, admits the expression*

$$\boldsymbol{S} = \boldsymbol{H} + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{H})\{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{H}) + \lambda \boldsymbol{I}\}^{-1}. \tag{5.5}$$

A proof of lemma 5.1 is given in Appendix 5.A. Similarly,

$$\boldsymbol{S} = \boldsymbol{H} + (\boldsymbol{I} - \boldsymbol{H})(\lambda^{-1}\boldsymbol{K})(\boldsymbol{I} - \boldsymbol{H})\{(\boldsymbol{I} - \boldsymbol{H})(\lambda^{-1}\boldsymbol{K})(\boldsymbol{I} - \boldsymbol{H}) + \boldsymbol{I}\}^{-1},$$

so that the degree of freedom may be canonically referenced as

$$df = df(\boldsymbol{H}, \lambda^{-1}\boldsymbol{K}),$$

where

$$df(\boldsymbol{H}, \boldsymbol{L}) = \mathrm{tr}(\boldsymbol{H}) + \mathrm{tr}[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{L}(\boldsymbol{I} - \boldsymbol{H})\{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{L}(\boldsymbol{I} - \boldsymbol{H}) + \boldsymbol{I}\}^{-1}].$$

We now establish that λ impacts $df(\boldsymbol{H}; \lambda^{-1}\boldsymbol{K})$ in a monotone fashion.

	$k_1(s,t)$	$\frac{d}{d\omega}K_\omega$
pth degree polynomial	$(1+s^\mathsf{T}t)^p$	$2\omega p K_{\omega,p-1} \odot (XX^\mathsf{T})$
Triangular	$1-\|s-t\|$	$-\Lambda$
Laplace	$\exp(-\|s-t\|)$	$-(K_\omega \odot \Lambda)$
Gaussian	$\exp(-\|s-t\|^2)$	$-2\omega(K_\omega \odot \Lambda \odot \Lambda)$

Table 5.1. *Expressions for $\frac{d}{d\omega}K_\omega$ for some commonly used kernels. The triangular kernel has domain $\mathcal{X} = \{x\colon \|x\| \leq 1\}$.*

Theorem 5.2. *The degrees of freedom, $df(H;\lambda^{-1}K)$, is monotonically decreasing in λ.*

Theorem 5.2 shows that a kernel machine may be parameterised not just in terms of λ, but in terms of the degrees of freedom. The degrees of freedom offers an intuitive parameterisation. In some contexts it is common to work with parametrisations such as $C = 1/\lambda$ where C is called the "cost" parameter. Theorem 5.2 immediately implies that $df(H;CK)$ is monotonically increasing in C. As well as monotonicity, higher order results are also obtained.

Theorem 5.3. *The degrees of freedom, $df(H;\lambda^{-1}K)$, is a convex function of λ.*

How is df impacted by other kernel parameters? The most common one is a scale parameter such as the γ appearing in the Gaussian kernel (5.1). More generally, for a given kernel K, we may consider the class of kernels

$$K_\omega(s,t) \equiv K(\omega s, \omega t), \quad \omega > 0,$$

corresponding to scaling of the inputs by ω. Taking the derivative of df via the smoother expression (5.5) leads to

$$(\partial/\partial\omega)df(H;\lambda^{-1}K_\omega) = \lambda \mathrm{tr}[\{(I-H)K_\omega(I-H)+\lambda I\}^{-2}(I-H)\tfrac{d}{d\omega}K_\omega(I-H)].$$

Simplification of $\frac{d}{d\omega}K_\omega$ depends on the functional form of k_1. Table 5.1 gives explicit expressions for this matrix for some common kernels. The following definitions apply in Table 5.1:

$$X \equiv [x_1 \cdots x_n]^\mathsf{T}, \quad \Lambda \equiv [\|x_i - x_j\|]_{1\leq i,j\leq n}, \quad [K_{\omega,p}]_{ij} \equiv \left\{1+(\omega x_i)^\mathsf{T}(\omega x_j)\right\}^p$$

and $A \odot B$ is the elementwise product of equal sized matrices A and B.

At the time of writing, we have not found an example of $(\partial/\partial\omega)df(H;\lambda^{-1}K_\omega)$ being negative for any of the kernels in Table 5.1. This leads to the conjecture that

$(\partial/\partial\omega)df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}_\omega) \geq 0$ for all $\omega > 0$ and positive definite k_1. We do have a proof for some special cases of kernels. *Dot product* kernels are defined to be those of the form

$$k(\boldsymbol{s},\boldsymbol{t}) = g(\boldsymbol{s}^\mathsf{T}\boldsymbol{t}),$$

where $g\colon \mathbb{R} \to \mathbb{R}$. Dot product kernels include the polynomial kernels, as well as, for $d = 1$,

$$g(s) = |s|^p, \quad \text{and} \quad g(s) = \text{sign}(s)|s|^p,$$

for any $p \in (0,\infty)$. The next theorem shows monotonicity of the degrees of freedom for all dot product kernels.

Theorem 5.4. *For a dot product kernel, the degrees of freedom, $df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}_\omega)$, is a monotonically increasing function of ω.*

It is not only the dot product kernels that have such a monotonicity property. The *translation invariant* kernels are defined to be those of the form

$$k(\boldsymbol{s},\boldsymbol{t}) = h(\boldsymbol{s}-\boldsymbol{t}),$$

where $h\colon \mathbb{R}^d \to \mathbb{R}$. Translation invariant kernels include the Gaussian, Laplacian and triangular kernels, as shown in Table 5.1.

Theorem 5.5. *For a triangular kernel, and $\mathcal{H}_0 = \mathbb{R}$ the RKHS of the null space, the degrees of freedom, $df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}_\omega)$, is a monotonically increasing function of ω.*

Extensive simulation have indicated that the monotonicity property also holds for Gaussian kernels. (For general d, monotonicity for Gaussian kernels implies monotonicity for Laplacian kernels (Schoenberg, 1935) and (Micchelli, 1986, Theorem A), as well as other translation invariant kernels.) An appreciation for the joint impact of kernel parameters on degrees of freedom can be obtained from Figure 5.1. It is based on Gaussian kernel smoothing of the fossil data described, for example, in Chaudhuri and Marron (1999) and Ruppert, Wand and Carroll (Section 3.6 of 2003). Note that Figure 5.1 uses $C = 1/\lambda$ so that the *df* surface is monotonically increasing in both directions. It also uses $\gamma = \omega^2$ corresponding to (5.1). The joint monotonic impact of both kernel parameters on effective degrees of freedom is apparent.

For a general kernel k with scale factor ω, an interesting question concerns that of choosing (λ,ω) for a fixed value, ℓ, of the effective degrees of freedom. This problem is more relevant for kernel machines with low-dimensional structure such as additivity. For least squares kernel machines a reasonable strategy is to base the choice on the

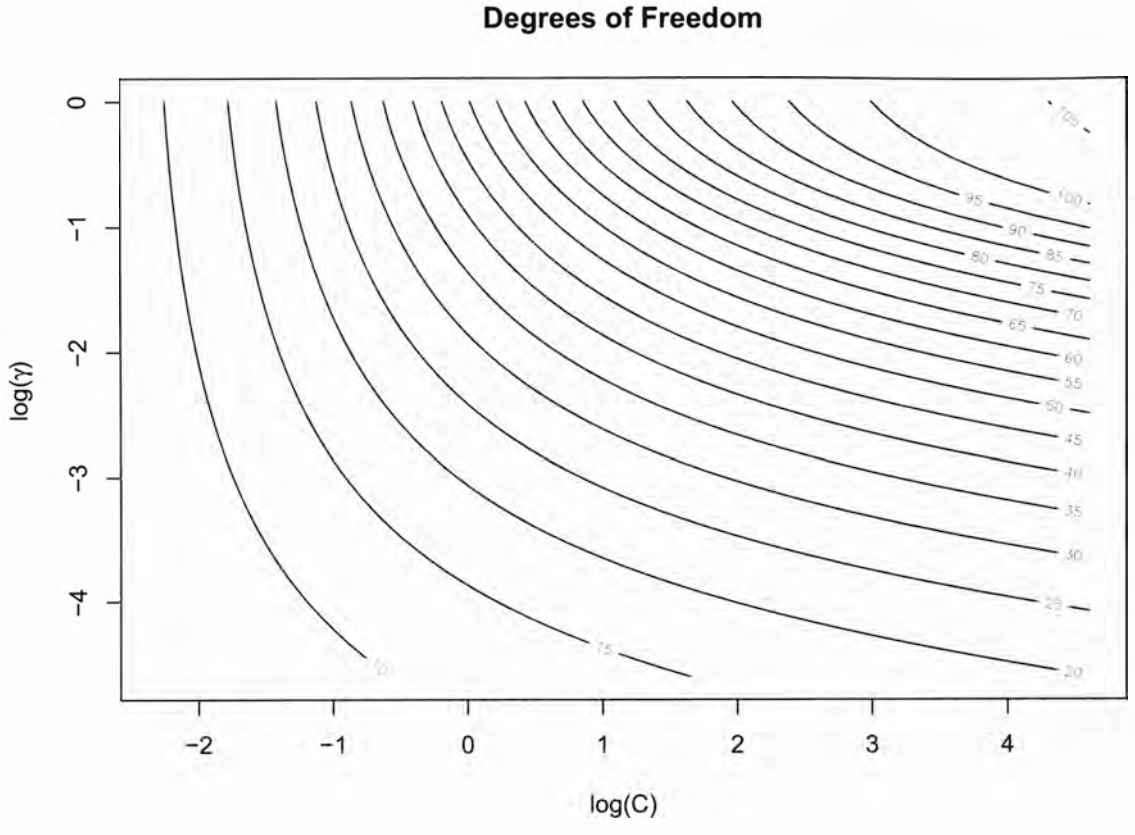


Figure 5.1. *Graph of df as a function of* (C, γ) *for the Gaussian kernel and motorcycle data. A logarithmic scale is used for the C and* γ *axes to aid visualisation. We have* $C = \lambda^{-1}$*, and* $\gamma = \omega^2$*.*

residual sum of squares $\text{RSS}(\lambda, \omega) \equiv \|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|^2$. We then choose the "best" (λ, ω) given by:

$$(\lambda^*, \omega^*)_\ell \quad \text{minimises RSS}(\lambda, \omega) \quad \text{subject to} \quad df(\boldsymbol{H}; \lambda^{-1}\boldsymbol{K}_\omega) = \ell.$$

The set of points given by $(\lambda^*, \omega^*)_\ell$ can be said to dominate all other points, and gives the minimum RSS for each value of the degrees of freedom.

Figure 5.2 shows the $(\lambda^*, \omega^*)_\ell$ path for the motorcycle data, as analysed by Fan and Gijbels (1996). A grid search was used to find the path. It is apparent that the path varies little in the ω direction as opposed to the λ direction. The optimal path remains within a small band of values for ω.

5.2.1 Extension to Multiple Kernel Penalisations

The degrees of freedom can be considered under a more general least squares setting. Here we extend the previous setting (5.2)–(5.3) to multiple kernel penalisations. The

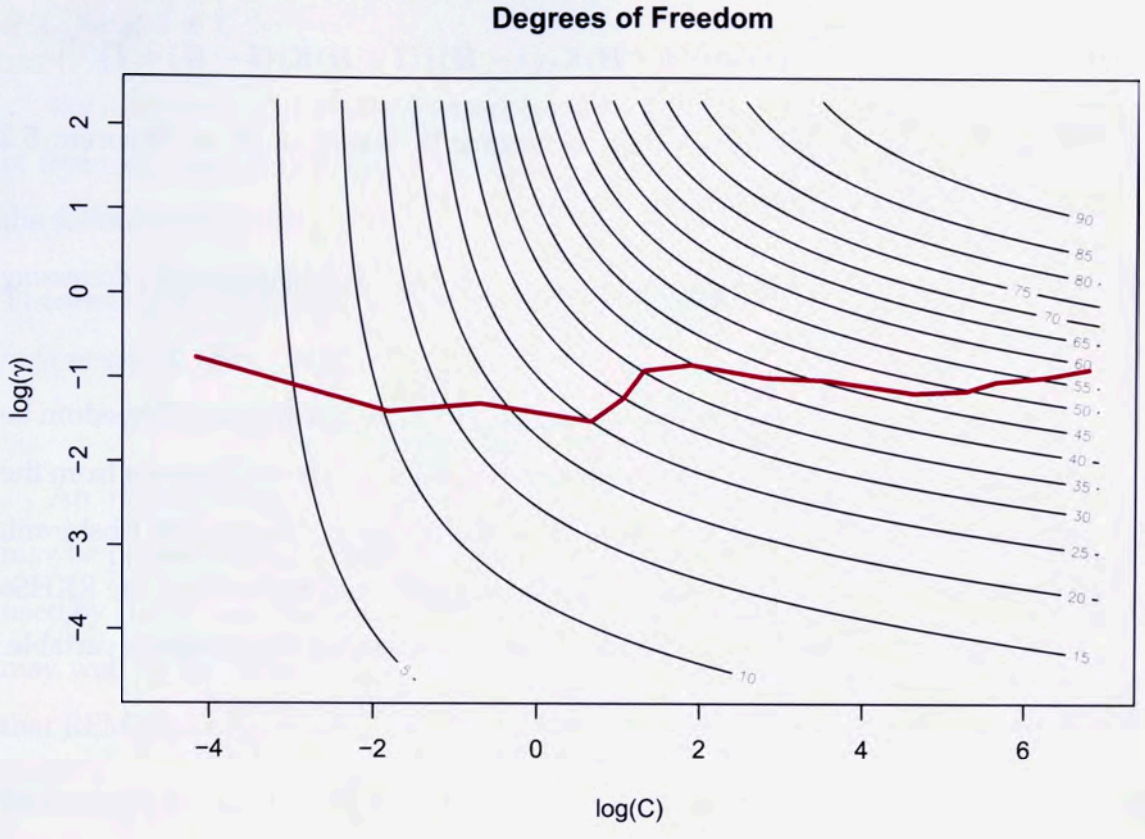


Figure 5.2. *The RSS-based $(C^*, \gamma^*)_\ell$ path for the motorcycle data. The path varies little in the γ or ω direction as opposed to the C or λ or direction.*

general additive model structure has:

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i,$$

for $f \in \mathcal{H}_k$, where

$$\mathcal{H}_k = \mathcal{H}_0 \oplus \cdots \oplus \mathcal{H}_r.$$

The least squares model is

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda_1 \|P_1 f\|^2_{\mathcal{H}_k} + \cdots + \lambda_r \|P_r f\|^2_{\mathcal{H}_k} \right\}. \tag{5.6}$$

Following Theorem 2.13, the fits are then

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{K}_\lambda \widehat{\boldsymbol{c}} = \boldsymbol{S}\boldsymbol{y}$$

where

$$\boldsymbol{K}_\lambda = \sum_{i=1}^{r} \boldsymbol{K}_i / \lambda_i,$$

and

$$\boldsymbol{S} = \boldsymbol{H} + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}_\lambda \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}_\lambda + \boldsymbol{I}\}^{-1}(\boldsymbol{I} - \boldsymbol{H}).$$

By Lemma 5.1 we may denote the degrees of freedom, $\mathrm{tr}(\boldsymbol{S})$, as

$$df(\boldsymbol{H}; \lambda_1^{-1}\boldsymbol{K}_1, \ldots, \lambda_r^{-1}\boldsymbol{K}_r) = p + \mathrm{tr}[(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_\lambda(\boldsymbol{I}-\boldsymbol{H})\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_\lambda(\boldsymbol{I}-\boldsymbol{H})+\boldsymbol{I}\}^{-1}].$$

The following theorem shows shows that monotonicity results such as Theorem 5.2 extend naturally to the more general setting (5.6).

Theorem 5.6. *The degrees of freedom,* $df(\boldsymbol{H}; \lambda_1^{-1}\boldsymbol{K}_1, \ldots, \lambda_r^{-1}\boldsymbol{K}_r)$*, is a monotonically decreasing function of* λ_i*, for all* $1 \leq i \leq r$*.*

A proof of Theorem 5.6 is given in the Appendix. How can the degrees of freedom be attributed to each of the $r+1$ kernels? A common method is to remove a kernel from the model, and see how much the degrees of freedom changes (e.g., Hastie and Tibshirani, 1990, page 128). It is usual within the additive model framework that each of the RKHSs $\mathcal{H}_1, \ldots, \mathcal{H}_r$ be associated with a particular effect, or component of the predictor variable. The change in degrees of freedom on removing effect j is then expressed:

$$\Delta_j df \equiv df(\boldsymbol{H}; \sum_{i=1}^{r} \lambda_i^{-1}\boldsymbol{K}_i) - df(\boldsymbol{H}; \sum_{\substack{i=1,\\ i\neq j}}^{r} \lambda_i^{-1}\boldsymbol{K}_i).$$

It is natural to attribute some p degrees of freedom to the null space. The following definition is a dual space version of that given by e.g., Wood (2006, page 171) and Ganguli and Wand (2007).

Definition 5.7. *The degrees of freedom attributed to effect* j*,* df_j*, is*

$$df_j \equiv \mathrm{tr}\left[(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{K}_j/\lambda_j)(\boldsymbol{I}-\boldsymbol{H})\left\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_\lambda(\boldsymbol{I}-\boldsymbol{H})+\boldsymbol{I}\right\}^{-1}\right].$$

for all $1 \leq j \leq r$*.*

An attractive aspect of df_j is that

$$df = p + \sum_{j=1}^{r} df_j.$$

That is, the degrees of freedom of the model may be attributed to the null space, and to each of the r effects. We now show that monotonicity results also apply to the degrees of freedom per effect.

Theorem 5.8. *The degrees of freedom attributed to effect* j *is a monotonically decreasing function of* λ_j*.*

The following theorem is somewhat more surprising, and shows an increasing monotonicity in λ_i, for $i \neq j$.

Theorem 5.9. *The degrees of freedom attributed to effect j is a monotonically increasing function of λ_i, for all $i \neq j$.*

We now make a brief comparison between df_j and $\Delta_j df$. It turns out that the degrees of freedom attributed to effect j is an upper bound for $\Delta_j df$. This is explicitly stated in the following theorem.

Theorem 5.10. *For degrees of freedom attributed to effect j, df_j, and change in degrees of freedom from removing effect j, $\Delta_j df$,*

$$df_j \geq \Delta_j df.$$

An attractive aspect of both df_j and $\Delta_j df$ is that the multiple kernel setting of (5.6) may be parameterised in terms of the degrees of freedom. Similar parameterisations are used by Hastie and Tibshirani (1990) and Wood (2006). The degrees of freedom per effect may well be the natural way for parameterising models. The following theorem shows that REML estimates may be recognised as being parameterised in such a manner.

Theorem 5.11. *Let f be the fit to* (5.6) *where the parameters $\lambda_1, \ldots, \lambda_r$ are chosen via REML with known variance or REML estimated variance σ^2. Then*

$$df_j = \frac{\|P_j f\|^2_{\mathcal{H}_k}}{\sigma^2},$$

for all $1 \leq j \leq r$.

Theorem 5.11 shows an intimate relationship the degrees of freedom per effect, the projected RKHS penalty, $\|P_j f\|^2_{\mathcal{H}_k}$ and the variance of REML fits.

5.3 Generalised Degrees of Freedom

There are many alternatives to least square loss. In the continuous response setting, these alternatives include the t-distribution loss, Huber's loss and ϵ-insensitive loss. For Bernoulli observations, alternatives include the Heaviside loss, hinge loss and Bernoulli log-likelihood loss. Such kernel machines do not conform to the linear smoother representation (5.4). It is of interest to examine how the concept of the degrees of freedom may be carried over to the general convex loss case. For simplicity, we restrict ourselves to the single penalty kernel machine,

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda \|P_1 f\|^2_{\mathcal{H}_k} \right\}, \tag{5.7}$$

where $\mathcal{L}$ is a convex loss.

As (5.7) does not conform to being a linear smoother, we do not have a smoother matrix from which we may define the degree of freedom. We consider possibilities for the extension of the degrees of freedom to kernel machines. There are two approaches in particular that are studied. Section 5.3.1 considers extensions to the *effective degrees of freedom*, as given by Ye (1998). Section 5.3.2 considers the degrees of freedom through the context of iteratively reweighted least squares applied to generalised linear models.

5.3.1 Effective Degrees of Freedom

In the familiar RKHS setting, for loss function, $\mathcal{L}(\cdot,\cdot)$, RKHS $\mathcal{H}_k$, and null space projection P_1, we have fit

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}\mathcal{L}(y_i,f(\boldsymbol{x}_i))+\lambda\left\|P_1 f\right\|_{\mathcal{H}_k}^2\right\}. \tag{5.8}$$

The fitted values are of the form

$$f(\boldsymbol{x})=\sum_{i=0}^{p}\beta_i\psi_i(x)+\sum_{i=1}^{n}c_i k(\boldsymbol{x},\boldsymbol{x}_i),$$

for some $\beta_i\in\mathbb{R}$, $0\leq i\leq p$ and $c_i\in\mathbb{R}$, $1\leq i\leq n$. Let $\widehat{y}_i=f(\boldsymbol{x}_i)$. For regression problems, consider the restriction

$$y_i\sim \mathrm{N}(\mu_i,\sigma^2). \tag{5.9}$$

The *effective degree of freedom, edf*, is then given by (e.g., Stein, 1981; Ye, 1998; Efron, 2004),

$$edf=\sum_{i=1}^{n}\mathrm{Cov}(\widehat{y}_i,y_i)/\sigma^2.$$

We wish to estimate the effective degrees of freedom. An unbiased estimate of the degrees of freedom may be made, using Stein's unbiased risk estimation (Stein, 1981). We say that a mapping is *almost differentiable* if each of its coordinates can be represented as a directional integral. Stein (1981) showed that

$$edf=\sum_{i=1}^{n}\mathrm{Cov}(\widehat{y}_i,y_i)/\sigma^2=\mathsf{E}\sum_{i=1}^{n}\frac{d\widehat{y}_i}{dy_i},$$

under the normality condition (5.9), continuity and piecewise differentiability of the mapping $\widehat{\boldsymbol{y}}=m(\boldsymbol{y})$, and almost everywhere derivatives $\frac{d\widehat{y}_i}{dy_i}$. As such, an unbiased estimate of the degrees of freedom is given by the *divergence*, v,

$$v=\sum_{i=1}^{n}\frac{d\widehat{y}_i}{dy_i}.$$

The divergence has been used for quantile regression (Koenker, 2005; Li, Liu and Zhu, 2007) and the lasso (Zou, Hastie and Tibshirani, 2007). For both of these examples, $\widehat{y}_i$ is not a differentiable function of y_i, though it is continuous and differentiable almost everywhere.

5.3.2 Iteratively Reweighted Least Squares

The method of iteratively reweighted least squares (IRLS) is a popular method for solving a variety of optimisation problems. In particular, IRLS is a commonly used method to find the fit to generalised linear models. As shown in Section Section 3.3.11, the fit given by generalised linear mixed models may be expressed as a special case of the kernel machine. The generalised linear mixed model involves the familiar optimisation problem

$$\max_{\boldsymbol{\beta},\boldsymbol{u}} \left[\exp\{\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}^{\mathsf{T}}b(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) - \tfrac{1}{2}\boldsymbol{u}^{\mathsf{T}}\boldsymbol{G}^{-1}\boldsymbol{u}\}\right], \tag{5.10}$$

with $\boldsymbol{G} = \lambda\boldsymbol{I}$ and b: $\mathbb{R} \to \mathbb{R}$. Following for example Holland and Welsch (1977) and Ganguli and Wand (2007), we now detail the IRLS method for the generalised linear model. For stability of IRLS, it is required that the function b have finite first and second derivatives.

i) Set λ.

ii) Obtain starting values for $\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix}$.

iii) Repeat

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} \leftarrow \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} - \left(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{W}_{\widehat{\eta}}\boldsymbol{C} + \begin{bmatrix} \mathbf{0} & \mathbf{0}^{\mathsf{T}} \\ \mathbf{0} & \lambda\boldsymbol{I} \end{bmatrix}\right)^{-1} \boldsymbol{C}\boldsymbol{W}_{\widehat{\eta}}\boldsymbol{y}_{adj},$$

where

$$\boldsymbol{y}_{adj} = \frac{\boldsymbol{y} - b'(\widehat{\eta})}{b''(\widehat{\eta})} \quad \text{and} \quad \boldsymbol{W}_{\widehat{\eta}} = \operatorname{diag}\{b''(\widehat{\eta})\},$$

with

$$\widehat{\eta} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{u}}.$$

Until convergence.

We note that is IRLS is simply Newton-Raphson optimisation applied to the natural logarithm of (5.10). At the convergence of IRLS, Hastie and Tibshirani (1990, Chapter 6) define the degrees of freedom as

$$df \equiv \operatorname{tr}\left\{\left(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{W}_{\widehat{\eta}}\boldsymbol{C}\right)\left(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{W}_{\widehat{\eta}}\boldsymbol{C} + \begin{bmatrix} \mathbf{0} & \mathbf{0}^{\mathsf{T}} \\ \mathbf{0} & \lambda\boldsymbol{I} \end{bmatrix}\right)^{-1}\right\}. \tag{5.11}$$

With the linear link function the usual smoother matrix result for the degree of freedom is obtained. Although (5.11) gives us a measure of the degrees of freedom suitable for generalised linear models, there are a number of extensions that we would like to make. These include:

i) To define the degrees of freedom without requiring IRLS.

ii) A kernelisation, to allow for a range of kernels beyond those that have a low-rank property.

iii) To allow for the use of kernel machine loss functions outside those that are equivalent to a generalised linear model.

For some $a_i \in \mathbb{R}$, for $1 \leq i \leq n$, denote by $f_{\boldsymbol{a}}$ the fit to the recentred kernel machine,

$$\min_{f_{\boldsymbol{a}} \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f_{\boldsymbol{a}}(\boldsymbol{x}_i) - a_i) + \lambda \left\| P_1 f_{\boldsymbol{a}} \right\|_{\mathcal{H}_k}^2 \right\}. \tag{5.12}$$

On setting $\boldsymbol{a} = \boldsymbol{0}$, we recover the usual kernel machine. The next theorem show that we may obtain the degrees of freedom without the need for the IRLS algorithm.

Theorem 5.12. *Let df be the degrees of freedom of the generalised linear mixed model* (5.11)*, with* $\boldsymbol{G} = \lambda \boldsymbol{I}$*, and* $b'' \colon \mathbb{R} \to [0, \infty)$*. Let* $\mathcal{L}$ *be a loss function with* $\mathcal{L}(s,t) = -2(st - b(t)) + C$ *for some constant, C. Furthermore, let* $\mathcal{H}_0$ *and* $\mathcal{H}_1$ *be orthogonal RKHSs with corresponding Gram matrices* $\boldsymbol{K}_0 = \boldsymbol{X}\boldsymbol{X}^\mathsf{T}$ *and* $\boldsymbol{K}_1 = \boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}$ *respectively, and* $f_{\boldsymbol{a}}$ *be the fit to the recentred kernel machine* (5.12)*. Then*

$$df = \sum_{i=1}^{n} \frac{d}{da_i} f_{\boldsymbol{a}}(\boldsymbol{x}_i)\Big|_{\boldsymbol{a}=\boldsymbol{0}}.$$

An outline of the proof is given in Appendix 5.A. Theorem 5.12 gives the degree of freedom for kernel machines that are equivalent to a class of generalised linear models with $b'' \colon \mathbb{R} \to [0, \infty)$. We would also like to have the degree of freedom for kernel machines outside of such class. Denote by $\mathcal{L}^{02}$ the second derivative of $\mathcal{L}$, with respect to the second term, i.e.,

$$\mathcal{L}^{02}(a,b) \equiv \frac{d^2 \mathcal{L}(a,b)}{db^2}.$$

We now make the following definition for the degree of freedom for kernel machines, with the restriction that $\mathcal{L}^{02}(a,b) \colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$. Such loss functions include both probit and logistic regression loss (e.g., Zhu and Hastie, 2005).

Definition 5.13. *Let* f *be the fit to the kernel machine* (5.8)*, where the loss* $\mathcal{L}$ *has the property* $\mathcal{L}^{02}(a,b) \colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$*. Then the degree of freedom of* f *is*

$$df \equiv \sum_{i=1}^{n} \frac{d}{da_i} f_{\boldsymbol{a}}(\boldsymbol{x}_i)\Big|_{\boldsymbol{a}=\boldsymbol{0}}$$

where $f_{\boldsymbol{a}}$ *is the fit to the recentred kernel machine* (5.12)*.*

The following definition is used by Steinwart (2007).

Definition 5.14. *A **distance-based loss** is a loss function with the translation invariance property:*

$$\mathcal{L}(a,b) = \mathcal{L}(a+c, b+c),$$

for all a, b, $c \in \mathbb{R}$.

Examples of regression loss are given in Table 3.1, and include least squares, Huber's loss and quantile regression loss. We now show that for some problems, df and v are equivalent.

Theorem 5.15. *Let* f *be the fit to a kernel machine with convex distance-based loss* $\mathcal{L}$ *such that* $\mathcal{L}^{02}(a,b)\colon \mathbb{R}\times\mathbb{R}\to[0,\infty)$. *Then*

$$df = v,$$

where df is the degree of freedom, and v the divergence.

Although we have defined the degree of freedom for a wide range of loss functions, these have included the requirement that $\mathcal{L}^{02}(a,b)\colon \mathcal{Y}\times\mathbb{R}\to[0,\infty)$. There are several important loss functions that do not have such a property. These loss functions include the quantile regression loss, ϵ-insensitive loss and hinge loss. Such loss functions may be approximated with second differentiable loss functions. Nychka *et al.* (1995), Yuan (2006) and Li *et al.* (2007) provide suitable approximations to the quantile regression loss. Lee and Mangasarian (2001) and Diehl (2004) provide smooth approximations to the hinge loss.

Index by $\mathcal{A}$ the observations that have finite $\mathcal{L}^{02}(y_i, f(\boldsymbol{x}_i))$, and $\mathcal{E}$ the complement of $\mathcal{A}$ over $\{1,\ldots,n\}$. For each $\mathcal{E}$, we wish to approximate $\mathcal{L}(y_i,\cdot)$ locally about $f(\boldsymbol{x}_i)$ by a convex function with finite first and second derivatives. The aim is to then gradually allow the second derivatives to approach ∞. It turns out that a unique degree of freedom arises. We have the following definition for the degrees of freedom for a kernel machine.

Definition 5.16. *Let* f *be a fit to the kernel machine with convex, continuous loss,* $\mathcal{L}$. *Index by* $\mathcal{A}$ *the observations that have* $\mathcal{L}^{02}(y_i, f(\boldsymbol{x}_i)) < \infty$, *and* $\mathcal{E}$ *the complement of* $\mathcal{A}$. *Then the degrees of freedom is defined as*

$$df \equiv \operatorname{rank}\left((\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})_{\mathcal{E}\mathcal{E}} + \boldsymbol{K}_{\mathcal{E}\mathcal{E}}\right) + \sum_{i\in\mathcal{A}} \frac{df_{\boldsymbol{a}}(\boldsymbol{x}_i)}{da_i}\Big|_{\boldsymbol{a}=\boldsymbol{0}}. \tag{5.13}$$

As a consequence of 5.16, for full-rank Gram matrices, the degree of freedom for support vector machines, and of quantile regression, is $df = |\mathcal{E}|$.

Figure 5.3 shows the degrees of freedom for a support vector classifier. The data was drawn from a 1–dimensional multimodal Gaussian, with some 200 observations in each

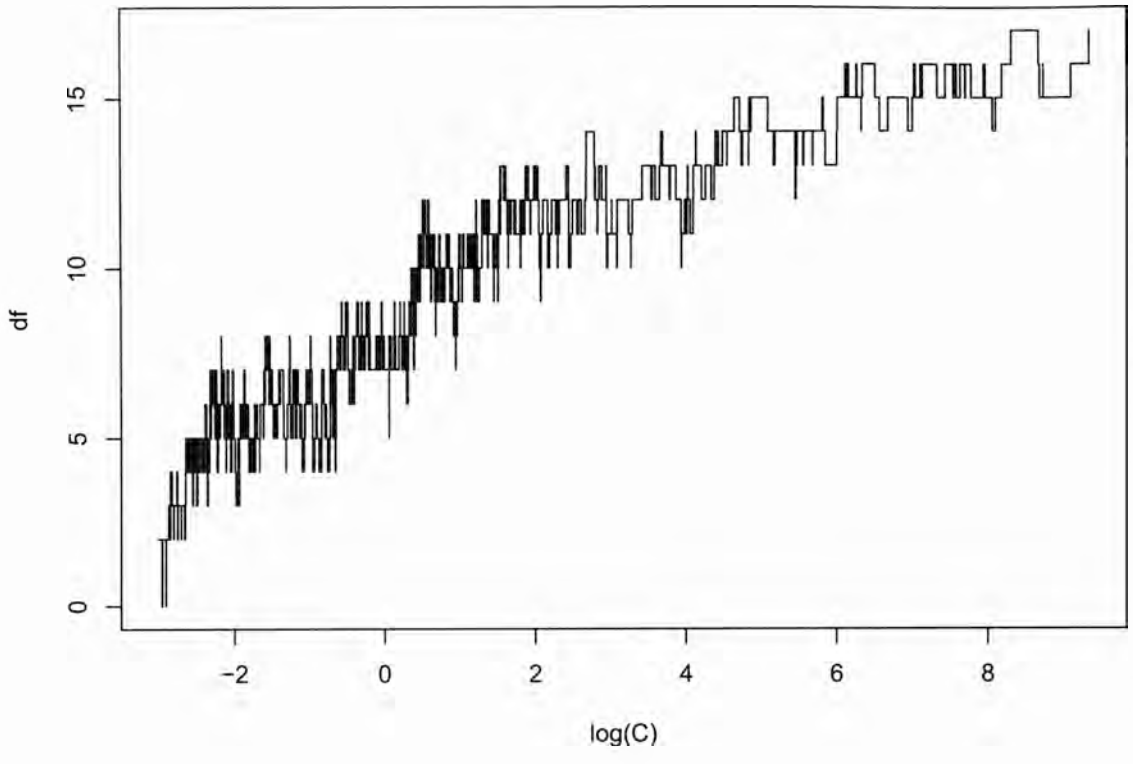


Figure 5.3. *Graph of df as a function of C for the Gaussian kernel and support vector machine. A logarithmic scale is used for the C to aid visualisation. The degrees of freedom is not monotonic. There is, however, a general increase in the degrees of freedom for larger values of C.*

class. Similar data was analysed by Hastie *et al.* (2004). Monotonicity in the degrees of freedom was not observed. There was however, a tendency toward higher degrees of freedom for larger C.

5.4 An Alternative for Classification

It is of interest if we can have an effective degrees of freedom for non-Gaussian distributions. Consider

$$df = \sum_{i=1}^{n} \mathrm{Cov}(\widehat{y}_i, y_i)/\sigma_i^2, \quad \text{where } \sigma_i^2 = \mathrm{Var}(y_i|\boldsymbol{x}_i) > 0. \tag{5.14}$$

It is easily seen that this matches the effective degrees of freedom under the normality assumption with fixed variance. In (5.14), we have a simple adjustment to allow for heterogeneous error distributions. Although it would appear that we now require knowledge of $\mathrm{Var}(y_i|\boldsymbol{x}_i)$, there is a special case whereby we do not: Binary classification.

For binary classification tasks, such as support vector machines, $y \in \{-1, 1\}$, and we have fits

$$f(\boldsymbol{x}) = \sum_{i=0}^{p} \beta_i \psi_i(x) + \sum_{i=1}^{n} c_i k(\boldsymbol{x}, \boldsymbol{x}_i),$$

for some $\beta_i \in \mathbb{R}$, $0 \leq i \leq p$ and $c_i \in \mathbb{R}$, $1 \leq i \leq n$. The corresponding fitted class values are

$$\widehat{y}_i = \text{sign}\left\{f(\boldsymbol{x}_i)\right\}, \quad \text{for all } 1 \leq i \leq n.$$

For classification problems, we define the effective degrees of freedom:

$$edf_c \equiv \mathsf{E} \sum_{i=1}^{n} \frac{\Delta \widehat{y}_i}{\Delta y_i}, \tag{5.15}$$

where

$$\frac{\Delta \widehat{y}_i}{\Delta y_i} = \frac{(\widehat{y}_i \mid y_i = 1) - (\widehat{y}_i \mid y_i = -1)}{2}.$$

We note that the degrees of freedom for classification matches that of (5.14). Definition 5.15 additionally allows for $\text{Var}(y_i|\boldsymbol{x}_i) = 0$. We give as classification deviance, v_c,

$$v_c = \sum_{i=1}^{n} \frac{\Delta \widehat{y}_i}{\Delta y_i}.$$

The deviance may be calculated by, in turn, changing the sign of y_i, and seeing if there is a change in $\widehat{y}_i$. Such calculations will be at least as computationally intensive as performing leave-one-out cross-validation.

5.4.1 Classification Example

For support vector machines, we would like to calculate the classification deviance. This involves, for each $1 \leq i \leq n$, changing the sign of y_i. Figure 5.4 shows the estimated degrees of freedom for a classification task. The data were drawn from a 1–dimensional multimodal Gaussian, with some 35 with $y_i = -1$ and 39 with $y_i = 1$. (Quite different results arise from having equal numbers in each class.) Unlike the least squares kernel machine, monotonicity in the degrees of freedom was not observed. There was however, a tendency toward higher degrees of freedom for larger λ, and a lowering of the degrees of freedom for larger γ.

