Contents

1 Introduction 2
1.1 Outline 2
1.2 Advantages of I-prior modelling 3
1.3 Relation with other work 4
1.4 Overview of paper 4

2 Function spaces with reproducing kernels 4
2.1 Definitions and basic properties 4
2.2 Some useful RKHSs 6
2.2.1 RKHS of constant functions 6
2.2.2 RKHSs over finite sets 6
2.2.3 RKHSs over a Hilbert space 6
2.3 Centering of an RKKS 7

3 Construction of RKKSs over product spaces 7
3.1 ANOVA RKHSs 8
3.2 Scale parameters for kernels 8
3.3 Application in regression modelling 9
3.3.1 Models with one covariate 9
3.3.2 Models with two covariates 10
3.3.3 Functional response model 11
3.3.4 Multi-class classification 11

4 The I-prior 12

5 Maximum marginal likelihood estimation of I-prior models 13
5.1 Marginal likelihood and posterior distribution of parameter estimates 14
5.2 Estimation of parameters using the EM algorithm 15

6 Application to data 16
6.1 Motivation for use of FBM RKHS 16
6.2 Regression with a functional covariate 17
6.3 Multilevel models 19
6.4 Multi-class classification 20
6.5 Longitudinal data analysis 20

7 Discussion and topics for further research 24

References 24

A ANOVA kernel construction 27

B OLD 28
B.1 A motivating example 28
B.2 The I-prior (old stuff, perhaps some for intro?) 29
B.3 Modelling with the I-prior 30
B.4 Comparison with Tikhonov regularization 31
B.5 Comparison with other methods 33
B.5.1 Random effects models 33
B.6 GPR comparison EM 34
B.6.1 Pearson RKHS 34

Regression modelling with I-priors

With applications to functional, multilevel, and longitudinal data

Wicher Bergsma and Haziq Jamil

December 16, 2019

Abstract

We introduce a methodology with the aim of unifying and generalizing a variety of regression methods and models, including multilevel, varying coefficient, and longitudinal models, and models with functional covariates. A natural space for the regression functions pertaining to such models is the reproducing kernel Krein space (RKKS). We introduce the I-prior over the RKKS of the model, defined as the maximizer of entropy subject to a suitable constraint involving the Fisher information on the regression function.

The I-prior is Gaussian with covariance kernel proportional to the Fisher information on the regression function, and the regression function is estimated by its posterior distribution under the I-prior. The I-prior has the intuitively appealing property that the more information is available on a linear functional of the regression function, the larger the prior variance, and the smaller the influence of the prior mean on the posterior distribution.

The methodology we introduce has some advantages compared to commonly used methods in terms of ease of estimation and model comparison. Firstly, a single methodology can be used for a variety of models, for which previously a different methods were used. Secondly, an EM algorithm with a simple E and M step for estimating the scale parameter of each covariate is available, facilitating estimation for complex models. Thirdly, we propose a novel parsimonious model formulation, requiring a single scale parameter for each covariate and no further parameters for interaction effects, allowing a semi-Bayes approach to the selection of interaction effects.

An R-package implementing our methodology is available (Jamil, 2019).

Keywords: reproducing kernel, RKHS, RKKS, Fisher information, objective prior, *g*-prior, empirical Bayes, regression, supervised learning, statistical learning, longitudinal data analysis, nonparametric regression, functional data analysis, multilevel models, classification, regularization, random effects.

1 Introduction

1.1 Outline

Consider a sample $(x_1, y_1), \ldots, (x_n, y_n)$, where y_i is a real-valued measurement on unit i, and $x_i = (x_{i1}, \ldots, x_{ip})$ is a row vector of p covariates, where each each x_{ik} belongs to some set $\mathcal{X}_k$ and may for example be real, categorical, multidimensional, or functional. To describe the dependence of the y_i on the x_i, we consider the regression model

$$y_i = f(x_i) + \varepsilon_i, \quad f \in \mathcal{F} \tag{1}$$

where $\mathcal{F}$ is a space of functions. We assume the errors have a multivariate normal distribution, i.e.,

$$(\varepsilon_1, \ldots, \varepsilon_n) \sim \text{MVN}(0, \Psi^{-1}), \tag{2}$$

where $\Psi = (\psi_{ij})$ is an $n \times n$ positive definite precision matrix. Here, Ψ is taken to be known up to a low dimensional parameter, e.g., $\Psi = \psi I_n$ ($\psi > 0$, I_n the $n \times n$ identity matrix), reflecting iid errors.

The function f is assumed to be partitioned into a sum of main effects and possible interactions. An example for $p = 2$ is

$$f(x) = f(x_1, x_2) = f_\emptyset + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2), \tag{3}$$

and for $p = 3$,

$$f(x) = f(x_1, x_2, x_3) = f_\emptyset + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2) + f_{23}(x_2, x_3).$$

Here, $f_\emptyset$ is a constant (called the intercept) and we assume that the functions f_k are in a reproducing kernel Hilbert space (RKHS) of functions over a covariate space $\mathcal{X}_k$. The functions f_{kl}, f_{klm}, etc. describing interaction effects are assumed to lie in the tensor product space of the corresponding main effect function spaces. An RKHS possesses a positive definite kernel $h_k(x, x')$, where $x, x' \in \mathcal{X}_k$, and this kernel is multiplied by a scale parameter $\lambda_k \in \mathbb{R}$ which may be negative. If one or more of the scale parameters are negative, the resulting kernel for the space of regression functions $\mathcal{F}$ is indefinite, so that $\mathcal{F}$ is a *reproducing kernel Krein space* (RKKS). If $p = 1$, however, the model only has an intercept and a main effect, so there is only one scale parameter and the RKHS framework suffices. This case was described in some detail by Bergsma (2019).

The I-prior for f is a Gaussian prior whose covariance kernel is the Fisher information for f. If the RKKS $\mathcal{F}$ has reproducing kernel h, then the Fisher information between $f(x)$ and $f(x')$ is given as

$$\mathcal{I}[f(x), f(x')] = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} h(x, x_i) h(x', x_j).$$

Hence, f follows an I-prior distribution if it can be written in the form

$$f(x) = f_0(x) + \sum_{i=1}^{n} h(x, x_j) w_j, \tag{4}$$

where $f_0 \in \mathcal{F}$ is the prior mean (typically set to zero), and the w_j are multivariate normal with zero mean and covariance matrix equal to Ψ, the inverse covariance matrix of the errors ε_i. As we show, the I-prior has a maximum entropy interpretation.

An intuitively attractive property of the I-prior is that if much information about a linear functional of f (e.g., a regression coefficient) is available, its prior variance is large, and the data have a relatively large influence on the posterior, while if little information about a linear functional is available, the posterior will be largely determined by the prior mean, which serves as a 'best guess' of f. The I-prior methodology consists of estimation of the regression function by its posterior distribution under the I-prior, where we take the posterior mean as the summary measure.

For simplicity three main classes of RKHSs will be used in this paper, allowing linear and smooth effects of Euclidean and functional covariates as well as the incorporation of categorical covariates: the *canonical* RKHS, consisting of linear functions of the covariates; the *fractional Brownian motion* (FBM) RKHS, consisting of smooth functions of the covariates; and the *canonical* or the *Pearson* RKHS for nominal categorical covariates. The FBM RKHS has smoothness parameter $\gamma \in (0, 1)$, called the Hurst coefficient.

1.2 Advantages of I-prior modelling

As mentioned, each covariate, which may be multidimensional or functions, is assigned a scale parameter, which we estimate using maximum marginal likelihood. An advantage of the I-prior methodology is the availability of an EM algorithm with simple E and M steps for estimating these scale parameters, which can greatly facilitate estimation. To compare, for Gaussian process regression there is no simple EM algorithm for estimating scale parameters.

Furthermore, in our approach scale parameters are only needed for each covariate, and no further parameters are needed for interaction effects. This is in contrast with the usual approach in the regularization and Gaussian process regression literature, where separate scale parameters are assigned to each interaction effect. Our parsimonious approach yields a simpler likelihood and hence simplifies estimation. In addition, this allows a semi-Bayes approach to the selection of interaction effects, potentially able to detect effects with smaller sample sizes than with existing approaches. The use of RKKSs implies there are no range restrictions for the (scalar) scale parameters. As a consequence, simple chi-square tests can be done for the significance of covariates, simplifying the chi-bar tests usually done in random effects modelling.

Note that any RKHS has an associated Gaussian process which has the reproducing kernel as its covariance kernel. The functions in an RKHS are typically smoother than the paths of the associated Gaussian process. For this reason, the FBM RKHS is quite useful in regression, containing functions of an appropriate smoothness range, whereas the FBM Gaussian process may have paths which are too rough for regression (see Bergsma (2019) for a detailed discussion).

The I-prior methodology has a theoretical advantage compared to Tikhonov regularization, described in more detail in Bergsma (2019). In particular, as is well-known, regularization can *systematically undersmooth*, whereas the I-prior does not. For example, with mean squared error loss and a penalty proportional to $\int \ddot{f}(x)^2 dx$, the regularizer is a cubic spline smoother, which is also the posterior mean under an integrated Brownian motion prior. However, with probability one, an integrated Brownian motion path does not have two derivatives, and as a result the regularizer is an inadmissable estimator (with respect to squared error loss) for the set of functions with two derivatives (Chakraborty & Panaretos, 2019).

For some commonly used models our methodology has some additional advantages. For example, in the standard approach, multilevel models require estimation of a latent covariance matrix for the random effects. If there are more than a few covariates this can be problematic due to the positive definiteness constraint and number of parameters to be estimated. In our approach, a smaller number of unrestricted scale parameters needs to be estimated.

1.3 Relation with other work

The present paper complements Bergsma (2019), which covers the case of a single, possibly multidimensional covariate. As mentioned, in this case there is a single scale parameter, and the RKHS framework suffices. In that paper, more details are given on the I-prior derivation, and generalization of I-priors to a broad class of statistical models is given. The relation with competing methods is outlined, including g-priors, Jeffreys and reference priors, and Fisher kernels. A detailed comparison with Tikhonov regularization is given, with particular detail on the relation with cubic spline smoothing. It is explained in detail how I-priors work when the regression functions are linear, or when they are assumed to lie in the FBM RKHS, which is a particularly attractive RKHS for I-prior modelling.

Jamil (2018) provides a number of extensions to the present methodology, including probit and logit models using a fully Bayes approach, Bayesian variable selection using I-priors, and Nyström approximations for speeding up the I-prior methodology. Furthermore, he contributed a user friendly R package `iprior` (Jamil, 2019), further described in Jamil and Bergsma (2019).

Ong, Mary, Canu, and Smola (2004) previously used RKKSs in the context of regularization. In particular, they considered a regularization framework, where the usual RKHS squared penalty norm $\|f\|^2_{\mathcal{F}}$ is replaced by the RKKS indefinite inner product $\langle f,f\rangle_{\mathcal{F}}$. As the latter may be negative, it does not make sense to minimize the "penalized" loss function, and instead they sought a saddle point. Their approach is very different from ours, firstly in that they considered very different RKKSs, and secondly by constructing a Gaussian prior over the RKKS the indefiniteness of the inner product becomes irrelevant.