To calculate the deviance we need to change the sign of each y_i for $1 \leq i \leq n$. This is equivalent to the removal, and subsequent addition of each observation. Our task is therefore similar in nature to performing leave-one-out cross-validation. We have, however, the additional requirement of re-learning the removed observation. There are a computationally efficient algorithms for both unlearning and learning observations.

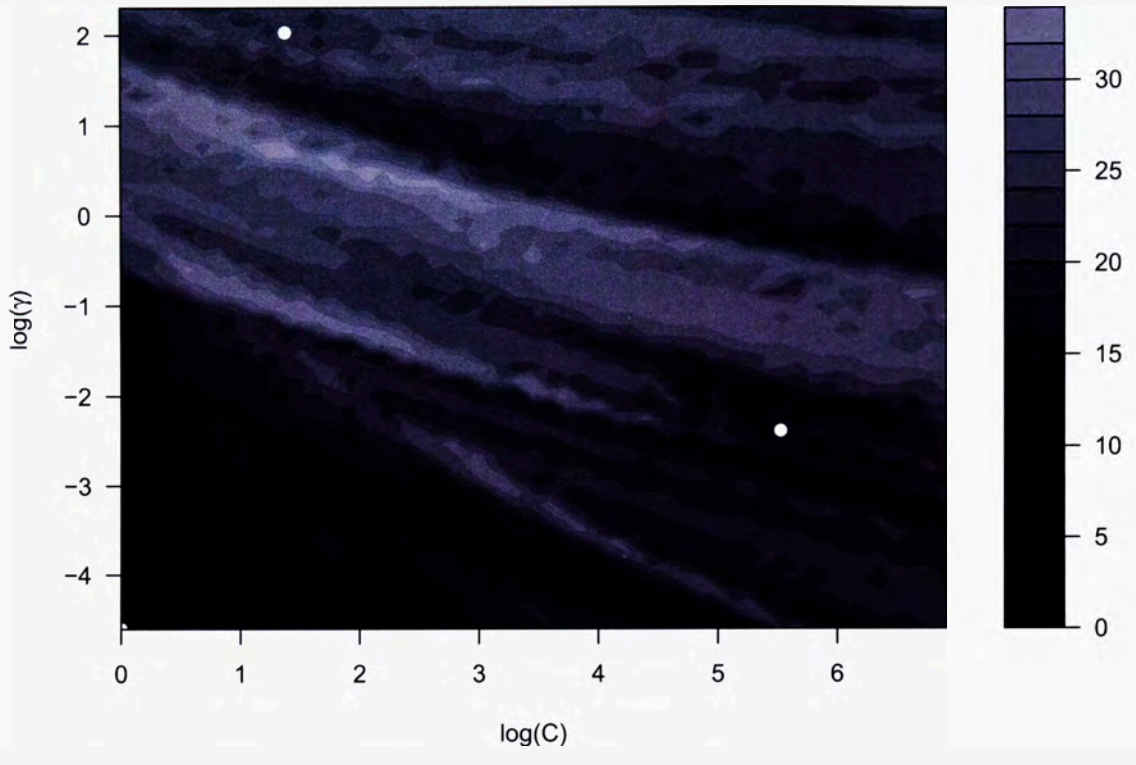


Figure 5.4. *Graph of v_c as a function of (C, γ) for the Gaussian kernel and support vector machine. A logarithmic scale is used for the C and γ axes to aid visualisation. The two white points indicated combinations (C, γ) that dominate many of the others. Only dominating points away from the plotted bounds on C and γ are shown. The classification deviance does not display monotonicity in either C or γ.*

For support vector machines, Cauwenberghs and Poggio (2001) detail a computationally efficient method. Similarly, there are accurate approximations available for leave-one-out cross-validation. In particular, the span bound of Vapnik and Chapelle (2000) may be adapted for an approximation to the deviance. Even for such efficient approximations, the degrees of freedom given in (5.13) will be simpler to calculate.

5.5 Discussion

This chapter has considered part of statistical folklore – that the degrees of freedom is a monotonic function of the smoothing parameter. Similar monotonicity results have been shown for kernel parameters. For least squares kernel machines, the degrees of freedom is a well established concept. The monotonicity in the smoothing parameter gives us

the ability to parameterise the model in terms of the degrees of freedom, as opposed to the less descriptive smoothing parameter. This also applies in attributing the degrees of freedom to various effects.

The IRLS algorithm can be used to calculate the degrees of freedom for general loss functions. Having instead derived the degrees of freedom without the need for IRLS broadens the scope of the degrees of freedom to instances such as quantile regression, whereby precise definitions are given.

The support vector machine is one of many algorithms for classification. Others, such as neural networks and logistic regression, have varying degrees of development in parameter selection. The classification deviance gives an approximation to the effective degrees of freedom across differing classification parameterisations and methods.

5.A Appendix

The proofs of this chapter require several properties of positive semidefinite matrices. Let us begin with a useful, though elementary Lemma. A proof of Lemma 5.17 is included for completeness.

Lemma 5.17. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be symmetric, positive semidefinite matrices of equal size. Then:*

i) $\boldsymbol{BAB}$ *is positive semidefinite.*

ii) $\operatorname{tr}(\boldsymbol{BA}) \geq 0$.

iii) $(\boldsymbol{A}+\boldsymbol{I})^{-1} - (\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{I})^{-1}$ *is positive semidefinite.*

Proof. i) Since $\boldsymbol{B}$ is symmetric, $\boldsymbol{BAB} = \boldsymbol{B}^{\mathsf{T}}\boldsymbol{AB}$ is positive semidefinite.

ii) Where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $n \times n$ matrices with ij terms a_{ij} and b_{ij} respectively,

$$\begin{aligned}\operatorname{tr}(\boldsymbol{BA}) &= \sum_{i,j=1}^{n} a_{ij}b_{ji} \\ &= \sum_{i,j=1}^{n} a_{ij}b_{ij} \\ &= \mathbf{1}^{\mathsf{T}}(\boldsymbol{A} \odot \boldsymbol{B})\mathbf{1} \\ &\geq 0,\end{aligned}$$

since the element-wise product, $\odot$, preserves positive semidefiniteness (e.g., Horn and Johnson, 1994, Theorem 5.2.1).

iii) The fundamental theorem of calculus gives

$$F(b) - F(a) = \int_a^b \left\{ \frac{d}{dt} F(t) \right\} dt.$$

On setting $F(t) = v^{\mathsf{T}}(A + tB + I)^{-1}v$, for any $v \in \mathbb{R}^n$, with $a = 0$ and $b = 1$, we have

$$v^{\mathsf{T}}(A+B+I)^{-1}v - v^{\mathsf{T}}(A+I)^{-1}v = \int_0^1 \left\{ \frac{d}{dt} v^{\mathsf{T}}(A+tB+I)^{-1}v \right\} dt.$$

Therefore,

$$\begin{aligned} v^{\mathsf{T}}\left[(A+I)^{-1} - (A+B+I)^{-1}\right] v &= -\int_0^1 v^{\mathsf{T}} \left\{ \frac{d}{dt}(A+tB+I)^{-1} \right\} v dt \\ &= \int_0^1 v^{\mathsf{T}}(A+tB+I)^{-1}B(A+tB+I)^{-1}v dt. \end{aligned}$$

By part *i)*, we know $(A+tB+I)^{-1}B(A+tB+I)^{-1}$ is positive semidefinite for all $t \geq 0$. Hence,

$$\begin{aligned} v^{\mathsf{T}}\left\{(A+I)^{-1} - (A+B+I)^{-1}\right\} v &= \int_0^1 v^{\mathsf{T}} \left\{ (A+tB+I)^{-1}B(A+tB+I)^{-1} \right\} v dt \\ &\geq 0. \end{aligned}$$

We therefore conclude that

$$(A+I)^{-1} - (A+B+I)^{-1}$$

is positive semidefinite. □

Proof of Lemma 5.1. We wish to show that

$$\begin{aligned} &H + (I-H)K\{(I-H)K + \lambda I\}^{-1}(I-H) \\ &\quad = H + (I-H)K(I-H)\{(I-H)K(I-H) + \lambda I\}^{-1} \end{aligned}$$

Let us begin by rearranging

$$\{(I-H)K(I-H) + \lambda I\} + \lambda(I+H),$$

in order to give

$$(I-H)\{(I-H)K(I-H) + \lambda I\} = \{(I-H)K + \lambda I\}(I-H). \tag{5.16}$$

From equation (5.16), we then apply left multiplication by $(I-H)K\{(I-H)K + \lambda I\}^{-1}$, and right multiplication by $\{(I-H)K(I-H) + \lambda I\}^{-1}$. We have

$$(I-H)K\{(I-H)K + \lambda I\}^{-1}(I-H) = (I-H)K(I-H)\{(I-H)K(I-H) + \lambda I\}^{-1}.$$

The addition of H to both sides then gives the required result. □

Proof of Theorem 5.2. The derivative of $df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K})$ is

$$\begin{aligned}\frac{d}{d\lambda}df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}) &= \frac{d}{d\lambda}\mathrm{tr}[(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}] \\ &= -\mathrm{tr}[\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\} \\ &\qquad \times\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}] \\ &\leq 0,\end{aligned}$$

since $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})$ is positive semidefinite. □

Proof of Theorem 5.3. The second derivative of $df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K})$ is

$$\begin{aligned}\frac{d^2}{d\lambda^2}df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}) &= \frac{d^2}{d\lambda^2}\mathrm{tr}\left[(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}\right] \\ &= -\frac{d}{d\lambda}\mathrm{tr}\left[\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\}\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-2}\right] \\ &= 2\frac{d}{d\lambda}\mathrm{tr}\left[\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\}\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-3}\right] \\ &\geq 0,\end{aligned}$$

where the inequality follows Lemma 5.17 *ii)*, as both the matrices $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})$ and $\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-3}$ are positive semidefinite. As the second derivative of $df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K})$ is non-negative, we conclude that $df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K})$ is convex. □

Proof of Theorem 5.4. The proof makes use of the following lemma, shown by FitzGerald, Micchelli and Pinkus (1995).

Lemma 5.18. *For dot product kernel, the matrix given by $\boldsymbol{K}_{\gamma+\varepsilon}-\boldsymbol{K}_{\gamma}$ is positive semidefinite for all γ, $\varepsilon\geq 0$.*

Letting γ, $\varepsilon\geq 0$,

$$\begin{aligned}&df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}_{\gamma+\varepsilon})-df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}_{\gamma}) \\ &\quad= -\mathrm{tr}\left\{(\boldsymbol{I}-\boldsymbol{H})\lambda^{-1}\boldsymbol{K}_{\gamma+\varepsilon}(\boldsymbol{I}-\boldsymbol{H})+\boldsymbol{I}\right\}^{-1}+\mathrm{tr}\left\{(\boldsymbol{I}-\boldsymbol{H})\lambda^{-1}\boldsymbol{K}_{\gamma}(\boldsymbol{I}-\boldsymbol{H})+\boldsymbol{I}\right\}^{-1} \\ &\quad= \mathrm{tr}\left[\left\{(\boldsymbol{I}-\boldsymbol{H})\lambda^{-1}\boldsymbol{K}_{\gamma}(\boldsymbol{I}-\boldsymbol{H})+\boldsymbol{I}\right\}^{-1}-\left\{(\boldsymbol{I}-\boldsymbol{H})\lambda^{-1}\boldsymbol{K}_{\gamma+\varepsilon}(\boldsymbol{I}-\boldsymbol{H})+\boldsymbol{I}\right\}^{-1}\right] \\ &\quad= \mathrm{tr}\left[(\boldsymbol{A}+\boldsymbol{I})^{-1}-(\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{I})^{-1}\right], \qquad (5.17)\end{aligned}$$

where

$$\boldsymbol{A}=(\boldsymbol{I}-\boldsymbol{H})\lambda^{-1}\boldsymbol{K}_{\gamma}(\boldsymbol{I}-\boldsymbol{H}),$$

and

$$\boldsymbol{B}=(\boldsymbol{I}-\boldsymbol{H})\lambda^{-1}(\boldsymbol{K}_{\gamma+\varepsilon}-\boldsymbol{K}_{\gamma})(\boldsymbol{I}-\boldsymbol{H}).$$

By Lemma 5.18, the matrix $\boldsymbol{B}$ is positive semidefinite. Applying Lemma 5.17 *iii)* to (5.17), we have

$$df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}_{\gamma+\varepsilon}) - df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}_{\gamma}) \geq 0.$$

Hence the degrees of freedom, $df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}_{\omega})$, monotonically increasing in ω. □

Proof of Theorem 5.5. We now show that for triangular kernel, and $\mathcal{H}_0 = \mathbb{R}$, that the degrees of freedom is a monotonically increasing function of ω. If $\mathcal{H}_0 = \mathbb{R}$, then, up to a multiplicative constant, $\boldsymbol{X} = \mathbf{1}$, and therefore $\boldsymbol{H} = \boldsymbol{I} - \mathbf{1}\mathbf{1}^{\mathsf{T}}/n$. We have the derivative,

$$\begin{aligned}
(d/d\omega)df(\boldsymbol{H};\lambda^{-1}\boldsymbol{K}_{\omega}) &= -(d/d\omega)\lambda\mathrm{tr}[\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_{\omega}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}] \\
&= -\lambda\mathrm{tr}\Big[\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_{\omega}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{H})\} \\
&\qquad \times \{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_{\omega}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}\Big] \\
&\geq 0,
\end{aligned}$$

where the inequality follows since the matrix given by

$$(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{H}) = -(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})$$

is negative semidefinite. For triangular kernel, the degrees of freedom is therefore a monotonically increasing function of ω. We note that this result may be generalised to any $\boldsymbol{X}$ with the property that $\mathbf{1}$ is in the column span of $\boldsymbol{X}$. □

Proof of Theorem 5.6. Taking the derivative of $df(\lambda_1,\ldots,\lambda_r;\boldsymbol{K}_1,\ldots,\boldsymbol{K}_r)$, with respect to λ_i, for some $1 \leq i \leq r$,

$$\begin{aligned}
&\frac{d}{d\lambda_i}df(\lambda_1,\ldots,\lambda_r;\boldsymbol{K}_1,\ldots,\boldsymbol{K}_r) \\
&\quad = \frac{d}{d\lambda_i}\Big(n+p-\mathrm{tr}[\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_{\lambda}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}]\Big) \\
&\quad = \mathrm{tr}[\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_{\lambda}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}\frac{d\boldsymbol{K}_{\lambda}}{d\lambda_i}\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_{\lambda}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}] \\
&\quad = -\lambda^{-2}\mathrm{tr}[\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_{\lambda}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}\boldsymbol{K}_i\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}_{\lambda}(\boldsymbol{I}-\boldsymbol{H})+\lambda\boldsymbol{I}\}^{-1}] \\
&\quad \leq 0.
\end{aligned}$$

Hence the degrees of freedom is monotonically decreasing in λ_i, for all $1 \leq i \leq r$. □

Proof of Theorem 5.8. We wish to show that degrees of freedom attributed to effect j is monotonically decreasing in λ_j. For the following, we denote by $\boldsymbol{M}$ the positive definite matrix:

$$\boldsymbol{M} = \left\{(\boldsymbol{I}-\boldsymbol{H})(\sum_{i\neq j}\lambda_i^{-1}\boldsymbol{K}_i)(\boldsymbol{I}-\boldsymbol{H}) + \boldsymbol{I}_n\right\}.$$

Deriving df_j with respect to λ_j, for some $1 \leq j \leq r$,

$$\begin{aligned}
\frac{d}{d\lambda_j} df_j &= \frac{d}{d\lambda_j}\left(\operatorname{tr}\left[\left\{(I-H)\lambda_j^{-1}K_j(I-H)\right\}\{(I-H)K_\lambda(I-H)+I\}^{-1}\right]\right) \\
&= \frac{d}{d\lambda_j}\left(\operatorname{tr}\left[\left\{M+(I-H)\lambda_j^{-1}K_j(I-H)-M\right\}\{(I-H)K_\lambda(I-H)+I\}^{-1}\right]\right) \\
&= \frac{d}{d\lambda_j}\left(\operatorname{tr}\left[I-M\{(I-H)K_\lambda(I-H)+I\}^{-1}\right]\right) \\
&= \frac{d}{d\lambda_j}\left(n-\operatorname{tr}\left[M\{(I-H)K_\lambda(I-H)+I\}^{-1}\right]\right) \\
&= \lambda_j^{-2}\frac{d}{d\lambda_j^{-1}}\left(\operatorname{tr}\left[M\{(I-H)K_\lambda(I-H)+I\}^{-1}\right]\right) \\
&= -\lambda_j^{-2}\operatorname{tr}\left[M\left\{\{(I-H)K_\lambda(I-H)+I\}^{-1}K_j\{(I-H)K_\lambda(I-H)+I\}^{-1}\right\}\right] \\
&\leq 0.
\end{aligned}$$

Where the inequality follows from Lemma 5.17 *ii)*, since

$$\{(I-H)K_\lambda(I-H)+I\}^{-1}K_j\{(I-H)K_\lambda(I-H)+I\}^{-1}$$

is positive semidefinite. □

Proof of Theorem 5.10. We wish to show that:

$$df_j \geq \Delta_j df.$$

Let us begin by

$$\begin{aligned}
&df_j - \Delta_j df \\
&\quad= \operatorname{tr}[(I-H)\lambda_j^{-1}K_j(I-H)\{(I-H)K_\lambda(I-H)+I\}^{-1}] \\
&\qquad - df(H;\sum_{i=1}^{r}\lambda_i^{-1}K_i) + df(H;\sum_{i\neq j}\lambda_i^{-1}K_i) \\
&\quad= \operatorname{tr}[(I-H)\lambda_j^{-1}K_j(I-H)\{(I-H)K_\lambda(I-H)+I\}^{-1}] \\
&\qquad - \operatorname{tr}[(I-H)K_\lambda(I-H)\{(I-H)K_\lambda(I-H)+I\}^{-1}] \\
&\qquad + \operatorname{tr}[(I-H)\left(K_\lambda-\lambda_j^{-1}K_j\right)(I-H)\{(I-H)\left(K_\lambda-\lambda_j^{-1}K_j\right)(I-H)+I\}^{-1}] \\
&\quad= -\operatorname{tr}[(I-H)\left(K_\lambda-\lambda_j^{-1}K_j\right)(I-H)\{(I-H)K_\lambda(I-H)+I\}^{-1}] \\
&\qquad + \operatorname{tr}[(I-H)\left(K_\lambda-\lambda_j^{-1}K_j\right)(I-H)\{(I-H)\left(K_\lambda-\lambda_j^{-1}K_j\right)(I-H)+I\}^{-1}] \\
&\quad= \operatorname{tr}\left[A\left\{(A+I)^{-1}-(A+B+I)^{-1}\right\}\right],
\end{aligned}$$

where

$$A = (I-H)\Big(\sum_{i\neq j}\lambda_i^{-1}K_i\Big)(I-H),$$

and

$$\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{H})(\lambda_j^{-1}\boldsymbol{K}_j)(\boldsymbol{I} - \boldsymbol{H}).$$

By Lemma 5.17 *iii)*,

$$\left\{(\boldsymbol{A} + \boldsymbol{I})^{-1} - (\boldsymbol{A} + \boldsymbol{B} + \boldsymbol{I})^{-1}\right\}$$

is positive definite, as $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semidefinite. Hence, by Lemma 5.17 *ii)*,

$$\begin{aligned} df_j - \Delta_j df &= \operatorname{tr}\left[\boldsymbol{A}\left\{(\boldsymbol{A} + \boldsymbol{I})^{-1} - (\boldsymbol{A} + \boldsymbol{B} + \boldsymbol{I})^{-1}\right\}\right] \\ &\geq 0. \end{aligned}$$

We may then conclude that

$$df_j \geq \Delta_j df,$$

as required. □

Proof of Theorem 5.11. The proof follows from a rearrangement of the REML equations. Using the notation of Chapter 4, REML equations give

$$\lim_{\sigma_\beta^2 \to \infty} \boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Pi}\boldsymbol{K}_i\boldsymbol{\Pi}\boldsymbol{y} = \lim_{\sigma_\beta^2 \to \infty} \operatorname{tr}(\boldsymbol{K}_i\boldsymbol{\Pi}),$$

for all $1 \leq i \leq r$. However,

$$\left\|P_j f\right\|_{\mathcal{H}_k}^2 = \lim_{\sigma_\beta^2 \to \infty} \boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Pi}\boldsymbol{K}_i\boldsymbol{\Pi}\boldsymbol{y}$$

and

$$\sigma^2 df_j = \lim_{\sigma_\beta^2 \to \infty} \operatorname{tr}(\boldsymbol{K}_i\boldsymbol{\Pi})$$

We therefore have

$$df_j = \frac{\left\|P_j f\right\|_{\mathcal{H}_k}^2}{\sigma^2}$$

as required. □

Proof of Theorem 5.12. It is shown in Chapter 3 that the generalised linear model has equivalent fits to the kernel machine. The proof follows by considering the second order Taylor series expansion of the recentred kernel machine objective function. The expansion is in terms of

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \\ \boldsymbol{a} \end{bmatrix} \quad \text{about} \quad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \\ \boldsymbol{0} \end{bmatrix},$$

and follows that of Diehl (2004). □

Proof of Theorem 5.15. It is clear that since $\mathcal{L}$ is a distance-based loss,

$$\mathcal{L}(y+\epsilon, f(x_i)) = \mathcal{L}(y, f(x_i)-\epsilon),$$

and hence, for $\epsilon \neq 0$,

$$\frac{\mathcal{L}(y+\epsilon, f(x_i)) - \mathcal{L}(y, f(x_i))}{\epsilon} = \frac{\mathcal{L}(y, f(x_i)-\epsilon) - \mathcal{L}(y, f(x_i))}{\epsilon}.$$

Letting $\epsilon \to 0$,

$$\frac{df(x_i)}{dy_i} = \frac{d}{da_i} f_a(x_i)\Big|_{a=0}$$

so that

$$v = \sum_{i=1}^{n} \frac{df(x_i)}{dy_i} = \sum_{i=1}^{n} \frac{d}{da_i} f_a(x_i)\Big|_{a=0} = df. \qquad \square$$

Chapter Six

Active Set Optimisation of Support Vector Machines

6.1 Introduction

The support vector machine[1] has emerged as an effective and elegant method for supervised learning, or classification, problems. There is a considerable amount of machine learning literature on training an SVM. Several implementations, such as Platt (1999), Joachims (1999) and Chang and Lin (2009), allow an SVM to be trained on large scale problems. When the SVM has special structure, such as a low-rank decomposition, interior point methods (Ormerod, Wand and Koch, 2008), or cutting plane methods (Joachims, 2006), allow for fast and accurate solutions. The optimisation task becomes more difficult on large-scale, full-rank problems.

There is a strong interest a strong interest in the Machine Learning community in such difficult optimisations. The most popular choice of kernel, the Gaussian, does not conform to a low-rank decomposition. It is the infinite dimensional, rich, reproducing kernel Hilbert spaces of Gaussian and Laplacian kernels that have attracted much interest (e.g., Steinwart, 2001). With such kernels the memory requirements of the algorithm is an important consideration. This chapter focuses on the fast and accurate training of support vector classification machines.

At the heart of an SVM lies a quadratic programming problem. This QP has simple box constraints, and a single linear constraint. Many existing QP algorithms have memory requirements that are not suitable for large scale problems. In this chapter, we implement a method that uses a combination of three different phases, an initialisation phase, a decomposition phase, and a conjugate gradient phase. We cycle through the phases until convergence is met. As demonstrated by numerical studies, the algorithm is fast and able to handle large training data sets.

[1]This chapter is based on material under submission for publication, to IEEE transactions on Neural Networks, as: N. D. Pearce and M. P. Wand, Active Set Optimization of Support Vector Machines.

In Section 6.2 we describe the SVM, and derive an appropriate QP. Optimality conditions will also be presented. Our algorithm is described in Section 6.3, including details of each phase. Computational results are reported in Section 6.4, and we conclude with a discussion in Section 6.5.

6.2 Support Vector Classifiers

Suppose we observe features, $\boldsymbol{x}_i \in \mathbb{R}^d, 1 \leq i \leq n$, together with corresponding classes $y_i \in \{-1, 1\}$. Let k be a kernel, and $\boldsymbol{K}$ be the corresponding Gram matrix, with entries $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The *dual* QP problem is

$$\min_{\boldsymbol{\alpha}} \tfrac{1}{2} (\boldsymbol{\alpha} \odot \boldsymbol{y})^\mathsf{T} \boldsymbol{K} (\boldsymbol{\alpha} \odot \boldsymbol{y}) - \mathbf{1}^\mathsf{T} \boldsymbol{\alpha}$$

$$\text{subject to } 0 \leq \alpha_i \leq c_i, \text{ for all } 1 \leq i \leq n, \text{ and } \boldsymbol{y}^\mathsf{T} \boldsymbol{\alpha} = 0.$$

With $c_i = C$, for all $1 \leq i \leq n$, we obtain the C-SVM formulation of Cortes and Vapnik (1995). The general form of the primal QP allows a different weight for each sample. Following Osuna, Freund and Girosi (1997a), a common approach is to set

$$c_i = \begin{cases} C_1, & y_i = 1, \\ C_{-1}, & y_i = -1, \end{cases}$$

allowing different weights to be given to each class.

Making the change of variable $\boldsymbol{a} = \boldsymbol{\alpha} \odot \boldsymbol{y}$, we denote the *canonical dual* QP problem as

$$\min_{\boldsymbol{a}} \tfrac{1}{2} \boldsymbol{a}^\mathsf{T} \boldsymbol{K} \boldsymbol{a} - \boldsymbol{y}^\mathsf{T} \boldsymbol{a}$$

$$\text{subject to } l_i \leq a_i \leq u_i, \text{ for all } 1 \leq i \leq n, \text{ and } \mathbf{1}^\mathsf{T} \boldsymbol{a} = 0,$$

where $\boldsymbol{l}$ and $\boldsymbol{u}$ are vectors with terms $l_i = \min\{0, c_i y_i\}$ and $u_i = \max\{0, c_i y_i\}$. We note that similar QPs arise from kernel quantile regression (Takeuchi, Le, Sears and Smola, 2006) and from support vector regression (Drucker, Burges, Kaufman, Smola and Vapnik, 1997). We denote the *canonical dual criterion* by

$$R(\boldsymbol{a}) = \tfrac{1}{2} \boldsymbol{a}^\mathsf{T} \boldsymbol{K} \boldsymbol{a} - \boldsymbol{y}^\mathsf{T} \boldsymbol{a}.$$

Similarly, the gradient, $g_i(\boldsymbol{a})$, is given by

$$g_i(\boldsymbol{a}) = \frac{\partial R(\boldsymbol{a})}{\partial a_i} = (\boldsymbol{K}\boldsymbol{a})_i - y_i, \quad \text{for all } 1 \leq i \leq n.$$

Via the Karush-Kuhn-Tucker (KKT) conditions it is easily shown (e.g., Schölkopf and Smola, 2002; Karush, 1939; Kuhn and Tucker, 1951) that for a solution to the canonical

1	$\mathbf{1}^{\mathsf{T}}\widehat{\boldsymbol{a}} = 0,$	
2	$l_i \leq \widehat{a}_i \leq u_i,$	for all $1 \leq i \leq n,$
3	$g_i(\widehat{\boldsymbol{a}}) + \widehat{b} = 0,$	for all $1 \leq i \leq n$ such that $l_i < \widehat{a}_i < u_i,$
4	$g_i(\widehat{\boldsymbol{a}}) + \widehat{b} \geq 0,$	for all $1 \leq i \leq n$ such that $\widehat{a}_i = l_i,$
5	$g_i(\widehat{\boldsymbol{a}}) + \widehat{b} \leq 0,$	for all $1 \leq i \leq n$ such that $\widehat{a}_i = u_i.$

Table 6.1. *Optimality conditions for the support vector machine.*

dual QP, $\widehat{\boldsymbol{a}}$, there exists $\widehat{b} \in \mathbb{R}$ such that the optimality conditions of Table 6.1 hold. The decision function is then $f(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{a}_i k(\boldsymbol{x}, \boldsymbol{x}_i) + \widehat{b}$, and the classifier takes the form $\operatorname{sign}\{f(\boldsymbol{x})\}$.

6.3 Active Set SVM

Our algorithm uses three distinct phases: an initialisation phase, a decomposition phase and a conjugate gradient phase. The phases may be used multiple times before convergence is met. The initialisation phase is a novel approach that allows us to move around the corners of the box constraints, in a manner similar to the Simplex method for linear programming (Dantzig, 1963). The initialisation phase varies only one variable at a time, a_i, for some $1 \leq i \leq n$. In doing so, we temporarily relax the constraint $\mathbf{1}^{\mathsf{T}}\boldsymbol{a} = 0$. As the initialisation phase moves around the corners of the box constraints, only in extreme cases be able to completely solve the QP. The initialisation phase requires only $O(n)$ memory and may be used multiple times during the optimisation.

The decomposition phase is based on decomposition methods for SVMs. Decomposition methods involve fixing some variables, and optimising with respect to the remaining variables, the so called "working set". A number of decomposition methods have been developed for the QP problem. An early reference is Platt (1999), whose *Sequential Minimization Optimisation* showed that some large scale SVMs can be solved quickly, and simply. Other solvers have included: Chang and Lin (2009), Fan, Chen and Lin (2005), Hush, Kelly, Scovel and Steinwart (2006), Joachims (1999), Keerthi, Shavade and Bhattacharyya (2001), List and Simon (2005), Simon (2004) and Vapnik (1998). These, and others, have become enormously popular and effective in solving SVMs. It is known, however, that some decomposition algorithms can have poor performance on various SVM tasks. If the problem is particularly large, or particularly difficult, as with a very large C, slow convergence can be experienced. The decomposition phase, like the ini-

Sample size n	Single precision storage space $2n(n+1)$	Double precision storage space $4n(n+1)$
100	20 KB	40 KB
1000	2 MB	4 MB
10000	200 MB	400 MB
100000	20 GB	40 GB
1000000	2 TB	4 TB

Table 6.2. *The relationship between sample size and required storage space for* $\boldsymbol{K}$.

tialisation phase, has small memory requirements. This is important for large scale problems.

In the third phase, the conjugate gradient phase, we attempt to optimise over all variables that are "free", that is, all variables not currently at their lower or upper bounds. The conjugate gradient phase is based on the active set method for QPs, for example, Wolfe (1959), Polyak (1969), Nocedal and Wright (1999) and Scheinberg (2006). The active set methods allow for highly accurate solutions to the QP, and is the traditional method for solving quadratic programs. Active set methods have been applied to SVMs with varying success. Both Wen, Edelman and Gorsich (2003) and Scheinberg (2006) detail active set algorithms that show a similar performance to that of SVM^{light} (Joachims, 1999). However, it appears that active set comparisons against LIBSVM (Chang and Lin, 2009; Fan *et al.*, 2005), have not been as favourable. Vogt and Kecman (2005) and Vishwanathan, Smola and Murty (2005) produced active set algorithms; they found LIBSVM to be many times faster. We argue that in combining characteristics from both decomposition and active set methods, a fast, and accurate algorithm can be created. Following Burges (1998), and Wen, Edelman and Gorsich (2003), optimisation is carried out through the use of projected conjugate gradients. The algorithm cycles though the initialisation, decomposition, and conjugate gradient phases. It is nowadays standard, amongst efficient algorithms, to make use of techniques such as kernel caching, selective pricing and sparsity handling. These too are employed to improve the speed of the algorithm. We call our algorithm AS-SVM, standing for active set support vector machine.

6.3.1 Initialisation Phase

Within Table 6.1, the optimality conditions (3)-(5) differentiate three groups of variables. Denote by $\mathcal{L}$ the indices of the variables at their lower bound, that is, for all $i \in \mathcal{L}, a_i = l_i$. Similarly let $\mathcal{U}$ denote the the indices of variables at their upper bound: for all $i \in \mathcal{U}, a_i = u_i$. The term *active set*, indexed by $\mathcal{A}$, refers to variables whose inequality bound is active, that is, the variables with corresponding indices in $\mathcal{L} \cup \mathcal{U}$. The complement of the active set is the set of *free* variables, with indices denoted by $\mathcal{S}$. We have $l_i < a_i < u_i$ for all $i \in \mathcal{S}$.

On typical classification problems, the majority of variables, at the solution to the QP, are in the active set. In turn, most of these variables are at 0. This can make the optimisation easier. Typically we would want to start the algorithm with $a_i = 0$ for all $1 \leq i \leq n$. There are other settings where we would want to use approximate solutions as our starting point, such as with incremental learning (Cauwenberghs and Poggio, 2001). Another important setting is when performing either k-fold cross-validation where we may not initially have $\mathbf{1}^\top \boldsymbol{a} = 0$.

During the initialisation phase, similar to Dantzig's Simplex algorithm (Dantzig, 1963), we directly push variables to their upper or lower bounds. Only a single variable is changed, to do this, we temporarily relax the condition $\mathbf{1}^\top \boldsymbol{a} = 0$. If $\mathbf{1}^\top \boldsymbol{a} \leq 0$, we set some a_j to its upper bound, and if $\mathbf{1}^\top \boldsymbol{a} > 0$, we set some a_j to its lower bound. In this manner, we maintain $\mathbf{1}^\top \boldsymbol{a} \approx 0$. It remains to decide which variable should be chosen. It is worth noting that there is an invariance in the canonical dual QP. If we were to replace the matrix $\boldsymbol{K}$ with matrix $\boldsymbol{Q}$, where $Q_{ij} = K_{ij} - \frac{1}{2}\left(K_{ii} + K_{jj}\right)$, Schölkopf (2001) shows this does not change the solution to the SVM. Taking this invariance into account, if $\mathbf{1}^\top \boldsymbol{a} \leq 0$, we set $a_j := u_j$ where

$$j = \operatorname*{argmin}_{i \notin \mathcal{U}} \left\{ g_i(\boldsymbol{a}) - \tfrac{1}{2} K_{ii} \mathbf{1}^\top \boldsymbol{a} \right\}.$$

Similarly, if $\mathbf{1}^\top \boldsymbol{a} > 0$, we set $a_j := l_j$, where

$$j = \operatorname*{argmax}_{i \notin \mathcal{L}} \left\{ g_i(\boldsymbol{a}) - \tfrac{1}{2} K_{ii} 1^\top \boldsymbol{a} \right\}.$$

By taking advantage of the invariance, we ensure that the canonical dual criterion is being maximally decreased, modulo $\mathbf{1}^\top \boldsymbol{a} = 0$. The gradient is updated with $O(n)$ time and memory requirements. Once a column of $\boldsymbol{K}$ is generated, it is stored in a kernel cache, as it may need to be reused.

Convergence is obtained when the same variable is selected twice in a row. To prevent cycling, in the case of ties, the variable with the lower index is chosen. On con-

Algorithm 6.1 Pseudo-code for the initialisation phase.

Require: $K(\cdot,\cdot)$, g, a, y

$\kappa \leftarrow \mathbf{1}^{\mathrm{T}}\boldsymbol{a}$
$j\_old \leftarrow -1, j \leftarrow -2$
while $j\_old \neq j$ **do**
  $j\_old \leftarrow j$
  if $\kappa \leq 0$ **then**
    $j \leftarrow \operatorname{argmin}_{i \notin \mathcal{U}} \left\{ g_i + \frac{1}{2}\kappa K_{ii} \right\}$
    $\Delta \leftarrow u_j - a_j$, $\kappa \leftarrow \kappa + \Delta$, and $a_j \leftarrow u_j$
  else
    $j \leftarrow \operatorname{argmax}_{i \notin \mathcal{L}} \left\{ g_i + \frac{1}{2}\kappa K_{ii} \right\}$
    $\Delta \leftarrow l_j - a_j$, $\kappa \leftarrow \kappa + \Delta$, and $a_j \leftarrow l_j$
  end if
  $g_i \leftarrow g_i + \delta K_{ij}$, for all $1 \leq i \leq n$
end while
$g_i \leftarrow g_i + \kappa K_{ij}$, for all $1 \leq i \leq n$
$a_j \leftarrow a_j + \kappa$

vergence, a small adjustment may be needed to ensure $\mathbf{1}^{\mathrm{T}}\boldsymbol{a} = 0$. Once this adjustment has been made, the canonical dual criterion is no larger than at the start of the initialisation phase. The initialisation phase will, in general, not solve the QP. In all but extreme cases, the solution to the QP requires variables to be free, the set $\mathcal{S}$ not being empty. Used multiple times during optimisation, the initialisation phase may converge early, with little or no reduction in the canonical dual criterion. The early convergence of the initialisation phase can be expected on problems with large values of C. Pseudo-code for the initialisation phase is given in Algorithm 6.1.

6.3.2 Decomposition Phase

The decomposition phase is based on the popular decomposition methods for SVMs. The Hessian of the QP, $\boldsymbol{K}$, may be too large to store in memory. With decomposition methods, a sequence of smaller QPs are solved. Algorithms such as Vapnik (1982), Osuna *et al.* (1997b), Joachims (1999) and Platt (1999) make use of this idea. Each iteration of the decomposition phase involves fixing some variables, and optimising with respect to the remaining variables, the so called "working set". Partitioning the set $\{1,\ldots,n\}$

into a working set, $\mathcal{W}$, and a non-working set, $\overline{\mathcal{W}}$, we minimise the canonical dual while holding the variables in the non-working set fixed. That is we solve

$$\min_{a_i a_i, i \in \mathcal{W}} \left(\tfrac{1}{2}\boldsymbol{a}^\mathsf{T}\boldsymbol{K}\boldsymbol{a} - \boldsymbol{y}^\mathsf{T}\boldsymbol{a}\right)$$

$$\text{subject to } l_i \leq a_i \leq u_i, \text{ for all } 1 \leq i \leq n, \text{ and } \mathbf{1}^\mathsf{T}\boldsymbol{a} = 0,$$

or equivalently,

$$\begin{aligned} &\min_{a_i, i \in \mathcal{W}} \left\{\tfrac{1}{2}\boldsymbol{a}_{\mathcal{W}}^\mathsf{T}\boldsymbol{K}_{\mathcal{W}\mathcal{W}}\boldsymbol{a}_{\mathcal{W}} - \boldsymbol{a}_{\mathcal{W}}^\mathsf{T}(\boldsymbol{y}_{\mathcal{W}} - \boldsymbol{K}_{\mathcal{W}\overline{\mathcal{W}}}\boldsymbol{a}_{\overline{\mathcal{W}}})\right\} \\ &\text{subject to } l_i \leq a_i \leq u_i, \text{ for all } i \in \mathcal{W}, \text{ and } \mathbf{1}^\mathsf{T}\boldsymbol{a}_{\mathcal{W}} = -\mathbf{1}^\mathsf{T}\boldsymbol{a}_{\overline{\mathcal{W}}}. \end{aligned} \tag{6.1}$$

With AS-SVM, the size of the decomposition is fixed at two, the minimum number required in order to strictly maintain the condition $\mathbf{1}^\mathsf{T}\boldsymbol{a} = 0$. The corresponding minimisation (6.1) can be achieved analytically, as was shown in Platt (1999). It is also known that the analytic solution is stable even when the matrix $\boldsymbol{K}_{\mathcal{W}\mathcal{W}}$ is not strictly positive definite (Lin, 2002). A number of decomposition methods have been developed for the QP problem. Since Platt (1999) these have mainly focused on working sets of size two. Even when the size of the decomposition is fixed at two, there are many alternatives for choosing the working set. A popular and particularly efficient decomposition algorithm is that of LIBSVM version 2.8, detailed in Chang and Lin (2009) and Fan *et al.* (2005). In the decomposition phase we use similar procedures for selecting the working set and convergence criteria. In particular, we follow Fan *et al.* (2005) in using second order information. Using second order information, (i, j) is selected where $i \in \operatorname{argmin}_{t \notin \mathcal{U}} g_t(\boldsymbol{a})$, and

$$j \in \operatorname*{argmax}_{t \notin \mathcal{L}} \left\{ \frac{\{g_t(\boldsymbol{a}) - g_i(\boldsymbol{a})\}^2}{\max\left(K_{ii} + K_{tt} - 2K_{it}, 10^{-12}\right)} \middle| g_t(\boldsymbol{a}) > g_i(\boldsymbol{a}) \right\}.$$

Once (i, j) is selected, the analytic solution given by Platt (1999) is used. A commonly used criteria for convergence is ε convergence,

$$\max_{i \notin \mathcal{L}} g_i(\boldsymbol{a}) - \min_{j \notin \mathcal{U}} g_j(\boldsymbol{a}) < \varepsilon,$$

where $\varepsilon > 0$ is some tolerance, typically 10^{-3}. On its own, the decomposition phase may take many iterations to achieve, and typically $O(1/\varepsilon)$. However, each iteration, at $O(n)$, is computationally inexpensive. We perform up to 1000 iterations during each decomposition phase. Pseudo-code for the Decomposition Phase is given in Algorithm 6.2.

Part of the Gram matrix is stored in a kernel cache (Joachims, 1999). On large data sets the cache may be quickly filled, various heuristics are used to reallocate memory.

Algorithm 6.2 Pseudo-code for the decomposition phase.

Require: $\boldsymbol{K}(\cdot,\cdot)$, g, a, l, u, ϵ,

$i \leftarrow \operatorname{argmax}_{t \notin \mathcal{L}} g_t$

$j \leftarrow \operatorname{argmin}_{t \notin \mathcal{U}} g_t$

$m \leftarrow 0$

while $g_i - g_j > \epsilon$ and $m < 1000$ **do**

  $m \leftarrow m + 1$

  $j \leftarrow \operatorname{argmin}_t \left\{ \dfrac{-(g_i - g_t)^2}{\max\left(K_{ii} + K_{tt} - 2K_{it}, 10^{-12}\right)} \middle| t \notin \mathcal{U}, g_i > g_t \right\}$

  if $g_i - g_j > (K_{ii} + K_{jj} - 2K_{ij}) \max(a_i - l_i, u_j - a_j)$ **then**

    $\Delta \leftarrow \max(a_i - l_i, u_j - a_j)$

  else

    $\Delta \leftarrow (g_i - g_j)/(K_{ii} + K_{jj} - 2K_{ij})$

  end if

  $a_i \leftarrow a_i - \Delta$, $a_j \leftarrow a_j + \Delta$

  $g_i \leftarrow g_i - \Delta K_{ij} + \Delta K_{ik}$ for all $1 \leq i \leq n$

end while

The columns of $\boldsymbol{K}$ corresponding to $\mathcal{S}$ are highly likely to be reused and are preferenced in the cache over columns corresponding to $\mathcal{A}$. Some details of the caching strategy are discussed in Section 6.3.4. As AS-SVM was coded in `Fortran` 77, dynamic memory allocation was not used, though it was still straightforward to dynamically allocate space within a large vector. When the cardinality of $\mathcal{S}$ is large, a smaller proportion of the columns corresponding to $\mathcal{S}$ may be stored in the cache, having a negative impact on the performance of the algorithm.

A difficulty identified by DeCoste and Schölkopf (2001), called the *intermediate support vector bulge*, is that for some problems many variables are required to pass from $\mathcal{L}$ to $\mathcal{U}$ or from $\mathcal{U}$ to $\mathcal{L}$. This creates a temporary bulge in $\mathcal{S}$, to the detriment of the time to convergence. Although the initialisation phase may be of some benefit here, the intermediate support vector bulge may still be seen on problems where the initialisation phase converged early. An example of an intermediate support vector bulge can be seen in Figure 6.1. During the training of the support vector machine, the intermediate cardinality of $\mathcal{S}$ can be significantly higher than the final cardinality at convergence. This causes the kernel caching to become less efficient, as less of $K_{\mathcal{SS}}$ may be stored in the cache.

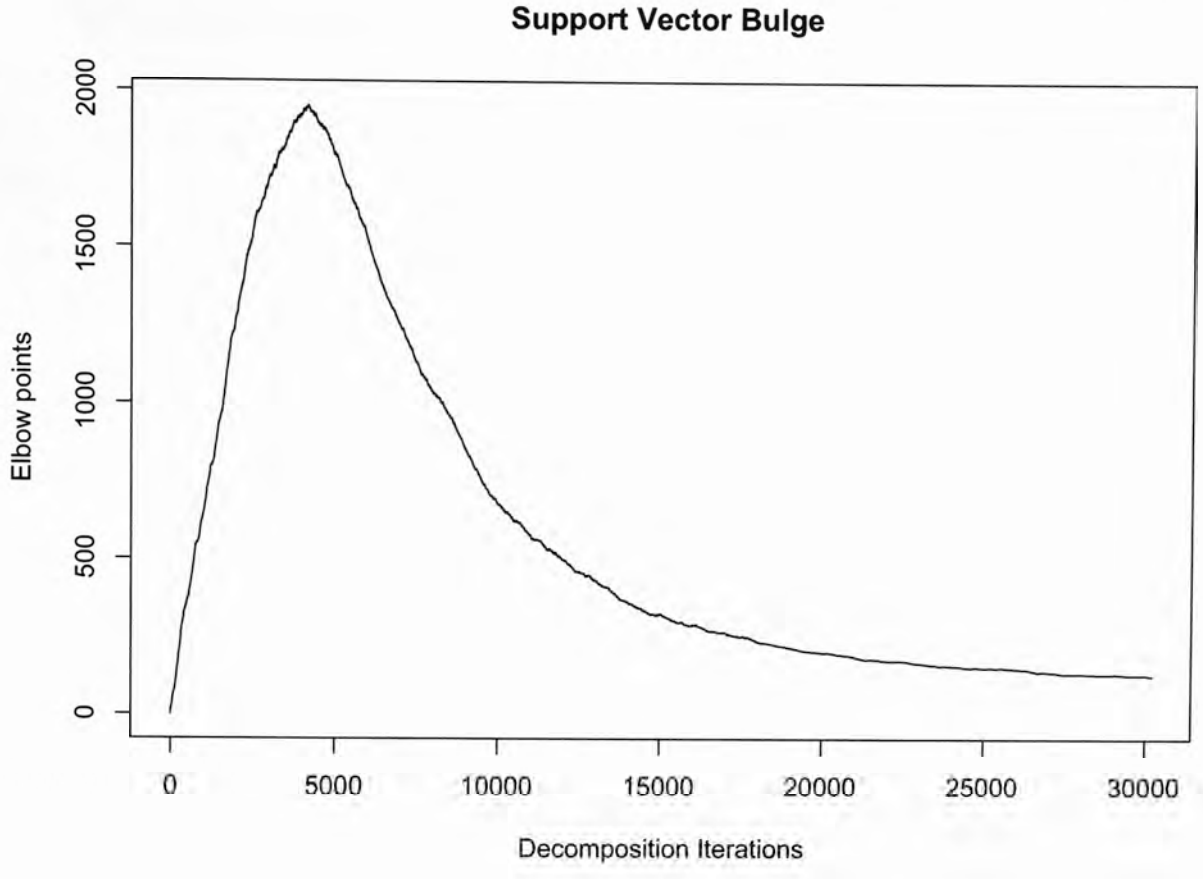


Figure 6.1. *Intermediate support vector bulge. Decomposion methods by themselves are known to experience an intermediate support vector bulge phenomenon. On the problem above, we find on convergence that there are only a couple of hundred free variables, or elbow points. Yet, during the course of optimisation, there may be many more free variables. The effect of the bulge is a diminished run-time performance.*

6.3.3 Conjugate Gradient Phase

If we had an estimate of the sets $\mathcal{L}$ and $\mathcal{U}$, then we would trivially know $a_{\mathcal{L}}$ and $a_{\mathcal{U}}$, and approximate values $a_{\mathcal{S}}$ could then be found by solving

$$\begin{aligned} &\min_{a} \left\{ \tfrac{1}{2} a_{\mathcal{S}}^{\mathsf{T}} K_{\mathcal{S}\mathcal{S}} a_{\mathcal{S}} - a_{\mathcal{S}}^{\mathsf{T}} \left(y_{\mathcal{S}} - K_{\mathcal{S}\mathcal{A}} a_{\mathcal{A}} \right) \right\} \\ &\text{subject to } l_i \leq a_i \leq u_i, \text{ for all } i \in \mathcal{S}, \text{ and } \mathbf{1}^{\mathsf{T}} a_{\mathcal{S}} = -\mathbf{1}^{\mathsf{T}} a_{\mathcal{A}}. \end{aligned} \tag{6.2}$$

Unlike decomposition methods, we do not solve (6.2) directly. As the cardinality of $\mathcal{S}$ may become large, directly solving (6.2) with say, an external QP solver, may not be viable due to either time or memory requirements. There are methods to gain an approximate the solution to (6.2). Traditional active set methods approximate the solution via the following steps:

i) Initialise $\boldsymbol{a}_{\mathcal{S}}$ such that $\mathbf{1}^{\mathsf{T}}\boldsymbol{a}_{\mathcal{S}} = -\mathbf{1}^{\mathsf{T}}\boldsymbol{a}_{\mathcal{A}}$ and $l_i \leq a_i \leq u_i$ for all $i \in \mathcal{S}$.

ii) Solve

$$\begin{aligned} &\min_{\boldsymbol{a}} \left\{ \tfrac{1}{2}\boldsymbol{a}_{\mathcal{S}}^{\mathsf{T}}\boldsymbol{K}_{\mathcal{S}\mathcal{S}}\boldsymbol{a}_{\mathcal{S}} - \boldsymbol{a}_{\mathcal{S}}^{\mathsf{T}}\left(\boldsymbol{y}_{\mathcal{S}} - \boldsymbol{K}_{\mathcal{S}\mathcal{A}}\boldsymbol{a}_{\mathcal{A}}\right) \right\} \\ &\text{subject to } \mathbf{1}^{\mathsf{T}}\boldsymbol{a}_{\mathcal{S}} = -\mathbf{1}^{\mathsf{T}}\boldsymbol{a}_{\mathcal{A}}. \end{aligned} \tag{6.3}$$

iii) Move toward the solution until we hit a box constraint, that is for some $i \in \mathcal{S}$, $a_i = l_i$ or $a_i = u_i$, or until a solution is reached.

iv) If we hit a box constraint, remove the observation from $\mathcal{S}$, update the vector $\boldsymbol{a}$ and set $\mathcal{A}$, and return to step *ii)*.

v) If $\boldsymbol{a}_{\mathcal{S}}$ satisfies the box constraints, exit.

These steps may provide a new estimate as to the partition of the sets $\mathcal{L}$, $\mathcal{U}$ and $\mathcal{S}$. Commonly, active set method QP solvers, such as Wolfe (1959) and Scheinberg (2006), select a single violating variable to add to the free variables. This can be slow if many variables are to be added. Scheinberg (2006) notes that adding multiple violating variables will not, in general, improve performance. The clustering of observations that exists on many data sets means that the violating variables may be close to each other. As such, adding the most violating variables to the set of free variables may have the same effect as simply adding the most violating variable. We overcome this by using the decomposition phase to choose new free variables. Using the decomposition phase has the added benefit of reducing the zigzagging that has an adverse effect on both active set methods, as well as on some decomposition methods that use large working sets.

The computationally expensive part of the active set phase is step (ii) and involves a system of linear equations. The KKT conditions may be represented as the linear equations:

$$\begin{pmatrix} \boldsymbol{K}_{\mathcal{S}\mathcal{S}} & \mathbf{1} \\ \mathbf{1}^{\mathsf{T}} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{a}_{\mathcal{S}} \\ b \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_{\mathcal{S}} - \boldsymbol{K}_{\mathcal{S}\mathcal{A}}\boldsymbol{a}_{\mathcal{A}} \\ -\boldsymbol{a}_{\mathcal{A}}^{\mathsf{T}}\mathbf{1} \end{pmatrix}. \tag{6.4}$$

There are several standard ways to solve such a system. Either Cholesky factorisation or a conjugate gradient method can accurately solve the system of equations, although on their own would take some $O(\text{card}\,(\mathcal{S})^3)$ operations, where $\text{card}(\mathcal{S})$ denotes the cardinality of $\mathcal{S}$. Often we are only adding or removing a small number of variables. Cholesky updates and downdates reduces this cost to $O(\text{card}(\mathcal{S})^2)$. Scheinberg (2006) uses Cholesky factorisation with updates and downdates to the Cholesky factor. A drawback with using Cholesky factorisations is that of memory requirements. A Cholesky factorisation requires some $O(\text{card}(\mathcal{S})^2)$ memory, for large, high dimensional data sets this may not be available.

Conjugate gradient methods, as an iterative approximation, can benefit from using current estimates of $\boldsymbol{a}$. Polyak (1969) uses the conjugate gradient method for general QPs. Both Burges (1998) and Wen *et al.* (2003) have applied the conjugate gradient method to SVMs. We do not need to precisely solve (6.3). Several conjugate gradient descent steps are sufficient for an approximate solution to the system of equations represented by (6.4). Due to the condition $\mathbf{1}^\mathsf{T}\boldsymbol{a} = 0$, the conjugate gradient directions are projected onto $\mathbf{1}^\mathsf{T}\boldsymbol{a} = 0$. Since the projection matrix $\boldsymbol{P} = \boldsymbol{I} - \mathbf{1}\mathbf{1}^\mathsf{T}/\operatorname{card}(\mathcal{S})$, the projection of a vector is cheap. The box constraint conditions that $l_i \leq a_i \leq u_i$ for all $1 \leq i \leq n$, are maintained at all times. The descent direction is clipped to the box constraints, this is particularly important as an infinite descent direction may occur. Pseudo-code is given in Algorithm 6.3.

A standard projected conjugate gradient algorithm offers the solution of step *ii)* in $O(\operatorname{card}(\mathcal{S})^3)$ flops (Nocedal and Wright, 1999). The conjugate gradient method offers some choices as to memory use. Inherently, the conjugate gradient method requires a minimum of only $O(\operatorname{card}(\mathcal{S}))$ memory, no more than the decomposition method. However, at a cost of $O(\operatorname{card}(\mathcal{S})^2)$ memory, $\boldsymbol{K}_{\mathcal{S}\mathcal{S}}$ may be stored for an improvement in speed. We thus have a trade off between speed and memory requirements, and a strategy to be chosen. Also to be considered is that the conjugate gradient method may be skipped altogether. The simple strategy we take is to skip the conjugate gradient phase if these memory requirements become large. This has the added advantage of ensuring simpler code than otherwise.

There may also be other reasons to avoid the conjugate gradient phase. Although the conjugate gradient phase offers superior convergence properties to the decomposition phase for ε convergence as $\varepsilon \to 0^+$, the benefits for say, $\varepsilon = 10^{-3}$ are more marginal. As shown in Figure 6.2, the conjugate gradient phase can greatly reduce $\operatorname{card}(\mathcal{S})$, the intermediate support vector bulge has been greatly lessened. Thus there are two areas where the conjugate gradient phase is particularly helpful: early on in the optimisation, to lessen the intermediate support vector bulge, and close to optimisation, when the active and free variables are close to being identified.

At the start of the conjugate gradient phase, the matrix $\boldsymbol{K}_{\mathcal{S}\mathcal{S}}$ is stored in packed storage within the kernel cache. This kernel cache is shared dynamically with other kernel evaluations. Pseudo-code for AS-SVM is given in Algorithm 6.4.

Algorithm 6.3 Pseudo-code for the conjugate gradient phase.

Require: $K_{\mathcal{SS}}$, g, a, l, u, $\mathcal{S}$, ϵ

$r \leftarrow g_{\mathcal{S}}$
$m \leftarrow 1$
while $\max\{r\} - \min\{r\} > \epsilon$ **do**
  $r \leftarrow Pr$
  repeat
    if $m = 1$ **then**
      $p \leftarrow r$
    else
      $p \leftarrow r + \frac{r^\top r}{w^\top w} p$
    end if
    $\Delta \leftarrow \max v$, subject to $l_{\mathcal{S}(i)} \leq a_{\mathcal{S}(i)} + v p_i \leq u_{\mathcal{S}(i)}$ for all $1 \leq i \leq \mathrm{card}(\mathcal{S})$
    $q \leftarrow PK_{\mathcal{SS}} p$
    if $r^\top r < \Delta p^\top q$ **then**
      $\kappa \leftarrow \frac{r^\top r}{p^\top q}$
    else
      $\kappa \leftarrow \Delta$
    end if
    $w \leftarrow r$, $r \leftarrow r - \kappa q$, and $m \leftarrow m + 1$
    $a_{\mathcal{S}(i)} \leftarrow a_{\mathcal{S}(i)} + \kappa p_i$ for all $1 \leq i \leq \mathrm{card}(\mathcal{S})$
  until $\kappa \geq \Delta$
  $m \leftarrow 1$
end while

6.3.4 Sparsity, Caching and Selective Pricing

Much of the computational burden is on kernel evaluations. For sparse data sets, where many of the features are 0, the kernel evaluations are performed to take advantage of the sparsity. It is often computationally much cheaper to re-use kernel evaluations than to recalculate them.

A kernel caching (Joachims, 1999), allows partial storage of the matrix K. The kernel cache consists of two components; some columns of K, as well as, possibly, the matrix $K_{\mathcal{SS}}$, stored in packed storage. On large data sets the cache may be quickly filled, various heuristics are used to reallocate memory. The columns of K corresponding to

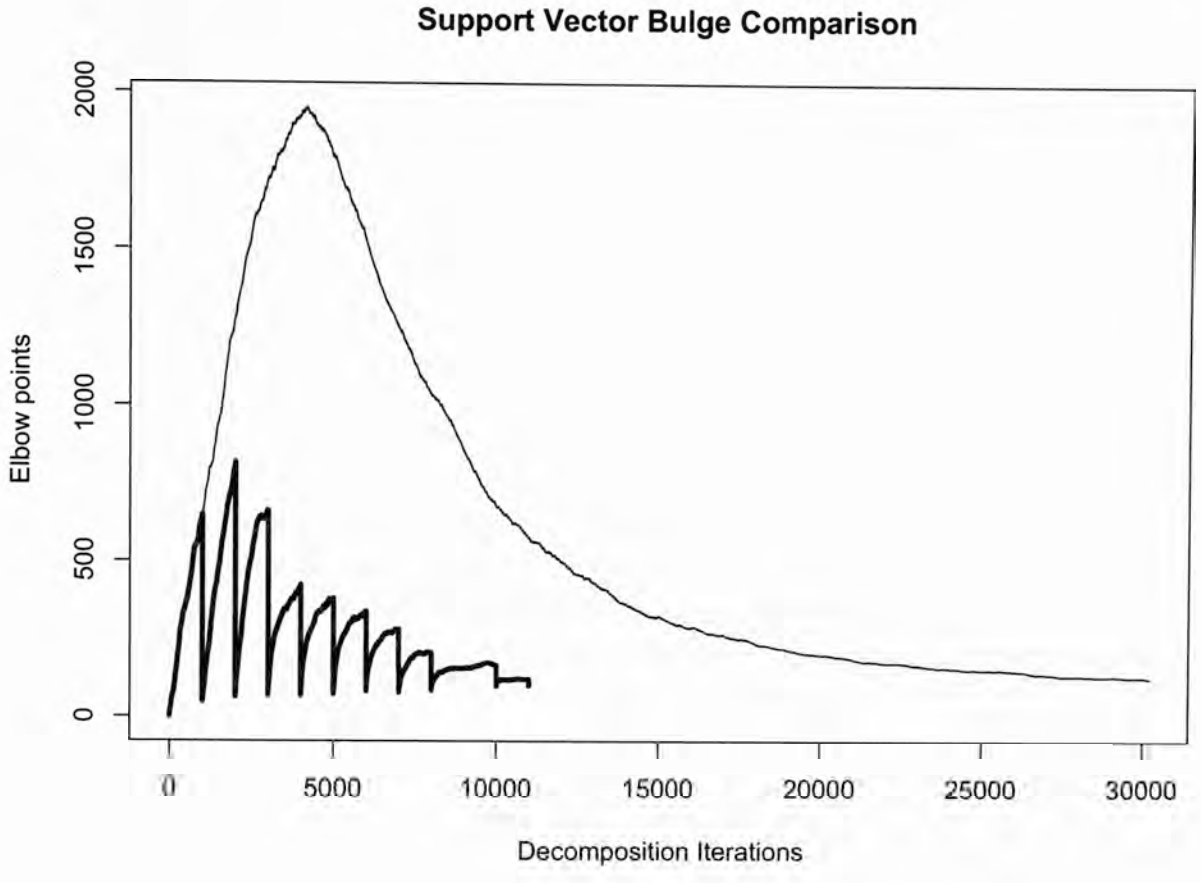


Figure 6.2. *Support vector bulge comparison. The darker, lower curve shows the number of free variables, or elbow points, at each iteration of the decomposition phase. At around every 1000 iterations, the conjugate gradient phase greatly reduces the number of free variables. This has significantly reduced the intermediate support vector bulge phenomenon. The conjugate gradient phase occurs some ten times; eight of those before the first unshrinking.*

$\mathcal{S}$ are highly likely to be reused, either in to the decomposition phase or the conjugate gradient phase. Accordingly, columns corresponding to $\mathcal{S}$ are preferenced within the cache over columns corresponding to $\mathcal{A}$.

Taking into consideration such a heuristic, we may still be required to choose to remove a column from $\mathcal{S}$ or choose a column to remove from $\mathcal{A}$. To this end, we adapted the minimal violating rule of Li *et al.* (2002). Although, as AS-SVM was coded in `Fortran` 77, dynamic memory allocation was not used, it was still straightforward to dynamically allocate space within a large vector. When the cardinality of $\mathcal{S}$ is large, a smaller proportion of the columns corresponding to $\mathcal{S}$ may be stored in the cache, having a negative impact on the performance of the algorithm. A trade-off exists between the memory available, and the speed of the algorithm.

Algorithm 6.4 Pseudo-code for AS-SVM.

Require: $K(\cdot,\cdot)$, x, y, l, u, ε

$a_i \leftarrow 0$ for all $1 \leq i \leq n$
$g \leftarrow y$
$\epsilon \leftarrow 10\varepsilon$
repeat
  Initialisation phase
  Decomposition phase
  if ϵ convergence **then**
    Unshrinking
    $\epsilon \leftarrow \varepsilon$
  else
    if not ε convergence **then**
      Shrinking
      Conjugate gradient phase
    end if
  end if
until ε convergence

In the optimisation literature, calculating $g_i(\boldsymbol{a})$ is referred to as "pricing". Partial pricing involves only calculating $g_i(\boldsymbol{a})$ for some i. The idea is to limit the computational burden, while perhaps only bringing about a slight increase in the number of iterations. A commonly used partial pricing strategy in the SVM literature is the *shrinking* strategy, whereby the number of variables are effectively 'shrunk' by ignoring those in the active set that are some distance from being violating variables. For choosing which variables to be shrunk we follow LIBSVM (Chang and Lin, 2009) Shrinking is also of benefit to freeing memory in the kernel cache. Not only can shrinking remove columns from the kernel cache, but it can also shorten the columns remaining in the cache. After each decomposition phase, shrinking is performed. This matches well with the conjugate gradient phase; memory is freed from the cache just in time. These shrunken variables must be rechecked again close to optimisation. We had also experimented an extra pricing technique, known as sprint (Forrest, 1989). Although widely used in active set algorithms (Bixby *et al.*, 1992; Scheinberg, 2006) we did not find sprint to be of benefit when used in conjuncture with the multiphase approach.

6.4 Computational Results

In this section, we give some computational results, comparing AS-SVM with LIBSVM version 2.81. A memory allowance of some 10^7 kernel evaluations was made available to both methods. Since LIBSVM uses single precision (4 byte) storage to store the kernel evaluations, the memory allowance is 40MB. AS-SVM uses double precision (8 byte) storage. Neither AS-SVM or LIBSVM have special treatment of the linear kernel, such as linear folding (see Platt, 1999). The following three data sets were used for our experiments:

Web: The Web data set, as preprocessed by Platt (1999). We have 300 binary features per observation, with an average of around 12 being non-zero. A subset of 24,692 training observations is used. As the data is sparse, the kernel evaluations take advantage of this sparsity. A Gaussian kernel is used, with a range of parameterisations. The Web data set is a commonly used test data set for SVM training algorithms.

MNIST: Handwritten upper-case letters. Contains handwritten text for zip code recognition. The data set contins 21,000 training observations, with 780 features per observation. These features were scaled to $[0, 1]$. We fitted three different models, MNIST(0) is recognising digit 0 versus all the other digits, and MNIST(9) is recognising digit 9 versus all the other digits. For MNIST(0-4), we recognise digits 0-4 versus digits 5-9. Of the 780 features, an average of 150 are non-zero per observation.

Adult: The Adult dataset. The data set looks at the census household income, in particular whether income is greater than \$50,000. There are 123 binary features per observation, on average around 14 being non-zero. We fit several classifiers using some 16,100 training examples. Similar classifiers were tested by Platt (1999) and Scheinberg (2006). Both Gaussian and linear kernels are used.

With time in seconds, a tolerance of $\epsilon = 10^{-3}$ was used for all C-SVM experiments. Results are shown in Table 6.3. On convergence, the column denoted card($\mathcal{S}$) gives the number of free variables on convergence. The number of bounded support vectors is given by 'bSV'. Both LIBSVM and AS-SVM were called through their respective `R` wrappers, times exclude the time taken to read in the data. The times are similar or better than those shown by LIBSVM, requiring less than half the time on some problems. Some of the largest proportions in time savings are on the Adult data sets, where there

Problem	$1/\gamma$	C	card($\mathcal{S}$)	bSV	LIBSVM	AS-SVM
Web	100	100	980	453	72	44
Web	40	10	1037	568	53	34
Web	40	100	1214	313	62	51
Web	100	10	679	835	47	24
MNIST(0)	780	100	576	0	83	80
MNIST(9)	780	100	1311	36	276	240
MNIST(0-4)	780	100	2926	95	866	610
Adult	100	1	97	5996	93	40
Adult	100	100	871	4823	185	116
Adult	200	1	168	5785	108	43
Adult	200	100	483	5219	133	68
Adult	50	10	615	5143	99	49
Adult	linear	1	211	375	70	43

Table 6.3. *Time comparison between LIBSVM and AS-SVM, shown in seconds. Performance of AS-SVM is similar or better than that of LIBSVM, with time savings of up to 60%. The time savings are most pronounced when there is a large number of bounded support vectors, bSV.*

are a large number of bounded support vectors. On the Adult data set with $\gamma = 1/100$ and $C = 1$ for example, over 90% of the time spent by AS-SVM is during the initialisation procedure. On the Adult data set with $\gamma = 1/100$ and $C = 100$, very little time is spent during the initialisation procedure, the time savings over LIBSVM can be attributed to the conjugate gradient phase, which ensures that the cardinality of $\mathcal{S}$ is kept low during the earlier stages of optimisation. On examples with a small number of bounded support vectors, such as MNIST(0), the performance of AS-SVM is similar to that of LIBSVM.

6.5 Discussion

A novel algorithm has been created by adopting a hybrid strategy. This allows for fast and accurate training of SVMs. The algorithm has a similar or better performance than LIBSVM on a range of standard SVM problems. The initialisation phase is a novel and particularly simple to implement. By using conjugate gradient methods, we avoid performing Cholesky factorisation. This also allows us a high accuracy, speed and ap-

plicability to large data sets.

We have avoided a rigorous convergence analysis; the convergence of the decomposition phase on its own is well known. For optimising an SVM, decomposition methods have been the method of choice in the machine learning literature. There are SVM settings where active set algorithms have held sway. Cauwenberghs and Poggio (2001) and Hastie *et al.* (2004) have provided active set algorithms to incremental learning, decremental learning and parameter tuning. In future, we hope to analyse the use of AS-SVM under similar settings.

Chapter Seven

On Model Validation and Selection

7.1 Introduction

This chapter is concerned with testing and fitting statistical models. Can we validate a given decision function? It is often assumed that we can fit a parametric model with mean zero errors. Can we test the appropriateness of this assumption? This chapter deals with these problems in novel manner - using kernel methods, and in doing so generalises and expands upon much of the literature on the topic.

A goal in supervised learning tasks is to use the data to produce a decision function that, for a given input x, can predict the response y. Typical in regression tasks is the least squares loss function $\mathcal{L}(a,b) = (a-b)^2$. The use of least squares loss corresponds to the modelling task of finding the conditional mean

$$f(x) = \mathsf{E}_{y|x}(y \mid x)\,.$$

Often the conditional mean is modelled under the assumption of homoscedasticity, that is, fixed variance errors, or indeed errors from some known distribution.

Parametric models assume that the functional form of f is known apart from a finite number of parameters. The unknown parameters are estimated, with statistical inference based on the resulting estimate. Incorrect parametric assumptions may lead to misleading inferences (Breiman, 2001). As such, parametric assumptions should be rigorously tested. Nonparametric analysis does not assume that the functional form of f has to have a finite number of parameters. In finding the functional form of f, we will first consider the related problem of model selection.

There is a huge literature associated with model selection. This literature includes a large body of research on testing a parametric null against a parametric or nonparametric alternative. Linear model selection criteria include AIC (Akaike, 1974), BIC (Schwarz, 1978), C_p (Mallows, 1973), FIC (Wei, 1992), GCV (Craven and Wahba, 1979), and PRESS (Allen, 1974). Such criteria have been used in the literature for choosing from two or more nested parametric models. The criteria are typically based on either finding

amongst models the lowest generalisation error, or in a similar vein, testing the statistical significance of linear predictor variables. The assumptions made and approximation techniques used will also vary between the various model selection criteria.

Ordinary least squares (OLS) is one of many methods of producing linear fits to data. Often, some form of regularisation is desired, or indeed needed, in order to improve the expected performance of our fitted model. Regularisation methods have been discussed in earlier chapters. Ridge regression (Hoerl and Kennard, 1970a,b), the LASSO (Alliney and Ruzinsky, 1994; Tibshirani, 1996), and least angle regression (Efron *et al.*, 2004) can be seen as regularised versions of OLS. Authors such as Allen (1974) have noted that the model selection for OLS can be seen as a special case of parameter selection for generalised ridge regression. The question remains as to whether there can be, with only our parametric or linear model, a fit sufficiently accurate enough to model the true mean response. The mean may not conform to a linear fit, even if one uses all the predictor variables.

Instead of testing a parametric null against a parametric alternative, we may be interested in testing a parametric null against a nonparametric alternative. We want to test whether a given parametric model will be adequate for the mean response. Such an alternative hypothesis may be of interest when we are considering if our hypothesis space is large enough to encompass the true mean. If the underlying regression model is indeed linear, then it is quite fair to expect that standard linear regression techniques, such as OLS, would provide rapid convergence. It is known, however, that OLS is often not optimal (James and Stein, 1961). It is also well known that an incorrectly specified model can give inaccurate and misleading conclusions (e.g., White, 1980). If the assumption of a parametric or linear mean is indeed correct, then this is a useful assumption to make. A key concern is whether our assumption is indeed correct, as this is then relied upon by further uses of the model.

As such, we are interested in testing the adequacy of a parametric model. Such testing procedures have featured strongly in both statistics and economics literature. Tests include those given by Andrews (1997); Bierens (1982, 1990); Bierens and Ploberger (1997); Chen and Fan (1999); Delgado (1993); Ellison and Ellison (2000); Fan and Li (1996, 1999, 2000); Gozalo (1993); Härdle and Mammen (1993); Hart (1997); Horowitz (2006); Horowitz and Härdle (1994); Horowitz and Spokoiny (2001); Kitamura (2005); Li and Wang (1998); Smith (2007); Stengos and Sun (2001); Tripathi and Kitamura (2003); Stinchcombe and White (1998) and Zheng (1996). There are some important differences

amongst these tests. The convergence rates can differ, many tests in the literature do not have the optimal rate of convergence. The computational cost can also vary between test statistics.

Our approach to testing model adequacy is novel. In particular, the existing literature has not yet explicitly made use of reproducing kernel Hilbert spaces. Using kernel methods, we develop a test that is consistent against alternative models. The test does not make strong assumptions about the alternative model, and has convergence rate that matches the lower bound. The asymptotic distribution of the test statistic is derived under both the null and alternate hypothesis.

Having decided to reject the null hypothesis, the question naturally arises as to what the fit should be under the alternative. Based on the test statistic, a new criterion is given for parameter selection, the "parameter information criterion" (PIC). As a parameter selection technique, the PIC requires the minimal assumption of i.i.d. random variables. A sharper alternative to the PIC is heuristically derived, the "curved information criterion", (CIC). Both mean squared error and mean squared predictive error versions of PIC and CIC are given. Extensive simulations show favourable results when compared with existing methods such as leave-one-out cross-validation and maximum-likelihood.

In Section 7.2 we derive test statistics for the functional form of the conditional mean. Criteria to fit a model under the alternative hypothesis are then given in 7.3. Section 7.4 draws comparisons with existing methods. Extensive simulations are given in Section 7.5 and we close with a discussion in Section 7.6.

7.2 The Mean Zero Hypothesis

A standard paradigm in supervised learning is that we have an unknown probability distribution $P_{x,y}$ over $\mathcal{X} \times \mathbb{R}$. Let Φ be a feature map $\Phi\colon \mathcal{X} \to \mathcal{F}$, with $\mathcal{F}$ a separable feature space with dimension $p \in \{\mathbb{N} \cup \infty\}$. The feature space may be a Banach space, denoted by $\mathcal{B}$, or even an RKHS, $\mathcal{H}_k$. With respect to probability distribution $P_{x,y}$, the *risk* of a measureable $f\colon \mathcal{X} \to \mathbb{R}$ is given by

$$\mathcal{R}_{x,y}(f) \equiv \mathsf{E}_{x,y}\{y - f(x)\}^2 .$$

The *Bayes risk*, denoted $\mathcal{R}^*_{x,y}$, is the minimal risk over all measureable $f\colon \mathcal{X} \to \mathbb{R}$,

$$\mathcal{R}^*_{x,y} \equiv \inf_f \left\{ \mathsf{E}_{x,y}\{y - f(x)\}^2 \mid f\colon \mathcal{X} \to \mathbb{R} \text{ measureable} \right\} .$$

Furthermore, denote by $\mathcal{R}^*_{x,y;\mathcal{F}}$ the minimal risk over $\mathcal{F}$,

$$\mathcal{R}^*_{x,y;\mathcal{F}} \equiv \inf_{f\in\mathcal{F}} \mathsf{E}_{x,y}\{y - f(x)\}^2.$$

The first problem we look at is the following.