1.4 Overview of paper

In Section 2, a summary of existing theory of RKHSs and RKSSs is given as needed for this paper. In Section 3, we describe the construction of RKKSs over product spaces. We describe how a number of well-known models, such as the varying intercept model, one-dimensional smoothing, and multidimensional (or functional) response models can be described using the RKKS framework. In Section 4, the I-prior is defined and its representation (4) is derived for model (1) with multivariate normal errors. In Section 5, the EM algorithm for estimating scale parameters is described. In Section 6, we apply the I-prior methodology to a number of data examples in the respective areas of multilevel modelling, functional data analysis, classification and longitudinal data analysis, illustrating some possible advantages over existing techniques, and showing competitive predictive performance.

2 Function spaces with reproducing kernels

This section summarizes existing theory as needed for this paper. In Section 2.1 we give the definition and some well-known basic properties of RKHSs and RKKSs. Section 2.2 briefly lists the RKHSs used in this paper. These RKHSs are used as building blocks to construct RKKSs over product spaces, called ANOVA RKKSs, which is the topic of the next Section 3. The RKHSs we use in this paper will normally be *centered*, i.e., the functions in the RKHS have zero mean, which is formally described in Section 2.3.

2.1 Definitions and basic properties

The first comprehensive treatment of RKHSs was given by Aronszajn (1950), and their usefulness for statistics was initially demonstrated by Parzen (1961) and further developed by

Kimeldorf and Wahba (1970). Some more recent overviews of RKHS theory with a view to application in statistics and machine learning are Wahba (1990b), Berlinet and Thomas-Agnan (2004), Steinwart and Christmann (2008, Chapter 4) and Hofmann, Schölkopf, and Smola (2008). Schwartz (1964) developed a general theory of Hilbertian subspaces of topological vector spaces which includes the theory of RKKSs. The first applications to statistics and machine learning of RKKSs were given by Ong et al. (2004) and Canu, Ong, and Mary (2009). A recent technical survey of the theory of RKKSs is given by Gheondea (2013). Below, we give a very brief overview of the theory as needed for this paper, more details can be found in the aforementioned literature.

We begin with the definition of the (possibly indefinite or negative definite) inner product.

Definition 1. *Let $\mathcal{F}$ be a vector space over the reals. A function $\langle\cdot,\cdot,\rangle_{\mathcal{F}} : \mathcal{F}\times\mathcal{F}\to\mathbb{R}$ is called an* inner product *on $\mathcal{F}$ if, for all $f, f', f'' \in \mathcal{F}$,*

- *(symmetry) $\langle f, f'\rangle_{\mathcal{F}} = \langle f', f\rangle_{\mathcal{F}}$*
- *(linearity) $\langle \alpha f + f', f''\rangle_{\mathcal{F}} = \alpha\langle f, f''\rangle_{\mathcal{F}} + \langle f', f''\rangle_{\mathcal{F}}$*
- *(nondegeneracy) $(\forall g \in \mathcal{F} : \langle f, g\rangle_{\mathcal{F}} = 0) \Rightarrow f = 0$*

If $\langle f, f\rangle_{\mathcal{F}} \geq 0$ for all $f \in \mathcal{F}$, the inner product is called positive definite *and $\|f\|_{\mathcal{F}} := \langle f, f\rangle_{\mathcal{F}}$ is called a* norm *on $\mathcal{F}$. If $\langle f, f\rangle_{\mathcal{F}} \leq 0$ for all $f \in \mathcal{F}$, the inner product is called* negative definite*. An inner product which is neither positive definite nor negative definite is called* indefinite.

Recall that a Hilbert space is a complete inner product space with a positive definite inner product. The more general notion of Krein space is defined as follows.

Definition 2. *A vector space $\mathcal{F}$ equipped with the inner product $\langle\cdot,\cdot,\rangle_{\mathcal{F}}$ is called a* Krein space *if there are two Hilbert spaces $\mathcal{F}_+$ and $\mathcal{F}_-$ spanning $\mathcal{F}$ such that*

- *All $f \in \mathcal{F}$ can be decomposed as $f = f_+ + f_-$ where $f_+ \in \mathcal{F}_+$ and $f_- \in \mathcal{F}_-$.*
- *For all $f, f' \in \mathcal{F}$, $\langle f, f'\rangle_{\mathcal{F}} = \langle f_+, f'_+\rangle_{\mathcal{F}_+} - \langle f_-, f'_-\rangle_{\mathcal{F}_-}$*

Note that any Hilbert space is a Krein space, which can be seen by taking $\mathcal{F}_- = \{0\}$.

We next define the notion of a *reproducing kernel*:

Definition 3. *Let $\mathcal{F}$ be a Krein space of functions over a set $\mathcal{X}$. A symmetric function $h : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a* reproducing kernel *of $\mathcal{F}$ if and only if*

(a) $h(x,\cdot) \in \mathcal{F}$ for all $x \in \mathcal{X}$

(b) $f(x) = \langle f, h(x,\cdot)\rangle_{\mathcal{F}}$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$.

A Hilbert space resp. Krein space is called a *reproducing kernel Hilbert space* (RKHS) resp. *reproducing kernel Krein space* (RKKS) if it possesses a reproducing kernel. Sometimes in this paper we will use the shorthand 'kernel' to refer to 'reproducing kernel'.

A function $h : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is said to be positive definite on $\mathcal{X}$ if $\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j h(x_i, x_j) \geq 0$ for all scalars $\alpha_1, \ldots, \alpha_n$ and all $x_1, \ldots, x_n \in \mathcal{X}$. From the definition of positive definite inner products it follows that the reproducing kernel of an RKHS is symmetric and positive definite. The reproducing kernel of an RKKS can be shown to be the difference of two positive definite kernels so need not be positive definite. The Moore-Aronszajn theorem states that every symmetric positive definite function defines a unique RKHS. Every RKSS also has a unique kernel, but a given kernel may have more than one RKKS associated with it (e.g., Alpay, 1991).

$\mathcal{X}$	RKHS	Functions $f(x)$	Kernel $h(x,x')$	Centered kernel
Any set	Constant	Constant functions	1	-
Finite set	Canonical	All functions	$\delta_{xx'}$	$\delta_{xx'} - p(x)p(x')$
Finite set	Pearson	All zero mean functions	$\delta_{xx'}/p(x) - 1$	$\delta_{xx'}/p(x) - 1$
Hilbert space	Canonical	$\langle x, \beta \rangle_{\mathcal{X}}$	$\langle x, x' \rangle_{\mathcal{X}}$	$\langle x - \overline{x}, x' - \overline{x} \rangle_{\mathcal{X}}$
$\mathbb{R}^p$	Mahalanobis	$x^\top \beta$	$x^\top S^{-1} x'$	$(x - \overline{x})^\top S^{-1}(x' - \overline{x})$
$\mathbb{R}$	Brownian motion	$\int_{-\infty}^{x} \beta(t)dt$	$\frac{1}{2}(\lvert x\rvert + \lvert x'\rvert - \lvert x - x'\rvert)$	Eq. (5) ($\gamma = 1/2$)
Hilbert space	Brownian motion	Hölder $\geq 1/2$	$\frac{1}{2}(\lVert x\rVert + \lVert x'\rVert - \lVert x - x'\rVert)$	Eq. (5) ($\gamma = 1/2$)
Hilbert space	FBM-γ	Hölder $\geq \gamma$	$\frac{1}{2}(\lVert x\rVert^{2\gamma} + \lVert x'\rVert^{2\gamma} - \lVert x - x'\rVert^{2\gamma})$	Eq. (5)

Table 1: List of RKHSs

2.2 Some useful RKHSs

Below we describe some RKHSs that we will use in this paper. A summary is given in Table 1.

2.2.1 RKHS of constant functions

The RKHS of constant functions with reproducing kernel given by $h(x,x') = 1$. For a constant function f with $f(x) = c$, $\|f\|_{\mathcal{F}} = |c|$. (The RKHS of constant functions will be an essential component in the construction of RKKSs over product spaces in Section 3.1.)

2.2.2 RKHSs over finite sets

Let $\mathcal{X}$ be a finite set. The *canonical* RKHS over $\mathcal{X}$ is the RKHS whose kernel is the Kronecker delta function, i.e., $h(x,x') = \delta_{xx'}$ consists of the set of all functions $f : \mathcal{X} \to \mathbb{R}$, with squared norm

$$\|f\|_{\mathcal{F}}^2 = \sum_{x \in \mathcal{X}} f(x)^2.$$

Note that, viewing f as a $|\mathcal{X}|$-dimensional vector, the canonical RKHS over $\mathcal{X}$ is just standard Euclidean space.

Alternatively, the Pearson RKHS over a finite probability space $(\mathcal{X}, p)$, defined as the RKHS with reproducing kernel $h(x,x') = p(x)^{-1}\delta_{xx'} - 1$, consisting of all functions with $\sum p(x)f(x) = 0$ and

$$\|f\|_{\mathcal{F}}^2 = \sum p(x)f(x)^2$$

2.2.3 RKHSs over a Hilbert space

Let $\mathcal{X}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{X}}$. The *canonical RKHS* over$\mathcal{X}$ is defined as its continuous dual space whose reproducing kernel is given by

$$h(x,x') = \langle x, x' \rangle_{\mathcal{X}}.$$

Functions in this space are of the form $f(x) = \langle x, \beta \rangle_{\mathcal{X}}$, with norm $\|f\|_{\mathcal{F}} = \|\beta\|_{\mathcal{X}}$.

A special case is the *Mahalanobis RKHS*, defined as the canonical RKHS over $\mathcal{X} = \mathbb{R}^p$ equipped with the Mahalanobis inner product; for a covariance matrix S, it is defined as

$$\langle x, x' \rangle_{\text{Mah}} = x^\top S^{-1} x'$$

The *Brownian motion RKHS* is defined as the RKHS over $\mathcal{X}$ whose reproducing kernel is the generalized Brownian motion covariance kernel

$$h(x,x') = -\frac{1}{2}\Big(\|x-x'\|_{\mathcal{X}} - \|x\|_{\mathcal{X}} - \|x'\|_{\mathcal{X}}\Big)$$

Functions in the Brownian motion RKHS are Hölder of degree at least 1/2 (see Bergsma (2019) for a proof). In the simplest nontrivial case, $\mathcal{X} = \mathbb{R}$, and the RKHS consists of functions with a square integrable derivative, whose norm is the L^2 norm of the derivative, i.e., every $f \in \mathcal{F}$ can be written as $f(x) = \int_{-\infty}^{x} \beta(t)dt$ for some square integrable β, and has norm $\int_{\mathbb{R}} \beta(t)^2 dt$.

The fractional Brownian motion (FBM) RKHS is the RKHS whose reproducing kernel is the generalized FBM covariance kernel

$$h(x,x') = -\frac{1}{2}\big(\|x-x'\|_{\mathcal{X}}^{2\gamma} - \|x\|_{\mathcal{X}}^{2\gamma} - \|x'\|_{\mathcal{X}}^{2\gamma}\big)$$

Functions in the FBM-γ RKHS are Hölder of degree at least γ (see Bergsma (2019) for a proof).