Problem 7.1. *Can we test if* $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y;\mathcal{F}}$*?*

Problem 7.1 is of interest when testing whether the conditional mean of the response variable can be explained in part through functions in the feature space. Problem 7.1 may be expressed as the hypothesis test

$$\begin{aligned} H_0&: \mathsf{E}_y(y^2) = \inf_{f\in\mathcal{F}} \mathsf{E}_{x,y}\left[\{y - f(x)\}^2\right] \text{ against} \\ H_1&: \mathsf{E}_y(y^2) > \inf_{f\in\mathcal{F}} \mathsf{E}_{x,y}\left[\{y - f(x)\}^2\right]. \end{aligned} \tag{7.1}$$

Under regularity conditions, addressed in later sections, we have the equivalent

$$\begin{aligned} H_0&: \mathsf{E}_{x,y}\{yf(x)\} = 0 \text{ for all } f \in \mathcal{F} \text{ against} \\ H_1&: \mathsf{E}_{x,y}\{yf(x)\} \neq 0 \text{ for some } f \in \mathcal{F}. \end{aligned}$$

We wish to simultaneously test the p hypotheses, $\mathsf{E}_{x,y}y\phi_j(x) = 0$, for each $1 \leq j \leq p$. Consider $\left(\mathsf{E}_{x,y}y\phi_1(x), \ldots, \mathsf{E}_{x,y}y\phi_p(x)\right)^\mathsf{T}$ as a vector in $\mathbb{R}^p$. In measuring the distance a vector is from the origin, we require some norm $\|\cdot\|_q$ on the vector space $\mathbb{R}^p$. An unbiased estimator of $\mathsf{E}_{x,y}y\phi_j(x)$ is given by its empirical estimate $n^{-1}\sum_{i=1}^n y_i\phi_j(x_i)$. In general, however,

$$n^{-1}\left\|\left(\sum_{i=1}^n y_i\phi_1(x_i), \ldots, \sum_{i=1}^n y_i\phi_p(x_i)\right)^\mathsf{T}\right\|_q$$

does not give an unbiased estimate of $\left\|\left(\mathsf{E}_{x,y}y\phi_1(x), \ldots, \mathsf{E}_{x,y}y\phi_p(x)\right)^\mathsf{T}\right\|_q$. Consider the sup-norm $\|\cdot\|_\infty$. A biased estimate of $\left\|\left(\mathsf{E}_{x,y}y\phi_1(x), \ldots, \mathsf{E}_{x,y}y\phi_p(x)\right)^\mathsf{T}\right\|_\infty$ is obtained by the Kolmogorov-Smirnov type statistic

$$n^{-1}\left\|\left(\sum_{i=1}^n y_i\phi_1(x_i), \ldots, \sum_{i=1}^n y_i\phi_p(x_i)\right)^\mathsf{T}\right\|_\infty.$$

Our focus is primarily on the Euclidean norm $\|\cdot\|_2$, and its generalisation, the RKHS norm. We identify two key benefits in using the RKHS norm:

i) An empirical approximation exists that is both simple and unbiased, with known asymptotic distributions under both the null and alternate hypotheses.

ii) A rich alternate hypothesis can be tested.

We show *i)* in Section 7.2.1 and *ii)* in Section 7.2.2.

7.2.1 Operator Norms on Banach Spaces

The term $\mathsf{E}_{x,y}yf(x)$ may be represented as

$$Af = \mathsf{E}_{x,y}yf(x) \tag{7.2}$$

where $A\colon \mathcal{B} \to \mathbb{R}$ is a linear functional and $f \in \mathcal{B}$. Following Royden (1968), a linear functional A is **bounded** if there is a constant M such that

$$|Af| \leq M \|f\|_{\mathcal{B}} \quad \text{for all } f \in \mathcal{B}.$$

For bounded linear functionals, the least such M is called the **operator norm** (e.g., Royden, 1968; Rudin, 1991). That is, denoting the operator norm as $\|A\|$, we have

$$\|A\| = \sup_{\|f\|_{\mathcal{B}} \leq 1} |Af|.$$

Bounded linear functionals and operators are fundamental concepts in functional analysis.

Definition 7.2. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, and $\mathcal{B}$ a Banach space on $\mathcal{X}$. Then the **operator norm criterion** (ONC) with respect to $\mathcal{B}$ and $P_{x,y}$ is defined by*

$$\text{ONC}(\mathcal{B}, P_{x,y}) = \sup_{\|f\|_{\mathcal{B}} \leq 1} \mathsf{E}_{x,y}yf(x),$$

when such a supremum exists.

It clear that the ONC is the operator norm of A given by (7.2). Hence, when A is bounded, $\sup_{\|f\|_{\mathcal{B}} \leq 1} \mathsf{E}_{x,y}yf(x) = 0$ if and only if $\mathsf{E}_{x,y}yf(x) = 0$ for all $f \in \mathcal{B}$. Further regularity conditions are required to ensure $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y;\mathcal{B}}$. The following theorem gives sufficient regularity conditions.

Theorem 7.3. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$ and $\mathcal{B}$ a Banach space such that*

$$\mathcal{R}_{x,y}(0) < \infty, \quad \sup_{\|f\|_{\mathcal{B}} \leq 1} \mathsf{E}_{x,y}\{yf(x)\} < \infty, \quad \textit{and} \quad \sup_{\|f\|_{\mathcal{B}} \leq 1} \mathsf{E}_{x}\{f(x)^2\} < \infty. \tag{7.3}$$

Then $\text{ONC}(\mathcal{B}, P_{x,y}) = 0$ *if and only if* $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y;\mathcal{B}}$.

Proof of Theorem 7.3 is given in Appendix 7.A.1. Theorem 7.3 contains three conditions in (7.3). The first condition is that the second moment of y is bounded. The second condition is equivalent to the continuity of A. The third condition is that $\mathcal{H}_k$ is continuously embedded in $L_2(P_x)$. In the next theorem, we show that the ONC may be simplified for a special type of Banach space - a reproducing kernel Hilbert space. Let us denote by (x, y) and (x', y') independent random variables from probability distribution $P_{x,y}$.

Theorem 7.4. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, and $\mathcal{H}_k$ be a separable RKHS on $\mathcal{X}$ such that $\mathsf{E}_{x,y,x',y'}\{yk(x,x')y'\} < \infty$. Then*

$$\mathrm{ONC}^2\left(\mathcal{H}_k, P_{x,y}\right) = \mathsf{E}_{x,y,x',y'}\left\{yk(x,x')y'\right\}.$$

A proof of Theorem 7.4 is given in Appendix 7.A.1. Theorem 7.4 ensures that $\mathrm{ONC}\left(\mathcal{H}_k, P_{x,y}\right)$ exists by Definition 7.2 whenever we have $\mathsf{E}_{x,y,x',y'}\{yk(x,x')y'\} < \infty$. We now specialise Theorem 7.3 to RKHSs.

Theorem 7.5. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, and $\mathcal{H}_k$ an RKHS on $\mathcal{X}$, such that*

$$\mathcal{R}_{x,y}(0) < \infty, \quad \mathsf{E}_{x,y,x',y'}\{yk(x,x')y'\} < \infty, \quad \textit{and} \quad \mathsf{E}_x\{k(x,x)\} < \infty.$$

Then $\mathrm{ONC}^2\left(\mathcal{H}_k, P_{x,y}\right) \geq 0$, with equality if and only if $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y;\mathcal{H}_k}$.

Proof of Theorem 7.5 is given in Appendix 7.A.1. Theorem 7.5 shows that under weak regularity conditions, $\mathrm{ONC}^2\left(\mathcal{H}_k, P_{x,y}\right) = 0$ if and only if $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y;\mathcal{H}_k}$. Of interest then is estimation of the ONC, and the hypothesis of whether the ONC is equal to zero. By the identity established in Theorem 7.4, we provide an unbiased empirical estimate of $\mathrm{ONC}^2\left(\mathcal{H}_k, P_{x,y}\right)$. For $1 \leq m \leq n$, denote by $\boldsymbol{i}^n_m$ the set of all collections of m indices chosen without replacement from $\{1,\ldots,n\}$. Furthermore, we denote the cardinality of this set by the falling sequential product, $(n)_m = \frac{n!}{(n-m!)}$. Theorem 7.6 then follows by application of Serfling (1980, Section 5.1.4).

Theorem 7.6. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, and $\mathcal{H}_k$ a separable RKHS with $\mathsf{E}_{x,y,x',y'}\left\{yk(x,x')y'\right\} < \infty$. Then given random i.i.d. samples $(x_i, y_i)_{i=1}^n$ from $P_{x,y}$ with $n \geq 2$, an unbiased empirical estimate of $\mathrm{ONC}^2\left(\mathcal{H}_k, P_{x,y}\right)$ is given by*

$$\mathrm{ONC}^2_u\left(\mathcal{H}_k, (x_i, y_i)_{i=1}^n\right) \equiv (n)_2^{-1} \sum_{i,j \in \boldsymbol{i}^n_2} y_i k(x_i, x_j) y_j. \tag{7.4}$$

Furthermore, for any $\mathcal{H}_k$, over the space of all distributions $P_{x,y}$ on $\mathcal{X} \times \mathbb{R}$, with $\mathsf{E}_{x,y,x',y'}\left\{yk(x,x')y'\right\} < \infty$, the empirical estimate in (7.4) is the minimum variance unbiased estimator of $\mathrm{ONC}^2\left(\mathcal{H}_k, P_{x,y}\right)$.

As the average of $y_i k(x_i, x_j) y_j$ over distinct samples (i, j), the empirical ONC^2_u is of the form of statistic known as a U-statistic (e.g., Hoeffding, 1963). Statistical properties of U-statistics are well-developed in the literature. The following theorem is drawn from those of Serfling (1980, Sections 5.5.1 and 5.5.2).

Theorem 7.7. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, and $\mathcal{H}_k$ a separable RKHS on $\mathcal{X}$ with $\mathsf{E}_{x,y}\{y^2 k(x,x)\} < \infty$. Then ONC_u^2 converges in distribution to a Gaussian according to*

$$\sqrt{n}\left\{\mathrm{ONC}_u^2 - \mathrm{ONC}^2(\mathcal{H}_k, P_{x,y})\right\} \to \mathrm{N}(0, \sigma^2), \tag{7.5}$$

where $\sigma^2 = \mathsf{E}_{x,y}\left[\left\{\mathsf{E}_{x',y'} yk(x,x')y'\right\}^2\right] - \left\{\mathsf{E}_{x,y,x',y'} yk(x,x')y'\right\}^2$. Moreover, if additionally we have $\mathrm{ONC}^2(\mathcal{H}_k, P_{x,y}) = 0$, then $\sigma^2 = 0$, and the distribution in (7.5) is degenerate. Furthermore, ONC_u^2 then converges in distribution according to

$$n\mathrm{ONC}_u^2 \to \sum_{i=1}^{\infty} \lambda_i \left(\xi_i^2 - 1\right),$$

where ξ_i^2 are i.i.d. χ_1^2 random variables, and λ_i are the solutions to the eigenvalue problem

$$\int_{\mathcal{X}\times\mathbb{R}} yk(x,x')y'\psi_i(x,y)dP_{x,y}(x,y) = \lambda_i \psi_i(x',y').$$

Theorem 7.7 gives the asymptotic distribution of ONC_u^2 under both the null and alternate hypothesis (7.1). Denote by $P_{(x_i,y_i)_{i=1}^n}$ the empirical distribution,

$$P_{(x_i,y_i)_{i=1}^n} \equiv n^{-1} \sum_{i=1}^{n} \delta_{x_i,y_i}.$$

For Banach spaces, a biased empirical estimate of the ONC is given by

$$\mathrm{ONC}\left(\mathcal{B}, P_{(x_i,y_i)_{i=1}^n}\right) = n^{-1} \sup_{\|f\|_{\mathcal{B}} \leq 1} \sum_{i=1}^{n} y_i f(x_i). \tag{7.6}$$

For RKHS, by Theorem 7.4, the biased statistic in (7.6) may be expressed as

$$\mathrm{ONC}^2\left(\mathcal{H}_k, P_{(x_i,y_i)_{i=1}^n}\right) = n^{-2} \sum_{i,j=1}^{n} y_i k(x_i, x_j) y_j.$$

As the average of $y_i k(x_i, x_j) y_j$ over all $1 \leq i, j \leq n$, the statistic $\mathrm{ONC}^2(\mathcal{H}_k, P_{(x_i,y_i)_{i=1}^n})$ is of the form of a V-statistic. With appropriate adjustments (e.g., Serfling, 1980; Gretton *et al.*, 2008a), Theorem 7.7 may be adapted for $\mathrm{ONC}^2(\mathcal{H}_k, P_{(x_i,y_i)_{i=1}^n})$.

Denote by $P_{y|(x_i)_{i=1}^n}$ the distribution generated by predictor data $x_1, \ldots, x_n$ and with respect to $P_{y|x_i}$ for all $1 \leq i \leq n$,

$$P_{y|(x_i)_{i=1}^n} = n^{-1} \sum_{i=1}^{n} P_{y|x_i} \delta_{x_i}. \tag{7.7}$$

Furthermore, we denote by σ_i^2 for $1 \leq i \leq n$, the conditional variance of y given x_i,

$$\sigma_i^2 = \mathrm{Var}_{y|x_i}\left(y \mid x_i\right). \tag{7.8}$$

When $\sigma_1^2 = \cdots = \sigma_n^2$, the errors are called *homoscedastic*, and we set $\sigma^2 \equiv \sigma_1^2$. The following theorem, proven in Appendix 7.A.1, gives an unbiased estimate of the operator norm over $P_{y|(x_i)_{i=1}^n}$.

Theorem 7.8. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, and $\mathcal{H}_k$ an RKHS with $\mathsf{E}_{x,y}\{y^2 k(x,x)\} < \infty$. For $(x_i, y_i)_{i=1}^n$ drawn independently form $P_{x,y}$, let $\sigma_1^2, \ldots, \sigma_n^2$ be given by (7.8). Then an unbiased empirical estimator of $\mathrm{ONC}^2(\mathcal{H}_k, P_{y|(x_i)_{i=1}^n})$ is given by*

$$n^{-2}\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{K}\boldsymbol{y} - \sum_{i=1}^n \sigma_i^2 K_{ii}\right).$$

For homoscedastic variance, σ^2, we denote the unbiased estimator in Theorem 7.8 as

$$\mathrm{ONC}_\sigma^2(\mathcal{H}_k, P_{y|(x_i)_{i=1}^n}, \sigma^2) \equiv n^{-2}\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{K}\boldsymbol{y} - \sigma^2 \sum_{i=1}^n K_{ii}\right).$$

We have considered the hypothesis test of Problem 7.1, over $\mathcal{H}_k$, corresponding to testing a set of some p alternative hypothesis. In the next section, the alternative hypothesis is specified by the set of measureable functions.

7.2.2 A Rich Alternate Hypothesis

Consider the following problem, which asks whether taking $f = 0$ achieves the Bayes risk.

Problem 7.9. *Can we test if $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y}$?*

For $\mathcal{R}_{x,y}(0) < \infty$, it is seen that Problem 7.9 is equivalent to testing whether we have $P_x\{\mathsf{E}_{y|x}(y \mid x) = 0\} = 1$. The following definition formalises the set of kernels suitable for testing the hypothesis of $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y}$.

Definition 7.10. *Let $\mathcal{H}_k$ be an RKHS on $\mathcal{X}$. Suppose that either*

i) $\mathrm{ONC}^2(\mathcal{H}_k, P_{x,y}) = 0$ *and* $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y}$ *or*

ii) $\mathrm{ONC}^2(\mathcal{H}_k, P_{x,y}) > 0$ *and* $\mathcal{R}_{x,y}(0) > \mathcal{R}^*_{x,y}$

for all probability distributions $P_{x,y}$ on $\mathcal{X} \times \mathbb{R}$, with

$$\mathcal{R}_{x,y}(0) < \infty, \quad \mathsf{E}_{x,y,x',y'}\{yk(x,x')y'\} < \infty, \quad \text{and} \quad \mathsf{E}_x\{k(x,x)\} < \infty. \tag{7.9}$$

*Then the kernel, k, is called **admissible** on $\mathcal{X}$.*

Subject to the regularity conditions in (7.9), if k is as admissible kernel, then $\mathrm{ONC}^2(\mathcal{H}_k, P_{x,y}) = 0$ implies $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y}$. The following theorem is helpful in determining if k is admissible.

Theorem 7.11. *Let $\mathcal{H}_k$ be a RKHS on $\mathcal{X}$. Then if*

$$\mathcal{R}^*_{x,y;\mathcal{H}_k} = \mathcal{R}^*_{x,y}$$

for all probability distributions $P_{x,y}$ on $\mathcal{X} \times \mathbb{R}$ with $\mathcal{R}_{x,y}(0) < \infty$, then k is admissible.

Proof of Theorem 7.11 is given in Appendix 7.A.1. We would like to know which kernels satisfy the conditions of Theorem 7.11. A version of Theorem 7.12, more general than we require, is proven by Steinwart and Christmann (2008, Theorem 4.26).

Theorem 7.12. *Let $\mathcal{X}$ be a measureable space, μ a σ-finite measure on $\mathcal{X}$, and $\mathcal{H}_k$ a separable RKHS on $\mathcal{X}$. Assume that $\int_{\mathcal{X}} k(x,x)d\mu(x) < \infty$. Let the operator $S\colon L_2(\mu) \to \mathcal{H}_k$ be defined by*

$$Sg(x) \equiv \int_{\mathcal{X}} k(x,x')g(x')d\mu(x') \quad \text{for all } g \in L_2(\mu),\ x \in \mathcal{X}.$$

Then $\mathcal{H}_k$ is dense in $L_2(\mu)$ if and only if $S\colon L_2(\mu) \to \mathcal{H}_k$ is injective.

Dense RKHSs on $\mathbb{R}^d$ include those generated by Laplacian and Gaussian kernels (Fukumizu *et al.*, 2008; Steinwart *et al.*, 2006). The following theorem follows from Steinwart and Christmann (2008, Theorem 5.31).

Theorem 7.13. *Let $P_{\mathsf{x},\mathsf{y}}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, with $\mathcal{R}_{\mathsf{x},\mathsf{y}}(0) < \infty$. Then, for every dense $\mathcal{F} \in L_2(P_{\mathsf{x}})$, we have*

$$\mathcal{R}^*_{\mathsf{x},\mathsf{y};\mathcal{F}} = \mathcal{R}^*_{\mathsf{x},\mathsf{y}}.$$

It is clear, by Theorem 7.11, that if $\mathcal{H}_k$ is dense in $L_2(P_{\mathsf{x}})$, then k is admissible. Kernels with dense RKHS are related to kernel types such as *strictly positive definite* and *universal*. The properties of dense RKHS and related kernel types have been considered by Steinwart (2001), Bach and Jordan (2002), Steinwart, Hush and Scovel (2006), Fukumizu *et al.* (2008) and Sriperumbudur *et al.* (2008) amongst others.

We have the following definition.

Definition 7.14. *A symmetric function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called* ***strictly positive definite*** *on $\mathcal{X}$ if, for all $n \in \mathbb{N}$, non-zero $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and all pairwise unique $x_1, \ldots, x_n \in \mathcal{X}$, we have*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i k(x_i, x_j)\alpha_j > 0.$$

It is easily shown that if $\mathcal{X}$ is a finite set and k strictly positive definite on $\mathcal{X}$, then k is admissible on $\mathcal{X}$. A stronger condition than strictly positive definite is required if $\mathcal{X}$ contains an open subset. The universal kernels were first given by Steinwart (2001).

Definition 7.15. *A continuous kernel k on a compact metric space $\mathcal{X}$ is called* ***universal*** *if the RKHS, $\mathcal{H}_k$, is dense in $\mathcal{C}_b(\mathcal{X})$ with respect to the infinity norm, $\|\cdot\|_\infty$. That is, for all $g \in \mathcal{C}_b(\mathcal{X})$ and $\varepsilon > 0$, there exists $f \in \mathcal{H}_k$ such that*

$$\|f - g\|_\infty \leq \varepsilon.$$

Name	$k(s,t)$	support ($\mathcal{X}$)
Gaussian	$\exp\left(-\gamma \|s-t\|^2\right)$	compact subset of $\mathbb{R}^d$
Laplacian	$\exp\left(-\gamma \|s-t\|\right)$	compact subset of $\mathbb{R}^d$
Exponential	$\exp\left(\gamma \langle s,t\rangle\right)$	compact subset of $\mathbb{R}^d$
Infinite Polynomial	$(1-\langle s,t\rangle)^{-\gamma}$	$\|x\|_2 < 1$

Table 7.1. *Examples of commonly used universal kernels, $\gamma > 0$. Each of the kernels is shown to be universal by Steinwart (2001).*

Some examples of universal kernels are given in Table 7.1. Each universal kernel in Table 7.1 admits a countably infinite orthogonal expansion. The following theorem is given by Steinwart and Christmann (2008, Corollary 5.29).

Theorem 7.16. *Let $\mathcal{X}$ be a compact metric space, $\mathcal{H}_k$ the RKHS of a universal kernel on $\mathcal{X}$ and $P_{x,y}$ a probability distribution on $\mathcal{X} \times \mathbb{R}$. Then we have*

$$\mathcal{R}^*_{x,y;\mathcal{H}_k} = \mathcal{R}^*_{x,y}.$$

By Theorem 7.11, universal kernels on compact metric spaces are admissible.

We now give a well-known example of the increasingly stronger required conditions for a positive definite kernel, strictly positive definite kernel, and universal kernel. Consider the polynomials

$$k(s,t) = \sum_{i=0}^{\infty} a_i (s^\mathsf{T} t)^i, \tag{7.10}$$

for $a_0, a_1, \ldots \in \mathbb{R}$, where the sum is assumed to converge. It is shown in Berg *et al.* (1984, page 159) that k is a kernel if and only if $a_i \geq 0$ for all $i \in \{0,1,\ldots\}$. The following theorem clarifies the necessary and sufficient conditions for a kernel of the form (7.10) to be either strictly positive definite or universal.

Theorem 7.17. *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact. Moreover, let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel of the form*

$$k(s,t) = \sum_{i=0} a_i (s^\mathsf{T} t)^i,$$

with $a_0 > 0$. Then k is:

i) Strictly positive definite if and only if

$$\sum_{a_{2i}>0} 1 = \sum_{a_{2i+1}>0} 1 = \infty.$$

ii) Universal if and only if

$$\sum_{a_{2i}>0} \frac{1}{a_{2i}} = \sum_{a_{2i+1}>0} \frac{1}{a_{2i+1}} = \infty.$$

Proof of *i)* is given by Pinkus (2004), who attributes *ii)* to the work of Dahmen and Micchelli (1987). It is well-known that all universal kernels on compact $\mathcal{X}$ are strictly positive definite. Theorem 7.17 shows that the converse is not true; not all strictly positive definite kernels are universal.

7.3 Modelling

The previous section considered the possibility that $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y;\mathcal{H}_k}$, that is, whether having $f = 0$ is the optimal choice for minimising the risk. More direct, and more difficult, is finding some $f \in \mathcal{H}_k$ so that $\mathcal{R}_{x,y}(f)$ is in some sense small. In this section, the ONC is used to guide the choice of f.

Following Chapter 2, for $\lambda \in (0, \infty)$, a fit to the mean response is given by

$$f_\lambda = \min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \|f\|^2_{\mathcal{H}_k} \right\}. \tag{7.11}$$

The domain of λ is taken to be the extended non-negative real number line $[0, \infty]$ by

$$f_\infty \equiv \lim_{\lambda \to \infty} f_\lambda = 0, \quad \text{and} \quad f_0 \equiv \lim_{\lambda \to 0^+} f_\lambda.$$

Hence $\text{ONC}(\mathcal{H}_k, P_{x,y}) = 0$ implies that $\mathcal{R}_{x,y}(f_\infty) = \mathcal{R}^*_{x,y;\mathcal{H}_k}$, and that $\lambda = \infty$ must be an optimal choice for $\lambda \in [0, \infty]$. We wish to further investigate the behaviour of f_λ, in particular for large values of λ. Make the change of variable $\tau = \lambda^{-1}$, which we express as $f_{(\tau)} \equiv f_{\tau^{-1}}$. That is,

$$f_{(\tau)} = \min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \tfrac{1}{\tau} \|f\|^2_{\mathcal{H}_k} \right\}, \quad \tau \in (0, \infty), \tag{7.12}$$

with

$$f_{(0)} = f_\infty, \quad \text{and} \quad f_{(\infty)} = f_0.$$

The dual solution to (7.12) is given by

$$f_{(\tau)}(x) = \sum_{i=1}^{n} \widehat{c}_i k(x, x_i),$$

where the $\widehat{c}_i$s have

$$\widehat{\boldsymbol{c}} = \left(\boldsymbol{K} + \tfrac{1}{\tau}\boldsymbol{I}\right)^{-1} \boldsymbol{y}.$$

Taking limits as $\tau \to 0^+$ yields the useful result

$$\lim_{\tau \to 0^+} f_{(\tau)}(x)/\tau = \sum_{i=1}^{n} k(x, x_i) y_i. \tag{7.13}$$

Equation (7.13) helps describe the behaviour of $f_{(\tau)}$ for small values of τ.

7.3.1 The Mean Squared Error

The aim of minimising the mean squared error (MSE) is standard paradigm of statistics and machine learning. Within the penalised framework of (7.12), we require some τ so that the data dependent $f_{(\tau)}$ has desirable properties. We therefore require a version of the risk not in terms of fit f, but in terms of parameterisation τ. In such a sense, the unobserved $f_{(\tau)}$ is a random variable over $P_{(\mathsf{x},\mathsf{y})^n}$, and dependent on the parameter τ. The MSE is then the average risk over $P_{(\mathsf{x},\mathsf{y})^n}$. This is made explicit in in following definition.

Definition 7.18. *Let $P_{(\mathsf{x},\mathsf{y})}$ be probability distribution on $\mathcal{X} \times \mathbb{R}$, $\mathcal{H}_k$ a separable RKHS on $\mathcal{X}$ and $f_{(\tau)}$ given by* (7.12) *with parameter τ and random sample drawn from $P_{(\mathsf{x},\mathsf{y})^n}$. Then the mean squared error is*

$$\mathrm{MSE}_n(\tau) \equiv \mathsf{E}_{(\mathsf{x},\mathsf{y})^n,\mathsf{x}',\mathsf{y}'}\left[\{y' - f_{(\tau)}(x')\}^2\right].$$

An estimate of the MSE is given by leave-one-out cross-validation, LOO,

$$\mathrm{LOO}(\tau) \equiv n^{-1}\sum_{i=1}^{n}\left\{y_i - f_{(\tau)}^{[-i]}(x_i)\right\}^2,$$

where $f_{(\tau)}^{[-i]}$ denotes the fit to (7.12), with the ith observation removed. It is commonplace to select τ (equivalently, λ), by searching for the minimiser of leave-one-out cross-validation,

$$\min_{\tau\in[0,\infty]} \mathrm{LOO}(\tau).$$

The following theorem shows the behaviour of $\mathrm{LOO}(\tau)$ for small values of τ.

Theorem 7.19. *Let $\mathcal{H}_k$ be an RKHS, and $(x_i, y_i)_{i=1}^n \in \mathcal{X} \times \mathbb{R}$ a data set with $n \geq 2$. Then*

$$\lim_{\tau\to 0^+}\frac{d}{d\tau}\mathrm{LOO}(\tau) = -2(n-1)\mathrm{ONC}_u^2(\mathcal{H}_k, (x_i, y_i)_{i=1}^n).$$

Proof of Theorem 7.19 is given in Appendix 7.A.2. Theorem 7.19 shows that if the minimiser of leave-one-out cross-validation gives $\tau = 0$, then $\mathrm{ONC}_u^2(\mathcal{H}_k, (x_i, y_i)_{i=1}^n) \leq 0$. It is well-known that leave-one-out cross-validation gives an almost unbiased estimate of the mean squared error, in the sense

$$\mathsf{E}_{\mathsf{x},\mathsf{y}}\{\mathrm{LOO}(\tau)\} = \mathrm{MSE}_{n-1}(\tau) \quad \text{for all } \tau \in [0,\infty].$$

We now show the ONC is related to the mean squared error on samples of size n.

Theorem 7.20. *Let $P_{\mathsf{x},\mathsf{y}}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$ and $\mathcal{H}_k$ a separable RKHS such that $\mathsf{E}_{\mathsf{x},\mathsf{y},\mathsf{x}',\mathsf{y}'}\{yk(x,x')y'\} < \infty$ and $\mathsf{E}_{\mathsf{x}}\{k(x,x)\} < \infty$. Then, for $n \in \mathbb{N}$,*

$$\lim_{\tau\to 0^+}\frac{d}{d\tau}\mathrm{MSE}_n(\tau) = -2n\mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{x},\mathsf{y}}).$$

Proof of Theorem 7.20 is given in Appendix 7.A.2. It follows from Theorem 7.20, that if $\mathrm{ONC}(\mathcal{H}_k, P_{\mathsf{x},\mathsf{y}}) > 0$, then for sufficiently small $\tau > 0$, we can improve the MSE. Alternatively, if $\mathrm{ONC}(\mathcal{H}_k, P_{\mathsf{x},\mathsf{y}}) = 0$, we should choose $\tau = 0$ in (7.12). That is, the fit would be $f(x) = 0$ for all $x \in \mathcal{X}$. An interesting aspect is that if the null hypothesis considered in (7.1) is false, we can always improve upon the MSE for sufficiently small $\tau > 0$. That is, for the MSE:

i) If $\mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{x},\mathsf{y}}) = 0$, then the optimal choice of τ is $\tau = 0$.

ii) If $\mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{x},\mathsf{y}}) > 0$, then the optimal choice of τ must be greater than zero.

We have an empirical quantity to guide whether to choose $\tau = 0$. For the MSE, Theorem 7.20 suggests that $\tau = 0$ should be selected if

$$(n)_2\mathrm{ONC}_u^2(\mathcal{H}_k, (x_i, y_i)_{i=1}^n) = \boldsymbol{y}^\mathsf{T}\boldsymbol{K}\boldsymbol{y} - \sum_{i=1}^n y_i^2 K_{ii} \leq 0.$$

We will return to investigating the optimal choice of τ for the MSE in Section 7.3.4.

7.3.2 The Mean Squared Prediction Error

The task of minimising the mean squared prediction error (MSPE) is a common alternative to the MSE.

Definition 7.21. *Let $P_{(\mathsf{x},\mathsf{y})}$ be probability distribution on $\mathcal{X} \times \mathbb{R}$, $\mathcal{H}_k$ a separable RKHS on $\mathcal{X}$ and $x_i \in \mathcal{X}$ for all $1 \leq i \leq n$. Moreover, let $f_{(\tau)}$ be given by (7.12) with parameter τ and random sample drawn as $y_i \sim P_{\mathsf{y}|x_i}$ for all $1 \leq i \leq n$. Then the mean squared predictive error is*

$$\mathrm{MSPE}(\tau) \equiv \sum_{i=1}^n \mathsf{E}_{\mathsf{y}|(x_j)_{j=1}^n, \mathsf{y}'|\mathsf{x}'=x_i} \left[\{y' - f_{(\tau)}(x_i)\}^2 \right].$$

Similar to the MSE case, we find a special relationship between the MSPE and the ONC.

Theorem 7.22. *Let $P_{\mathsf{x},\mathsf{y}}$ be a probability distribution of the form (7.7), with errors having homoscedastic variance σ^2, and $\mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{y}|(x_i)_{i=1}^n}) < \infty$. Then*

$$\lim_{\tau \to 0^+} \frac{d}{d\tau} \mathrm{MSPE}(\tau) = -2n^2 \mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{y}|(x_i)_{i=1}^n}).$$

Theorem 7.22 is proven in Appendix 7.A.2. Stein's unbiased risk estimate (SURE) is given by

$$\mathrm{SURE}(\tau) = \sum_{i=1}^n \{y_i - f_{(\tau)}(x_i)\}^2 + 2\sigma^2 \sum_{i=1}^n \frac{df_{(\tau)}(x_i)}{dy_i}. \tag{7.14}$$

Given by Stein (1981), SURE is an unbiased estimate of the mean squared predictive error, under the assumpion that the errors $y - \mathsf{E}_{y|x_i}(y \mid x_i)$ are normally distributed with mean zero and homoscedastic variance σ^2. The assumption of normality, however, is not a requirement for the following theorem.

Theorem 7.23. *With homoscedastic variance σ^2, let* SURE(τ) *be Stein's unbiased risk estimator of* (7.14)*, with $f_{(\tau)}$ given by* (7.12)*. Then*

$$\lim_{\tau \to 0^+} \frac{d}{d\tau} \text{SURE}(\tau) = -2n^2 \text{ONC}_\sigma^2 \left(\mathcal{H}_k, (x_i, y_i)_{i=1}^n, \sigma^2 \right).$$

Proof of Theorem 7.23 is given in Appendix 7.A.2. The following theorem is an extension of Theorem 7.5.

Theorem 7.24. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, with inputs $(x_i)_{i=1}^n$ such that $\mathsf{E}_{y|x_i}(y^2 \mid x_i) < \infty$ for all $1 \leq i \leq n$. Then* $\text{ONC}^2(\mathcal{H}_k, P_{y|(x_i)_{i=1}^n}) \geq 0$*, with equality if and only if*

$$\sum_{i=1}^n \mathsf{E}_{y|x_i}(y^2 \mid x_i) = \inf_{f \in \mathcal{H}_k} \sum_{i=1}^n \mathsf{E}_{y|x_i}\left[\{y - f(x_i)\}^2 \mid x_i \right].$$

Proof of Theorem 7.24 is given in Appendix 7.A.2. As such, for the mean squared predictive error:

i) If $\text{ONC}^2(\mathcal{H}_k, P_{y|(x_i)_{i=1}^n}) = 0$, then the optimal choice of τ is $\tau = 0$.

ii) If $\text{ONC}^2(\mathcal{H}_k, P_{y|(x_i)_{i=1}^n}) > 0$, then the optimal choice of τ must be greater than zero.

Empirically , for the MSPE, if

$$n^2 \text{ONC}_\sigma^2 \left(\mathcal{H}_k, (x_i, y_i)_{i=1}^n, \sigma^2 \right) = \boldsymbol{y}^\mathsf{T} \boldsymbol{K} \boldsymbol{y} - \sigma^2 \text{tr}(\boldsymbol{K}) \leq 0,$$

then the selection of $\tau = 0$ is suggested by Theorems 7.22 and 7.23. For both the MSE and MSPE there still remains the question of how best to choose τ when $\tau > 0$.

7.3.3 Alternative Parameterisations

The fitted model, (7.11), requires the choice of a smoothing parameter, λ. Typically λ is chosen in a data dependent way. There are alternative styles of parameterisation, including (7.12). For some λ_N, consider the optimisation problem

$$\min_{f \in \mathcal{B}} \left\{ \sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda_N \|f\|_{\mathcal{B}} \right\}. \tag{7.15}$$

It is known that there exists a monotonic relationship between λ and λ_N. That is, for each $\lambda \geq 0$, there exists λ_N such that f_λ is the solution to (7.15). For the Hilbert space

norm $\|\cdot\|_2$, we see that (7.11) is the most widely used parameterisation. Conversely, for the Banach space norm $\|\cdot\|_1$, we find that (7.15) is the typical choice of parameterisation, for example Donoho and Johnstone (1994) and Tibshirani (1996).

7.3.4 Residual-based Fits

We may consider the residuals of a fitted model. The residuals are given by $y_i - \widehat{y}_i$ where $\widehat{y}_i = f(x_i)$, for all $1 \leq i \leq n$. If the residuals appear to not have conditional mean zero, then that would indicate that the model is underfitted. As a manner of choosing the smoothing of a model, we may choose the smoothing parameter so that the errors appear to be mean zero. For example, in the MSPE case, if

$$\boldsymbol{y}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{y} - \sigma^2 \mathrm{tr}(\boldsymbol{K}) \leq 0 \tag{7.16}$$

we choose $f = 0$. Alternatively, if

$$\boldsymbol{y}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{y} - \sigma^2 \mathrm{tr}(\boldsymbol{K}) > 0,$$

we may choose τ in (7.12) such that

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) - \sigma^2 \mathrm{tr}(\boldsymbol{K}) = 0. \tag{7.17}$$

As $(\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{y}})$ is both continuous and monotonic in τ, we are simply finding the minimum $\tau \in [0, \infty]$, such that

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) \leq \sigma^2 \mathrm{tr}(\boldsymbol{K}).$$

We call the fit to (7.16)–(7.17) the **parameter information criterion** (PIC) with respect to the MSPE. For given $\mathcal{H}_k$, data $(x_i, y_i)_{i=1}^n$ and variance σ^2, we denote it by $\mathrm{PIC}_{\mathrm{MSPE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n, \sigma^2)$. An attractive aspect of the PIC is that the residuals of the fit, $\boldsymbol{y} - \widehat{\boldsymbol{y}}$, appear to have zero conditional mean, in that

$$\mathrm{ONC}_\sigma^2(\mathcal{H}_k, (x_i, y_i - \widehat{y}_i)_{i=1}^n, \sigma^2) \leq 0.$$

The following theorem shows that $\mathrm{PIC}_{\mathrm{MSPE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n, \sigma^2)$ may be expressed as the solution to a convex optimisation problem, with a specific parameterisation.

Theorem 7.25. *Let f be the solution to the optimisation problem*

$$\min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_N \|f\|_{\mathcal{H}_k} \right\}, \tag{7.18}$$

where

$$\lambda_N = 2\sigma\sqrt{\mathrm{tr}(\boldsymbol{K})}. \tag{7.19}$$

Then if

$$\boldsymbol{y}^\mathsf{T}\boldsymbol{K}\boldsymbol{y} - \sigma^2\mathrm{tr}(\boldsymbol{K}) \leq 0,$$

we have $f = 0$. *Otherwise* $f \neq 0$, *and*

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}})^\mathsf{T}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \sigma^2\mathrm{tr}(\boldsymbol{K}).$$

Proof is given in Appendix 7.A.2. With λ_N given by (7.19), Theorem 7.25 shows that $\mathrm{PIC}_{\mathrm{MSPE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n, \sigma^2)$ is the solution to the convex optimisation problem (7.18). The penalty term is proportional to $\|f\|_{\mathcal{H}_k}$, the RKHS norm, as opposed to the squared RKHS norm $\|f\|_{\mathcal{H}_k}^2$. Though without specifying the parameterisation (7.19), the RKHS penalisation has attracted some interest of late. Yuan and Lin (2006) and Bach (2008) proposed models that include the convex optimisation (7.18). Recently, Steinwart (2009) and Steinwart *et al.* (2009) have rigorously shown the broad adaptability of the RKHS penalisation (7.18).

There is also the mean squared error case. If

$$\boldsymbol{y}^\mathsf{T}\boldsymbol{K}\boldsymbol{y} \leq \sum_{i=1}^n y_i^2 K_{ii}, \tag{7.20}$$

we select $f = 0$. If (7.20) does not hold true, we select the minimum τ such that

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}})^\mathsf{T}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \sum_{i=1}^n (y - \widehat{y})_i^2 K_{ii}. \tag{7.21}$$

Equivalently, by the continuity of $\widehat{\boldsymbol{y}}$, we seek the minimum $\tau \in [0, \infty]$ such that

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}})^\mathsf{T}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) \leq \sum_{i=1}^n (y - \widehat{y})_i^2 K_{ii}.$$

For given $\mathcal{H}_k$ and data $(x_i, y_i)_{i=1}^n$, we denote the fit to (7.20)–(7.21) by the parameter information criterion, $\mathrm{PIC}_{\mathrm{MSE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n)$. Taking limits as $\tau \to \infty$,

$$\lim_{\tau\to\infty}\left\{(\boldsymbol{y} - \widehat{\boldsymbol{y}})^\mathsf{T}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{y}})\right\} = 0 \leq \lim_{\tau\to\infty}\left\{\sum_{i=1}^n (y - \widehat{y})_i^2 K_{ii}\right\},$$

suggesting that $\mathrm{PIC}_{\mathrm{MSE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n)$ is well-defined. For translation invariant kernels, the following theorem is helpful in characterising $\mathrm{PIC}_{\mathrm{MSE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n)$.

Theorem 7.26. *Let* k *be a translation invariant kernel, with* $k(x, x) > 0$ *and* $\boldsymbol{y}^\mathsf{T}\boldsymbol{y} \neq 0$. *Then*

$$\frac{(\boldsymbol{y} - \widehat{\boldsymbol{y}})^\mathsf{T}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{y}})}{\sum_{i=1}^n (y - \widehat{y})_i^2 K_{ii}}$$

is a monotonically increasing function of $\lambda > 0$. *Furthermore, if* $\boldsymbol{y}$ *is not an eigenvector of* $\boldsymbol{K}$, *then the monotonicity is strict. Finally, we have*

$$\lim_{\lambda \to 0^+} \frac{(\boldsymbol{y} - \widehat{\boldsymbol{y}})^\mathsf{T} \boldsymbol{K} (\boldsymbol{y} - \widehat{\boldsymbol{y}})}{\sum_{i=1}^n (\boldsymbol{y} - \widehat{\boldsymbol{y}})_i^2 K_{ii}} = \begin{cases} \frac{\boldsymbol{y}^\mathsf{T} \boldsymbol{K}^- \boldsymbol{y}}{\sum_{i=1}^n (\boldsymbol{y}^\mathsf{T} \boldsymbol{K}^-)_i^2 K_{ii}}, & \boldsymbol{y} \in \operatorname{span}(\boldsymbol{K}), \\ 0, & \boldsymbol{y} \notin \operatorname{span}(\boldsymbol{K}), \end{cases} \tag{7.22}$$

where $\operatorname{span}(\boldsymbol{K})$ *is the span of the columns of* $\boldsymbol{K}$, *and* $\boldsymbol{K}^-$ *the Moore-Penrose inverse of* $\boldsymbol{K}$.

Proof of Theorem 7.26 is given in Appendix 7.A.2. Translation invariant kernels include the important Gaussian and Laplacian kernels. Theorem 7.26 shows that, for such translation invariant kernels, the search for a solution to (7.21) is equivalent to finding the zero of a monotonic function. Finding the zero of a monotonic function is closely related to convex optimisation. Also, in (7.22) we find the characterisation of when $\mathrm{PIC}_{\mathrm{MSE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n)$ has $\tau = \infty$ (that is, $\lambda = 0$). Explicitly, if $\boldsymbol{y} \in \operatorname{span}(\boldsymbol{K})$ with $\boldsymbol{y}^\mathsf{T} \boldsymbol{K}^- \boldsymbol{y} > \sum_{i=1}^n (\boldsymbol{y}^\mathsf{T} \boldsymbol{K}^-)_i^2 K_{ii}$, then $\mathrm{PIC}_{\mathrm{MSE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n)$ has $\tau = \infty$.

The residuals of $\mathrm{PIC}_{\mathrm{MSE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n) \neq 0$ have

$$\mathrm{ONC}_u^2(\mathcal{H}_k, (x_i, y_i - \widehat{y}_i)_{i=1}^n) = 0,$$

and the residuals of $\mathrm{PIC}_{\mathrm{MSPE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n) \neq 0$ have

$$\mathrm{ONC}_\sigma^2\left(\mathcal{H}_k, (x_i, y_i - \widehat{y}_i)_{i=1}^n, \sigma^2\right) = 0.$$

Both lie on the cusp between overfitting and underfitting by the operator norm criterion. Little by the way of assumptions has been made in the derivation of the $\mathrm{PIC}_{\mathrm{MSE}}$ and $\mathrm{PIC}_{\mathrm{MSPE}}$. In particular, we have done without the assumption of normality of the errors, or for the $\mathrm{PIC}_{\mathrm{MSE}}$, even knowledge of the variance. This contrasts with such methods such as SURE and ML. We can expect that $\mathrm{PIC}_{\mathrm{MSE}}$ and $\mathrm{PIC}_{\mathrm{MSPE}}$ to be therefore robust to model violations of SURE or of ML. It turns out that $\mathrm{PIC}_{\mathrm{MSPE}}$ tends toward more conservative fits, with higher regularisation, than the comparative SURE or ML fits. Based on some special cases, the next section puts forward a heuristic that often gives similar fits to those of SURE and of ML. The heuristic is an aid to the interpretation of the PIC.

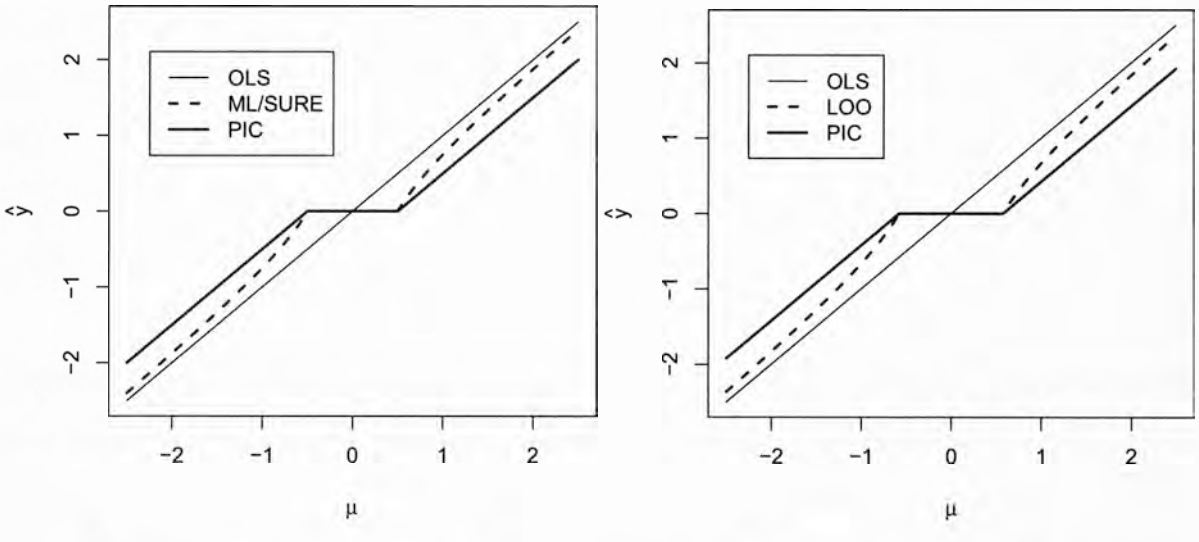
7.3.5 Allowing for Curvature

It is common practice to choose the smoothing parameter by such methods as SURE or LOO. For example, we could choose $\tau = \lambda^{-1}$ by searching for a minimiser,

$$\widehat{\tau} = \operatorname*{argmin}_\tau \mathrm{SURE}(\tau) \quad \text{or} \quad \widehat{\tau} = \operatorname*{argmin}_\tau \mathrm{LOO}(\tau).$$

A drawback to such methods is that the resulting optimisation is not guaranteed to be convex. There will typically be multiple local minima, and non-convex optimisation techniques such as grid search are to be employed. There are, however, some cases whereby the minimiser of either SURE or LOO is simply calculable. Some simple situations show that the minimisers of SURE and LOO tend to give somewhat less regularised fits than $\mathrm{PIC}_{\mathrm{MSPE}}$ and $\mathrm{PIC}_{\mathrm{MSE}}$ respectively.

Figure 7.1 shows the fitted values for a toy example with $n = 4$. We have Gram matrix with terms $K_{ij} = 1$ for $1 \leq i, j \leq 4$, and $\boldsymbol{y} = [\mu - 1, \mu - 1, \mu + 1, \mu + 1]^{\mathsf{T}}$, $\mu \in \mathbb{R}$. Moreover, for the MSPE case, we set $\sigma^2 = 1$. The figure shows the fitted $\widehat{y}_1$ as a function of μ. There is a visible curvature to the fits for SURE, ML and for LOO.



(a) *Mean squared predictive error fits*

(b) *Mean squared error fits*

Figure 7.1. *Comparison of fits given by various procedures. A simple case was chosen whereby the fits are tractable. A noticeable curvature exists in the fits for some of the procedures. Left: Mean squared prediction error based fits of maximum likelihood (ML), Stein's unbiased risk estimator (SURE) and the parameter information criterion (PIC) of* (7.17)*. Ordinary least squares is also included for comparison. Right: Comparison of leave-one-out cross validation (LOO) and the parameter information criterion (PIC) for mean squared error.*

Both (7.17) and (7.21) give residuals that appear to neither overfit or underfit. However, it is typical of fits made by SURE and LOO that the residuals show signs of fitting. Can an adjustment be made to ONC style fits in order to give a similar curvature? Although it is clear at where should have $f = 0$, we may want sharper, less smooth fits

otherwise.

Mean squared predictive error

Situations are identified whereby the fits given by minimising SURE are tractable. Consider the block diagonal Gram matrix

$$\boldsymbol{K} = \begin{bmatrix} \mathbf{1}_{m\times m} & \mathbf{0}_{m\times m} & \cdots & \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times m} & \mathbf{1}_{m\times m} & \cdots & \mathbf{0}_{m\times m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \cdots & \mathbf{1}_{m\times m} \end{bmatrix}. \tag{7.23}$$

For $\boldsymbol{K}$ given by (7.23), it is straightforward to calculate the fits given by either ML or SURE. For $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} \geq \sigma^2\mathrm{tr}(\boldsymbol{K})$, both ML and SURE have

$$(\boldsymbol{y}-\widehat{\boldsymbol{y}})^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} = \left\{\sigma^2\mathrm{tr}(\boldsymbol{K})\right\}^2. \tag{7.24}$$

Such fits are given by the optimisation problem

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}\left\{y_i - f(x_i)\right\}^2 + \lambda_N \left\|f\right\|_{\mathcal{H}_k}\right\}, \tag{7.25}$$

where

$$\lambda_N = \frac{2\sigma^2\mathrm{tr}(\boldsymbol{K})}{\sqrt{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y}}}. \tag{7.26}$$

The accuracy of parameter selection based on (7.25) may be compared with such procedures as ML and SURE. The fits to (7.25) have the distinctive curvature shown by ML and SURE. We call the fit to (7.25)–(7.26) the **curved information criterion** (CIC) with respect to the MSPE. With respect to kernel, $\mathcal{H}_k$, data $(x_i, y_i)_{i=1}^n$ and variance, σ^2, we denote the fit by, $\mathrm{CIC}_{\mathrm{MSPE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n, \sigma^2)$.

Mean squared error

Some situations are identified whereby the fits given by minimising LOO are tractable. When $\boldsymbol{K}$ is given by (7.23), the fit given by leave-one-out has, for $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} > \sum_{i=1}^n y_i K_{ii}$,

$$\left\{\frac{(\boldsymbol{y}-\widehat{\boldsymbol{y}})^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})}{\sum_{i=1}^n (y_i-\widehat{y}_i)^2 K_{ii}} - m^{-1}\right\}\left\{\frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y}}{\sum_{i=1}^n y_i^2 K_{ii}} - m^{-1}\right\} = \left\{1 - m^{-1}\right\}^2, \tag{7.27}$$

The functional form of (7.27) suggests that we may provide a curvature for general Gram matrices. Somewhat unfortunate is the existence of the m^{-1} components of (7.27). There are a range of entities that are equal to m^{-1} for the special case Gram matrix (7.23). These include

$$\frac{\mathrm{tr}(\boldsymbol{K})}{\mathbf{1}^{\mathrm{T}}\boldsymbol{K}\mathbf{1}}, \quad \frac{\mathrm{tr}(\boldsymbol{K}^2)}{\mathbf{1}^{\mathrm{T}}\boldsymbol{K}^2\mathbf{1}}, \quad \text{and} \quad \frac{\mathrm{tr}(\boldsymbol{K})^2}{n\sum_{i=1}^n c_i^2},$$

where c_i are the eigenvalues of $\boldsymbol{K}$, for $1 \leq i \leq n$. For simplicity, we approximate $m^{-1} \approx 0$ in (7.27), yielding

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathrm{T}} \boldsymbol{K} (\boldsymbol{y} - \widehat{\boldsymbol{y}}) \boldsymbol{y}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{y} = \left\{ \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 K_{ii} \right\} \left\{ \sum_{i=1}^{n} y_i^2 K_{ii} \right\}. \tag{7.28}$$

We note that (7.28) tends toward more regularised fits than those of (7.27). An attractive aspect of (7.28) is the close semblance to the $\mathrm{CIC}_{\mathrm{MSPE}}$ curvature of (7.24). If

$$\boldsymbol{y}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{y} \leq \sum_{i=1}^{n} y_i^2 K_{ii},$$

we select $f = 0$. Alternatively, for

$$\boldsymbol{y}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{y} > \sum_{i=1}^{n} y_i^2 K_{ii},$$

we select the maximum λ such that (7.28) holds. With respect to kernel, $\mathcal{H}_k$, and data, $(x_i, y_i)_{i=1}^n$, we denote the corresponding fit by the curved information criterion, $\mathrm{CIC}_{\mathrm{MSE}}(\mathcal{H}_k, (x_i, y_i)_{i=1}^n)$.

By a rearrangement of (7.28), for $\{\sum_{i=1}^n (y_i - \widehat{y}_i)^2 K_{ii}\} > 0$,

$$\frac{(\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathrm{T}} \boldsymbol{K} (\boldsymbol{y} - \widehat{\boldsymbol{y}})}{\sum_{i=1}^n (y_i - \widehat{y}_i)^2 K_{ii}} = \frac{\sum_{i=1}^n y_i^2 K_{ii}}{\boldsymbol{y}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{y}}.$$

For translation invariant kernels, Theorem 7.26 shows that calculating $\mathrm{CIC}_{\mathrm{MSE}}$ is equivalent to finding the zero of a monotonic function. Both $\mathrm{PIC}_{\mathrm{MSE}}$ and $\mathrm{CIC}_{\mathrm{MSE}}$ choose $f = 0$ if and only if $\mathrm{ONC}_u^2 \leq 0$. Otherwise, fits using $\mathrm{CIC}_{\mathrm{MSE}}$ have less smoothing than those with $\mathrm{PIC}_{\mathrm{MSE}}$.

Figure 7.2 shows fits for both $\mathrm{PIC}_{\mathrm{MSE}}$ and $\mathrm{CIC}_{\mathrm{MSE}}$, as applied to the "motorcycle" data of Eubank (1999). Both methods appear to give quite sensible fits, with the PIC fit being smoother. The accuracy of parameter selection based on the PIC, or the CIC, may be compared with such procedures as leave-one-out cross-validation. Experiments on the accuracy of fits are delayed until Section 7.5.

7.3.6 A Broader Null Hypothesis

For RKHS, $\mathcal{H}_k$, with null space $\mathcal{H}_0$, we wish to choose the parameter λ in the kernel machine,

$$f_\lambda = \min_{f \in \mathcal{H}_k} \left\{ \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \left\| P_1 f \right\|_{\mathcal{H}_k}^2 \right\}, \tag{7.29}$$

where P_1 is the projection onto $\mathcal{H}_0^{\perp}$, and $\lambda \in (0, \infty)$. As was the case without null space, we take the extended domain of λ, by taking $f_0 \equiv \lim_{\lambda \to 0^+} f_\lambda$, and $f_\infty \equiv \lim_{\lambda \to \infty} f_\lambda$.

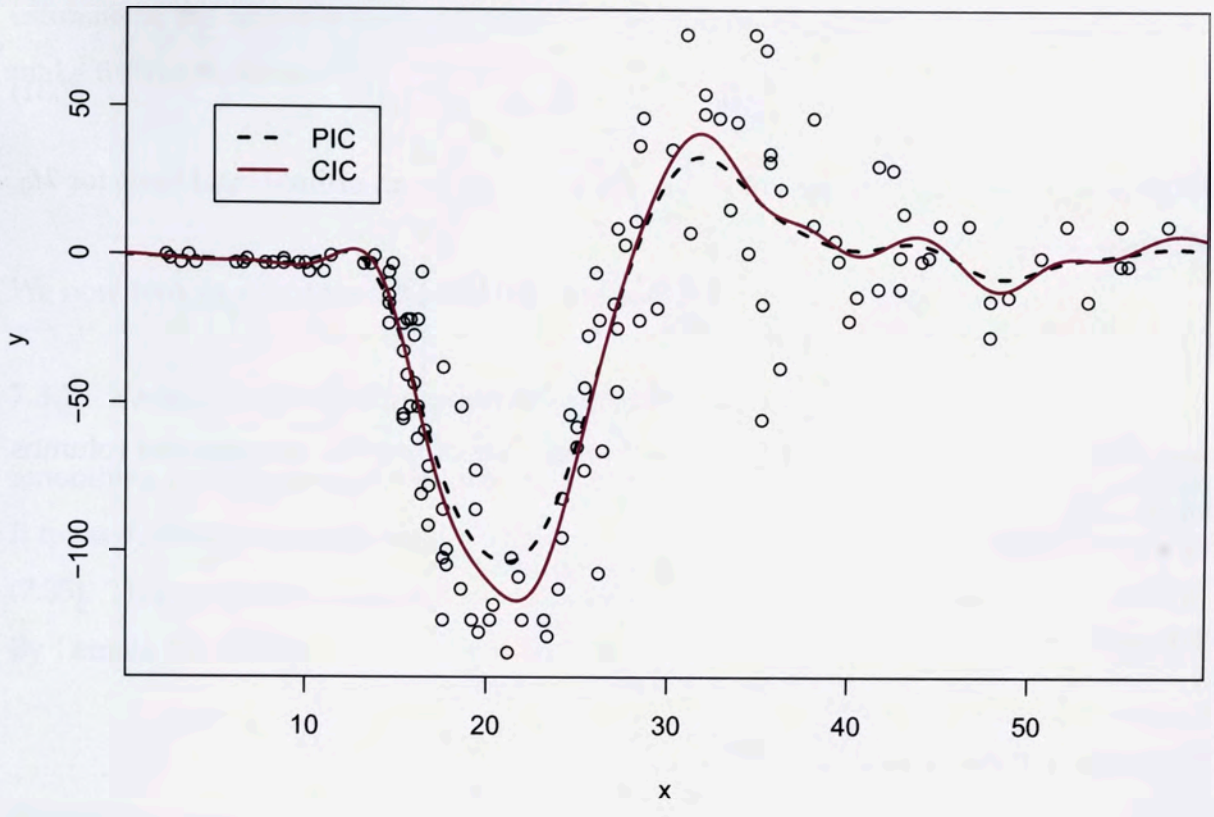


Figure 7.2. *Mean squared error PIC and CIC fits to the motorcycle data analysed by Eubank (1999). The x values are the time measurements in milliseconds after a simulated motorcycle accident, and the y values are measurements of head acceleration. A Gaussian kernel is used, with* $\gamma = 1/s^2$*, where* s^2 *is the sample estimate of the variance. Both the PIC and CIC provide sensible fits to the data, with the CIC the less smooth of the two fits.*

To begin with, we would like to test

$$H_0 \colon \mathcal{R}^*_{\mathsf{x},\mathsf{y};\mathcal{H}_0} = \mathcal{R}^*_{\mathsf{x},\mathsf{y};\mathcal{H}_k} \text{ against}$$
$$H_1 \colon \mathcal{R}^*_{\mathsf{x},\mathsf{y};\mathcal{H}_0} > \mathcal{R}^*_{\mathsf{x},\mathsf{y};\mathcal{H}_k}$$

Let $g^* \in \mathcal{H}_0$ be a solution to $\mathcal{R}_{\mathsf{x},\mathsf{y}}(g^*) = \mathcal{R}^*_{\mathsf{x},\mathsf{y};\mathcal{H}_0}$. Then the operator norm criertia, with respect to $\mathcal{H}_k$, $\mathcal{H}_0$ and P is

$$\mathrm{ONC}(\mathcal{H}_k, \mathcal{H}_0, P_{\mathsf{x},\mathsf{y}}) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathsf{E}_{\mathsf{x},\mathsf{y}} \left\{ (y - g^*(x)) f(x) \right\}. \tag{7.30}$$

We wish to find an empirical estimator for $\mathrm{ONC}(\mathcal{H}_k, \mathcal{H}_0, P_{\mathsf{x},\mathsf{y}})$. For some special cases of $\mathcal{H}_0$, such as $\mathcal{H}_0 = \{0\}$ and $\mathcal{H}_0 = \mathbb{R}$, unbiased empirical estimators are derived (e.g., Theorem 7.4 and furthermore Theorem 7.35). We now put forward an estimator of (7.30) for general $\mathcal{H}_0$.

By application of Theorem 7.4, if some g^* is known, then an unbiased estimator of $\mathrm{ONC}^2(\mathcal{H}_k, \mathcal{H}_0, P_{x,y})$ would be given by

$$(n)_2^{-1} \sum_{(i,j)\in i_2^n} (y_i - g^*(x_i))K_{ij}(y_j - g^*(x_j)). \tag{7.31}$$

In general, however, g^* will be unknown. Let $\psi_0, \ldots, \psi_q$ be an orthonormal basis for $\mathcal{H}_0$. We then set

$$\boldsymbol{X} = \begin{bmatrix} \psi_0(x_1) & \cdots & \psi_q(x_1) \\ \vdots & \ddots & \vdots \\ \psi_0(x_n) & \cdots & \psi_q(x_n) \end{bmatrix}.$$

Recall from Section 4.3.3 that $\boldsymbol{X}^*$ is a matrix made up of linearly independent columns of $\boldsymbol{X}$, with

$$\mathrm{rank}(\boldsymbol{X}^*) = \mathrm{rank}(\boldsymbol{X}).$$

For some $\boldsymbol{X}^*$, let

$$\boldsymbol{H} = \boldsymbol{X}^*(\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\mathsf{T}}.$$

Denote by $\widehat{g}$ the empirical risk minimiser,

$$\widehat{g} = \min_{g\in\mathcal{H}_0} \sum_{i=1}^{n} \{y_i - g(x_i)\}^2.$$

Then, for each $1 \leq i \leq n$,

$$y_i - \widehat{g}(x_i) = \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\}_i.$$

Replacing g^* with $\widehat{g}$ in (7.31), we have an estimate of $\mathrm{ONC}^2(\mathcal{H}_k, \mathcal{H}_0, P_{x,y})$ in

$$(n)_2^{-1} \sum_{(i,j)\in i_2^n} \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\}_i K_{ij} \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\}_j. \tag{7.32}$$

There is reason to believe that the estimator in (7.32) may be improved upon. Inspired by the REML derivation of Patterson and Thompson (1971), we now make an adjustment to (7.32). Consider the estimator given by

$$(n)_2^{-1} \sum_{(i,j)\in i_2^n} \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\}_i \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{H})\}_{ij} \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\}_j. \tag{7.33}$$

Unlike the estimator in (7.32), we find that (7.33) is invariant through changing of scale of the norm on $\mathcal{H}_0$. In particular, (7.32) is dependent on the empirical

$$\boldsymbol{X}^{*\mathsf{T}}(\boldsymbol{y} - \boldsymbol{v}), \tag{7.34}$$

where $v_i = g^*(x_i)$. As (7.34) gives us no information on (7.30), the estimator given by (7.33) is to be preferred. We hence denote (7.33) as the empirical null operator norm criterion, ONC_n^2,

$$\mathrm{ONC}_n^2(\mathcal{H}_k, \mathcal{H}_0, (x_i, y_i)_{i=1}^n) = (n)_2^{-1} \sum_{(i,j)\in i_2^n} \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\}_i \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{H})\}_{ij} \{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\}_j.$$

The empirical null operator norm criterion may be seen as an extension of the unbiased estimate of the operator norm criterion. In calculating ONC_n^2 as opposed to ONC_u^2, we make the replacements

$$\begin{aligned} \boldsymbol{K} &\leftarrow (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H}) \quad \text{and,} \\ \boldsymbol{y} &\leftarrow (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}. \end{aligned} \tag{7.35}$$

We now look to adapt the PIC and CIC to allow for a broader null space.

7.3.7 Parameter Selection with Parametric Null

Smoothing parameter selection may be achieved by adaptation of the above argument. It turns out that we simply need to replace $\boldsymbol{K}$ and $\boldsymbol{y}$ with their corresponding projections (7.35). This approach is applied in the selection of λ, either for the MSE or the MSPE. By Lemma 5.1, the optimised fit with null space (7.29) obeys $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$, where

$$\boldsymbol{S} = \boldsymbol{H} + (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H}) + \lambda\boldsymbol{I}\}^{-1}.$$

Therefore, we have residuals to (7.29),

$$\boldsymbol{y} - \widehat{\boldsymbol{y}} = \left(\boldsymbol{I} + (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H}) + \lambda\boldsymbol{I}\}^{-1}\right)(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}.$$

These residuals may be compared to the residuals of the fit without a null space (7.11),

$$\boldsymbol{y} - \widehat{\boldsymbol{y}} = \left(\boldsymbol{I} + \boldsymbol{K}\{\boldsymbol{K} + \lambda\boldsymbol{I}\}^{-1}\right)\boldsymbol{y}.$$

For a null space, it is clear that the residuals may be calculated by the use of the replacements in (7.35). Theorem 7.27 is a generalisation of Theorem 7.23 to allow for null space.

Theorem 7.27. *With homoscedastic variance σ^2, let* $\mathrm{SURE}(\tau)$ *be Stein's unbiased risk estimator of* (7.14)*, with f given by* (7.29)*. Then*

$$\lim_{\tau\to 0^+} \frac{d}{d\tau}\mathrm{SURE}(\tau) = -2\left\{\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y} - \sigma^2\mathrm{tr}\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\}\right\}.$$

Proof of Theorem 7.27 is given in Appendix 7.A.2. The following Theorem, an extension of Theorem 7.23, is proven in Appendix 7.A.2.

Theorem 7.28. *Let $\mathcal{H}_k$ be an RKHS with null space $\mathcal{H}_0$, and linear operator P_1 the projection onto $\mathcal{H}_0^{\perp}$. Moreover, let f be the solution to the optimisation problem*

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda_N \left\|P_1 f\right\|_{\mathcal{H}_k}\right\}, \tag{7.36}$$

where

$$\lambda_N = 2\sigma\sqrt{\mathrm{tr}\{(I-H)K(I-H)\}}, \tag{7.37}$$

with H *given by* (7.3.6). *Then if*

$$y^\top(I-H)K(I-H)y - \sigma^2\mathrm{tr}\{(I-H)K(I-H)\} \leq 0,$$

we have $f = f_\infty$. *Otherwise* $f \neq f_\infty$, *and*

$$(y-\widehat{y})^\top(I-H)K(I-H)(y-\widehat{y}) = \sigma^2\mathrm{tr}\{(I-H)K(I-H)\}.$$

For RKHS $\mathcal{H}_k$, null space $\mathcal{H}_0$ and data $(x_i, y_i)_{i=1}^n$, we denote the parametric criterion, $\mathrm{PIC}_{\mathrm{MSPE}}(\mathcal{H}_k, \mathcal{H}_0, (x_i, y_i)_{i=1}^n)$ as the optimiser of (7.36)–(7.37). Similarly, we denote the curved parametric criterion, $\mathrm{CIC}_{\mathrm{MSPE}}(\mathcal{H}_k, \mathcal{H}_0, (x_i, y_i)_{i=1}^n)$ as the optimiser of (7.36) with

$$\lambda_N = \frac{2\sigma^2\mathrm{tr}\{(I-H)K(I-H)\}}{\sqrt{y^\top(I-H)K(I-H)y}}.$$

It is clear that in allowing for a null space, we find λ by replacement of K with $(I-H)K(I-H)$, and of y with $(I-H)y$.

We now consider the MSE case. Denote by $\mathrm{PIC}_{\mathrm{MSE}}(\mathcal{H}_k, \mathcal{H}_0, (x_i, y_i)_{i=1}^n)$ the parameter information criterion with respect to RKHS, $\mathcal{H}_k$, null space $\mathcal{H}_0$ and data $(x_i, y_i)_{i=1}^n$. Similarly, denote the curved information criterion, $\mathrm{CIC}_{\mathrm{MSE}}(\mathcal{H}_k, \mathcal{H}_0, (x_i, y_i)_{i=1}^n)$ For both PIC and CIC, if $\mathrm{ONC}_n^2(\mathcal{H}_k, \mathcal{H}_0, (x_i, y_i)_{i=1}^n) \leq 0$, we take $\lambda = \infty$. For the PIC, we seek a maximum $\lambda \in [0, \infty]$ such that

$$(y-\widehat{y})^\top K(y-\widehat{y}) \leq \sum_{i=1}^n (y-\widehat{y})_i^2\{(I-H)K(I-H)\}_{ii}.$$

For the CIC, we seek maximum λ such that

$$\begin{aligned}&(y-\widehat{y})^\top K(y-\widehat{y})y^\top(I-H)K(I-H)y \\ &\quad\leq \left\{\sum_{i=1}^n (y_i-\widehat{y}_i)^2\{(I-H)K(I-H)\}_{ii}\right\}\left\{\sum_{i=1}^n \{(I-H)y\}_i^2\{(I-H)K(I-H)\}_{ii}\right\}.\end{aligned}$$

For both the PIC and CIC, if $\mathrm{ONC}_n^2(\mathcal{H}_k, \mathcal{H}_0, (x_i, y_i)_{i=1}^n) \leq 0$, we have $\lambda = \infty$.

7.3.8 Computational Issues

For our experiments we would like to solve a variety of optimisation problems. The optimisation of the RKHS norm penalisation problem (7.18) has been considered in Yuan and Lin (2006). Here we provide some simple algorithms for calculating PIC and CIC; more sophisticated algorithms with guaranteed run time and accuracy are beyond the scope of this chapter.

For PIC_{MSPE}, we search for λ such that

$$(\boldsymbol{y}-\widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})=\sigma^2\text{tr}(\boldsymbol{K}).$$

Rearranging gives,

$$\lambda = d(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a})^{-1/2},$$

where $d=\sigma\left\{\text{tr}(\boldsymbol{K})\right\}^{1/2}$, and $\boldsymbol{a}=(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y}$. On initialising λ, the successive approximation method then gives

$$\begin{aligned}\boldsymbol{a} &\leftarrow (\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y}\\ \lambda &\leftarrow d(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a})^{-1/2}\end{aligned} \tag{7.38}$$

until convergence in $\boldsymbol{a}$. We may not obtain convergence in λ, due to the possibility that $\lambda\rightarrow\infty$. For CIC_{MSPE}, we search for λ such that

$$(\boldsymbol{y}-\widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})\boldsymbol{y}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{y}=\left\{\sigma^2\text{tr}(\boldsymbol{K})\right\}^2.$$

That is,

$$\lambda = d(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a})^{-1/2},$$

where $d=\sigma^2\text{tr}(\boldsymbol{K})\left\{\text{tr}(\boldsymbol{K})\right\}^{-1/2}$. The successive approximations of (7.38) then follow.

For PIC_{MSE}, we seek λ such that

$$(\boldsymbol{y}-\widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})=\sum_{i=1}^{n}y_i^2K_{ii}.$$

That is,

$$\lambda \leftarrow d\left\{\frac{\sum_{i=1}^{n}(\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2K_{ii}}{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}\right\}^{1/2}, \tag{7.39}$$

where $d=1$. For PIC_{MSE}, we seek

$$(\boldsymbol{y}-\widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})\boldsymbol{y}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{y}=\left\{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2K_{ii}\right\}\left\{\sum_{i=1}^{n}y_i^2K_{ii}\right\}.$$

Rearrangement yields (7.39) with

$$d\leftarrow\left\{\frac{\boldsymbol{y}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{y}}{\sum_{i=1}^{n}(y_i^2K_{ii})}\right\}^{1/2}.$$

Successive approximation yields the recursive

$$\begin{aligned}\boldsymbol{a} &\leftarrow (\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y}\\ \lambda &\leftarrow d\left\{\frac{\sum_{i=1}^{n}(\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2K_{ii}}{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}\right\}^{1/2}\end{aligned}$$

until convergence in $\boldsymbol{a}$.

Name	$K_1(u)$
Uniform	$\frac{1}{2}\,1_{(\lvert u\rvert\le 1)}$
Triangle	$(1-\lvert u\rvert)\,1_{(\lvert u\rvert\le 1)}$
Epanechnikov	$\frac{3}{4}(1-u^2)\,1_{(\lvert u\rvert\le 1)}$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$
Laplacian	$\frac{1}{\sqrt{2}}\exp\left(-\lvert u\rvert\right)$

Table 7.2. *Examples of commonly used smoother kernels, with dimension* $d = 1$.

The extension to a broad null hypothesis is achieved by the replacement of $\boldsymbol{K}$ and $\boldsymbol{y}$ with $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})$ and $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$ in the successive approximations. On convergence, we have appropriate λ for the optimisation (7.29). Appendix 7.A.4 contains brief pseudo-code for optimisation. Algorithm 7.1 gives the pseudo-code for both the $\mathrm{PIC}_{\mathrm{MSPE}}$ and $\mathrm{CIC}_{\mathrm{MSPE}}$, and Algorithm 7.2 gives the pseudo-code for the MSE alternatives.

7.4 Relationship to Existing Methods

In this section, we examine some of the tests based on the estimation of the regression function.

7.4.1 Smoother Kernel-Based Tests

Known in the literature as kernel-based tests, smoother kernel-based tests employ the use of smoother kernels

Definition 7.29. *A function* $K_h(\cdot)\colon \mathbb{R}^d \to \mathbb{R}$ *is called a* ***smoother kernel*** *if it is bounded with*

$$K_1(u) = K_1(-u), \quad K_1(u) \ge 0, \text{ for all } u \in \mathbb{R}^d, \quad \text{and } \int_{u\in\mathbb{R}^d} K_1(u)du = 1.$$

The smoother kernels have parameter, h, called the ***bandwidth****, with* $K_h(u) = h^{-d}K_1(uh^{-1})$.

With $d = 1$, examples of smoother kernels are given in Table 7.2. We note that the Gaussian and Laplacian smoother kernels have analogs in the positive definite Gaussian and Laplacian kernels. That is, if K_h is either the Gaussian or Laplacian smoother kernel, and if $k(s,t) = K_h(s-t)$, then k is a Gaussian or Laplacian kernel, respectively. Smoother kernels are widely used in kernel density estimation (e.g., Silverman, 1986; Wand and Jones, 1995). Several authors, such as Härdle and Mammen (1993); Fan and Li (1996)

and Zheng (1996) have used kernel density estimation as an intermediate step toward producing model adequacy criteria.

We now briefly review two tests based on smoother kernels:

i) The test in Zheng (1996).

ii) The test in Fan and Li (2000).

The test for H_0 versus H_1 established in Zheng (1996) employs the Nadaraya-Watson kernel estimator of $g(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^d$ given by

$$\widehat{g}(\boldsymbol{x}) = \frac{(1/n)\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i)\, y_i}{(1/n)\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i)},$$

where $K(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a smoother kernel and $h = h_n \to 0$ is a bandwidth parameter.