2.3 Centering of an RKKS

We say a function space $\mathcal{F}$ over $\mathcal{X}$ is *centered* with respect to a data set $\{x_1,\ldots,x_n\} \subset \mathcal{X}$ if

$$\sum_{i=1}^{n} f(x_i) = 0 \quad \forall f \in \mathcal{F}$$

It can be verified that an RKKS $\mathcal{F}$ is centered if and only if its kernel h is centered, in the sense that $\sum_{i=1}^{n} h(x,x_i) = 0$ for all $x \in \mathcal{X}$. If h is a kernel, then h_{cent} defined as follows is centered:

$$h_{\text{cent}}(x,x') = h(x,x') - \frac{1}{n}\sum_{j=1}^{n} h(x,x_j) - \frac{1}{n}\sum_{i=1}^{n} h(x_i,x') + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(x_i,x_j)$$

Table 1 gives a list of kernels discussed in Section 2.2 and their centered versions. The centered FBM RKKS has kernel

$$h_{\text{cent}}(x,x') = -\frac{1}{2}\Bigg(\|x-x'\|_{\mathcal{X}}^{2\gamma} - \frac{1}{n}\sum_{j=1}^{n}\|x-x_j\|_{\mathcal{X}}^{2\gamma} - \frac{1}{n}\sum_{i=1}^{n}\|x_i-x'\|_{\mathcal{X}}^{2\gamma} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i-x_j\|_{\mathcal{X}}^{2\gamma}\Bigg) \tag{5}$$

where the Brownian motion RKHS is obtained if $\gamma = 1/2$.

3 Construction of RKKSs over product spaces

ANOVA constructions of RKHSs over product spaces are a natural tool for formulating regression models and were introduced for this purpose by Wahba (1990a) and Gu and Wahba (1993). In Section 3.1 we describe ANOVA RKKSs, an immediate extension of ANOVA RKHSs which are needed in this paper. In Section 3.2, we describe what as far as we are aware is a novel approach to use scale parameters parsimoniously in the ANOVA construction. In Section 3.3 we show how the framework is useful in regression, as, for example, it can be used to easily formulate multilevel and varying coefficient models.

3.1 ANOVA RKHSs

An ANOVA decomposition of a function f over product space $\mathcal{X}_1 \times \mathcal{X}_2$ is given by

$$f(x_1, x_2) = f_\emptyset + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2)$$

where the components are orthogonal in some way. To formalize this, let us first define the tensor product of RKHSs. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ by two RKHSs over $\mathcal{X}_1$ resp. $\mathcal{X}_2$. For $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$, the tensor product $f_{12} = f_1 \otimes f_2$ is defined by $f_{12}(x_1, x_2) = f_1(x_1)f_2(x_2)$. The tensor product of $\mathcal{F}_1$ and $\mathcal{F}_2$ is denoted as $\mathcal{F}_1 \otimes \mathcal{F}_2$ and is defined as the closure of the set of functions $\{f_1 \otimes f_2 | f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ equipped with the inner product

$$\langle f_1 \otimes f_2, f_1' \otimes f_2' \rangle_{\mathcal{F}_1 \otimes \mathcal{F}_2} = \langle f_1, f_1' \rangle_{\mathcal{F}_1} \langle f_2, f_2' \rangle_{\mathcal{F}_2}.$$

The tensor product of RKKSs is defined analogously, the closure being defined with respect to the corresponding positive definite inner product.

Let $\mathcal{C}_k$ be the RKHS of constant functions over $\mathcal{X}_k$ with kernel $c_k(x, x') = 1$ and let $\mathcal{F}_k$ be an RKKS over $\mathcal{X}_k$ with kernel h_k ($k = 1, 2$). An ANOVA RKKS over $\mathcal{X}_1 \times \mathcal{X}_2$ is given as

$$\mathcal{F} = \mathcal{C}_1 \otimes \mathcal{C}_2 + \mathcal{F}_1 \otimes C_2 + \mathcal{C}_1 \otimes \mathcal{F}_2 + \mathcal{F}_1 \otimes \mathcal{F}_2$$

with kernel h given by

$$h((x_1, x_2), (x_1', x_2')) = 1 + h_1(x_1, x_1') + h_2(x_2, x_2') + h_1(x_1, x_1')h_2(x_2, x_2')$$

In this paper we assume the components h_k are centered relative to data $\{x_{1k}, \dots, x_{nk}\} \subset \mathcal{X}_k$ (Section 2.3). Hence, the $\mathcal{C}_k$ and $\mathcal{F}_k$ are orthogonal in the sense that

$$\sum_{i=1}^{n} c_k(x_{1k}) f_k(x_{1k}) = \sum_{i=1}^{n} f_k(x_{1k}) = 0$$

for any $c_k \in \mathcal{C}_k$ and $f_k \in \mathcal{F}_k$.

With p covariates, the ANOVA model with all interactions can be written succinctly as

$$\mathcal{F} = \bigoplus_{k=1}^{p} \big(\mathcal{C}_k \otimes \mathcal{F}_k\big) \tag{6}$$

with reproducing kernel

$$h(x, x') = \prod_{k=1}^{p} \big(1 + h_k(x_k, x_k')\big) \tag{7}$$

More general ANOVA kernels are described in Appendix A.

3.2 Scale parameters for kernels

In practice, the length of a vector in an RKHSs is measured on an arbitrary scale, and this can be taken into account by multiplying the kernel by a real-valued scale parameter which is to be estimated. In the ANOVA case, we can use component kernels $\mu_k h_k$ and $\tau_k c_k$ for real-valued μ_k and τ_k, giving the ANOVA kernel

$$h_{\mu,\tau}((x_1, x_2), (x_1', x_2')) = \tau_1\tau_2 + \mu_1\tau_2 h_1(x_1, x_1') + \tau_1\mu_2 h_2(x_2, x_2') + \mu_1\mu_2 h_1(x_1, x_1')h_2(x_2, x_2')$$

Model: $y_i = \alpha + f_1(x_i) + \varepsilon_i$, $x_i \in \mathcal{X}_1$, $f_1 \in \mathcal{F}_1$			
$\mathcal{X}_1$	RKHS $\mathcal{F}_1$	Model name	Usual notation
A finite set	Pearson	One-way ANOVA/Varying intercept model	$y_{ij} = \alpha + \beta_j + \varepsilon_{ij}$
$\mathbb{R}$	Canonical	Simple regression	$y_i = \alpha + x_i\beta + \varepsilon_i$
$\mathbb{R}$	FBM	Smoothing spline model	$y_i = \alpha + \int_0^{x_i} \beta(t)dt + \varepsilon_i$
An RKHS	Canonical	Functional linear regression	$y_i = \alpha + \int x_i(t)\beta(t)d\mu(t) + \varepsilon_i$
An RKHS	FBM	Smooth functional regression	$y_i = \alpha + f(x_i) + \varepsilon_i$

Table 2: Some models with one, possibly multidimensional covariate

This expression is overparameterized, and setting $\lambda_0 = \tau_1\tau_2$ and $\lambda_k = \mu_k/\tau_k$ (assuming $\tau_k \neq 0$), we obtain the identified parameterization

$$h_\lambda((x_1, x_2), (x_1', x_2')) = \lambda_0\big\{1 + \lambda_1 h_1(x_1, x_1') + \lambda_2 h_2(x_2, x_2') + \lambda_1\lambda_2 h_1(x_1, x_1')h_2(x_2, x_2')\big\} \quad (8)$$

Typically, the kernels h_k will be positive definite, so that the corresponding $\mathcal{F}_k$ are RKHSs. Then if at least one of the lambda parameters is negative, a function space with h_λ as its kernel will be an RKKS.

In the literature, a less parsimonious construction than (8) has been used, namely

$$h_\upsilon((x_1, x_2), (x_1', x_2')) = \upsilon_0 + \upsilon_1 h_1(x_1, x_1') + \upsilon_2 h_2(x_2, x_2') + \upsilon_{12} h_1(x_1, x_1')h_2(x_2, x_2') \quad (9)$$

(e.g., Wahba, 1990b, Section 10.2, Berlinet & Thomas-Agnan, 2004, Section 10.2, Gu, 2013, Section 2.4.5). We refer to the corresponding RKKS as the *extended ANOVA RKKS*. Here, each of the four terms has a separate scale parameter, and is thus less parsimonious than our approach which only requires three scale parameters. For models with all interactions, our approach has $p + 1$ scale parameters, while 2^p parameters are required if every interaction is assigned a separate parameter.

3.3 Application in regression modelling

We show how some well-known models which can be written in the form (1) can be formulated using the ANOVA function space construction. In Sections 10 and 12 we consider models with one and two covariates respectively (see also Tables 2 and 3). In Section 3.3.3 we consider functional responses, and in Section 3.3.4 we consider multi-class classification models.

3.3.1 Models with one covariate

We consider examples of regression models (1) where $\mathcal{F}$ is of the form (6) with $p = 1$. We can write

$$f(x) = \alpha + f_1(x), \quad f_1 \in \mathcal{F}_1, x \in \mathcal{X}_1 \quad (10)$$

where $\mathcal{F}_1$ is in the centered RKHS over a set $\mathcal{X}_1$ with kernel h_1. Recall that the centering means $\sum_{i=1}^n f(x_i) = 0$ for all $f \in \mathcal{F}_1$.

If $\mathcal{X}_1$ is a finite set, then (10) is known as a one-way ANOVA model or varying intercept model. The usual notation for this model is

$$y_{jk} = \alpha + f_1(k) + \varepsilon_{jk} \quad f_1 \in \mathcal{F}_1, k \in \mathcal{X}_1 \quad (11)$$

Model: $y_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + f_{12}(x_{1i}, x_{2i}) + \varepsilon_i,\ x_{ki} \in \mathcal{X}_k,\ f_k \in \mathcal{F}_k,\ f_{12} \in \mathcal{F}_1 \otimes \mathcal{F}_2$					
$\mathcal{X}_1$	RKHS $\mathcal{F}_1$	$\mathcal{X}_2$	RKHS $\mathcal{F}_2$	Model name	Usual notation
Finite set	Pearson	$\mathbb{R}$	Canonical	Varying slope model	$y_{ij} = \alpha + \alpha_j + x_{ij}\beta + x_{ij}\beta_j + \varepsilon_{ij}$
$\mathbb{R}^p$	FBM	$\mathbb{R}^p$	Canonical	Varying coefficient model	$y_i = \alpha + x_{2i}^\top \beta(x_{1i}) + \varepsilon_i$
$\mathbb{R}^p$	FBM	$\mathbb{R}^p$	Canonical	Multivariate regression	$y_{ij} = \alpha + x_i\beta_j + \varepsilon_{ij}$
$\mathbb{R}^p$	FBM	$\mathbb{R}^p$	Canonical	Functional response model	$y_i(t) = f(x_i, t) + \varepsilon_i$

Table 3: Some models with two, possibly multidimensional covariates

where y_{jk} is the kth y_i-value for which $x_i = j$. Suitable RKHSs $\mathcal{F}_1$ are the canonical for which $\|f\|_{\mathcal{F}}^2 = \sum f(j)^2$ or the Pearson for which $\|f\|_{\mathcal{F}}^2 = \sum p(j)f(j)^2$, where $p(j)$ is the number of x_is equal to j (Section 2.3).

If $\mathcal{X}_1$ is a subset of a Hilbert space, a flexible range of models is obtained by taking $\mathcal{F}_1$ to be the canonical RKHS, the Brownian motion RKHS, or more generally the FBM RKHS. In the first case, we obtain

$$y_i = \alpha + \langle x_i, \beta \rangle_{\mathcal{X}} + \varepsilon_i$$

If if $\mathcal{X} = \mathbb{R}^q$ we obtain the special case

$$y_i = \alpha + \sum_{k=1}^{q} x_{iq}\beta_q + \varepsilon_i$$

or if $\mathcal{X} = L^2(\mathcal{X}, \mu)$ we obtain the functional linear model (e.g., Yao, Müller, and Wang (2005))

$$y_i = \alpha + \int_{\mathcal{X}} x_i(t)\beta(t)d\mu(t) + \varepsilon_i$$

A smooth dependence model is obtained if $\mathcal{F}_1$ is the Brownian motion or FBM RKHS. A special case is the smoothing spline model

$$y_i = \alpha + \int_{-\infty}^{x_i} \beta(t)dt + \varepsilon_i$$

when $\mathcal{X}_1 = \mathbb{R}$. However, there are many other potentially useful linear or smooth dependence models, such as the model where $\mathcal{X}$ is a Brownian motion RKHS.

3.3.2 Models with two covariates

We next consider examples of regression models (1) where $\mathcal{F}$ is of the form (6) with $p = 2$. We can write

$$f(x_1, x_2) = \alpha + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2), \quad f_k \in \mathcal{F}_k, x_k \in \mathcal{X}_k, k = 1, 2, \tag{12}$$

where $\mathcal{F}_k$ is in the centered RKHS over a set $\mathcal{X}_k$ with kernel h_k, and $f_{12} \in \mathcal{F}_1 \otimes \mathcal{F}_2$.

First consider the case that $\mathcal{X}_1 = \mathbb{R}$ and $\mathcal{X}_2$ is a finite set. Taking $\mathcal{F}_1$ to be the canonical RKHS yields the *varying slope* model, that is, for each element of $\mathcal{X}_2$, we have a linear dependence model. The usual representation of this model is as a two-level regression model asserting that the y_{ij} depend both on the covariate x_{ij} and on the cluster j,

$$y_{ij} = f(j, x_{ij}) + \varepsilon_{ij} \qquad i = 1, \dots, n_j,\ j = 1, \dots, m,\ x_{ij} \in \mathcal{X}. \tag{13}$$

where

$$f(j, x_{ij}) = \alpha + \beta_{1,j} + x_{ij}\beta_2 + x_{ij}\beta_{12,j}, \tag{14}$$

Here, $\alpha + \beta_{1,j}$ is called the intercept for cluster j and $\beta_2 + \beta_{12,j}$ its slope.

Next consider the case $\mathcal{X}_1 = \mathcal{X}_2 = \mathbb{R}$. The canonical RKHSs give the model

$$y_i = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{1i}x_{2i}\beta_{12} + \varepsilon_i$$

Note that this approach is suitable if the x_{1i} and x_{2i} are measured on different scales, such as height and weight. If the x_{ik} are measured on the same scale, e.g., height measured at two different time points, it may be better to consider the pairs (x_{i1}, x_{2i}) as a single covariate in $\mathbb{R}^2$ and use the approach in Section 3.3.1.

If $\mathcal{F}_1$ is the canonical RKHS and $\mathcal{F}_2$ the Brownian motion RKHS, we obtain

$$y_i = \alpha + x_{1i}\beta_1 + f_2(x_{2i}) + x_{1i}\beta(x_{2i}) + \varepsilon_i$$

This has been called the *varying coefficient model* (Hastie & Tibshirani, 1993), where $\beta(x_{2i})$ is the varying regression coefficient for the x_{1i}.

3.3.3 Functional response model

We now consider the case that, rather than scalars, the y_i are real-valued functions over a set $\mathcal{T}$. A regression model then be formulated as

$$y_i(t) = f(x_i, t) + \varepsilon_{it} \qquad i = 1, \ldots, n, t \in \mathcal{T} \tag{15}$$

where

$$f(x_i, t) = \alpha + f_1(x_i) + f_2(t) + f_{12}(x_i, t) \quad f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2, f_{12} \in \mathcal{F}_1 \otimes \mathcal{F}_2$$

If $\mathcal{T}$ is a finite set, y_i can be viewed as a vector in $\mathbb{R}^{|\mathcal{T}|}$. Then if additionally $\mathcal{F}_1$ is the canonical RKHS over $\mathbb{R}^p$, we obtain the usual multivariate regression model

$$y_{it} = \alpha + \alpha_t + x_i^\top \beta_t + \varepsilon_{it}$$

where $y_{it} = y_i(t)$ and $\alpha_t = f_2(t)$.

In practice, we do not observe y_i entirely but rather a finite set of evaluations at index points $u_{i1}, \ldots, u_{im_i}$, i.e., we observe $y_i(u_{i1}), \ldots, y_i(u_{im_i})$. For example, in a repeated measurements setting, both the number of measurements m_i and the times of measurement may be different for different units i. Then (15) becomes

$$y_i(u_{is}) = f(x_i, u_{is}) + \varepsilon_{is} \quad i = 1, \ldots, n, s = 1, \ldots, m_i$$

Note that this can be viewed as an instance of model (1).

In Section 6.5 we apply this model, as well as an extension with an extra covariate, to a longitudinal data set, taking $\mathcal{F}_2$ the FBM RKHS.

3.3.4 Multi-class classification

Consider a multi-class classification problem where, with $\mathcal{C}$ a finite set of classes, we have observations $(x_1, c_1), \ldots, (x_n, c_n)$ for $x_i \in \mathcal{X}$ and $c_i \in \mathcal{C}$. The aim is to find a prediction function to predict the class $c \in \mathcal{C}$ for a future observation $x \in \mathcal{X}$. We can use the present framework as follows. Let $y_{ij} = 1$ if $c_i = j$ and let $y_{ij} = 0$ otherwise. We may now consider the model

$$y_{ij} = f(x_i, j) + \varepsilon_{ij}, \quad i = 1, \ldots, n, j \in \mathcal{C}$$

where the y_{ij} satisfy the restriction

$$\sum_{j \in \mathcal{C}} y_{ij} = 1 \quad \forall i$$

Assuming $\sum_{j \in \mathcal{C}} \varepsilon_{ij} = 0$, we then have $\sum_{j \in \mathcal{C}} f(x_i, j) = 1$ and we may consider a decomposition

$$f(x_i, j) = |\mathcal{C}|^{-1} + f_2(j) + f_{12}(x_i, j) + \varepsilon_{ij} \quad f_2 \in \mathcal{F}_2, f_{12} \in \mathcal{F}_1 \otimes \mathcal{F}_2 \tag{16}$$

subject to the identifying restrictions $\sum_{j \in \mathcal{C}} f_2(j) = 0$, $\sum_{j \in \mathcal{C}} f_{12}(x_i, j) = 0$ for all i, and $\sum_{i=1}^n f_{12}(x_i, j) = 0$ for all j. Hence, we may assume $\mathcal{F}_2$ is the centered canonical RKHS, and $\mathcal{F}_1$ is any appropriate RKKS. Note that a main effect for x_i is not needed in (16).

Of course, more than one covariate can be incorporated, for example with $p = 2$ we may take $\mathcal{F}_1 = \mathcal{F}_{11} + \mathcal{F}_{12} + \mathcal{F}_{11} \otimes \mathcal{F}_{12}$ so that

$$f_{12}(x_i, j) = f_{12}(x_{1i}, x_{2i}, j) = g_1(x_{1i}, j) + g_2(x_{2i}, j) + g_{12}(x_{1i}, x_{2i}, j)$$

In this paper we naively assume the errors are iid normal conditional on $\sum_j \varepsilon_{ij} = 0$. Though this is unrealistic, the real data examples in Section 6 show competitive performance of this approach.

4 The I-prior

Consider model (1) subject to (2), where $\mathcal{F}$ is an RKKS with reproducing kernel h. In this section we derive a prior for the regression function f based on the Fisher information on f.

As shown by Bergsma (2019), the Fisher information on f is given by

$$I[f](x, x') = \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(x, x_i) h(x, x_j)$$

$I[f]$ is positive definite and hence induces an RKHS over $\mathcal{X}$, which we denote by $\mathcal{F}_n$. Note that $\mathcal{F}_n$ is a finite dimensional subspace of $\mathcal{F}$, consisting of functions of the form $f(x) = \sum_{i=1}^n h(x, x_i) w_i$. With $\hat{f}$ an unbiased estimator of the true regression function f, the Crámer-Rao inequality implies that for any $g \in \mathcal{F}$

$$\operatorname{var}\left(\langle g, \hat{f} \rangle_{\mathcal{F}}\right) \geq \|g\|^2_{\mathcal{F}_n}$$

By standard weighted least squares theory, equality is achieved if $\hat{f}$ is a maximum likelihood estimator of f.

We can write any $f \in \mathcal{F}$ as $f = f_n + r_n$, where $f_n \in \mathcal{F}_n$ and $r_n(x_1) = \ldots = r_n(x_n) = 0$. Then $r_n \in \mathcal{F}_n^\perp$, where $\mathcal{F}_n^\perp$ is the orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$. The likelihood for f does not depend on r_n, i.e., the data contain no information on r_n, and we can replace r_n by a ‘best guess’. In this paper, we set $r_n = 0$.

We define the I-prior as a maximum entropy prior as follows. Let ν be volume measure induced by $\|\cdot\|_{\mathcal{F}_n}$. The entropy of a prior π over $\mathcal{F}_n$ relative to ν is

$$\mathcal{E}(\pi) = -\int_{\mathcal{F}_n} \pi(f) \log \pi(f) \nu(\mathrm{d}f).$$

We define the I-prior for f as the prior π maximizing entropy subject to the constraint

$$E_{f\sim\pi} \|f\|^2_{\mathcal{F}_n} = \text{constant}$$

Variational calculus shows that an I-prior for f is the Gaussian variable with mean $r_0 = 0$ and covariance kernel proportional to the Fisher information on f, i.e.,

$$\operatorname{cov}_\pi(f(x), f(x')) = \tau^2 \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(x, x_i) h(x', x_j)$$

for some $\tau > 0$. Equivalently, under the I-prior, f can be written in the form

$$f(x) = \tau \sum_{i=1}^n h(x, x_i) w_i, \qquad (w_1, \ldots, w_n) \sim \text{MVN}(0, \Psi), \tag{17}$$

Note that the ANOVA kernel described in Section 3 has a scale parameter λ_0 which makes the parameter τ superfluous, and we will omit τ in further developments.

As shown in Bergsma (2019), if $\mathcal{F}$ consists of functions $f(x) = x^\top \beta$ ($x, \beta \in \mathbb{R}^p$) with norm $\|f\|_{\mathcal{F}} = \|\beta\|_{\mathbb{R}^p}$, then under the I-prior, $\beta \sim \text{MVN}(0, X^\top \Psi X)$, where X is the $n \times p$ matrix with ith row x_i. If $\mathcal{F}$ is the Brownian motion RKHS, i.e., $f \in \mathcal{F}$ possesses a square integrable derivative, then the I-prior for f is an integrated discrete Brownian bridge, and the posterior is similar to a cubic spline smoother.

5 Maximum marginal likelihood estimation of I-prior models

In this section we describe an EM algorithm for estimating the hyperparameters of model (1) subject to (2). In this paper we use ANOVA kernels as described in Section 3, so the hyperparameters are the lambda scale parameters (of which there are $p + 1$ for p possibly multidimensional covariates) and the parameters of the error precision matrix Ψ. For example, if errors are iid, $\Psi = \psi I$ for some $\psi > 0$ which is to be estimated, or if the errors are AR(1) two real-valued parameters need to be estimated.

In Section 5.1, we give the marginal likelihood of the hyperparameters, and in Section 5.2 we describe an efficient EM algorithm for estimating them. In particular, the E step is in closed form and the M steps is either also in closed form or, with d scale parameters, requires d polynomials in d unknowns to be solved, which can be done very efficiently. We may compare this with Gaussian process regression, for which it is well-known that the M step is as complex as maximizing the marginal likelihood directly.