Zheng (1996) suggests the test statistic

$$U_h = \frac{1}{n(n-1)} \sum_{i \neq j} u_i K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)\, u_j,$$

where K_h is a kernel density smoother with bandwidth parameter, h. Zheng (1996) derives U_h as a biased estimate of $\mathsf{E}_{x,y}\{y \mathsf{E}_{y|x}(y|x) f_x(x)\}$, and requires $h \to 0^+$ to ensure consistency. The rate of convergence of U_h is $O((nh)^{-1/2})$, and is slower than the optimal rate of $O(n^{-1/2})$. Fan and Li (2000) shows that, with a special class of smoother kernels, the test statistic of Zheng (1996) does not require $h \to 0^+$. The use of so-called "fixed bandwidth" smoother kernels dates back to Anderson, Hall and Titterington (1994). Fan and Li (2000, Lemma 2.2) implies the following theorem.

Theorem 7.30. *For the test based on U_h with a fixed h to be consistent, the Fourier transform of $K_h(\cdot)$, denoted $\bar{K}_h(\cdot)$, must be such that there exists a compact subset Θ of $\mathbb{R}^d$, containing the origin such that $\bar{K}_h(th)$ vanishes outside Θ and the set $\{t \in \Theta : \bar{K}_h(th) \leq 0\}$ has Lebesgue measure zero.*

We recognise the conditions of Theorem 7.30 as being sufficient for $k(s,t) \equiv K_h(s-t)$ to be a universal kernel on compact $\mathcal{X}$ (Micchelli, Xu and Zhang, 2006, Proposition 16). Some differences are identified between the use of the ONC and the tests of Fan and Li (2000). These include:

i) The allowance for kernels that do not conform to kernel smoothers.

ii) The ability to test against parametric alternative hypotheses.

iii) Testing on general, separable domains $\mathcal{X}$.

iv) The invariance adjustment of (7.33).

Many of these benefits come from the use of kernel methods. There have been other recent applications of kernel methods to some traditional statistical problems.

7.4.2 Testing for Covariance

In this section, we show a relationship between the ONC and some recent developments in the Machine Learning literature. Borgwardt *et al.* (2006) and Gretton *et al.* (2008a) use kernel methods hypothesis testing in the two sample problem. Kernel methods are used for independence testing by Song *et al.* (2007); Song (2008) and Gretton *et al.* (2008b).

It is often desired that our model be invariant under changes of location. Like Problems 7.1 and 7.9, we have the following.

Problem 7.31. *Given a feature space, $\mathcal{F}$, consisting of functions $\mathcal{X} \to \mathbb{R}$, does*

$$\mathrm{Cov}(y, f(x)) = 0 \quad \textit{for all } f \in \mathcal{F}?$$

Problem 7.31 is of interest when testing whether the conditional mean of the response variable, y, can be explained in part through functions in the function space, $\mathcal{F}$. A similar hypothesis test is

$$\begin{aligned} H_0 &: \mathrm{Var}_y(y^2) = \min_{f \in \mathcal{F}} \mathrm{Var}_{x,y} \left\{ y - f(x) \right\}^2, \text{ against,} \\ H_1 &: \mathrm{Var}_y(y^2) > \min_{f \in \mathcal{F}} \mathrm{Var}_{x,y} \left\{ y - f(x) \right\}^2. \end{aligned} \tag{7.40}$$

We will focus here on Problem 7.31, as opposed to the hypothesis test in (7.40).

Problem 7.32. *Can we test if*

$$\mathrm{Cov}(y, f(x)) = 0 \quad \textit{for all measureable } f \colon \mathcal{X} \to \mathbb{R}?$$

In answering Problems 7.31 and 7.32, we define the following extension of the ONC.

Definition 7.33. *The **operator norm covariance criterion** (ONCC) of the joint distribution, $P_{x,y}$, with respect to the Banach space, $\mathcal{B}$ is*

$$\mathrm{ONCC}\left(\mathcal{B}, P_{x,y}\right) = \sup_{\|f\|_{\mathcal{B}} \leq 1} \mathsf{E}_{x,y} \left\{ (y - \mathsf{E}_y y) f(x) \right\},$$

where such a supremum exists.

A biased estimate of the $\mathrm{ONCC}\left(\mathcal{B}, P_{x,y}\right)$ is then given by

$$\mathrm{ONCC}\left(\mathcal{B}, P_{(x_i,y_i)_{i=1}^n}\right) = \sup_{\|f\|_{\mathcal{B}} \leq 1} n^{-1} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} y_j / n \right) f(x_i).$$

Assuming the Banach space is a reproducing kernel Hilbert space, we have the following simplification.

Theorem 7.34. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$ and $\mathcal{H}_k$ a separable RKHS on $\mathcal{X}$. Then if $\mathsf{E}_{x,y,x',y',y'',y'''}\{(y-y'')k(x,x')(y'-y''')\} < \infty$, we have*

$$\mathrm{ONCC}^2(\mathcal{H}_k, P_{x,y}) = \mathsf{E}_{x,y,x',y',y'',y'''}\{(y-y'')k(x,x')(y'-y''')\}.$$

Proof of Theorem 7.34 is given in Appendix 7.A.3. Where u is a random variable with $u = y - \mathsf{E}_y(y)$, the ONCC is simply related to the ONC via

$$\mathrm{ONCC}(\mathcal{H}_k, P_{x,u}) = \mathrm{ONC}(\mathcal{H}_k, P_{x,y}). \tag{7.41}$$

As with the operator norm criterion, we have an unbiased estimator of the operator norm covariance criterion.

Theorem 7.35. *Suppose that $\mathsf{E}_{x,y,x',y',y'',y'''}\{(y-y'')k(x,x')(y'-y''')\} < \infty$. Then given random i.i.d. samples $(x,y)_{i=1}^n$, with $n \geq 4$, an unbiased empirical estimate of $\mathrm{ONCC}^2(\mathcal{H}_k, P_{x,y})$ is given by*

$$\begin{aligned}&\mathrm{ONCC}_u^2(\mathcal{H}_k, (x_i,y_i)_{i=1}^n)\\ &\quad= (n)_2^{-1}\sum_{(i,j)\in i_2^n} y_i k(x_i,x_j) y_j - 2(n)_3^{-1}\sum_{(i,j,k)\in i_3^n} y_i k(x_i,x_j) y_k + (n)_4^{-1}\sum_{(i,j,k,l)\in i_4^n} y_k k(x_i,x_j) y_l.\end{aligned}$$

Proof of Theorem 7.35 is given in Appendix 7.A.3. Denote by $\overline{y}$ the vector of length n with terms $y^\top \mathbf{1}/n$. The following theorem shows that ONCC_u^2 may be easily calculated.

Theorem 7.36. *Denote by $\widetilde{K}$ the $n \times n$ matrix with $\widetilde{K}_{ij} = K_{ij}$ for $i \neq j$ and $\widetilde{K}_{ii} = 0$ for all $1 \leq i,j \leq n$. Then the unbiased empirical estimate, ONCC_u^2, may be calculated in $O(n^2)$ by*

$$\mathrm{ONCC}_u^2(\mathcal{H}_k, (x_i,y_i)_{i=1}^n) = \frac{1}{n(n-3)}\left\{(y-\overline{y})^\top \widetilde{K}(y-\overline{y}) - \frac{1}{n-2}\mathbf{1}^\top \widetilde{K} v\right\},$$

where $v = -2(y-\overline{y}) \odot (y-\overline{y}) + \mathbf{1}\frac{(y-\overline{y})^\top (y-\overline{y})}{n-1}$.

Theorem 7.36 is proven in Appendix 7.A.3. Knowledge of the ONCC can provide answers to Problems 7.31 and 7.32. As the ONCC is an operator norm, we have the following.

Theorem 7.37. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, and $\mathcal{H}_k$ an RKHS, such that $\mathsf{E}_{x,y,x',y',y'',y'''}\{(y-y'')k(x,x')(y'-y''')\} < \infty$. Then*

$$\mathrm{Cov}(y, f(x)) = 0 \quad \textit{for all } f \in \mathcal{H}_k$$

if and only if

$$\mathrm{ONCC}(\mathcal{H}_k, P_{x,y}) = 0.$$

The hypothesis tests associated with Problems 7.31 and 7.32 may then be given as hypothesis tests concerning $\mathrm{ONCC}(\mathcal{H}_k, P_{x,y}) = 0$. Of interest then is the empirical quantity ONCC_u^2.

For the case $y_i \in \{-1, 1\}$, with $-1 < \mathsf{E}_{x,y}(y) < 1$, the maximum mean discrepancy (MMD), given by Borgwardt *et al.* (2006), is

$$\mathrm{MMD}(\mathcal{H}_k, P_{x,y}) \equiv \left[\mathsf{E}_{\substack{x|y=-1, \\ x'|y'=-1}} \{k(x,x')\} - 2\mathsf{E}_{\substack{x|y=-1, \\ x'|y'=1}} \{k(x,x')\} + \mathsf{E}_{\substack{x|y=1, \\ x'|y'=1}} \{k(x,x')\} \right]^{1/2}.$$

We have the following theorem showing a close relationship between the maximum mean discrepancy and the operator norm covariance criterion.

Theorem 7.38. *Let $P_{x,y}$ be a probability distribution on $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{Y} = \{-1, 1\}$, and $-1 < \mathsf{E}_{x,y}(y) < 1$. Then the maximum mean discrepancy and the operator norm covariance criterion satisfy*

$$\mathrm{ONCC}(\mathcal{H}_k, P_{x,y}) = \frac{\mathrm{Var}_y(y)}{2} \mathrm{MMD}(\mathcal{H}_k, P_{x,y}).$$

Proof of Theorem 7.38 is given in Appendix 7.A.3. For universal k, it was shown by Borgwardt *et al.* (2006) that MMD $= 0$ if and only if $P_{x|y=1} = P_{x|y=-1}$.

The Hilbert Schmidt independence criterion (HSIC) of Gretton *et al.* (2008b) may be derived via the maximum mean discrepancy. With respect to RKHSs $\mathcal{H}_k$ on $\mathcal{X}$ and $\mathcal{H}_l$ on $\mathcal{Y}$, and to probability distribution $P_{x,y}$ on $\mathcal{X} \times \mathcal{Y}$, the Hilbert-Schmidt independence criterion, $\mathrm{HSIC}(\mathcal{H}_k, \mathcal{H}_l, P_{x,y})$ is given by

$$\mathrm{HSIC}(\mathcal{H}_k, \mathcal{H}_l, P_{x,y}) = \left[\mathsf{E}_{x,y,x',y',y'',y'''} k(x,x') \{l(y,y') - 2l(y,y'') + l(y'',y''')\} \right]^{1/2}.$$

For universal kernels k and l, Gretton *et al.* (2008b) shows that $\mathrm{HSIC}(\mathcal{H}_k, \mathcal{H}_l, P_{x,y}) = 0$ if and only if x and y are independent. For $\mathcal{Y} = \mathbb{R}$ and linear kernel, $l(y,y') = yy'$, it is clear that $\mathrm{ONCC}(\mathcal{H}_k, P_{x,y}) = \mathrm{HSIC}(\mathcal{H}_k, \mathcal{H}_l, P_{x,y})$.

7.5 Experiments

We have conducted an extensive set of simulation experiments using data obtained from a variety of source distributions. The experiments used Gaussian, Laplacian and polynomial kernels. The mean structures were generated using Gaussian processes. We tested Gaussian errors, double exponential, centred exponential and Rademacher random variables. We also studied heteroscedasticity (the violation of homoscedasticity), and changes in the signal-to-noise ratio. Sample sizes of both 20 and 100 were used.

Broader null hypothesis are tested with $\mathcal{H}_0 = \mathbb{R}$. A colloquialism of Machine Learning literature is that fits made with such null space are referred to as having "offset", b. Finally, we considered the performance of the algorithms under both the null and alternative hypotheses.

Comparisons were made with the algorithms LOO, SURE, ML and REML. Both LOO and SURE were calculated with grid search, followed by a finer grid search around the optimal. ML and REML were calculated using the "Gaussian dual optimisation" of Algorithm 4.3.

7.5.1 Experimental Setup

For Gaussian processes, the response variables may be simulated from a multivariate normal distribution. The covariance structure is given by

$$\operatorname*{Cov}_{\substack{y|x \\ y'|x'}}(y, y') = k(x, x'), \quad \text{and} \quad \operatorname{Var}_{y|x}(y) = k(x, x) + \sigma^2.$$

Equivalently, we may have $\boldsymbol{y} \sim \boldsymbol{w} + \boldsymbol{\varepsilon}$, for $1 \leq i \leq n$, where

$$\operatorname*{Cov}_{\substack{w|x \\ w'|x'}}(w, w') = k(x, x'), \quad \text{and} \quad \operatorname{Var}_{w|x}(w) = k(x, x), \tag{7.42}$$

and

$$\varepsilon_i \overset{\text{ind.}}{\sim} \mathrm{N}(0, \sigma^2). \tag{7.43}$$

Figure 7.3 shows a simulation of the mean structure given by (7.42). The structure of (7.42)–(7.43) allows for a non-Gaussian error structure. For example, with Rademacher errors we have

$$P(\varepsilon_i = -1) = P(\varepsilon_i = 1) = 0.5, \quad \text{independently for all } 1 \leq i \leq n,$$

We also use double exponential and recentred (mean zero) exponential errors.

Table 7.3 gives the particulars of the models used in the experiments. Some models (a–j) test the performance of model criteria under the null hypothesis, i.e., $\boldsymbol{w} = \mathbf{0}$. Other models (k–o) are given in the the Gaussian processes in (7.42)–(7.43). Others have non-Gaussian errors (f,g,h,p,q,r), or heteroscedasticity (i,j,s,t). For the heteroscedastic models, the MSPE criteria were calculated with the average error variance, setting $\sigma^2 = 1$.

7.5.2 Mean Squared Prediction Error

For the MSPE, the results are displayed as

$$n^{-1}\mathrm{MSPE}.$$

model	kernel	γ	errors	$\mathrm{Var}(y\|x)$	hypothesis
a	Gaussian	1	Normal	1	H_0
b	Gaussian	5	Normal	1	H_0
c	Gaussian	25	Normal	1	H_0
d	Laplacian	5	Normal	1	H_0
e	Cubic Poly.	1/6	Normal	1	H_0
f	Gaussian	5	Double Exp.	1	H_0
g	Gaussian	5	Exponential	1	H_0
h	Gaussian	5	Rademacher	1	H_0
i	Gaussian	5	Normal	x	H_0
j	Gaussian	5	Normal	$x^2/\sqrt{3}$	H_0
k	Gaussian	1	Normal	1	H_1
l	Gaussian	5	Normal	1	H_1
m	Gaussian	25	Normal	1	H_1
n	Laplacian	5	Normal	1	H_1
o	Cubic Poly.	1/6	Normal	1	H_1
p	Gaussian	5	Double Exp.	1	H_1
q	Gaussian	5	Exponential	1	H_1
r	Gaussian	5	Rademacher	1	H_1
s	Gaussian	5	Normal	x	H_1
t	Gaussian	5	Normal	$x^2/\sqrt{3}$	H_1
u	Gaussian	5	Normal	0.01	H_1
v	Gaussian	5	Normal	100	H_1

Table 7.3. *Gaussian processes and other models used in the simulation: (a-f) Independent, identically distributed Gaussian noise with a variety of kernels; (f) double exponential; (g) exponential noise, with mean zero and variance one; (h) Rademacher noise; (i-j) heteroscedastic noise; (k-o) Gaussian processes, with a variety of covariance structures; (p-r) alternative error distributions; (s-t) Gaussian processes with heteroscedastic errors; (u) high signal-to-noise ratio (v) low signal-to-noise ratio.*

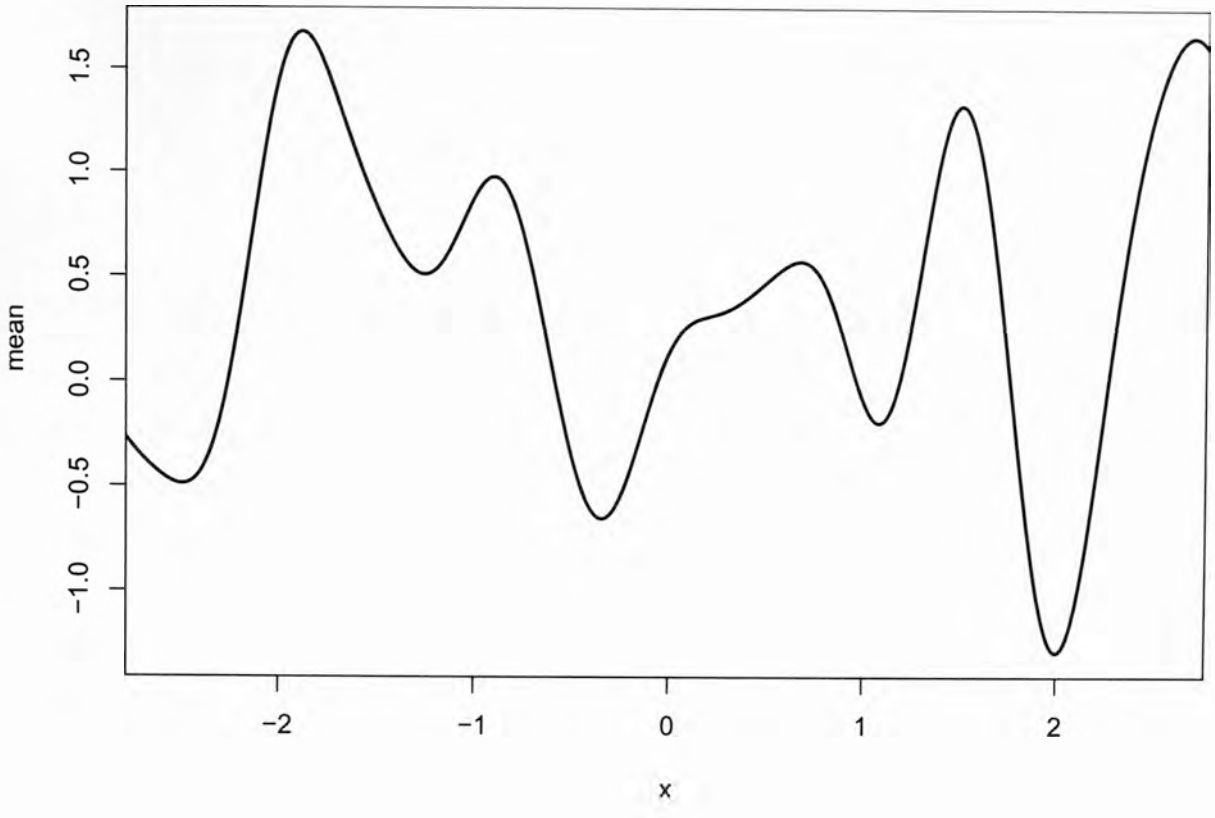


Figure 7.3. *Sample mean path for a Gaussian process. We have Gaussian kernel with* $\gamma = 5$. *The covariance structure is given by* (7.42).

The scaling by n^{-1} was made to ease comparisons across differing sample sizes. The results are shown in Table 7.4 and in Table 7.5. Table 7.5 incorporates the offset.

Under H_0, it is seen that PIC provides a very clear and significant improvement over SURE, ML/REML, and CIC. For the Gaussian processes (a–e), the performance of CIC is similar to that of ML/REML, similar or better under the non-Gaussian errors (g–h), and clearly better when there are violations of homoscedasticity (i–j). Under the H_1 models, we see that ML outperforms the alternatives for all of (k–r). For heteroscedasticity (s–t), we see a preference for PIC, then CIC. Model (v) shows a preference toward PIC. SURE was consistently outperformed by both ML/REML and by CIC. Such results were consistent across changes in sample size, and with and without offset.

We now specifically compare ML/REML and CIC. For Gaussian processes they have a similar performance, albeit with a slight preference for ML/REML. The largest differences in performance are seen with the heteroscedastic models (g,h,s,t). With each, we see a strong preference toward CIC.

model	SURE	ML	PIC	CIC
a	0.040	0.019	**0.006**	0.018
b	0.032	0.021	**0.007**	0.020
c	0.025	0.023	**0.007**	0.023
d	0.023	0.023	**0.007**	0.023
e	0.030	0.019	**0.007**	0.018
f	0.097	0.043	**0.012**	0.034
g	0.136	0.055	**0.018**	0.048
h	0.013	0.017	**0.006**	0.018
i	0.201	0.130	**0.024**	0.067
j	1.875	1.634	**0.670**	1.272
k	0.243	**0.200**	0.248	0.206
l	0.317	**0.286**	0.342	0.293
m	0.402	**0.377**	0.431	0.383
n	0.436	**0.430**	0.479	0.441
o	0.157	**0.144**	0.209	0.149
p	0.411	**0.318**	0.362	0.319
q	0.409	**0.294**	0.351	0.297
r	0.295	**0.286**	0.355	0.293
s	0.475	0.408	**0.363**	0.386
t	2.228	1.993	**1.023**	1.725
u	0.006	**0.006**	0.015	0.006
v	5.298	3.520	**1.696**	3.477

model	SURE	ML	PIC	CIC
a	0.011	0.006	**0.002**	0.006
b	0.008	0.006	**0.002**	0.006
c	0.007	0.007	**0.002**	0.007
d	0.007	0.007	**0.002**	0.008
e	0.007	0.006	**0.002**	0.005
f	0.010	0.007	**0.002**	0.007
g	0.016	0.007	**0.002**	0.006
h	0.006	0.005	**0.002**	0.005
i	0.249	0.071	**0.005**	0.015
j	1.969	1.302	**0.238**	0.582
k	0.069	**0.057**	0.103	0.059
l	0.115	**0.106**	0.164	0.107
m	0.178	**0.172**	0.242	0.174
n	0.265	**0.259**	0.340	0.276
o	0.037	**0.035**	0.076	0.036
p	0.120	**0.109**	0.164	0.111
q	0.134	**0.107**	0.157	0.108
r	0.106	**0.102**	0.160	0.104
s	0.327	0.188	**0.153**	0.176
t	2.050	1.421	**0.419**	1.052
u	0.002	**0.002**	0.008	0.002
v	1.567	1.426	**1.026**	1.449

Table 7.4. *Comparison of model performance for mean squared prediction error case, without offset. Left:* $n = 20$. *Right:* $n = 100$.

model	SURE	REML	PIC	CIC
a	0.089	0.069	**0.055**	0.068
b	0.075	0.070	**0.055**	0.070
c	0.074	0.071	**0.054**	0.071
d	0.069	0.072	**0.054**	0.072
e	0.080	0.066	**0.053**	0.063
f	0.142	0.090	**0.061**	0.083
g	0.182	0.102	**0.067**	0.095
h	0.058	0.063	**0.052**	0.065
i	0.240	0.188	**0.086**	0.147
j	1.918	1.724	**0.894**	1.502
k	0.247	**0.212**	0.233	0.216
l	0.332	**0.302**	0.341	0.308
m	0.420	**0.397**	0.440	0.403
n	0.449	**0.446**	0.483	0.454
o	0.161	**0.149**	0.172	0.151
p	0.413	0.327	0.349	**0.326**
q	0.418	**0.308**	0.342	**0.308**
r	0.305	**0.295**	0.342	0.301
s	0.496	0.438	**0.378**	0.421
t	2.263	2.057	**1.203**	1.855
u	0.006	**0.006**	0.014	0.006
v	9.569	8.378	**6.690**	8.318

model	SURE	REML	PIC	CIC
a	0.021	0.016	**0.013**	0.016
b	0.019	0.017	**0.013**	0.017
c	0.017	0.018	**0.013**	0.018
d	0.017	0.018	**0.013**	0.018
e	0.017	0.016	**0.013**	0.015
f	0.021	0.017	**0.012**	0.017
g	0.026	0.016	**0.011**	0.015
h	0.015	0.016	**0.013**	0.017
i	0.255	0.090	**0.019**	0.037
j	1.972	1.333	**0.338**	0.737
k	0.070	**0.060**	0.087	0.061
l	0.116	**0.107**	0.154	0.109
m	0.180	**0.174**	0.238	0.176
n	0.267	**0.261**	0.333	0.276
o	0.037	**0.036**	0.058	0.037
p	0.121	**0.110**	0.155	0.112
q	0.137	**0.109**	0.151	0.111
r	0.107	**0.104**	0.151	0.106
s	0.328	0.197	**0.154**	0.185
t	2.050	1.443	**0.500**	1.093
u	0.002	**0.002**	0.007	0.002
v	2.350	2.232	**1.858**	2.245

Table 7.5. *Comparison of model performance for mean squared prediction error case, with offset. Left:* $n = 20$. *Right:* $n = 100$.

7.5.3 Mean Squared Error

For the MSE case, there are optimal fits to which we may compare the MSE. If we knew the w in (7.42), we would be able to make optimal fits, that would on average have lower MSE than any fit based on y. The results are displayed as

$$\text{MSE} - \text{MSE}_{\text{optimal}},$$

where $\text{MSE}_{\text{optimal}}$ is the mean squared error of the optimal fit using knowledge of w, and where appropriate, the offset. For models (a–j), we have $\text{MSE}_{\text{optimal}} = 1$, the Bayes optimal error rate.

Results for the MSE case are given in Tables 7.7 and 7.6. Under H_0, with sample size 20, we see a clear ordering of the performances from PIC to CIC to ML/REML and to LOO. For sample sizes 100, we see again a consistent preference for PIC. For the Gaussian processes (k–o), there are similar performances by ML/REML and CIC.

For $n = 20$, we see that CIC tends towards more conservative models than those of ML/REML. This is apparent in the strong performance of CIC under H_0, and the comparative degradation in performance in the low-noise model (u). Like with MSPE, we see the largest disparity between ML/REML and CIC is with heteroscedasticity (i,j,s,t). Both PIC and CIC perform well in such situations.

7.6 Discussion

By using universal kernels we have generalised the results of both the nonparametric model estimation, and nuisance parameters methods. Methods such as Bierens (1982); Zheng (1996); Bierens and Ploberger (1997) and Fan and Li (2000), amongst others, are special cases of our tests. We have also extended the range of admissible kernels by using bivariate, universal kernels. Moreover, in choosing a low-rank kernel, we may test a parametric alternative hypothesis.

By ensuring that model residuals appear to be mean zero, the parametric criterion emerges naturally. The operator norm criterion allows for heteroscedasticity, and under such heteroscedasticity we see an improvement in the performance of the PIC in comparison to ML/REML. The CIC offers sharper, less smooth fits than that of PIC. There is an overall strong performance by the CIC. The robustness of the PIC has been promising; the derivation of the PIC required very little assumptions. In light of the demonstrated convexity, we find good reason to recommend their widespread use.

model	LOO	ML	PIC	CIC
a	6.744	0.038	**0.012**	0.029
b	0.104	0.041	**0.013**	0.033
c	0.104	0.047	**0.014**	0.034
d	0.055	0.047	**0.014**	0.036
e	0.189	0.031	**0.010**	0.023
f	0.232	0.044	**0.011**	0.028
g	0.354	0.053	**0.015**	0.036
h	0.831	0.041	**0.014**	0.033
i	0.418	0.170	**0.018**	0.043
j	1.351	0.752	**0.070**	0.135
k	0.867	**0.245**	0.307	0.249
l	0.596	0.379	0.429	**0.368**
m	0.616	0.423	0.446	**0.406**
n	0.304	0.283	0.285	**0.271**
o	0.491	**0.243**	0.345	0.259
p	0.578	0.384	0.433	**0.375**
q	1.390	0.385	0.415	**0.363**
r	0.724	**0.367**	0.433	0.373
s	1.140	0.414	0.369	**0.343**
t	2.569	1.043	**0.483**	0.582
u	0.116	**0.037**	0.100	0.053
v	17.49	5.215	**2.044**	3.984

model	LOO	ML	PIC	CIC
a	0.014	0.006	**0.002**	0.006
b	0.007	0.006	**0.002**	0.006
c	0.007	0.007	**0.002**	0.007
d	0.008	0.007	**0.002**	0.008
e	0.014	0.005	**0.002**	0.005
f	0.010	0.005	**0.002**	0.005
g	0.054	0.010	**0.003**	0.007
h	0.006	0.007	**0.002**	0.007
i	0.191	0.107	**0.003**	0.009
j	1.018	0.971	**0.016**	0.044
k	1.110	**0.064**	0.116	0.068
l	0.143	0.116	0.177	**0.116**
m	0.215	**0.202**	0.283	**0.202**
n	0.225	**0.219**	0.295	0.224
o	0.054	**0.047**	0.109	0.055
p	0.149	**0.117**	0.178	0.118
q	1.719	**0.119**	0.185	0.121
r	0.152	**0.119**	0.182	0.121
s	0.772	0.215	0.167	**0.160**
t	4.554	1.394	**0.247**	0.407
u	0.019	**0.007**	0.021	0.009
v	2.146	1.276	**0.964**	1.284

Table 7.6. *Comparison of model performance for mean squared error case, without offset. Left:* $n = 20$. *Right:* $n = 100$.

model	LOO	REML	PIC	CIC
a	26.96	0.050	**0.017**	0.040
b	0.129	0.052	**0.015**	0.038
c	0.094	0.063	**0.016**	0.040
d	0.057	0.078	**0.018**	0.045
e	0.237	0.049	**0.016**	0.031
f	0.445	0.057	**0.014**	0.036
g	0.359	0.072	**0.018**	0.042
h	0.233	0.052	**0.018**	0.041
i	0.408	0.271	**0.032**	0.072
j	2.000	1.399	**0.235**	0.430
k	1.089	0.262	0.277	**0.255**
l	0.611	0.410	0.430	**0.397**
m	0.650	0.476	0.476	**0.453**
n	0.343	0.339	**0.310**	0.311
o	0.469	0.262	0.279	**0.258**
p	0.567	0.426	0.440	**0.412**
q	1.206	0.413	0.421	**0.391**
r	0.627	**0.403**	0.437	0.409
s	0.703	0.486	0.405	**0.388**
t	3.144	1.291	**0.643**	0.719
u	0.098	**0.047**	0.103	0.062
v	25.87	10.99	**7.453**	9.605

model	LOO	REML	PIC	CIC
a	51.96	0.009	**0.003**	0.008
b	0.008	0.007	**0.003**	0.007
c	0.007	0.007	**0.002**	0.007
d	0.009	0.008	**0.003**	0.008
e	0.016	0.007	**0.002**	0.005
f	0.014	0.007	**0.002**	0.006
g	0.096	0.014	**0.002**	0.006
h	0.006	0.007	**0.002**	0.007
i	0.241	0.181	**0.004**	0.011
j	2.596	2.223	**0.045**	0.134
k	1.160	**0.067**	0.100	0.070
l	0.145	**0.118**	0.167	0.119
m	0.219	0.207	0.279	**0.206**
n	0.230	**0.222**	0.288	0.228
o	0.057	**0.051**	0.085	0.057
p	0.153	**0.120**	0.170	0.121
q	0.227	**0.125**	0.180	0.126
r	0.146	**0.121**	0.173	0.123
s	0.835	0.229	0.179	**0.164**
t	5.176	1.527	**0.285**	0.375
u	0.026	**0.008**	0.020	0.009
v	2.931	2.047	**1.762**	2.053

Table 7.7. *Comparison of model performance for mean squared error case, with offset. Left:* $n = 20$. *Right:* $n = 100$.

For simplicity, our analysis has been limited to the use of reproducing kernel Hilbert spaces, as well as to the use of least squares loss. Some further research has so far been promising. Current research includes the use of operator norms for parameter selection for the "lasso", as well as for quantile regression and other alternative loss functions.

7.A Appendix

7.A.1 Proofs for Section 7.2

Proof of Theorem 7.3. The first part of the proof shows that $\mathrm{ONC}(\mathcal{B}, P_{x,y}) = 0$ implies $\mathcal{R}_{x,y}(0) = \inf_{f\in\mathcal{B}} \mathsf{E}_{x,y}\{y - f(x)\}^2$. For any $\varepsilon > 0$,

$$\begin{aligned}
\mathrm{ONC}(\mathcal{B}, P_{x,y}) &= \sup_{\|f\|_{\mathcal{B}}\le 1} \mathsf{E}_{x,y}\{yf(x)\} \\
&= \sup_{\|f\|_{\mathcal{B}}\le \varepsilon} \varepsilon^{-1}\mathsf{E}_{x,y}\{yf(x)\} \\
&\ge (2\varepsilon)^{-1} \sup_{\|f\|_{\mathcal{B}}\le \varepsilon} \mathsf{E}_{x,y}\{2yf(x) - f(x)^2\} \\
&= (2\varepsilon)^{-1}\left[\mathsf{E}_y y^2 - \inf_{\|f\|_{\mathcal{B}}\le\varepsilon} \mathsf{E}_{x,y}\{y - f(x)\}^2\right].
\end{aligned}$$

Hence, if $\mathrm{ONC}(\mathcal{B}, P_{x,y}) = 0$, taking $\varepsilon \to \infty$ yields

$$\left[\mathsf{E}_y y^2 - \inf_{f\in\mathcal{B}} \mathsf{E}_{x,y}\{y - f(x)\}^2\right] \le 0.$$

Noting that, clearly, $\inf_{f\in\mathcal{B}} \mathsf{E}_{x,y}\{y - f(x)\}^2 \le \mathsf{E}_y y^2$, we obtain

$$\mathsf{E}_y y^2 - \inf_{f\in\mathcal{B}} \mathsf{E}_{x,y}\{y - f(x)\}^2 = 0,$$

and $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y;\mathcal{B}}$.

With the second part of the proof, we wish to show that $\mathrm{ONC}(\mathcal{B}, P_{x,y}) > 0$ implies $\mathsf{E}_y y^2 > \inf_{f\in\mathcal{B}} \mathsf{E}_{x,y}\{y - f(x)\}^2$. Since $\sup_{\|f\|_{\mathcal{B}}\le 1} \mathsf{E}_x\{f(x)^2\} < \infty$, we may set $C = \sup_{\|f\|_{\mathcal{B}}\le 1} \mathsf{E}_x\{f(x)^2\}$. Moreover, since $\mathcal{R}_{x,y}(0) < \infty$, we have, for $0 < \varepsilon < 1$,

$$\begin{aligned}
\left[\inf_{\|f\|_{\mathcal{B}}\le 1} \mathsf{E}_{x,y}\{y - f(x)\}^2\right] - \mathsf{E}_y y^2 &= \inf_{\|f\|_{\mathcal{B}}\le 1} \mathsf{E}_{x,y}\{-2yf(x) + f(x)^2\} \\
&= -\sup_{\|f\|_{\mathcal{B}}\le 1} \mathsf{E}_{x,y}\{2yf(x) - f(x)^2\} \\
&\le -\sup_{\|f\|_{\mathcal{B}}\le 1} \mathsf{E}_{x,y}\{2yf(x) - C\} \\
&= -\sup_{\|f\|_{\mathcal{B}}\le \varepsilon} \mathsf{E}_{x,y}\left\{2\varepsilon^{-1}yf(x) - \varepsilon^{-2}C\right\} \\
&= -2\varepsilon^{-1}\mathrm{ONC}(\mathcal{B}, P_{x,y}) + \varepsilon^{-2}C.
\end{aligned}$$

For $\varepsilon = C\{\mathrm{ONC}(\mathcal{B}, P_{x,y})\}^{-1}$, we have $-2\varepsilon^{-1}\mathrm{ONC}(\mathcal{B}, P_{x,y}) + \varepsilon^{-2}C < 0$, and hence

$$\mathsf{E}_y y^2 > \inf_{f \in \mathcal{B}} \mathsf{E}_{x,y}\{y - f(x)\}^2],$$

and $\mathcal{R}_{x,y}(0) > \mathcal{R}^*_{x,y;\mathcal{B}}$. □

Proof of Theorem 7.4. As k is a kernel, there exists a feature map $\Phi\colon \mathcal{X} \to \mathcal{H}$ such that for all $x, x' \in \mathcal{X}$,

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle. \tag{7.44}$$

By the reproducing property, for $f \in \mathcal{H}_k$,

$$f(x) = \langle \Phi(x), f \rangle_{\mathcal{H}_k}. \tag{7.45}$$

Moreover, for $g \in \mathcal{H}_k$,

$$\sup_{\|f\|^2_{\mathcal{H}_k} \le 1} \langle g, f \rangle^2_{\mathcal{H}_k} = \langle g, g \rangle_{\mathcal{H}_k}. \tag{7.46}$$

Hence,

$$\begin{aligned}
\sup_{\|f\|^2_{\mathcal{H}_k} \le 1} [\mathsf{E}_{x,y}\{y f(x)\}]^2 &= \sup_{\|f\|^2_{\mathcal{H}_k} \le 1} \left[\mathsf{E}_{x,y}\left[y \langle \Phi(x), f \rangle_{\mathcal{H}_k}\right]\right]^2 && (7.47)\\
&= \sup_{\|f\|^2_{\mathcal{H}_k} \le 1} \left[\langle \mathsf{E}_{x,y}\{y\,\Phi(x)\}, f \rangle_{\mathcal{H}_k}\right]^2 \\
&= \langle \mathsf{E}_{x,y}\{y\,\Phi(x)\}, \mathsf{E}_{x,y}\{y\,\Phi(x)\} \rangle_{\mathcal{H}_k} && (7.48)\\
&= \langle \mathsf{E}_{x,y}\{y\,\Phi(x)\}, \mathsf{E}_{x',y'}\{y'\Phi(x')\} \rangle_{\mathcal{H}_k} \\
&= \mathsf{E}_{x,y,x',y'}\left\{\langle y\,\Phi(x), y'\Phi(x') \rangle_{\mathcal{H}_k}\right\} \\
&= \mathsf{E}_{x,y,x',y'}\left\{y \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}_k} y'\right\} \\
&= \mathsf{E}_{x,y,x',y'}\{y k(x, x') y'\}, && (7.49)
\end{aligned}$$

where (7.47), (7.48) and (7.49) are due to (7.44), (7.46) and (7.45), respectively. □

Proof of Theorem 7.5. The essence of the proof is to show that $\mathsf{E}_x\{k(x,x)\} < \infty$ implies $\sup_{\|f\|_{\mathcal{H}_k} \le 1} \mathsf{E}_x\{f(x)^2\} < \infty$. Following e.g., Dunford and Schwartz (1963), Serfling (1980, page 196) and Schölkopf *et al.* (1998), let $\lambda_1 \ge \ldots \ge 0$ be the solutions to the eigenvalue problem

$$\int_{\mathcal{X}} k(x, x')\psi_i(x) dP_x(x) = \lambda_i \psi_i(x').$$

Then k has an orthogonal expansion in $k(x, x') = \sum_i \Phi_i(x)\Phi_i(x')$, where we have $\Phi_i(x) = \lambda_i^{1/2}\psi_i(x)$. For f in the pre-Hilbert space given by $\mathrm{span}_i\{\Phi_i(x)\}$, we have $f = \sum_i a_i \Phi_i(x)$. Then $\mathsf{E}_x\{f(x)^2\} = \sum_i a_i^2 \lambda_i$, and $\|f\|_{\mathcal{H}_k} = \sum_i a_i^2$. Hence,

$$\sup_{\|f\|_{\mathcal{H}_k} \le 1} \mathsf{E}_x\{f(x)^2\} = \sup_{\sum_i a_i^2 \le 1} \sum_i a_i^2 \lambda_i = \lambda_1.$$

However, $\mathsf{E}_x\{k(x,x)\} = \mathsf{E}_x\{\lambda_i\psi_i(x)^2\} = \sum_i \lambda_i \geq \lambda_1 = \sup_{\|f\|_{\mathcal{H}_k}\leq 1} \mathsf{E}_x\{f(x)^2\}$. Therefore have $\mathsf{E}_x\{k(x,x)\} < \infty$ implies $\sup_{\|f\|_{\mathcal{H}_k}\leq 1} \mathsf{E}_x\{f(x)^2\} < \infty$, and under the conditions set out in Theorem 7.5,

$$\mathcal{R}_{x,y}(0) < \infty, \quad \sup_{\|f\|_{\mathcal{H}_k}\leq 1} \mathsf{E}_{x,y}\{yf(x)\} < \infty, \quad \text{and} \quad \sup_{\|f\|_{\mathcal{H}_k}\leq 1} \mathsf{E}_x\{f(x)^2\} < \infty.$$

The result then follows by application of Theorem 7.3, with $\mathcal{B} = \mathcal{H}_k$. □

Proof of Theorem 7.8. We consider an expression for $\mathrm{ONC}^2(\mathcal{H}_k, P_{y|(x_i)_{i=1}^n})$. We have

$$\begin{aligned}
&\mathrm{ONC}^2(\mathcal{H}_k, P_{y|(x_i)_{i=1}^n}) \\
&\quad = \mathsf{E}_{y\sim y|(x_i)_{i=1}^n, y'\sim y|(x_i)_{i=1}^n} yk(x,x')y' \\
&\quad = n^{-2}\sum_{i,j} \mathsf{E}_{y|x=x_i, y'\sim y|x=x_j} yk(x_i,x_j)y' \\
&\quad = n^{-2}\left[\left(\sum_{i,j\in i_2^n} \mathsf{E}_{y|x=x_i,y'\sim y|x=x_j} yk(x_i,x_j)y'\right) + \left(\sum_{i=1}^n \mathsf{E}_{y|x=x_i,y'\sim y|x=x_i} yk(x_i,x_i)y'\right)\right] \\
&\quad = n^{-2}\left[\sum_{i,j=1}^n \mathsf{E}_{y|x=x_i,y'\sim y|x=x_j} yk(x_i,x_j)y' - \sum_{i=1}^n \sigma_i^2 k(x_i,x_i)\right].
\end{aligned}$$

Hence, we find an unbiased estimator of $\mathrm{ONC}^2(\mathcal{H}_k, P_{y|(x_i)_{i=1}^n})$ is given by the empirical $n^{-2}\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{K}\boldsymbol{y} - \sum_{i=1}^n \sigma_i^2 K_{ii}\right)$. □

Proof of Theorem 7.11. By Theorem 7.5, we know that either

i) $\mathrm{ONC}^2(\mathcal{H}_k, P_{x,y}) = 0$ and $\mathcal{R}_{x,y}(0) = \mathcal{R}^*_{x,y;\mathcal{H}_k}$ or

ii) $\mathrm{ONC}^2(\mathcal{H}_k, P_{x,y}) > 0$ and $\mathcal{R}_{x,y}(0) > \mathcal{R}^*_{x,y;\mathcal{H}_k}$

for all probability distributions $P_{x,y}$ on $\mathcal{X}\times\mathbb{R}$, with

$$\mathcal{R}_{x,y}(0) < \infty, \quad \mathsf{E}_{x,y,x',y'}\{yk(x,x')y'\} < \infty, \quad \text{and} \quad \mathsf{E}_x\{k(x,x)\} < \infty.$$

Now assume that

$$\mathcal{R}^*_{x,y;\mathcal{H}_k} = \mathcal{R}^*_{x,y}$$

for all probability distributions $P_{x,y}$ with $\mathcal{R}_{x,y}(0) < \infty$. On substituting $\mathcal{R}^*_{x,y;\mathcal{H}_k}$ with $\mathcal{R}^*_{x,y}$ in *i)* and *ii)*, we obtain the required result. □

7.A.2 Proofs for Section 7.3

Proof of Theorem 7.19. We have, for $n \geq 2$,

$$\begin{aligned}
\lim_{\tau \to 0^+} \frac{d}{d\tau} \mathrm{LOO}(\tau) &= \lim_{\tau \to 0^+} \frac{d}{d\tau} n^{-1} \sum_{i=1}^{n} \left\{ y_i - f_{(\tau)}^{[-i]}(x_i) \right\}^2 \\
&= -2n^{-1} \lim_{\tau \to 0^+} \sum_{i=1}^{n} \left\{ \frac{d}{d\tau} f_{(\tau)}^{[-i]}(x_i) \right\} \left\{ y_i - f_{(\tau)}^{[-i]}(x_i) \right\} \\
&= -2n^{-1} \sum_{i=1}^{n} \left\{ \sum_{j \neq i} K_{ij} y_j \right\} y_i \qquad (7.50) \\
&= -2n^{-1} \sum_{i,j \in i_2^n} y_i K_{ij} y_j \\
&= -2(n-1) \mathrm{ONC}_u^2(\mathcal{H}_k, (x_i, y_i)_{i=1}^n),
\end{aligned}$$

where (7.50) is due to the result $\lim_{\tau \to 0^+} f_{(\tau)}(x)/\tau = \sum_{i=1}^{n} k(x, x_i) y_i$. □

Proof of Theorem 7.20. We know from Theorem 7.19 that, for $n \geq 2$,

$$\lim_{\tau \to 0^+} \frac{d}{d\tau} \mathrm{LOO}(\tau) = -2(n-1) \mathrm{ONC}_u^2(\mathcal{H}_k, (x_i, y_i)_{i=1}^n). \qquad (7.51)$$

Subject to the regularity conditions $\mathsf{E}_{\mathsf{x},\mathsf{y},\mathsf{x}',\mathsf{y}'} \{ y k(x, x') y' \} < \infty$ and $\mathsf{E}_{\mathsf{x}} \{ k(x, x) \} < \infty$, the expected value of either side of (7.51) yields

$$\lim_{\tau \to 0^+} \frac{d}{d\tau} \mathrm{MSE}_{n-1}(\tau) = -2(n-1) \mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{x},\mathsf{y}}).$$

As $\mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{x},\mathsf{y}})$ is not dependent on the sample size, replacing $n-1$ with n, we have

$$\lim_{\tau \to 0^+} \frac{d}{d\tau} \mathrm{MSE}_n(\tau) = -2n \mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{x},\mathsf{y}}),$$

where $n \geq 1$. □

Proof of Theorem 7.22. We have seen in (7.13) that

$$\lim_{\tau \to 0^+} f_{(\tau)}(x)/\tau = \sum_{i=1}^{n} k(x, x_i) y_i.$$

Hence,

$$\begin{aligned}
\lim_{\tau \to 0^+} \frac{d}{d\tau} \mathrm{MSPE}(\tau) &= \lim_{\tau \to 0^+} \frac{d}{d\tau} \sum_{i=1}^{n} \mathsf{E}_{\mathsf{y},\mathsf{y}'|x_i} \left\{ (y' - f_{(\tau)}(x_i))^2 \right\} \\
&= \lim_{\tau \to 0^+} \frac{d}{d\tau} \mathsf{E}_{\mathsf{y},\mathsf{y}'|x_i} \left\{ (\boldsymbol{y}' - \tau \boldsymbol{K} \boldsymbol{y})^{\mathsf{T}} (\boldsymbol{y}' - \tau \boldsymbol{K} \boldsymbol{y}) \right\} \\
&= -2 \mathsf{E}_{\mathsf{y},\mathsf{y}'|x_i} \left\{ \boldsymbol{y}'^{\mathsf{T}} \boldsymbol{K} \boldsymbol{y} \right\} \\
&= -2 \mathsf{E}_{\mathsf{y}|x_i} \left\{ \boldsymbol{y}^{\mathsf{T}} \boldsymbol{K} \boldsymbol{y} - \left(\sum_{i=1}^{n} \sigma_i^2 K_{ii} \right) \right\} \\
&= -2n^2 \mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{y}|(x_i)_{i=1}^n}).
\end{aligned}$$

□

Proof of Theorem 7.23. We have seen that

$$\lim_{\tau\to 0^+} f_{(\tau)}(x)/\tau = \sum_{i=1}^{n} k(x, x_i) y_i. \tag{7.52}$$

Substituting (7.52) into (7.14) and taking the derivative at 0^+ gives

$$\begin{aligned}
\lim_{\tau\to 0^+} \frac{d}{d\tau}\mathrm{SURE}(\tau) &= \lim_{\tau\to 0^+} \frac{d}{d\tau}\left\{\sum_{i=1}^{n}(y_i - f_{(\tau)}(x_i))^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{df_{(\tau)}(x_i)}{dy_i}\right\} \\
&= \lim_{\tau\to 0^+} \frac{d}{d\tau}\left\{(\boldsymbol{y} - \tau\boldsymbol{K}\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{y} - \tau\boldsymbol{K}\boldsymbol{y}) + 2\sigma^2 \sum_{i=1}^{n} \frac{d(\tau\boldsymbol{K}\boldsymbol{y})_i}{dy_i}\right\} \\
&= \lim_{\tau\to 0^+} \left\{-2(\boldsymbol{K}\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{y} - \tau\boldsymbol{K}\boldsymbol{y}) + 2\sigma^2 \sum_{i=1}^{n} K_{ii}\right\} \\
&= -2\left\{\boldsymbol{y}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{y} - \sigma^2 \mathrm{tr}(\boldsymbol{K})\right\} \\
&= -2n^2 \mathrm{ONC}_{\sigma}^2\left(\mathcal{H}_k, (x_i, y_i)_{i=1}^n, \sigma^2\right).
\end{aligned}$$

□

Proof of Theorem 7.24. As the theorem only consideres the probability distribution $P_{\mathsf{x},\mathsf{y}}$ through its conditional, $P_{\mathsf{y}|(x_i)_{i=1}^n}$, without loss of generality, $P_{\mathsf{x},\mathsf{y}} = P_{\mathsf{y}|(x_i)_{i=1}^n}$. As $\mathcal{H}_k$ is an RKHS, we have $k(x,x) < \infty$ for all $x \in \mathcal{X}$, and we have $\mathsf{E}_{\mathsf{x}}\{k(x,x)\} = \sum_{i=1}^n k(x_i, x_i) < \infty$. Assuming $\mathsf{E}_{\mathsf{y}}(y^2) < \infty$, we obtain the similar expression $\mathsf{E}_{\mathsf{x},\mathsf{y},\mathsf{x}',\mathsf{y}'}\{yk(x,x')y'\} < \infty$. Moreover, applying Theorem 7.5, we have $\mathrm{ONC}^2\left(\mathcal{H}_k, P_{\mathsf{x},\mathsf{y}}\right) \geq 0$, with equality if and only if $\mathsf{E}_{\mathsf{y}} y^2 = \inf_{f\in\mathcal{H}_k} \mathsf{E}_{\mathsf{x},\mathsf{y}}\left\{y - f(x)\right\}^2$. On replacing $P_{\mathsf{x},\mathsf{y}}$ with the equivalent $P_{\mathsf{y}|(x_i)_{i=1}^n}$, we have $\mathrm{ONC}^2(\mathcal{H}_k, P_{\mathsf{y}|(x_i)_{i=1}^n}) \geq 0$, with equality if and only if

$$\sum_{i=1}^{n} \mathsf{E}_{y|x_i}\left(y^2 \mid x_i\right) = \inf_{f\in\mathcal{H}_k} \sum_{i=1}^{n} \mathsf{E}_{y|x_i}\left[\{y - f(x_i)\}^2 \mid x_i\right].$$

□

Proof of Theorem 7.25. Let f be the solution to the optimisation problem

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda_N \|f\|_{\mathcal{H}_k}\right\}, \tag{7.53}$$

as well as the equivalent, for some $\lambda \geq 0$,

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_k}^2\right\}. \tag{7.54}$$

The task is to find λ, given λ_N, such that the minimisers of (7.53) and (7.54) are identical. By the Representer Theorem, we have $f(x) = \sum_{i=1}^n k(x, x_i) a_i$. The optimisation problem (7.53) becomes

$$\min_{\boldsymbol{a}}\left\{(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{a})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{a}) + \lambda_N\sqrt{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}\right\}. \tag{7.55}$$

Differentiating the objective function of (7.55) with respect to $\boldsymbol{a}$,

$$\frac{d}{d\boldsymbol{a}}\left\{(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{a})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{a}) + \lambda_N\sqrt{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}\right\} = -2\boldsymbol{K}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{a}) + \frac{\lambda_N \boldsymbol{K}\boldsymbol{a}}{\sqrt{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}},$$

for $a^T K a > 0$. Setting the derivative equal to zero, while temporarily assuming that such a solution a exists,

$$a = \left(K + \frac{\lambda_N}{2\sqrt{a^T K a}} I \right)^{-1} y.$$

However, as f is the minimiser of (7.54),

$$a = (K + \lambda I)^{-1} y.$$

On equating alternative expressions for a,

$$\lambda = \frac{\lambda_N}{2\sqrt{a^T K a}} = \frac{\lambda_N}{2 \|f\|_{\mathcal{H}_k}},$$

which rearranges to give $\lambda^2 \|f\|_{\mathcal{H}_k}^2 = \frac{\lambda_N^2}{4}$. However,

$$\begin{aligned}(y - \hat{y})^T K (y - \hat{y}) &= (y - K(K + \lambda I)^{-1} y)^T K (y - K(K + \lambda I)^{-1} y) \\ &= \lambda^2 a^T K a = \lambda^2 \|f\|_{\mathcal{H}_k}^2.\end{aligned}$$

Recalling that $\lambda_N = 2\sigma\sqrt{\mathrm{tr}(K)}$, we have

$$(y - \hat{y})^T K (y - \hat{y}) = \lambda^2 \|f\|_{\mathcal{H}_k}^2 = \frac{\lambda_N^2}{4} = \sigma^2 \mathrm{tr}(K).$$

We have required the assumption that such an a exists with $a^T K a > 0$. As a unique solution exists to (7.53), we must now only consider the special case of the differentiability of the objective function of (7.55) at $a^T K a = 0$.

As $\lambda \to \infty$, we have $a^T K a \to 0^+$. Using the notation $\tau = \lambda^{-1}$, we check the boundary point $\tau = 0$ for optimality,

$$\begin{aligned}\lim_{\tau \to 0^+} \frac{d}{d\tau} \left\{ (y - Ka)^T (y - Ka) + \lambda_N \sqrt{a^T K a} \right\} &= -2 y^T K y + \lambda_N \sqrt{y^T K y} \\ &= -2 y^T K y + 2\sqrt{\sigma^2 \mathrm{tr}(K)} \sqrt{y^T K y}. \quad (7.56)\end{aligned}$$

The right-hand-side of (7.56) is non-negative if and only if

$$y^T K y - \sigma^2 \mathrm{tr}(K) \leq 0,$$

in which case $\tau = 0$ is optimal, and we have $f = 0$. □

Proof of Theorem 7.26. As K is real symmetric, we have

$$K = Q^T \Lambda Q,$$

where Λ is a diagonal matrix, with the diagonal made up of the ordered eigenvalues $c_1 \geq \ldots \geq c_n \geq 0$. For corresponding eigenvectors $v_1, \ldots, v_n$ the matrix Q is orthogonal,

$$Q = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}.$$

We have

$$(\boldsymbol{y}-\widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}}) = (\boldsymbol{y}-\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y}-\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y})$$
$$= \lambda^2\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y}$$

and

$$\sum_{i=1}^{n}(\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2 K_{ii} = \lambda^2\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\mathrm{diag}(K_{11},\ldots,K_{nn})(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y}.$$

As k is a translation invariant kernel, we have $K_{11} = \ldots = K_{nn} > 0$. Hence,

$$\sum_{i=1}^{n}(\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2 K_{ii} = \lambda^2 K_{11}\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-2}\boldsymbol{y},$$

and we have

$$\frac{(\boldsymbol{y}-\widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})}{\sum_{i=1}^{n}(\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2 K_{ii}} = \frac{\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y}}{K_{11}\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-2}\boldsymbol{y}}. \tag{7.57}$$

Let $\boldsymbol{a} = \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{y}$, then

$$\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{a} = \sum_{i=1}^{n} a_i\boldsymbol{v}_i.$$

Since

$$(\boldsymbol{K}+\lambda\boldsymbol{I})^2\sum_{i=1}^{n}\frac{a_i}{(c_i+\lambda)^2}\boldsymbol{v}_i = \sum_{i=1}^{n}\frac{a_i}{(c_i+\lambda)^2}(\boldsymbol{K}+\lambda\boldsymbol{I})^2\boldsymbol{v}_i$$
$$= \sum_{i=1}^{n}\frac{a_i}{(c_i+\lambda)^2}(c_i+\lambda)^2\boldsymbol{v}_i$$
$$= \boldsymbol{y},$$

it is seen that

$$(\boldsymbol{K}+\lambda\boldsymbol{I})^{-2}\boldsymbol{y} = \sum_{i=1}^{n}\frac{a_i}{(c_i+\lambda)^2}\boldsymbol{v}_i.$$

Therefore, on the denominator of the right hand side of (7.57), we have

$$K_{11}\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-2}\boldsymbol{y} = K_{11}\sum_{i=1}^{n}\frac{a_i^2}{(c_i+\lambda)^2}. \tag{7.58}$$

Similarly, we find that

$$(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y} = \sum_{i=1}^{n}\frac{c_i a_i}{(c_i+\lambda)^2}\boldsymbol{v}_i,$$

and hence

$$\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y} = \sum_{i=1}^{n}\frac{c_i a_i^2}{(c_i+\lambda)^2}. \tag{7.59}$$

Substituting (7.58) and (7.59) into (7.57),

$$\frac{(\boldsymbol{y}-\widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})}{\sum_{i=1}^{n}(\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2 K_{ii}} = \frac{\sum_{i=1}^{n}\frac{c_i a_i^2}{(c_i+\lambda)^2}}{K_{11}\sum_{i=1}^{n}\frac{a_i^2}{(c_i+\lambda)^2}}. \tag{7.60}$$

We now show that the right hand side of (7.60) is monotonically decreasing in $\lambda > 0$. We have

$$\frac{d}{d\lambda}\left\{\frac{\sum_{i=1}^{n}\frac{c_i a_i^2}{(c_i+\lambda)^2}}{K_{11}\sum_{i=1}^{n}\frac{a_i^2}{(c_i+\lambda)^2}}\right\} = \frac{\left\{\frac{d}{d\lambda}\sum_{i=1}^{n}\frac{c_i a_i^2}{(c_i+\lambda)^2}\right\}\left\{\sum_{i=1}^{n}\frac{a_i^2}{(c_i+\lambda)^2}\right\} - \left\{\sum_{i=1}^{n}\frac{c_i a_i^2}{(c_i+\lambda)^2}\right\}\left\{\frac{d}{d\lambda}\sum_{i=1}^{n}\frac{a_i^2}{(c_i+\lambda)^2}\right\}}{K_{11}\left\{\sum_{i=1}^{n}\frac{a_i^2}{(c_i+\lambda)^2}\right\}^2}. \quad (7.61)$$

As the denominator of (7.61) is positive, we need further consider only the numerator. We have

$$\begin{aligned}
&\left\{\frac{d}{d\lambda}\sum_{i=1}^{n}\frac{c_i a_i^2}{(c_i+\lambda)^2}\right\}\left\{\sum_{i=1}^{n}\frac{a_i^2}{(c_i+\lambda)^2}\right\} - \left\{\sum_{i=1}^{n}\frac{c_i a_i^2}{(c_i+\lambda)^2}\right\}\left\{\frac{d}{d\lambda}\sum_{i=1}^{n}\frac{a_i^2}{(c_i+\lambda)^2}\right\} \\
&= \left\{-2\sum_{i=1}^{n}\frac{c_i a_i^2}{(c_i+\lambda)^3}\right\}\left\{\sum_{j=1}^{n}\frac{a_j^2}{(c_j+\lambda)^2}\right\} - \left\{\sum_{j=1}^{n}\frac{c_j a_j^2}{(c_j+\lambda)^2}\right\}\left\{-2\sum_{i=1}^{n}\frac{a_i^2}{(c_i+\lambda)^3}\right\} \\
&= -2\left\{\sum_{i,j}\frac{c_i a_i^2 a_j^2}{(c_i+\lambda)^3(c_j+\lambda)^2}\right\} + 2\left\{\sum_{i,j}\frac{c_j a_i^2 a_j^2}{(c_i+\lambda)^3(c_j+\lambda)^2}\right\} \\
&= -2\left\{\sum_{i,j}\frac{a_i^2 a_j^2(c_i-c_j)\left\{(c_i+\lambda)^{-1}-(c_j+\lambda)^{-1}\right\}}{(c_i+\lambda)^2(c_j+\lambda)^2}\right\} \\
&\geq 0,
\end{aligned}$$

with equality if and only if

$$a_i a_j(c_i - c_j) = 0 \quad \text{for all } 1 \leq i, j \leq n. \quad (7.62)$$

As such, (7.60) is monotonically increasing.

We now show that (7.62) will hold true if and only if $\boldsymbol{y}$ is an eigenvector of $\boldsymbol{K}$. If $a_1 = \cdots = a_n = 0$, then $\boldsymbol{K}\boldsymbol{y} = 0\boldsymbol{y} = \boldsymbol{0}$, and $\boldsymbol{y}$ is an eigenvector of $\boldsymbol{K}$. Hence, without loss of generality, we may assume $a_\ell \neq 0$ for some $1 \leq l \leq n$. The conditions of (7.62) are then equivalent to $c_i = c_\ell$ for all i such that $a_i \neq 0$. We have

$$\begin{aligned}
\boldsymbol{K}\boldsymbol{y} &= \boldsymbol{K}\Big(\sum_{a_i \neq 0} c_i \boldsymbol{v}_i\Big) \\
&= c_\ell \sum_{a_i \neq 0} c_i \boldsymbol{v}_i \\
&= c_\ell \boldsymbol{y},
\end{aligned}$$

and $\boldsymbol{y}$ is an eigenvector of $\boldsymbol{K}$.

We now consider the limit as $\lambda \to 0^+$. By (7.60), we have

$$\lim_{\lambda\to 0^+} \frac{(\boldsymbol{y}-\widehat{\boldsymbol{y}})^\mathsf{T}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})}{\sum_{i=1}^n (\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2 K_{ii}} = \lim_{\lambda\to 0^+} \frac{\sum_{i=1}^n \frac{c_i a_i^2}{(c_i+\lambda)^2}}{K_{11}\sum_{i=1}^n \frac{a_i^2}{(c_i+\lambda)^2}}$$
$$= \lim_{\lambda\to 0^+} \frac{\sum_{a_i,c_i\neq 0} a_i/c_i}{K_{11}\sum_{a_i\neq 0} a_i^2/(c_i+\lambda)^2} \tag{7.63}$$

The right hand side of (7.63) is equal to zero if and only if $a_i \neq 0$ and $c_i = 0$ for some $1 \leq i \leq n$. Equivalently, we have equality in (7.63) if and only if $\boldsymbol{y}$ is not in the column span of $\boldsymbol{K}$. Hence,

$$\lim_{\lambda\to 0^+} \frac{(\boldsymbol{y}-\widehat{\boldsymbol{y}})^\mathsf{T}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})}{\sum_{i=1}^n (\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2 K_{ii}} = \begin{cases} \dfrac{\sum_{c_i>0} a_i^2/c_i}{K_{11}\sum_{c_i>0} a_i^2/c_i^2}, & \boldsymbol{y}\in \operatorname{span}(\boldsymbol{K}), \\ 0, & \boldsymbol{y}\notin \operatorname{span}(\boldsymbol{K}). \end{cases}$$

Similar to the discussion around (7.58), it can be shown that, for $\boldsymbol{y} \in \operatorname{span}(\boldsymbol{K})$,

$$\frac{\sum_{c_i>0} a_i^2/c_i}{K_{11}\sum_{c_i>0} a_i^2/c_i^2} = \frac{\boldsymbol{y}^\mathsf{T}\boldsymbol{K}^-\boldsymbol{y}}{\sum_{i=1}^n (\boldsymbol{y}^\mathsf{T}\boldsymbol{K}^-)_i^2 K_{ii}}.$$

We then have

$$\lim_{\lambda\to 0^+} \frac{(\boldsymbol{y}-\widehat{\boldsymbol{y}})^\mathsf{T}\boldsymbol{K}(\boldsymbol{y}-\widehat{\boldsymbol{y}})}{\sum_{i=1}^n (\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2 K_{ii}} = \begin{cases} \dfrac{\boldsymbol{y}^\mathsf{T}\boldsymbol{K}^-\boldsymbol{y}}{\sum_{i=1}^n (\boldsymbol{y}^\mathsf{T}\boldsymbol{K}^-)_i^2 K_{ii}}, & \boldsymbol{y}\in \operatorname{span}(\boldsymbol{K}), \\ 0, & \boldsymbol{y}\notin \operatorname{span}(\boldsymbol{K}), \end{cases}$$

as required. □

Proof of Theorem 7.27. With parametric null, we have $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$, with

$$\boldsymbol{S} = \boldsymbol{H} + (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H}) + \lambda\boldsymbol{I}\}^{-1}.$$

Recall Stein's unbiased risk estimate,

$$\begin{aligned} \text{SURE}(\tau) &= \sum_{i=1}^n \{y_i - f_{(\tau)}(x_i)\}^2 + 2\sigma^2 \sum_{i=1}^n \frac{df_{(\tau)}(x_i)}{dy_i} \\ &= \sum_{i=1}^n \{y_i - \widehat{y}_i\}^2 + 2\sigma^2 \operatorname{tr}(\boldsymbol{S}). \end{aligned}$$

Let $\boldsymbol{K}^* = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})$, $\boldsymbol{y}^* = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$. Furthermore, let $\text{SURE}^*(\tau)$ denote Stein's unbiased risk estimate with Gram matrix $\boldsymbol{K}^*$, response vector $\boldsymbol{y}^*$ and variance σ^2.

$$\text{SURE}^*(\tau) = \sum_{i=1}^n \{y_i^* - \widehat{y}_i^*\}^2 + 2\sigma^2 \operatorname{tr}(\boldsymbol{K}^*(\boldsymbol{K}^* + \lambda\boldsymbol{I})^{-1})$$

By the argument preceding Theorem 7.34, $y_i - \widehat{y}_i = y_i^* - \widehat{y}_i^*$ and we have

$$\text{SURE}(\tau) = \text{SURE}^*(\tau) + \sigma^2 \operatorname{tr}(\boldsymbol{H}).$$

Applying Theorem 7.23,

$$\begin{aligned}\lim_{\tau\to 0^+}\frac{d}{d\tau}\mathrm{SURE}(\tau) &= \lim_{\tau\to 0^+}\frac{d}{d\tau}\mathrm{SURE}^*(\tau)\\ &= -2\left\{\boldsymbol{y}^{*\mathsf{T}}\boldsymbol{K}^*\boldsymbol{y}^* - \sigma^2\mathrm{tr}\{\boldsymbol{K}^*\}\right\}\\ &= -2\left\{\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y} - \sigma^2\mathrm{tr}\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{H})\}\right\}\end{aligned}$$ □

Proof of Theorem 7.28. The proof essentially follows that of Theorem 7.25. Let f be the solution to the optimisation problem

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda_N \|P_1 f\|_{\mathcal{H}_k}\right\}, \tag{7.64}$$

as well as the equivalent, for some $\lambda \geq 0$,

$$\min_{f\in\mathcal{H}_k}\left\{\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \|P_1 f\|^2_{\mathcal{H}_k}\right\}. \tag{7.65}$$

The task is to find λ, given λ_N, such that the minimisers of (7.64) and (7.65) are identical. By the Representer Theorem, we have $f(x_i) = (\boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{K}\boldsymbol{a})_i$. The optimisation problem (7.64) becomes

$$\min_{\boldsymbol{a},\boldsymbol{\beta}}\left\{(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{a})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{a}) + \lambda_N\sqrt{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}\right\}. \tag{7.66}$$

Differentiating the objective function of (7.66) with respect to $\boldsymbol{a}$,

$$\frac{d}{d\boldsymbol{a}}\left\{(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{a})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{a}) + \lambda_N\sqrt{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}\right\} = -2\boldsymbol{K}(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{a}) + \frac{\lambda_N\boldsymbol{K}\boldsymbol{a}}{\sqrt{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}},$$

for $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a} > 0$. Temporarily assuming that such a solution $\boldsymbol{a}$ and $\boldsymbol{\beta}$ exists,

$$\boldsymbol{a} = \left(\boldsymbol{K} + \frac{\lambda_N}{2\sqrt{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}}\boldsymbol{I}\right)^{-1}(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}).$$

As f is the minimiser of (7.65),

$$\boldsymbol{a} = (\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}).$$

On equating alternative expressions for $\boldsymbol{a}$,

$$\lambda = \frac{\lambda_N}{2\sqrt{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a}}} = \frac{\lambda_N}{2\|f\|_{\mathcal{H}_k}}.$$

Hence, noting $(\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathsf{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{y}})$, we have

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathsf{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \lambda^2\|f\|^2_{\mathcal{H}_k} = \frac{\lambda_N^2}{4} = \sigma^2\mathrm{tr}\{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{H})\}.$$

It is then straightforward to check the optimality at $\boldsymbol{a} = \boldsymbol{0}$ with $\widehat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$. □

7.A.3 Proofs for Section 7.4

Proof of Theorem 7.34. Applying Theorem 7.4 to probability distribution $P_{\mathrm{x,y}-\mu}$,

$$\sup_{\|f\|_{\mathcal{H}_k}\leq 1} \mathsf{E}_{\mathrm{x,y}}\left\{(y-\mathsf{E}_{\mathrm{y}}y)\,f(x)\right\} = \mathsf{E}_{\mathrm{x,y,x',y'}}\left\{y-\mu)k(x,x')(y'-\mu)\right\}$$

Hence we have

$$\mathrm{ONCC}^2\left(\mathcal{H}_k, P_{\mathrm{x,y}}\right) = \mathsf{E}_{\mathrm{x,y,x',y',y'',y'''}}\left\{(y-y'')k(x,x')(y'-y''')\right\}. \qquad \square$$

Proof of Theorem 7.35. For $n \geq 2$, we have

$$(n)_2^{-1}\mathsf{E}_{(\mathrm{x,y})^n}\sum_{(i,j)\in i_2^n} y_i k(x_i,x_j)y_j = \mathsf{E}_{\mathrm{x,y,x',y',y'',y'''}}\left\{(y-y'')k(x,x')(y'-y''')\right\}, \tag{7.67}$$

for $n \geq 3$,

$$(n)_3^{-1}\mathsf{E}_{(\mathrm{x,y})^n}\sum_{(i,j,k)\in i_3^n} y_i k(x_i,x_j)y_k = \mathsf{E}_{\mathrm{x,y,x',y',y'',y'''}}\left\{(y-y'')k(x,x')(y'-y''')\right\}, \tag{7.68}$$

and for $n \geq 4$,

$$(n)_4^{-1}\mathsf{E}_{(\mathrm{x,y})^n}\sum_{(i,j,k,l)\in i_4^n} y_k k(x_i,x_j)y_l = \mathsf{E}_{\mathrm{x,y,x',y',y'',y'''}}\left\{(y-y'')k(x,x')(y'-y''')\right\}. \tag{7.69}$$

Recall that

$$\begin{aligned}&\mathrm{ONCC}_u^2\left(\mathcal{H}_k,(x_i,y_i)_{i=1}^n\right)\\ &\quad\equiv (n)_2^{-1}\sum_{(i,j)\in i_2^n} y_i k(x_i,x_j)y_j - 2(n)_3^{-1}\sum_{(i,j,k)\in i_3^n} y_i k(x_i,x_j)y_k + (n)_4^{-1}\sum_{(i,j,k,l)\in i_4^n} y_k k(x_i,x_j)y_l.\end{aligned}$$

On substituting equations (7.67)–(7.69), we have,

$$\begin{aligned}&\mathsf{E}_{(x,y)^n}\left\{\mathrm{ONCC}_u^2\left(\mathcal{H}_k,(x_i,y_i)_{i=1}^n\right)\right\}\\ &\quad= \mathsf{E}_{(x,y)^n}\left\{(n)_2^{-1}\sum_{(i,j)\in i_2^n} y_i k(x_i,x_j)y_j - 2(n)_3^{-1}\sum_{(i,j,k)\in i_3^n} y_i k(x_i,x_j)y_k + (n)_4^{-1}\sum_{(i,j,k,l)\in i_4^n} y_k k(x_i,x_j)y_l\right\}\\ &\quad= \mathsf{E}_{\mathrm{x,y,x',y'}}\left\{yk(x,x')y'\right\} - 2\mathsf{E}_{\mathrm{x,y,x',y''}}\left\{yk(x,x')y''\right\}s + \mathsf{E}_{\mathrm{x,x',y'',y'''}}\left\{y''k(x,x')y'''\right\}\\ &\quad= \mathsf{E}_{\mathrm{x,y,x',y',y'',y'''}}\left\{(y-y'')k(x,x')(y'-y''')\right\}\\ &\quad= \mathrm{ONCC}^2\left(\mathcal{H}_k,P_{\mathrm{x,y}}\right),\end{aligned}$$

hence $\mathrm{ONCC}_u^2\left(\mathcal{H}_k,(x_i,y_i)_{i=1}^n\right)$ is an unbiased estimator of $\mathrm{ONCC}^2\left(\mathcal{H}_k,P_{\mathrm{x,y}}\right)$. $\square$

Proof of Theorem 7.36. We have

$$\sum_{(i,j)\in i_2^n} y_i k(x_i,x_j)y_j = \boldsymbol{y}^{\mathsf{T}}\widetilde{\boldsymbol{K}}\boldsymbol{y},$$

and

$$\sum_{(i,j,k)\in i_3^n} y_i k(x_i, x_j) y_k = \left(\mathbf{1}^\top \widetilde{K} y y^\top \mathbf{1} - y^\top \widetilde{K} y\right).$$

$$\begin{aligned}\sum_{(i,j,k)\in i_3^n} y_i k(x_i, x_j) y_k &= \sum_{\substack{(i,j)\in i_2^n,\\ k}} y_i k(x_i, x_j) y_k - \sum_{\substack{(i,j)\in i_2^n,\\ k=j}} y_i k(x_i, x_j) y_k - \sum_{\substack{(i,j)\in i_2^n,\\ k=i}} y_i k(x_i, x_j) y_k \\ &= \mathbf{1}^\top \widetilde{K} y y^\top \mathbf{1} - y^\top \widetilde{K} y - \mathbf{1} K (y \odot y).\end{aligned}$$

Moreover,

$$\begin{aligned}\sum_{(i,j,k,l)\in i_4^n} y_k k(x_i, x_j) y_l &= \sum_{\substack{(i,j)\in i_2^n\\ k,l}} y_k k(x_i, x_j) y_l - 4 \sum_{\substack{(i,j,k)\in i_3^n\\ l=k}} y_k k(x_i, x_j) y_l - 2 \sum_{\substack{(i,j)\in i_2^n\\ (k,l)=(i,j)}} y_k k(x_i, x_j) y_l \\ &\quad - \sum_{\substack{(i,j)\in i_2^n\\ k=l}} y_k k(x_i, x_j) y_l \\ &= \sum_{\substack{(i,j)\in i_2^n\\ k,l}} y_k k(x_i, x_j) y_l - 4 \sum_{(i,j,k)\in i_3^n} y_k k(x_i, x_j) y_j - 2 \sum_{(i,j)\in i_2^n} y_i k(x_i, x_j) y_j \\ &\quad - \sum_{\substack{(i,j)\in i_2^n\\ k}} y_k k(x_i, x_j) y_k \\ &= \mathbf{1}^\top \widetilde{K} \mathbf{1} (y^\top \mathbf{1})^2 - 4 \left\{\mathbf{1}^\top \widetilde{K} y y^\top \mathbf{1} - y^\top \widetilde{K} y - \mathbf{1} \widetilde{K} (y \odot y)\right\} - 2 y^\top \widetilde{K} y \\ &\quad - \mathbf{1}^\top \widetilde{K} \mathbf{1} y^\top y \\ &= \mathbf{1}^\top \widetilde{K} \mathbf{1} (y^\top \mathbf{1})^2 - 4 \mathbf{1}^\top \widetilde{K} y y^\top \mathbf{1} + 2 y^\top \widetilde{K} y + 4 \mathbf{1} \widetilde{K} (y \odot y) - \mathbf{1}^\top \widetilde{K} \mathbf{1} y^\top y.\end{aligned}$$