In this section, for readability, we switch to boldface notation for finite-dimensional vectors and matrices. Hence the matrix Ψ in (2) will be denoted by the boldface $\boldsymbol{\Psi}$.

5.1 Marginal likelihood and posterior distribution of parameter estimates

Denote $\mathbf{y} = (y_1, \dots, y_n)^\top$, $\mathbf{f} = (f(x_1), \dots, f(x_n))^\top$, $\mathbf{f}_0 = (f_0(x_1), \dots, f_0(x_n))^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$, $\mathbf{w} = (w_1, \dots, w_n)^\top$. Then (1) implies $\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}$. Let $\mathbf{H}_{\boldsymbol{\lambda}}$ be the $n \times n$ matrix with (i, j)th coordinate $h_{\boldsymbol{\lambda}}(x_i, x_j)$, where $h_{\boldsymbol{\lambda}}$ is an ANOVA reproducing kernel with scale parameter vector $\boldsymbol{\lambda}$ (see Section 3).

Under the I-prior, $\mathbf{f} \sim \text{MVN}(\mathbf{f}_0, \mathbf{H}_{\boldsymbol{\lambda}}\boldsymbol{\Psi}\mathbf{H}_{\boldsymbol{\lambda}})$, so the marginal distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \text{MVN}(\mathbf{f}_0, \mathbf{V}_{\mathbf{y}}) \tag{18}$$

where the marginal covariance is given as

$$\mathbf{V}_{\mathbf{y}} = \mathbf{H}_{\boldsymbol{\lambda}}\boldsymbol{\Psi}\mathbf{H}_{\boldsymbol{\lambda}} + \boldsymbol{\Psi}^{-1} \tag{19}$$

Thus, the marginal log likelihood of $(\boldsymbol{\lambda}, \Psi)$ is

$$L(\boldsymbol{\lambda}, \boldsymbol{\Psi}|y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_{\mathbf{y}}| - \frac{1}{2}(\mathbf{y} - \mathbf{f}_0)^\top \mathbf{V}_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{f}_0). \tag{20}$$

The maximum likelihood (ML) estimate $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\Psi}})$ of $(\boldsymbol{\lambda}, \boldsymbol{\Psi})$ maximizes $L(\boldsymbol{\lambda}, \boldsymbol{\Psi}|\mathbf{y})$, and its asymptotic distribution can be found from the Fisher information. In particular, assume $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\theta})$ and $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\theta})$ are sufficiently smooth functions of a vector $\boldsymbol{\theta}$. Then straightforward calculations give the well-known result that the Fisher information matrix $\mathbf{U}$ for $\boldsymbol{\theta}$ has (i, j)th coordinate

$$u_{ij} = \frac{1}{2}\,\text{tr}\left(\mathbf{V}_{\mathbf{y}}^{-1}\frac{\partial \mathbf{V}_{\mathbf{y}}}{\partial \theta_i}\mathbf{V}_{\mathbf{y}}^{-1}\frac{\partial \mathbf{V}_{\mathbf{y}}}{\partial \theta_j}\right),$$

where the derivatives are applied to each coordinate of the matrix. Now under well-known conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ has an asymptotic multivariate normal distribution with mean zero and covariance matrix $\mathbf{U}^{-1}$.

The next lemma gives the posterior distribution of f in (1) under the I-prior. A proof is provided in Bergsma (2019).

Lemma 1. *The posterior distribution of f in (1) subject to (2) given $\mathbf{y}$ under the I-prior π is Gaussian with mean given by*

$$E_\pi\big[f(x)|\mathbf{y}\big] = f_0(x) + \sum_{i=1}^n h(x, x_i)\hat{w}_i$$

where

$$\tilde{\mathbf{w}} = \boldsymbol{\Psi}\mathbf{H}_{\boldsymbol{\lambda}}\mathbf{V}_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{f}_0) \tag{21}$$

and covariance kernel given by

$$\text{cov}_\pi\big(f(x), f(x')|y_1, \dots, y_n\big) = \sum_{i=1}^n \sum_{j=1}^n h(x, x_i)h(x', x_j)(\mathbf{V}_{\mathbf{y}}^{-1})_{ij}$$

5.2 Estimation of parameters using the EM algorithm

We now describe the EM algorithm for estimating the scale parameter $\boldsymbol{\lambda}$ of $\mathbf{H}_{\boldsymbol{\lambda}}$, as well as parameters of the precision matrix Ψ (often it is just assumed the errors are iid $N(0, \psi^{-1})$, i.e., $\Psi = \psi \mathbf{I}_n$). For estimating these parameters, EM turns out to be particularly efficient. There can be other unknown parameters as well, e.g., the Hurst coefficient for the FBM RKHS, but for this parameter EM is computationally much less attractive, and we will not go into this.

The EM algorithm has, as is well-known, guaranteed convergence under conditions often satisfied in practice. In the present case, the E-step is in closed form while the M-step is typically not, but in practice the M-step is computationally easy to carry out.

With g denoting the density function related to its argument and using (4), the complete data log likelihood is

$$\begin{aligned}
L(\boldsymbol{\lambda}, \boldsymbol{\Psi}|\mathbf{y}, \mathbf{w}) &= \log g(\mathbf{y}|\mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\Psi}) + \log g(\mathbf{w}|\boldsymbol{\Psi}) \\
&= c + \cancel{\frac{1}{2}\log|\Psi|} - \frac{1}{2}(\mathbf{y}-\mathbf{f})^\top \boldsymbol{\Psi}(\mathbf{y}-\mathbf{f}) - \cancel{\frac{1}{2}\log|\Psi|} - \frac{1}{2}\mathbf{w}^\top \boldsymbol{\Psi}^{-1}\mathbf{w} \\
&= c - \frac{1}{2}(\mathbf{y}-\mathbf{f})^\top \boldsymbol{\Psi}(\mathbf{y}-\mathbf{f}) - \frac{1}{2}\mathbf{w}^\top \boldsymbol{\Psi}^{-1}\mathbf{w} \\
&= c - \frac{1}{2}(\mathbf{y}-\mathbf{f}_0 - \mathbf{H}_{\boldsymbol{\lambda}}\mathbf{w})^\top \boldsymbol{\Psi}(\mathbf{y}-\mathbf{f}_0 - \mathbf{H}_{\boldsymbol{\lambda}}\mathbf{w}) - \frac{1}{2}\mathbf{w}^\top \boldsymbol{\Psi}^{-1}\mathbf{w} \\
&= c - \frac{1}{2}(\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}(\mathbf{y}-\mathbf{f}_0) - \frac{1}{2}\mathbf{w}^\top \mathbf{H}_{\boldsymbol{\lambda}}^\top \boldsymbol{\Psi}\mathbf{H}_{\boldsymbol{\lambda}}\mathbf{w} + (\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}\mathbf{H}_{\boldsymbol{\lambda}}\mathbf{w} - \frac{1}{2}\mathbf{w}^\top \boldsymbol{\Psi}^{-1}\mathbf{w} \\
&= c - \frac{1}{2}(\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}(\mathbf{y}-\mathbf{f}_0) - \frac{1}{2}\mathbf{w}^\top \mathbf{V}_{\mathbf{y}}\mathbf{w} + (\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}\mathbf{H}_{\boldsymbol{\lambda}}\mathbf{w} \\
&= c - \frac{1}{2}(\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}(\mathbf{y}-\mathbf{f}_0) - \frac{1}{2}\operatorname{tr}\left[\mathbf{V}_{\mathbf{y}}\mathbf{w}\mathbf{w}^\top\right] + (\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}\mathbf{H}_{\boldsymbol{\lambda}}\mathbf{w},
\end{aligned}$$

where c is a constant. Write

$$\tilde{\mathbf{W}} = E\big(\mathbf{w}\mathbf{w}^\top \big| \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\Psi}\big) = \tilde{\mathbf{V}}_{\mathbf{w}} + \tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top,$$

where $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{V}}_{\mathbf{w}}$ are given by (21) and (19). Let $\tilde{\mathbf{w}}^{(0)}$ and $\tilde{\mathbf{W}}^{(0)}$ be $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{W}}$ with $\boldsymbol{\Psi}$ and $\boldsymbol{\lambda}$ replaced by $\boldsymbol{\Psi}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$. The E-step consists of computing

$$\begin{aligned}
Q(\boldsymbol{\lambda}, \boldsymbol{\Psi}) &= E\left\{L(\boldsymbol{\lambda}, \boldsymbol{\Psi}|\mathbf{y}, \mathbf{w})\big| \mathbf{y}, \boldsymbol{\lambda}^{(0)}, \boldsymbol{\Psi}^{(0)}\right\} \\
&= c - \frac{1}{2}(\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}(\mathbf{y}-\mathbf{f}_0) - \frac{1}{2}\operatorname{tr}\left[\mathbf{V}_{\mathbf{y}}\tilde{\mathbf{W}}^{(0)}\right] + (\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}\mathbf{H}_{\boldsymbol{\lambda}}\tilde{\mathbf{w}}^{(0)}. \qquad (22)
\end{aligned}$$

The M-step entails maximizing $Q(\boldsymbol{\lambda}, \boldsymbol{\Psi})$. We assume the global maximum can be found by differentiating, equating to zero, and solving. Supposing $\boldsymbol{\Psi}$ but not $\mathbf{H}_{\boldsymbol{\lambda}}$ depends on a parameter ψ and $\mathbf{H}_{\boldsymbol{\lambda}}$ but not $\boldsymbol{\Psi}$ depends on a parameter λ, the derivatives are given by

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\lambda}, \boldsymbol{\Psi})}{\partial \lambda} &= -\operatorname{tr}\left[\frac{\partial \mathbf{H}_{\boldsymbol{\lambda}}}{\partial \lambda}\boldsymbol{\Psi}\mathbf{H}_{\boldsymbol{\lambda}}\tilde{\mathbf{W}}^{(0)}\right] + (\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}\frac{\partial \mathbf{H}_{\boldsymbol{\lambda}}}{\partial \lambda}\tilde{\mathbf{w}}^{(0)} \\
\frac{\partial Q(\boldsymbol{\lambda}, \boldsymbol{\Psi})}{\partial \psi} &= -\frac{1}{2}(\mathbf{y}-\mathbf{f}_0)^\top \frac{\partial \boldsymbol{\Psi}}{\partial \psi}(\mathbf{y}-\mathbf{f}_0) - \frac{1}{2}\operatorname{tr}\left[\frac{\partial \mathbf{V}_{\mathbf{y}}}{\partial \psi}\tilde{\mathbf{W}}^{(0)}\right] + (\mathbf{y}-\mathbf{f}_0)^\top \frac{\partial \boldsymbol{\Psi}}{\partial \psi}\mathbf{H}_{\boldsymbol{\lambda}}\tilde{\mathbf{w}}^{(0)}.
\end{aligned}$$

For the examples in Section 6, the errors are iid, so $\boldsymbol{\Psi} = \psi \mathbf{I}_n$ for a scalar ψ, and $\mathbf{H}_{\boldsymbol{\lambda}}$ represents an ANOVA kernel as described in Section 3, so it is of the form $\mathbf{H}_{\boldsymbol{\lambda}} = \sum_{s=1}^k g_s(\boldsymbol{\lambda})\mathbf{H}_s$, where g_s is a polynomial function. Then

$$\frac{\partial \boldsymbol{\Psi}}{\partial \psi} = \mathbf{I}_n \qquad \frac{\partial \mathbf{H}_{\boldsymbol{\lambda}}}{\partial \lambda} = \sum_{s=1}^k \frac{\partial g_s(\boldsymbol{\lambda})}{\partial \lambda}\mathbf{H}_s$$

The partial derivatives of Q set to zero can then normally be solved very quickly numerically for the purposes of this paper. For example, if $g_s(\boldsymbol{\lambda}) = \lambda_s$ (which is the case, e.g., if there are no interactions in the model) then the equations have a closed form solution. In general, with iid errors, and $\mathcal{V}$ the set of (possibly multidimensional) variables involved, $2|\mathcal{V}| + 1$ polynomial equations in $2|\mathcal{V}| + 1$ unknowns need to be solved, which we found can be done very rapidly using built in solvers in R and Mathematica. The computational bottleneck is not the M step, but the E step which is $O(n^3)$.

We did sometimes encounter accuracy problems, making it impossible to obtain full convergence of the EM algorithm (this problem was encountered and mentioned in Section 6.4). This is not a problem just for EM, as the marginal likelihood is difficult to estimate as well, see Bergsma (2019) for a visualization of this problem. For prediction purposes, the lack of convergence did not seem to matter, but the lack of an accurate maximal value of the marginal likelihood can make model comparison difficult.

6 Application to data

We reanalyze some well-known data sets in the respective areas of multilevel modelling, functional data analysis, classification, and longitudinal data analysis. Whereas in the literature different methods are typically used in different areas, and often more than method per area, we fit all models using the single method introduced in this paper. In all cases we obtain a performance competitive with existing techniques in terms of mean squared prediction error for test data. Furthermore, our methodology is flexible in that it is easy to incorporate extra covariates and interaction effects using the parsimonious ANOVA framework given in Section 3.

This section is organized as follows. In Section 6.3 we fit the standard varying intercept and varying slope models using the I-prior. The purpose of this section is mainly to illustrate the differences between the I-prior approach and the standard random effects approach. Here, the I-prior method has an estimation advantage in that there are no positive definiteness restriction on a latent covariance matrix. In Section 6.4 we look at multi-class classification, both with high and low dimensional covariates. The latter two sections illustrate the potentially good predictive performance of the I-prior methodology compared to other methods. Furthermore, they illustrate the ease with which high-dimensional smoothing can be done using the I-prior. In Section 6.5 we do a longitudinal data analysis with I-priors. Here, we treat the longitudinal response curves as ‘functional’ responses. In contrast to standard approaches, we do not need to specify a covariance structure for the longitudinal responses, instead we merely need to specify an appropriate class of functions, e.g., a class of smooth functions.

Throughout this section, we will use the parsimonious ANOVA approach described in Section 3.2 for which interaction effects do not require extra parameters. In Section 6.5, we will compare this approach with the ‘classical’ non-parsimonious extended ANOVA one (also under the I-prior) where each interaction effect does take an extra parameter.

6.1 Motivation for use of FBM RKHS

As explained in more detail in Bergsma (2019), the use of the I-prior methodology is particularly attractive if $\mathcal{F}$ is a fractional Brownian motion (FBM) RKHS over a Euclidean space (which has as a special case the aforementioned centered Brownian motion). FBM *process paths* are non-differentiable and, having Hölder smoothness ranging between 0 and 1, an FBM process prior for the regression function may be too rough for many applications. In contrast, functions in the FBM *RKHS* with Hurst coefficient γ are (weakly) differentiable if the Hurst $\gamma \geq 1/2$ and

RKHS/RKKS	Name of effect
Centered canonical	Linear
Pearson	Nominal
Centered FBM	Smooth
Polynomial	Polynomial
ANOVA	Interaction

Table 4: Terminology for regression modelling, allowing a description of regression models without referring to RKKSs and RKHSs. For example, we say a (possibly multidimensional) covariate has a linear effect on the response if the corresponding regression function lies in the centered canonical RKKS.

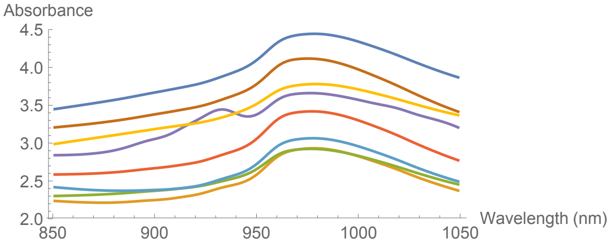


Figure 1: Sample of spectrometric curves used to predict fat content of meat

have minimum Hölder smoothness 2γ. This wide range of smoothnesses make it an attractive general purpose function space for nonparametric regression. Another advantage is that it allows us to do multivariate smoothing with just one or two parameters to be estimated: either only the scale parameter λ, while using a default setting of, say, 1/2 for the Hurst coefficient, or both the scale parameter and the Hurst coefficient. This is in contrast with standard kernel based smoothing methods, which require a scale parameter and at least one kernel hyperparameter to be estimated. For example, if we use the exponential kernel

$$r(x,x') = \exp\Big(-\frac{\|x-x'\|^{2\xi}}{2\sigma^2}\Big), \tag{23}$$

the scale parameter λ, the smoothness parameter ξ (somewhat analogous to the Hurst coefficient), and a 'variance' parameter σ^2 need to be estimated. Default settings $\xi = 1$ or $\xi = 2$ could be used to reduce the number of free parameters to two. Furthermore, the functions in the squared exponential RKHS are analytical, which is too smooth for many applications.

6.2 Regression with a functional covariate

We illustrate the prediction of a real valued response when one of the covariates is a function using a widely analysed data set used for quality control in the food industry. The data consist of measurements on a sample of 215 pieces of finely chopped meat. The response variable is fat content, and the covariate is light absorbance for 100 different wavelengths. The absorbance curve can be considered a 'functional' variable (see a sample of such curves plotted in Figure 1). For more details see http://lib.stat.cmu.edu/datasets/tecator and Thodberg (1996). Our aim is to predict fat content from the 100 measurements of absorbance. The first 172 observations in the data set are used as a training sample, and the remaining 43 observations are used as a test sample (following Thodberg's original recommendation).

Method	RMSE	
	Training	Test
Global constant model	12.50	13.3
Neural network (Vila et al., 2000)		0.34
Kernel smoothing (Ferraty & Vieu, 2006, Section 7.2)		1.85
Double index model (Chen, Hall, & Müller, 2011)		1.58
Single index model (Goia & Vieu, 2014)		1.18
Sliced inverse regression (Lian & Li, 2014)		0.90
MARS (Zhu, Yao, & Zhang, 2014)		0.88
Partial least squares (Zhu et al., 2014)		1.01
CSEFAM (Zhu et al., 2014)		0.85
Tikhonov regularization (linear)	3.32	3.54
Tikhonov regularization (FBM-1/2 kernel)	4.32	4.54
I-prior (linear)	2.82	3.15
I-prior (FBM RKHS with $\gamma = 0.5$)	0.00	0.67
I-prior (FBM RKHS with $\hat{\gamma} = 0.98$)	0.00	0.57
I-prior (squared exponential RKHS, $\hat{\sigma} = 0.0079$)	0.35	0.58

Table 5: RMSEs for predicting fat content from spectrometric functional covariate (see Figure 1): previously published results, Tikhonov regularization, and I-prior methodology.

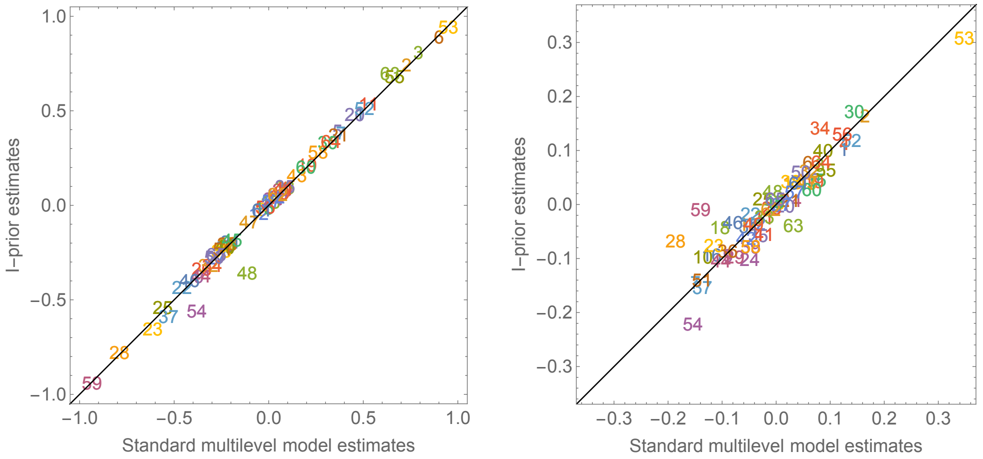
Many different methods have been applied in the literature to the data set, estimating a model using the training sample and evaluating its performance using the test sample. One of the best results was achieved early on by Thodberg (1996), who used neural networks on the first 10 principal components and achieved a test mean squared error of 0.36. The best test error performance we found was by Vila, Wagner, and Neveu (2000) who achieved an error rate of 0.34, also using neural networks on the principal components. More recently various other statistical models have been tried on the data set, see Table 5 for a summary. In spite of their lesser performance compared to neural networks, the interest of these methods is that they do not rely on an a priori data reduction in terms of the main principal components.

The ith spectral curve is denoted x_i, with $x_i(t)$ denoting the absorbance for wavelength t. We assume $x_i \in \mathcal{X}$, where $\mathcal{X}$ is a set of functions over $\mathbb{R}$ and is equipped with an appropriate inner product. From Figure 1 it appears the curves are differentiable, so it seems reasonable to assume the $\mathcal{X}$ is the Brownian motion RKHS over $\mathbb{R}$ with squared norm

$$\|x\|_{\mathcal{X}}^2 = \int_{\mathbb{R}} \dot{x}(t)^2 dt.$$

Note that the L^2 inner product for a set of functions will rarely be appropriate.

Since $\mathcal{X}$ is a Hilbert space, a linear effect of the spectral curve on fat content can now be modelled using the canonical RKHS over $\mathcal{X}$. We see in Table 5 that both Tikhonov regularization and the I-prior give a poor performance, with test RMSEs of 3.54 and 2.89, respectively. Next we fitted a smooth dependence of fat content on spectrometric curve using the FBM RKHS over $\mathcal{X}$. As seen in the table, Tikhonov regularization performs very poorly. We tried various values of the Hurst coefficient, but all give worse results than the linear model. On the other hand, the I-prior performs rather well for different RKHSs, including the FBM and the squared exponential ones. We had some convergence problems so could not get the ML estimator of γ, the Hurst coefficient for the FBM RKHS, so instead estimated it by minimizing the cross-validation error (10-fold cross-validation gave $\hat{\gamma} = 0.98$). For the squared exponential RKHS we did manage to find the ML estimator $\hat{\sigma}$ of σ, and it is given in Table 5.



(a) Intercepts (varying intercept model)

(b) Slopes (varying slope model)

Figure 2: Estimated intercepts and slopes for school achievement data under varying intercept and varying slope model. The numbers are the school indices. In contrast to the I-prior, the standard random effects multilevel model assumes iid slopes and intercepts across schools. It is seen that I-prior estimated intercepts are bigger in absolute value than the standard ones for small schools (48 and 54; see text).

Instead of fat content, protein content can be predicted from the spectral curve. With the I-prior based on a smooth dependence of protein content on the spectral curve we obtained an RMSE of 0.52, using a local (non-global) maximum likelihood estimate of the Hurst coefficient, $\hat{\gamma} = 0.997$. This improves on Zhu et al. (2014) who obtained an RMSE of 0.85.