Gathering like terms gives

$$\begin{aligned}&\mathrm{ONCC}_u^2 (\mathcal{H}_k, (x_i, y_i)_{i=1}^n) \\ &= (n)_2^{-1} \sum_{(i,j)\in i_2^n} y_i k(x_i, x_j) y_j - (n)_3^{-1} \sum_{(i,j,k)\in i_3^n} y_i k(x_i, x_j) y_k + (n)_4^{-1} \sum_{(i,j,k,l)\in i_4^n} y_i k(x_j, x_k) y_l \\ &= (n)_2^{-1} y^\top \widetilde{K} y - (n)_3^{-1} \left\{\mathbf{1}^\top \widetilde{K} y y^\top \mathbf{1} - y^\top \widetilde{K} y - \mathbf{1} \widetilde{K} (y \odot y)\right\} \\ &\quad + (n)_4^{-1} \left\{\mathbf{1}^\top \widetilde{K} \mathbf{1} (y^\top \mathbf{1})^2 - 4 \mathbf{1}^\top \widetilde{K} y y^\top \mathbf{1} + 2 y^\top \widetilde{K} y + 4 \mathbf{1} \widetilde{K} (y \odot y) - \mathbf{1}^\top \widetilde{K} \mathbf{1} y^\top y\right\} \\ &= \frac{1}{n(n-3)} \left[y^\top \widetilde{K} y - \frac{1}{n-2} \mathbf{1}^\top \widetilde{K} \left\{-2(y \odot y) + 2 y y^\top \mathbf{1} + \mathbf{1} \frac{y^\top y - (y^\top \mathbf{1})^2}{n-1}\right\}\right].\end{aligned}$$

Since $\overline{y} = \mathbf{1} \frac{y^\top \mathbf{1}}{n}$, we have

$$\begin{aligned}&(y - \overline{y})^\top \widetilde{K} (y - \overline{y}) - \frac{1}{n-2} \mathbf{1}^\top \widetilde{K} \left\{-2(y - \overline{y}) \odot (y - \overline{y}) + \mathbf{1} \frac{(y - \overline{y})^\top (y - \overline{y})}{n-1}\right\} \\ &\quad = y^\top \widetilde{K} y - \frac{1}{n-2} \mathbf{1}^\top \widetilde{K} \left\{-2(y \odot y) + 2 y y^\top \mathbf{1} + \mathbf{1} \frac{y^\top y - (y^\top \mathbf{1})^2}{n-1}\right\}.\end{aligned}$$

Hence,

$$\mathrm{ONCC}_u^2\left(\mathcal{H}_k, (x_i, y_i)_{i=1}^n\right) = \frac{1}{n(n-3)}\left\{(y-\overline{y})^\mathsf{T}\widetilde{K}(y-\overline{y}) - \frac{1}{n-2}\mathbf{1}^\mathsf{T}\widetilde{K}v\right\},$$

where $v = -2(y-\overline{y}) \odot (y-\overline{y}) + \mathbf{1}\frac{(y-\overline{y})^\mathsf{T}(y-\overline{y})}{n-1}$. It is clear that $\mathrm{ONCC}_u^2\left(\mathcal{H}_k, (x_i, y_i)_{i=1}^n\right)$ may then be calculated in $O(n^2)$. □

Proof of Theorem 7.38. Let us begin by considering the MMD,

$$\mathrm{MMD}^2 = \mathsf{E}_{\substack{x|y=-1,\\ x'|y'=-1}}\left\{k(x,x')\right\} - 2\mathsf{E}_{\substack{x|y=-1,\\ x'|y'=1}}\left\{k(x,x')\right\} + \mathsf{E}_{\substack{x|y=1,\\ x'|y'=1}}\left\{k(x,x')\right\}. \tag{7.70}$$

Let $p = P(y=1)$. Bayes' theorem gives

$$\mathsf{E}_{\substack{x|y=-1,\\ x'|y'=-1}}\left\{k(x,x')\right\} = \frac{\mathsf{E}_{x,x'}\left\{k(x,x')I_{\substack{y=-1,\\ y'=-1}}\right\}}{(1-p)^2}, \tag{7.71}$$

$$\mathsf{E}_{\substack{x|y=-1,\\ x'|y'=1}}\left\{k(x,x')\right\} = \frac{\mathsf{E}_{x,x'}\left\{k(x,x')I_{\substack{y=-1,\\ y'=1}}\right\}}{p(1-p)}, \tag{7.72}$$

and

$$\mathsf{E}_{\substack{x|y=1,\\ x'|y'=1}}\left\{k(x,x')\right\} = \frac{\mathsf{E}_{x,x'}\left\{k(x,x')I_{\substack{y=1,\\ y'=1}}\right\}}{p^2}. \tag{7.73}$$

Since $\mathrm{Var}(y) = p(1-p)$, substituting (7.71)–(7.73) into (7.70) gives

$$\mathrm{Var}(y)^2\mathrm{MMD}^2 = \mathsf{E}_{x,x'}\left[k(x,x')\left\{p^2 I_{\substack{y=-1,\\ y'=-1}} - 2p(1-p)I_{\substack{y=-1,\\ y'=1}} + (1-p)^2 I_{\substack{y=1,\\ y'=1}}\right\}\right].$$

Now considering the ONCC, since $\mathsf{E}_y y = 1-2p$,

$$\begin{aligned}\mathrm{ONCC}^2 &= \mathsf{E}_{x,y,x',y',y'',y'''}\left\{(y-y'')k(x,x')(y'-y''')\right\}\\ &= \mathsf{E}_{x,y,x',y'}\left\{(y-1+2p)k(x,x')(y'-1+2p)\right\}\\ &= \mathsf{E}_{x,y,x',y'}\left\{k(x,x')(y-1+2p)(y'-1+2p)\right\}.\end{aligned} \tag{7.74}$$

However, since $\mathcal{Y} = \{-1,1\}$,

$$(y-1+2p)(y'-1+2p) = 4\left\{p^2 I_{\substack{y=-1,\\ y'=-1}} - 2p(1-p)I_{\substack{y=-1,\\ y'=1}} + (1-p)^2 I_{\substack{y=1,\\ y'=1}}\right\}. \tag{7.75}$$

Substituting (7.75) into (7.74), we have

$$\mathrm{ONCC}^2 = \left\{\frac{\mathrm{Var}(y)}{2}\right\}^2 \mathrm{MMD}^2,$$

and the stated result follows. □

7.A.4 Pseudo-code for ONC-based Modelling

Algorithm 7.1 Pseudo-code for PIC_{MSPE} and CIC_{MSPE}.

Require: $\boldsymbol{K}$, $\boldsymbol{y}$, σ^2, CIC

$\lambda \leftarrow \sigma^2$
if CIC **then**
  $d \leftarrow \sigma^2 \text{tr}(\boldsymbol{K})(\boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{y})^{-1/2}$
else
  $d \leftarrow \{\sigma^2 \text{tr}(\boldsymbol{K})\}^{1/2}$
end if
repeat
  $\boldsymbol{a} \leftarrow (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$
  $\lambda \leftarrow d\{\boldsymbol{a}^\top \boldsymbol{K}\boldsymbol{a}\}^{-1/2}$
until convergence in $\boldsymbol{a}$
$\widehat{\boldsymbol{y}} \leftarrow \boldsymbol{K}\boldsymbol{a}$
return λ, $\boldsymbol{a}$, $\widehat{\boldsymbol{y}}$

Algorithm 7.2 Pseudo-code for PIC_{MSE} and CIC_{MSE}.

Require: $\boldsymbol{K}$, $\boldsymbol{y}$, CIC

$\lambda \leftarrow 1$
if CIC **then**
  $d \leftarrow \left\{ \frac{\boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{y}}{\sum_{i=1}^{n}(y_i^2 K_{ii})} \right\}^{1/2}$
else
  $d \leftarrow 1$
end if
repeat
  $\boldsymbol{a} \leftarrow (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$
  $\widehat{\boldsymbol{y}} \leftarrow \boldsymbol{K}\boldsymbol{a}$
  $\lambda \leftarrow d \left\{ \frac{\sum_{i=1}^{n}(\boldsymbol{y}-\widehat{\boldsymbol{y}})_i^2 K_{ii}}{\boldsymbol{a}^\top \boldsymbol{K}\boldsymbol{a}} \right\}^{1/2}$
until convergence in $\boldsymbol{a}$
return λ, $\boldsymbol{a}$, $\widehat{\boldsymbol{y}}$

Notation and Symbols

$\mathbb{R}$	the set of reals
$\mathbb{N}$	the set of natural numbers, $\mathbb{N} = \{1, 2, \ldots\}$
$\mathcal{X}$	input space
x_i	inputs
$\mathcal{Y}$	response domain
y_i	responses
n	number of training examples
i, j	indices, by default running over $\{1, \ldots, n\}$
ε_i	errors
$\boldsymbol{X}$	matrix corresponding to unpenalised component
$\boldsymbol{X}^*$	full rank matrix corresponding to unpenalised component
$\boldsymbol{Z}$	matrix corresponding to penalised component
$\boldsymbol{\beta}$	coefficients to unpenalised component $\boldsymbol{X}$
$\boldsymbol{u}$	coefficients to penalised component $\boldsymbol{Z}$
$(\cdot)_+$	$\max(0, \cdot)$
κ_i	knot points
$\mathcal{B}$	Banach space
$\mathcal{H}$	Hilbert space
k	kernel
$\mathcal{H}_k$	RKHS with kernel k

$\mathcal{F}$	a family of functions
f	a function $\mathcal{F} \to \mathbb{R}$
δ_x	Dirac functional
Φ	feature map, $\Phi\colon \mathcal{X} \to \mathcal{H}$
a_i, c_i	dual form coefficients
b	constant offset
$\boldsymbol{K}$	Gram matrix
K	smoother kernel
$\mathcal{M}$	indicies of the non–bounded support vectors
$\mathcal{L}$	indicies of the variables at their lower bound; overloaded with the loss function
$\mathcal{U}$	indicies of the variables at their upper bound
$\mathcal{A}$	indicies of the variables in the active set
$\mathcal{S}$	indicies of the variables in the free set (Section 6.3.1)
$\mathsf{E}(\xi)$	expectation of a random variable ξ
$P(C)$	probability of a set (event) C
$p(x)$	density evaluated at x
$N(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
(μ, σ^2)	approximately normally distributed with mean μ and variance σ^2
ε	parameter of the ε-insensitive loss function
α_i	Lagrange multiplier or expansion coefficient
βi	Lagrange multiplier
$\boldsymbol{\alpha}$	vectors of Lagrange multipliers
ξ_i	slack variables
$\boldsymbol{A}^{-1}$	inverse matrix
$\boldsymbol{A}^{-}$	Moore-Penrose generalised inverse
$\boldsymbol{A}^{\mathsf{T}}$	transposed matrix (or vector)
$\det \boldsymbol{A}$	matrix determinant
$\langle x, x' \rangle$	dot product between x and x'
$\boldsymbol{A} \odot \boldsymbol{B}$	element-wise product between equalsized matricies $\boldsymbol{A}$ and $\boldsymbol{B}$

$\lvert\cdot\rvert$	absolute value
$\lVert\cdot\rVert$	norm, such as the 2-norm $\lVert x\rVert \equiv \sqrt{\langle x, x'\rangle}$ or the operator norm
$\lVert\cdot\rVert_p$	p-norm, $\lVert x\rVert_p \equiv \left(\sum_i \lvert x_i\rvert^p\right)^{1/p}$
$\lVert\cdot\rVert_\infty$	∞-norm , $\lVert\cdot\rVert_\infty \equiv \sup_i \lvert x_i\rvert$
$\lVert\cdot\rVert_{\mathcal{B}}$	Banach space norm
$\lVert\cdot\rVert_{\mathcal{H}_k}$	RKHS norm
log	natural logarithm
C	regularisation parameter in front of the empirical risk term
λ	regularisation parameter in front of the regulariser
$x \in [a, b]$	interval $a \leq x \leq b$
$x \in (a, b]$	interval $a < x \leq b$
$x \in (a, b)$	interval $a < x < b$
L_p	function spaces with finite p-norm
I_A	characteristic (or indicator) function on a set A
$\mathbf{0}$	zero vector of unspecified length
$\mathbf{0}_n$	zero vector of length n
$\mathbf{1}$	unit vector of unspecified length
$\mathbf{1}_n$	unit vector of length n
$\mathrm{card}(C)$	cardinality of a set C
H_0	null hypothesis
H_1	alternate hypothesis
$\mathcal{C}_b(\mathcal{X})$	the set of continuous, bounded functions on $\mathcal{X} \to \mathbb{R}$
$O(g(n))$	a function $f(n)$ is $O(g(n))$ if there exists constants C and n_0 such that $\lvert f(n)\rvert \leq Cg(n)$ for all $n > n_0$
$\mathrm{Cov}(x, y)$	covariance between x and y, $\mathrm{Cov}(x, y) = \mathsf{E}_{yx}(xy) - \mathsf{E}_x(x)\,\mathsf{E}_y(y)$
$\lhd$	the end of an example
$\square$	the end of a proof

Abbreviations

AIC	Akaike's an information criterion
a.s.	almost surely
AS-SVM	active set support vestor machine
BIC	Bayesian information criterion
BLUP	best linear unbiased predictor
BRUTO	an adaptive back fitting algorithm
EBLUP	estimated best linear unbiased predictor
CIC	curved information criterion
FIC	Fisher information criterion
GCV	generalised cross validation
GEE	generalised estimating equations
GLM	generalised linear model
GLMM	generalised linear mixed model
i.i.d.	independent and identically distributed
KKT	Karush-Kuhn-Tucker
LASSO	least absolute shrinkage and selection operator
LOO	leave-on-out cross-validation
MARS	multivariate adaptive regression splines
ML	maximum likelihood
MNIST	a handwritten upper-case letter database
MSE	mean squared error

MSPE	mean squared prediction error
LIBSVM	a library for support vector machines
OLS	ordinary least squares
ONC	operator norm criterion
ONCC	operator norm covariance criterion
PIC	parameter information criteria
PRESS	prediction error sum of squares
PQL	penalised quasi-likelihood
QP	quadratic program
REML	restricted maximum likelihood
RKHS	reproducing kernel Hilbert space
SURE	Stein's unbiased risk estimator
SVC	support vector classifier
SVM	support vector machine

Bibliography

Aizerman, M. A., Braverman, E. M. and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, **25**, 821–837.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.

Albert, J. and Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, **82**, 747–769.

Albrecht, J., Bjrklund, A. and Vroman, S. (2003). Is there a glass ceiling in Sweden? *Journal of Labor Economics*, **21**, 145–177.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.

Alliney, S. and Ruzinsky, S. A. (1994). An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation. *IEEE Transactions on Signal Processing*, **42**, 618–627.

Anderson, N. H., Hall, P. and Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, **50**, 41–54.

Andrews, D. W. K. (1997). A conditional Kolmogorov test. *Econometrica*, **65**, 1097–1128.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations (with discussion). *Journal of the American Statistical Association*, **96**, 939–967.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, **68**, 337–404.

Attias, H. (2000). A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*, pages 209–215, MIT Press, Cambridge, MA.

Bach, F. R. (2008). Consistency of the group Lasso and multiple kernel learning. *The Journal of Machine Learning Research*, **9**, 1179–1225.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *The Journal of Machine Learning Research*, **37**, 1–48.

Bachrach, L. K., Hastie, T., Wang, M. C., Narasimhan, B. and Marcus, B. (1999). Bone mineral acquisition in healthy asian, hispanic, black, and caucasian youth: A longitudinal study. *Journal of Clinical Endocrinology & Metabolism*, **84**, 4702–4712.

Bareiss, E. H. (1968). Sylvester's identity and multistep integer-preserving Gaussian elimination. *Mathematics of Computation*, **22**, 565–578.

Barrodale, I. (1968). L1 approximation and the analysis of data. *Applied Statistics*, **17**, 51–57.

Berg, C., Christensen, J. P. R. and Ressel, P. (1984). *Harmonic Analysis on Semigroups*. Springer-Verlag, New York, NY.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers.

Berry, S. M., Carroll, R. J. and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, **97**, 160–169.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics*, **20**, 105–134.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica*, **58**, 1129–1152.

Bierens, H. J. and Ploberger, W. (1997). Asymptotic theorey of integrated conditional moment tests. *Econometrica*, **65**, 1129–1151.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Singapore.

Bixby, R. E., Gregory, J. W., Lustig, I. J., Marsten, R. E. and Shanno, D. F. (1992). Very large-scale linear programming: A case study in combining interior point and simplex methods. *Operations Research*, pages 885–897.

Blei, D. M. and Jordan, M. I. (2005). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, **1**, 121–144.

Borgwardt, K. M., Gretton, A., Rasch, M., Kriegel, H.-P., Schölkopf, B. and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, **22**, 1–9.

Boser, B. E., Guyon, I. and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992*, pages 144–152. ACM Press.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, **16**, 199–215.

Breiman, L. and Peters, S. (1992). Comparing automatic smoothers (a public service enterprise). *International Statistical Review*, **60**, 271–290.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brumback, B. A., Ruppert, D. and Wand, M. P. (1999). Comment on Shively, Kohn and Wood. *Journal of the American Statistical Association*, **94**, 794–797.

Buchinsky, M. (1994). Changes in the US wage structure 1963-1987: Application of quantile regression. *Econometrica*, **62**, 405–458.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, **17**, 453–510.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.

Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, **1**, 412–420.

Callen, H. B. (1985). *Thermodynamics and an Introduction to Thermostatistics*. John Wiley & Sons, New York, NY.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, **35**, 2313–2351.

Cauwenberghs, G. and Poggio, T. (2001). Incremental and decremental support vector machine learning. In *Advances in Neural Information Processing Systems 13*, pages 409–415, MIT Press, Cambridge, MA.

Chambers, J. M. and Hastie, T. (1991). *Statistical Models in S*. CRC Press, Boca Raton, FL.

Chang, C.-C. and Lin, C.-J. (2009). *LIBSVM: a library for support vector machines*.

Chaudhuri, P., Doksum, K. and Samarov, A. (1997). On average derivative quantile regression. *The Annals of Statistics*, **25**, 715–744.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807–808.

Chen, X. and Fan, Y. (1999). Consistent hypothesis testing in semiparametric and nonparametric models for econometric time series. *Journal of Econometrics*, **91**, 373–401.

Christmann, A. and Steinwart, I. (2008). Consistency of kernel-based quantile regression. *Applied Stochastic Models in Business and Industry*, **24**, 171–183.

Cortes, A. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273–297.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**.

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

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York, NY.

Cristianini, N. and Shawe-Taylor, J. (2000). *Support Vector Machines*. Cambridge University Press, Cambridge, UK.

Dahmen, W. and Micchelli, C. A. (1987). Some remarks on ridge functions. *Approximation Theory and its Applications*, **3**, 139–143.

Dantzig, G. B. (1963). *Linear Programming and Extensions*. Princeton University Press, Princeton, NJ.

DeCoste, D. and Schölkopf, B. (2001). Training invariant support vector machines. *Machine Learning*, **46**, 161–190.

Delgado, M. A. (1993). Testing the equality of nonparametric curves. *Statistics and Probability Letters*, **17**, 199–204.

Diehl, C. P. (2004). Approximate leave-one-out error estimation for learning with smooth, strictly convex margin loss functions. In *Machine Learning for Signal Processing: Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*, pages 63–72.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data (Second edition)*. Oxford University Press, Oxford, UK.

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1995). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J. and Vapnik, V. (1997). Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, pages 155–161, MIT Press, Cambridge, MA.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge, UK.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–88.

Dunford, N. and Schwartz, J. T. (1963). Linear operators, part II: Spectral Theory, self adjoint operators in Hilbert space. *Wiley Intercience*.

Efron, B. (2004). The Estimation of Prediction Error: Covariance Penalties and Cross-Validation. *Journal of the American Statistical Association*, **99**, 619–632.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**, 407–499.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

Ellison, G. and Ellison, S. F. (2000). A simple framework for nonparametric specification testing. *Journal of Econometrics*, **96**, 1–23.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, **22**, 367–382.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing (Second edition)*. Marcel Dekker, New York, NY.

Evgeniou, T., Pontil, M. and Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, **13**, 1–50.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall/CRC, London, UK.

Fan, R. E., Chen, P. H. and Lin, C.-J. (2005). Working set selection using second order information for training SVM. *The Journal of Machine Learning Research*, **6**, 1889–1918.

Fan, Y. and Li, Q. (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica*, **64**, 865–890.

Fan, Y. and Li, Q. (1999). Central limit theorem for degenerate U-statistics of absolutely regular processes with applications to model specification tests. *Journal of Nonparametric Statistics*, **10**, 245–271.

Fan, Y. and Li, Q. (2000). Consistent model specification tests. *Econometric Theory*, **16**, 1016–1041.

Feynman, R. P. (1972). *Statistical Mechanics*. Benjamin, Reading, MA.

Fine, S. and Scheinberg, K. (2001). Efficient SVM training using low-rank kernel representations. *The Journal of Machine Learning Research*, **2**, 243–264.

FitzGerald, C. H., Micchelli, C. A. and Pinkus, A. (1995). Functions that preserve families of positive semidefinite matrices. *Linear Algebra and its Applications*, **221**, 83–102.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, New York, NY.

Forrest, J. J. (1989). Mathematical programming with a library of optimisation subroutines. Presentation. ORSA/TIMS Joint National Meeting New York.

French, J. L., Kammann, E. E. and Wand, M. P. (2001). Comment on Ke and Wang. *Journal of the American Statistical Association*, **96**, 1285–1288.

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J. and Penny, W. (2006). Variational free energy and the Laplace approximation. *NeuroImage*, **34**, 220–234.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G. and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: theory. *NeuroImage*, **16**, 465–483.

Fukumizu, K., Gretton, A., Sun, X. and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, MIT Press, Cambridge, MA.

Ganguli, B. and Wand, M. P. (2007). Feature significance in generalized additive models. *Statistics and Computing*, **17**, 179–192.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972–985.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelfand, I. M. and Fomin, S. V. (2000). *Calculus of Variations*. Dover Publications.

Gianola, D., Fernando, R. L. and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, **173**, 1761–1776.

Gibbs, M. N. and MacKay, D. J. C. (2000). Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks*, **11**, 1458–1464.

Gilks, W. R. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman Hall/CRC.

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, **43**, 169–169.

Girolami, M. and Rogers, S. (2006). Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, **18**, 1790–1817.

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural computation*, **10**, 1455–1480.

Gozalo, P. L. (1993). A consistent model specification test for nonparametric estimation of regression function models. *Econometric Theory*, **9**, 451–477.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall/CRC, London, UK.

Green, P. J. and Yandell, B. (1985). Semi-parametric generalized linear models. In *Proceedings of the 2nd International GLIM Conference*, pages 44–55, Springer, Berlin.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B. and Smola, A. J. (2008a). A kernel method for the two-sample-problem. *The Journal of Machine Learning Research*, **1**, 1–10.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B. and Smola, A. J. (2008b). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592, MIT Press, Cambridge, MA.

Härdle, W. and Mammen, E. (1993). Comparing parametric versus nonparametric regression fits. *The Annals of Statistics*, **21**, 1926–1947.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-Of-Fit Tests*. Springer, New York, NY.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–338.

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

Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2004). The entire regularization path for the support vector machine. *The Journal of Machine Learning Research*, **5**, 1391–1415.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC, London, UK.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **55**, 757–796.

Hastie, T. and Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics*, **5**, 329–340.

Hastie, T., Tibshirani, R. and Freidman, J. H. (2001). *The Elements of Statistical Learning*. Springer, New York, NY.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**, 13–30.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, **6**, 813–827.

Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, UK.

Horowitz, J. T. (2006). Testing a parametric model against a nonparametric alternative with identification through instrumental variables. *Econometrica*, **74**, 521–538.

Horowitz, J. T. and Härdle, W. (1994). Testing a parametric model against a semiparametric alternative. *Econometric Theory*, **10**, 821–848.

Horowitz, J. T. and Spokoiny, V. G. (2001). An adaptive, rate-optimal test of a parametric model against a nonparametric alternative. *Econometrica*, **69**, 599–631.

Huber, P. J. and Wiley, J. (1981). *Robust Statistics*. John Wiley & Sons, New York, NY.

Hush, D., Kelly, P., Scovel, C. and Steinwart, I. (2006). QP algorithms with guaranteed accuracy and run time for support vector machines. *The Journal of Machine Learning Research*, **7**, 733–769.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, **10**, 25–37.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 533–550.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379.

Joachims, T. (1999). Making large-scale SVM learning practical. In Schölkopf, B., Burges, C. J. C. and Smola, A. J., editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184, MIT Press, Cambridge, MA.

Joachims, T. (2006). Training linear SVMs in linear time. In *Proceedings of the Twelth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226, ACM Press, New York, NY.

Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, New York, NY.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401–407.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, **37**, 183–233.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Applied Statistics*, **52**, 1–18.

Karush, W. (1939). *Minima of functions of several variables with inequalities as side conditions*. PhD thesis, University of Chicago.

Keerthi, S. S., Shavade, S. K. and Bhattacharyya, C. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. Control Division Technical Report CD-99-14, Department of Mechanical and Production Engineering, National University of Singapore, Singapore.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, **33**, 82–95.

Kitamura, Y. (2005). Empirical likelihood methods in econometrics: theory and practice. In *Invited paper presented at the Econometric Society World Congress, U.C.L.*, London, UK.

Knight, C. A. and Ackerly, D. D. (2002). Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters*, **5**, 66–76.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge, UK.

Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, **71**, 265–283.

Kolmogorov, A. N. (1941). Stationary sequences in Hilbert spaces. *University of Moscow Mathematics Bulletin*, **2**, 1–40.

Korn, G. A. and Korn, T. M. (2000). *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review*. Courier Dover Publications.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, **84**, 881–896.

Lee, Y., Kim, Y., Lee, S. and Koo, J. Y. (2006). Structured multicategory support vector machines with analysis of variance decomposition. *Biometrika*, **93**, 555–571.

Lee, Y., Lin, Y. and Wahba, G. (2004). Multicategory support vector machines. *Journal of the American Statistical Association*, **99**, 67–81.

Lee, Y. and Nelder, J. (2003). Extended-REML estimators. *Journal of Applied Statistics*, **30**, 845–856.

Lee, Y. J. and Mangasarian, O. L. (2001). SSVM: A smooth support vector machine for classification. *Computational optimization and Applications*, **20**, 5–22.

Li, J., Zhang, B. and Lin, F. (2002). A new cache replacement algorithm in SMO. In *SVM '02: Proceedings of the First International Workshop on Pattern Recognition with Support Vector Machines*, pages 342–353, Springer, London, UK.

Li, Q. and Wang, S. (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics*, **87**, 145–165.

Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, **102**, 255–268.

Lin, C.-J. (2002). Asymptotic convergence of an SMO algorithm without any assumptions. *IEEE Transactions on Neural Networks*, **13**, 248–250.

Lin, Y., Lee, Y. and Wahba, G. (2002). Support vector machines for classification in nonstandard situations. *Machine Learning*, **46**, 191–202.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, **34**, 2272–2297.

List, N. and Simon, H. U. (2005). General polynomial time decomposition algorithms. In *Proceedings of The 18th Annual Conference on Learning Theory, COLT 2005*, pages 308–322. Springer.

Liu, D., Lin, X. and Ghosh, D. S. (2007). Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machines and linear mixed models. *Biometrics*, **63**, 1079–1088.

Luenberger, D. G. and Yinyu, Y. (2003). *Linear and Nonlinear Programming*. Springer, New York, NY.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, **28**, 193–209.

Matérn, B. (1960). *Spatial Variation: Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations*. Springer, Stockholm.

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York, NY.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**, 1087–1091.

Micchelli, C. A. (1986). Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, **2**, 11–22.

Micchelli, C. A., Xu, Y. and Zhang, H. (2006). Universal kernels. *The Journal of Machine Learning Research*, **7**, 2651–2667.

Minsky, M. L. and Papert, S. (1988). *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York, NY.

Müller, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian journal of statistics*, **32**, 223–240.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Springer, New York, NY.

Nychka, D., Gray, G., Haaland, P., Martin, D. and O'Connell, M. (1995). A nonparametric regression approach to syringe grading for quality improvement. *Journal of the American Statistical Association*, **90**, 1171–1178.

Opper, M. and Winther, O. (2000). Gaussian processes and SVM: Mean field and leave-one-out. In Smola, A. J., Bartlett, P. L., Schölkopf, B. and Schuurmans, D., editors, *Advances in large margin classifiers*, pages 311–326. MIT Press, Cambridge, MA.

Ormerod, J. T. and Wand, M. P. (2009). Explaining Variational Approximations. *preprint*.

Ormerod, J. T., Wand, M. P. and Koch, I. (2008). Penalised spline support vector classifiers: computational issues. *Computational Statistics*, **23**, 623–641.

Osuna, E., Freund, R. and Girosi, F. (1997a). Support vector machines: Training and applications. AI Memo 1602, Massachusetts Institute of Technology, Cambridge, MA.

Osuna, E., Freund, R. and Girosi, F. (1997b). Training support vector machines: an application to face detection. In Principe, J., Giles, L., Morgan, N. and Wilson, E., editors, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 24, pages 276–285.

Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley, Redwood City, CA.

Parker, R. L. and Rice, J. A. (1985). Discussion of "Some aspects of the spline smoothing approach to nonparametric regression curve fitting" by B. W. Silverman. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **47**, 40–42.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Pearce, N. D. and Wand, M. P. (2006). Penalized splines and reproducing kernel methods. *The American Statistician*, **60**, 233–240.

Pearce, N. D. and Wand, M. P. (2009). Explicit connections between longitudinal data analysis and kernel machines. *Electronic Journal of Statistics*, **3**, 797–823.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**, 339–348.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects Models in S and S-PLUS*. Springer Verlag, New York, NY.

Pinkus, A. (2004). Strictly positive definite functions on a real inner product space. *Advances in Computational Mathematics*, **20**, 263–271.

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B., Burges, C. J. C. and Smola, A. J., editors, *Advances in Kernel Methods - Support Vector Learning*, pages 185–208, MIT Press, Cambridge, MA.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, **78**, 1481–1497.

Polyak, B. T. (1969). The conjugate gradient method in extremal problems. *USSR Computational Mathematics and Mathematical Physics*, **9**, 94–112.

Rabe-Hesketh, S. and Skrondal, A. (2008). Generalized linear mixed-effects models. In Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenbergs, G., editors, *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, pages 79–108. Chapman & Hall/CRC.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. MIT Press, Cambridge, MA.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, NY.

Robinson, G. K. (1991). That BLUP is a good thing - the estimation of random effects. *Statistical Science*, **6**, 15–51.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ.

Royden, H. L. (1968). *Real Analysis*. Macmillan, New York, NY.

Rudin, W. (1991). *Functional Analysis*. McGraw Hill, New York, NY.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.

Scheinberg, K. (2006). An efficient implementation of an active set method for SVM. *The Journal of Machine Learning Research*, **7**, 2237–2257.

Schoenberg, I. J. (1935). Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert". *Annals of Mathematics*, **36**, 724–732.

Schoenberg, I. J. (1969). Monosplines and quadrature formulae. In Greville, T. N. E., editor, *Theory and Application of Spline Functions*, pages 157–207. Academic Press, New York, NY.

Schölkopf, B. (2001). The kernel trick for distances. In *Advances in Neural Information Processing Systems 13*, pages 301–307, MIT Press, Cambridge, MA.

Schölkopf, B., Herbrich, R., Smola, A. J. and Williamson, R. C. (2001). A generalized representer theorem. *Lecture Notes in Computer Science*, **2111**, 416–426.

Schölkopf, B., Smola, A. and Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, **10**, 1299–1319.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Seeger, M. W., Kakade, S. M. and Foster, D. P. (2008). Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory*, **54**, 2376–2382.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, NY.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **43**, 310–313.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC.

Simmons, G. F. (1963). *Introduction to Topology and Modern Analysis*. McGraw-Hill, New York, NY.

Simon, H. U. (2004). On the complexity of working set selction. In *Proceedings of the 15th International Conference on Algorithmic Learning Theory*, pages 324–337.

Smith, R. J. (2007). Efficient information theoretic inference for conditional moment restrictions. *Journal of Economics*, **138**, 430–460.

Smola, A. J. and Schölkopf, B. (2000). Sparse greedy matrix approximation for machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 911–918, Morgan Kaufmann Publishers, San Francisco, CA.

Song, L. (2008). *Learning via Hilbert Space Embedding of Distributions*. PhD thesis, University of Sydney.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M. and Bedo, J. (2007). Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 823–830. ACM.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G. and Schölkopf, B. (2008). Injective Hilbert space embeddings of probability measures. In *Proceedings of The 21st Annual Conference on Learning Theory, COLT 2008*, pages 111–122. Omnipress.

Staudenmayer, J., Lake, E. E. and Wand, M. P. (2009). Robustness for general design mixed models using the t-distribution. *Statistical Modelling*, in press.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, **9**, 1135–1151.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, NY.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research*, **2**, 67–93.

Steinwart, I. (2007). How to compare different loss functions and their risks. *Constructive Approximation*, **26**, 225–287.

Steinwart, I. (2009). Oracle inequalities for support vector machines that are based on random entropy numbers. *Journal of Complexity*, **25**, 437–454.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer, New York, NY.

Steinwart, I., Hush, D. R. and Scovel, C. (2005). Density level detection is classification. In *Advances in Neural Information Processing Systems 17*, pages 1337–1344, MIT Press, Cambridge, MA.

Steinwart, I., Hush, D. R. and Scovel, C. (2006). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, **52**, 4635.

Steinwart, I., Hush, D. R. and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of The 22nd Annual Conference on Learning Theory, COLT 2009*, pages 387–396. Omnipress.

Stengos, T. and Sun, Y. (2001). A consistent model specification test for a regression function based on nonparametric wavelet estimation. *Econometric Reviews*, **20**, 41–60.

Stinchcombe, M. B. and White, H. (1998). Consistent specification testing with unidentified nuisance parameters using duality and Banach space limit. *Econometric Theory*, **14**, 295–324.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B. and Vandewalle, J. (2002). *Least Squares Support Vector Machines*. World Scientific, Singapore.

Takeuchi, I., Le, Q. V., Sears, T. D. and Smola, A. J. (2006). Nonparametric quantile estimation. *The Journal of Machine Learning Research*, **7**, 1231–1264.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**, 267–288.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Tripathi, G. and Kitamura, Y. (2003). Testing conditional moment restrictions. *The Annals of Statistics*, **31**, 2059–2095.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer, New York, NY.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons, New York, NY.

Vapnik, V. and Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural Computation*, **12**, 2013–2036.

Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, **24**, 774–780.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York, NY.

Vishwanathan, S. V. N., Smola, A. J. and Murty, M. N. (2005). Simple and SimplerSVM. NIPS Workshop on Large Scale Kernel Machines.

Vogt, M. and Kecman, V. (2005). Active-set methods for support vector machines. In *Support Vector Machines: Theory and Applications*, pages 133–158. Springer, Berlin.

Wahba, G. (1969). Estimating derivatives from outer space. Research Report AD0703190, Defense Technical Information Center.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, pages 1378–1402.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall/CRC, London, UK.

Wei, C. Z. (1992). On predictive least squares principles. *The Annals of Statistics*, **20**, 1–42.

Welham, S. J. and Thompson, R. (2009). A note on bimodality in the log-likelihood function for penalized spline mixed models. *Computational Statistics and Data Analysis*, **53**, 920–931.

Wen, T., Edelman, A. and Gorsich, D. (2003). A fast projected conjugate gradient algorithm for training support vector machines. *AMS Contemporary Mathematics*, **323**, 245–263.

White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Review*, **21**, 149–170.

Williams, C. K. I. and Seeger, N. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688, MIT Press, Cambridge, MA.

Wipf, D., Palmer, J., Rao, B. and Kreutz-Delgado, K. (2007). Performance analysis of latent variable models with sparse priors. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Wolfe, P. (1959). The simplex method for quadratic programming. *Econometrica*, **27**, 382–398.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. CRC Press, Boca Raton, FL.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association: Theory and Methods*, **93**, 120–131.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228–237.

Yuan, M. (2006). GACV for quantile smoothing splines. *Computational Statistics and Data Analysis*, **50**, 813–829.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49–67.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, **75**, 263–298.

Zhu, J. and Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, **14**, 185–205.

Zhu, J., Kosset, S., Hastie, T. and Tibshirani, R. (2004). 1-norm support vector machines. In *Advances in Neural Information Processing Systems 16*, pages 49–56, MIT Press, Cambridge, MA.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, **35**, 2173–2192.

FNLB