6.3 Multilevel models

The purpose of this section is to illustrate the differences between the standard random effects and the I-prior approaches for estimating varying intercept and varying slope models. We consider a data set which accompanies the MLwiN software (Rasbash, Steele, Browne, & Goldstein, 2012) on school achievement of 4059 pupils at 65 inner-London schools. The response variable is the GCSE score at age 16.

First we consider the varying intercept model (11). The ‘covariate’ is the nominal variable school (ranging from 1 to 65), which we assume to have a nominal effect on GCSE score, i.e., we assume that the regression function f in (11) is in the Pearson RKHS. A standard approach to fitting this model is the random intercept model, which is based on the assumption that the intercepts are iid normal with zero mean. In Figure 2(a), the posterior means of the intercepts are plotted for the standard (random effects) model and the I-prior. It can be seen the estimates are in broad agreement, with conspicuously different estimates for schools 48 (-0.11 vs. -0.36) and 54 (-0.38 vs. -0.56), the I-prior giving the largest estimate in absolute value in both cases. The reason for the relatively large I-prior estimates can be found from the prior variance formula ... which implies that the smaller the sample size for a particular school, the larger the prior variance of the intercept for that school. Indeed, schools 48 and 54 have the smallest sample sizes of all schools, namely 2 and 8 (the next smallest school is number 37, with 22 students).

Next we consider the varying slope model (14), which regresses, for each school, the GCSE score on the result of the London reading test (LRT), taken at age 11. A standard approach to fitting this model is the random intercept/slope model, which is based on the assumption that the intercept/slope pairs are iid bivariate normal with zero mean. To obtain an I-prior, we assume as above a nominal effect of school, and a linear effect of LRT (i.e., $\mathcal{F}_2$ in (**??**) is the canonical RKHS). In Figure 2(b), the posterior means of the slopes obtained using the standard random effects model are plotted against the ones obtained using the I-prior. Again we see broad agreement of the estimates, but much less so than for the varying intercept model.

A limited cross-validation study yielded on average a small advantage of the standard random effects approach in terms of mean squared error, in the order of half a percent, indicating the iid assumption is reasonable. An advantage of the I-prior is that no a priori assumption about the distribution of the parameters need to be made. Furthermore, as discussed in Section 1, our approach is more parsimonious and allows potentially simpler estimation and testing.

6.4 Multi-class classification

We apply the model described in Section 3.3.4 to two data sets that have received widespread attention in the literature, one for which the most important covariate is low dimensional and one for which it is high dimensional. In both cases, the classes $1, \ldots, T$ are unordered. The effect of x is first assumed to be linear, then smooth, and the results are compared. We show that the classifier obtained from the I-prior is competitive with the other classifiers that we were able to find in the literature. Bergsma (2019) previously considered I-prior modelling for two-class classification.

The first problem was originally presented by Ramaswamy et al. (2001), and concerns the prediction of cancer type based on 16,063 gene expression measurements. A training set of 144 patients with 14 different types of cancer is available, as well as a test set of 54 patients which can be used to assess the performance of a classifier. Assuming a linear dependence on class of the covariate vector of 16,063 gene expressions, we obtained 0 training errors and test 12 errors, which competes well with other methods (see Table 6). In particular our training error rate far outperforms competing methods. Modelling a smooth dependence of class on the covariate vector gave some numerical problems and we were unable to obtain full convergence of the EM algorithm. Fortunately, the classification errors did not seem to be affected by this, so we are confident these are accurate (the maximum of the likelihood we obtained, not given here, was inaccurate). For $0.72 < \gamma < 0.86$ the training/test errors were 0/10, which was the best we could achieve.

The second problem concerns vowel recognition based on a 10-dimensional vector computed from a voice recording. The training and test sets are based on recordings of 8 resp. 7 people, each of whom spoke 11 different vowels 6 times. As seen in Table 7, if a linear effect of the covariates is assumed, the I-prior gives no advantage compared to ordinary least squares. The reason is that there are only a small number of predictors. Again we had some numerical difficulties fitting a smooth model, but the MLE seems to be $\hat{\gamma} = 0.652$ giving training/test error rates 0/0.35, improving on the results given in Hastie, Tibshirani, and Friedman (2009).

6.5 Longitudinal data analysis

We consider a balanced longitudinal data set consisting of weights and 60 cows, 30 of which are randomly assigned to a treatment group A and 30 to a treatment group B. Weight was measured 11 times over a 133-day period, at two-week intervals, except for the last measurement, which was taken one week after the preceding measurement. As the response variable of

Method	Training errors Out of 144	Test errors Out of 54
Nearest neighbors	41	26
L^2-penalized discriminant analysis	25	12
Support vector classifier	26	14
Lasso	30.7	12.5
L^1 penalized multinomial	17	13
Elastic net penalized multinomial	22	11.8
SCRDA (Guo, Hastie, & Tibshirani, 2007)	24	8
Scout (Witten & Tibshirani, 2011)	21	8
I-prior (linear)	0	12
I-prior (smooth, $\gamma = 0.8$)	0	10

Table 6: Comparison of classifier performance for the 14 cancer classification problem. The covariate is high-dimensional, consisting of a vector of 16,063 gene expressions. The top six lines are taken from Hastie et al. (2009), Table 18.1. With the I-prior, high-dimensional smoothing is as straightforward as fitting a linear effect, while giving slightly better test performance. We did have some numerical problems with maximum likelihood estimation of the Hurst coefficient γ, which we took to be 0.8.

Method	Error rates	
	Training	Test
Nearest neighbours		0.44
OLS regression (linear effects)	0.48	0.67
Linear discriminant analysis	0.32	0.56
Neural network (Gaussian nodes)		0.45
FDA/BRUTO	0.06	0.44
FDA/MARS (best reduced dimension)	0.13	0.39
I-prior (linear)	0.48	0.67
I-prior (smooth, $\hat{\gamma} \approx 0.652$)	0	0.35

Table 7: Comparison of classifier performance for the vowel classification problem (first six lines taken from Hastie et al. (2009)). Since the dimension of the covariate is low (equal to 10), the I-prior gives no advantage over ordinary least squares (OLS) if the covariate effect is linear.

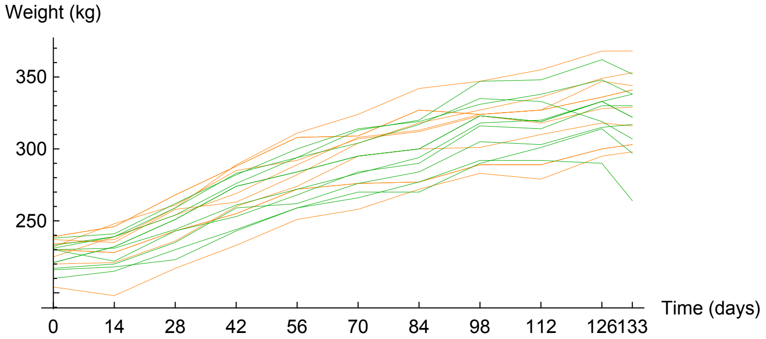


Figure 3: Sample of 20 (out of available 60) growth curves of cows. The two colors green and orange represent the two treatments received. The question of interest is whether and how treatment affects growth.

interest we take the weight growth curve. In Figure 3 a sample of growth curves is shown. Of interest is investigating whether a treatment effect is present, and if it is, to assess its nature.

The usual approach to analyze a longitudinal data set such as this one is to assume that the observed growth curves are realizations of a Gaussian process. For example, Kenward (1987) assumed a so-called ante-dependence structure of order k, which assumes an observation depends on the previous k observations, but given these is independent of any preceding observations. Various other process families have been considered (Núñez-Antón & Zimmerman, 2000; Pourahmadi, 2000; Pan & Mackenzie, 2003; Zhang, Leng, & Tang, 2014), see the latter for an overview. Steele (2008) points out the connections with multilevel and structural equation modelling.

Using the I-prior, it is not necessary to assume the growth curves were drawn randomly, instead, it suffices to assume they lie in an appropriate function class. In this section we assume this function class is the FBM RKHS, i.e., we assume a 'smooth' effect of time on weight (see Table 4). The growth curves form a multidimensional (or functional) response, so we can use the multidimensional/functional response model of Section 3.3.3. In the present case we have two covariates potentially influencing growth, namely cow index (C, indexed by i) and treatment (X), and we can write the regression model as

$$y_{it} = f(i, x_i, t) + \varepsilon_{it},$$

where y_{it} is weight of cow i at time t, and $x_i \in \{A, B\}$ is the treatment group of cow i. As discussed in Section 3.3.3, we can also interpret the model as a unidimensional response model, where the unidimensional response is weight, and covariates are cow index (C), treatment (X), and time (T). A main effect of C on (multidimensional) *growth* is then equivalent to an interaction effect of T and C on (unidimensional) *weight*.

Covariate	Range	Effect	RKHS
Time (T)	$0 - 133$ days	Smooth	Centered FBM ($\gamma = 0.3$)
Cow index (C)	$\{1, \ldots, 60\}$	Nominal	Pearson
Treatment (X)	$\{A, B\}$	Nominal	Pearson

Table 8: Covariates used for modelling cow data

We assume iid errors and in addition to a smooth effect of time we assume a nominal effect of both C and X (see Table 8). We can estimate the Hurst coefficient γ of the FBM RKHS using the maximum likelihood estimator, but a difficulty is that it varies a lot across models (for the present data set, we found estimates between 0.2 and 0.4). To make model comparison easier, we took as a compromise a fixed value $\gamma = 0.3$ for all models; we found that substantive conclusions were not greatly affected by the choice of γ. The model asserting that the growth curve does not vary with treatment or among cows is denoted $\{\}$ (no effect of either C or X on growth), and can be written as

$$y_{it} = \alpha + f(t) + \varepsilon_{it}.$$

The model asserting that the growth curve depends on cow but without a treatment effect is denoted $\{C\}$, and can be written as

$$y_{it} = \alpha + \beta_i + f(t) + f_i^C(t) + \varepsilon_{it},$$

subject to the identifying constraints $\sum_t f(t) = \sum_t f_i^C(t) = 0$ and $\sum_{i=1}^{60} f_i^C(t) = 0$ for all t. Here t ranges over $0, 14, \dots, 133$. Further, f and the f_i^C are functions in the centered FBM RKHS. The model which includes a treatment effect on growth, but without an interaction effect of treatment and cow index, is denoted $\{C, X\}$ and can be formulated as

$$y_{it} = \alpha + \beta_i + \xi_{x_i} + f(t) + f_i^C(t) + f_{x_i}^X(t) + \varepsilon_{it},$$

where $f_{x_i}^X$ also lies in the FBM RKHS. It can be seen that, since treatment x_i is determined by the cow index i, the model is not identified in the sense that model $\{C\}$ contains exactly the same functions as model $\{C, X\}$ (and similarly, model $\{CX\}$ also contains the same functions as model $\{C\}$). However, the inner products of these ANOVA RKHSs differ, so they also have different I-priors, allowing us to perform model selection.

In Section 3.2, the parsimonious ANOVA RKHS and the less parsimonious extended ANOVA RKHS were described. The former requires a parameter for each covariate which has at least a main effect, but no extra parameters for interaction effects; the latter requires one parameter for each main effect and each interaction effect. The two approaches are different, as can be seen in the summary in Table 9, but reassuringly yield the same substantive conclusions.

Let us first consider the ANOVA approach. We can test for a treatment effect by testing model $\{C, X\}$ against the alternative that $\{C\}$ is true. The log-likelihood ratio is $(-2242.30) - (-2266.39) = 24.09$ and has an asymptotic chi-square distribution with $3 - 2 = 1$ degree of freedom if $\{C, X\}$ is true. The p-value is less than 10^{-6} so we conclude that model $\{C, X\}$ is significantly better than model $\{C\}$. We can next investigate whether the treatment effect differs among cows by comparing models $\{CX\}$ and $\{C, X\}$. As these two models have the same number of ANOVA (λ) parameters, we can simply choose the one with the highest likelihood, which is model $\{C, X\}$. We conclude that treatment has an effect on growth, but there is no reason to assume the effect of treatment differs among cows.

The extended ANOVA approach resembles classical least squares model selection since there are extra parameters for interaction effects. In this case model $\{C, X\}$ has five parameters, two more than model $\{C\}$, because X has a non-time-varying main effect and a time-varying effect on weight. Using likelihood ratio testing the best model is found to be $\{C, X\}$, giving the same result as in the parsimonious approach. Interestingly, $\{C, X\}$ and $\{CX\}$ give exactly the same fit for this data set, so we choose the simplest model $\{C, X\}$, which is the same result as for the ANOVA approach.

Time varying covariates were not available for the present data set but could easily be added. As the data set is balanced (the cows were weighted at the same time points), fitting

Type of RKKS	Model	Log-likelihood	Error standard deviation	Number of λ parameters
ANOVA	$\{\}$	−2792.78	16.3	1
	$\{X\}$	−2792.73	16.3	2
	$\{C\}$	−2266.39	2.7	2
	$\{C, X\}$	−2242.30	2.5	3
	$\{CX\}$	−2251.30	3.3	3
Extended ANOVA	$\{\}$	−2792.78	16.3	2
	$\{X\}$	−2792.65	16.3	3
	$\{C\}$	−2250.31	3.6	3
	$\{C, X\}$	−2226.32	3.3	5
	$\{CX\}$	−2226.32	3.3	7

Table 9: Goodness of fit for cow data, see Table 8 for the covariates used. The model consists of the highest order effects on the growth curve, e.g., model $\{C, X\}$ means the growth curve depends on cow (C) and treatment (X), and there is no interaction meaning that the treatment effect is the same for all cows. In the ANOVA models, no extra parameters are needed for interaction effects, and among nested models with the same number of parameters the one with the highest likelihood can be chosen. For the extended ANOVA models, each interaction requires at least one extra parameter and model selection is via the likelihood ratio test.

can be done more efficiently without time varying covariates, however, we have 60 cows and 11 time points, giving $60 \times 11 = 660$ data points, which is easy to handle in any case.

7 Discussion and topics for further research

The most important reasons for using the I-prior methodology are that, as we aimed to show, it is relatively easy to use, flexible, and yields good predictive power. In particular, the I-prior methodology, consisting of an estimation method and a modelling method, can be used as a single solution for a wide range of applications where a large variety of methods have been used in the literature. For some well studied data sets we show that predictive performance is competitive with existing methods. The use of a full likelihood framework facilitates inference.

References

Alpay, D. (1991). Some remarks on reproducing kernel Krein spaces. *Rocky Mountain J. Math.*

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, *68*(3), 337–404.

Bartholomew, D., Steele, F., Moustaki, I., & Galbraith, J. (2008). *Analysis of multivariate social science data.* Routledge.

Bergsma, W. (2019). Regression with i-priors. *Econometrics and Statistics*.

Berlinet, A., & Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics.* Kluwer Academic.

Bernardo, J. M. (2005). Reference analysis. *Handbook of statistics*, *25*, 17–90.

Canu, S., Ong, C. S., & Mary, X. (2009). Splines with non positive kernels. In W. S. Publishing (Ed.), *More progress in analysis* (pp. 163–173). Singapore.

Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, *2*, 1281–1299.

Chakraborty, A., & Panaretos, V. M. (2019, 08). Hybrid regularisation and the (in)admissibility of ridge regression in infinite dimensional Hilbert spaces. *Bernoulli*, *25*(3), 1939–1976. Retrieved from `https://doi.org/10.3150/18-BEJ1041` doi: 10.3150/18-BEJ1041

Chen, D., Hall, P., & Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Stat.*, *39*(3), 1720–1747.

Choi, T., & Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, *98*(10), 1969–1987.

Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Stat.*, 903–923.

Cressie, N. A. (1993). *Statistics for spatial data, revised edition.* Wiley, New York.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1–32.

Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.) , Vol. III, European Mathematical Society, Zurich, 595-622*.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis.* Springer Science + Business Media.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

Gheondea, A. (2013). A survey on reproducing kernel Krein spaces. *arXiv preprint arXiv:1309.2393*.

Goia, A., & Vieu, P. (2014). Some advances on semi-parametric functional data modelling. *Contributions in infinite-dimensional statistics and related topics*, 135.

Goldstein, H. (2011). *Multilevel statistical models.* Wiley.

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *In: algorithmic learning theory: 16th international conference; Springer*, *1*, 63-78.

Gu, C. (2013). *Smoothing spline anova models.* Springer.

Gu, C., & Wahba, G. (1993). Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society. Series B (Methodological)*, 353–368.

Guo, Y., Hastie, T., & Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, *8*(1), 86–100.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. B*, 757–796.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* New York: Springer.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Ann. Stat.*, 1171–1220.

Jamil, H. (2018). *Regression modelling using priors depending on Fisher information covariance kernels (i-priors)* (Unpublished doctoral dissertation). The London School of Economics and Political Science (LSE).

Jamil, H. (2019). iprior: Regression modelling using i-priors [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=iprior` (R package version 0.7.3)

Jamil, H., & Bergsma, W. (2019). iprior: An R package for regression modelling using I-priors. *https://arxiv.org/abs/1912.01376*.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical review*, *106*(4), 620.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. ii. *Physical review*, *108*(2), 171.

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics*, 296–308.

Kimeldorf, G. S., & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.*, *41*(2), 495–502.

Knapik, B., van der Vaart, A., & van Zanten, J. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Stat.*, *39*(5), 2626–2657.

Lancaster, H. O. (1969). *The chi-squared distribution.* New York: Wiley.

Lian, H., & Li, G. (2014). Series expansion for functional sufficient dimension reduction. *Journal of Multivariate Analysis*, *124*, 150–165.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in applied probability*, 439–468.

Núñez-Antón, V., & Zimmerman, D. L. (2000). Modeling nonstationary longitudinal data. *Biometrics*, *56*(3), 699–705.

Ong, C. S., Mary, X., Canu, S., & Smola, A. J. (2004). Learning with non-positive kernels. In *ICML 21* (p. 639-646).

Pan, J., & Mackenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, *90*(1), 239–244.

Parzen, E. (1961). An approach to time series analysis. *Ann. Math. Stat.*, 951–989.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, *87*(2), 425–435.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., ... others (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, *98*(26), 15149–15154.

Ramsay, J. O., & Dalzell, C. (1991). Some tools for functional data analysis (with discussion). *J. Roy. Statist. Soc. B*, 539–572.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2012). A user's guide to MLwiN, v2.26. *Centre for Multilevel Modelling, University of Bristol*.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Stat. Soc. B*, *71*(2), 319–392.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press.

Schwartz, L. (1964). Sous-espaces Hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'analyse mathématique*, *13*(1), 115–256.

Seeger, M., Kakade, S. M., & Foster, D. P. (2008). Information consistency of nonparametric Gaussian process methods. *Information Theory, IEEE Transactions on*, *54*(5), 2376–2382.

Sejdinovic, D., Gretton, A., & Bergsma, W. (2013). A kernel test for three-variable interactions. In *Advances in neural information processing systems* (pp. 1124–1132).

Shi, J. Q., & Choi, T. (2011). *Gaussian process regression analysis for functional data.* CRC Press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman and Hall.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Sage Publications Limited.

Steele, F. (2008). Multilevel models for longitudinal data. *J. Roy. Statist. Soc. A*, *171*(1), 5–19.

Steinwart, I., & Christmann, A. (2008). *Support vector machines.* Springer.

Thodberg, H. H. (1996). A review of Bayesian neural networks with an application to near infrared spectroscopy. *Neural Networks, IEEE Transactions on*, *7*(1), 56–72.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *The journal of machine learning research*, *1*, 211–244.

van der Vaart, A. W., & van Zanten, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Stat.*, 1435–1463.

van der Vaart, A. W., & van Zanten, J. H. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh* (pp. 200–222). Institute of Mathematical Statistics.

Vila, J.-P., Wagner, V., & Neveu, P. (2000). Bayesian nonlinear model selection and neural networks: a conjugate prior approach. *Neural Networks, IEEE Transactions on*, *11*(2), 265–278.

Wahba, G. (1990a). Multivariate model building with additive interaction and tensor product thin plate splines. In *Curves and surfaces* (pp. 491–504). Elsevier.

Wahba, G. (1990b). *Spline models for observational data* (Vol. 59). Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Wasserman, L. (2006). *All of nonparametric statistics.* New York: Springer.

Witten, D. M., & Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *J. Roy. Statist. Soc. B*, *73*(5), 753–772.

Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, *33*(6), 2873–2903.

Zhang, W., Leng, C., & Tang, C. Y. (2014). A joint modelling approach for longitudinal studies. *J. Roy. Statist. Soc. B*.

Zhu, H., Yao, F., & Zhang, H. H. (2014). Structured functional additive regression in reproducing kernel Hilbert spaces. *J. Roy. Statist. Soc. B*, *76*(3), 581–603.

A ANOVA kernel construction

With $\mathcal{V}$ a finite set of variables and any $v \in \mathcal{V}$, let $\mathcal{X}_v$ be nonempty set, let $\mathcal{F}_v$ be the RKKS over $\mathcal{X}_v$ with kernel h_v, and let $\mathcal{C}_v$ be the RKHS of constant functions over $\mathcal{X}_v$ with kernel c_v defined by $c_v(x, x') = 1$. Denote the power set of a set A by $\mathbb{P}(A)$, and define the power set of a set of subsets $\mathcal{A}$ as $\mathbb{P}(\mathcal{A}) = \cup_{A \in \mathcal{A}} \mathbb{P}(A)$. Let $\mathcal{A}$ be a Sperner family of $\mathcal{V}$, i.e., $\mathcal{A}$ consists of subsets of $\mathcal{V}$ such that no set in $\mathcal{A}$ contains another set in $\mathcal{A}$. Define the RKKS $\mathcal{F}_\mathcal{A}$ over $\times_{v \in \mathcal{V}} \mathcal{X}_v$ as

$$\mathcal{F}_\mathcal{A} = \sum_{A \in \mathbf{P}(\mathcal{A})} \bigotimes_{v \in A} \mathcal{F}_v \bigotimes_{v \in V \setminus A} \mathcal{C}_v. \tag{24}$$

Then $f \in \mathcal{F}_\mathcal{A}$ are of the form

$$f(x) = \sum_{A \in \mathbf{P}(\mathcal{A})} f_A(x_A)$$

for $x \in \times_{v \in \mathcal{V}} \mathcal{X}_v$, with x_A the subvector of x which retains the coordinates corresponding to A. The reproducing kernel of $\mathcal{F}_\mathcal{A}$ is given as

$$h_\mathcal{A} = \sum_{A \in \mathbf{P}(\mathcal{A})} \bigotimes_{v \in A} h_v \bigotimes_{v \in V \setminus A} c_v.$$