Regression with I-priors

LSE Research Online URL for this paper: http://eprints.lse.ac.uk/102136/

Version: Accepted Version

Article:

Bergsma, Wicher (2019) Regression with I-priors. Econometrics and Statistics. ISSN 2452-3062 (In Press)

Regression with I-priors

Wicher P. Bergsma

London School of Economics, Houghton Street, London, WC2A 2AE, United Kingdom

Abstract

The problem of estimating a parametric or nonparametric regression function in a model with normal errors is considered. For this purpose, a novel objective prior for the regression function is proposed, defined as the distribution maximizing entropy subject to a suitable constraint based on the Fisher information on the regression function. The prior is named I-prior. For the present model, it is Gaussian with covariance kernel proportional to the Fisher information, and mean chosen a priori (e.g., 0).

The I-prior has the intuitively appealing property that the more information is available about a linear functional of the regression function, the larger its prior variance, and, broadly speaking, the less influential the prior is on the posterior. Unlike the Jeffreys prior, it can be used in high dimensional settings. The I-prior methodology can be used as a principled alternative to Tikhonov regularization, which suffers from well-known theoretical problems which are briefly reviewed.

The regression function is assumed to lie in a reproducing kernel Hilbert space (RKHS) over a low or high dimensional covariate space, giving a high degree of generality. Analysis of some real data sets and a small-scale simulation study show competitive performance of the I-prior methodology, which is implemented in the R-package `iprior`.

Keywords: reproducing kernel, RKHS, Fisher information, maximum entropy, objective prior, g-prior, empirical Bayes, regression, nonparametric regression, functional data analysis, classification, Tikhonov regularization.
2010 MSC: 62G08, 62C12

1. Introduction

Consider a sample $(x_1, y_1), \ldots, (x, y_n)$, where y_i is a real-valued measurement on unit i, and x_i lies in a set $\mathcal{X}$ and represents some characteristic or collection of characteristics, numerical or otherwise, of unit i. Furthermore, let $\Psi = (\psi_{ij})$ be an $n \times n$ positive definite matrix, h a symmetric positive definite kernel over $\mathcal{X}$, and $\mathcal{F}$ a set of real-valued functions over $\mathcal{X}$. In this paper we consider the

regression model

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

where

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

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 i.i.d. errors. We shall further assume that $\mathcal{F}$ is a reproducing kernel Hilbert space, i.e., $\mathcal{F}$ possesses a reproducing kernel h (see Section 2.1 for more details). For further reference, we write

$$\mathcal{F} \text{ is an RKHS over } \mathcal{X} \text{ with reproducing kernel } h. \tag{3}$$

An RKHS is a Hilbert space of functions for which point evaluation is a continuous linear functional, i.e., functions which are sufficiently close in norm are also pointwise close. It follows that a normwise consistent estimator of f is also pointwise consistent. The assumption that $\mathcal{F}$ is an RKHS has the further benefit that the Fisher information on f in (1) subject to (2) exists. Our proposed methodology is general in that essentially arbitrary covariate spaces $\mathcal{X}$ and RKHSs $\mathcal{F}$ can be used.

If the dimension of $\mathcal{F}$ is high compared to n, the maximum likelihood (ML) estimator of f is typically of little use, for example, it may interpolate the data. Only one generally applicable and 'automatic' (i.e., requiring no additional user choices) estimation method for f in (1) subject to (2) and (3) appears to have been described in the literature, namely *Tikhonov regularization*. The Tikhonov regularizer can be defined as the minimizer of the function from $\mathcal{F}$ to $\mathbb{R}$ defined by the mapping

$$f \mapsto \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} \big(y_i - f(x_i) \big) \big(y_j - f(x_j) \big) + \lambda^{-1} \| f - f_0 \|_{\mathcal{F}}^2, \tag{4}$$

where $\lambda > 0$ is a scale (or smoothness) parameter (usually estimated using cross-validation), f_0 is a prior 'best guess' of f, and the first term on the right hand side is minus two times the log-likelihood of f up to a constant. However, Chakraborty and Panaretos (2019) showed that the Tikhonov regularizer is essentially inadmissible with respect to squared error loss if $\mathcal{F}$ is infinite dimensional.

To overcome this problem, it seems reasonable to use a Bayes or empirical Bayes approach, assigning a prior to f whose support is a subset of $\mathcal{F}$: Wald's complete class theorem then ensures admissibility. In this paper we propose a

novel objective prior for f, where by 'objective' we mean automatically generated without further user input after the model has been chosen. Our prior is called *I-prior*, where the 'I' refers to (Fisher) information, and is defined as a maximizer of entropy subject to a suitable constraint based on the Fisher information. Following Jaynes (2003) the I-prior can thus be thought of as 'least informative'. For the present problem the I-prior is Gaussian, with prior mean f_0, and covariance kernel proportional to the Fisher information on f. Under the I-prior, f has the simple representation

$$f(x) = f_0(x) + \lambda \sum_{i=1}^{n} h(x, x_i) w_i, \qquad (w_1, \ldots, w_n) \sim \text{MVN}(0, \Psi), \lambda > 0. \qquad (5)$$

Since $h(\cdot, x_i) \in \mathcal{F}$ (see Section 2.1), this representation immediately shows that I-prior realizations are in $\mathcal{F}$.

From an intuitive perspective, the I-prior is reasonable because if the Fisher information on a linear functional of f is high, the linear functional will have a high prior variance, and the posterior mean may be largely determined by the data; if on the other hand little Fisher information is available for a particular linear functional, the prior variance will be small, and the posterior mean may be largely determined by the prior mean.

It can be seen that the I-prior depends on the data $x_1, \ldots, x_1$ and its support is only a subset of $\mathcal{F}$, which is justified in Sections 2.3 and 2.5.

An alternative approach to estimating f in (1) is for the user to choose a prior over the space of functions $\mathcal{F}$ and compute the posterior distribution. Such a choice can be made based on subjective beliefs, prior knowledge, or theoretical or practical considerations. If the chosen prior is Gaussian, this method is called Gaussian process regression (GPR). A general class of Gaussian and Lévy process priors over an RKHS was characterized by Pillai et al. (2007).

In the I-prior approach, a number of hyperparameters remain undetermined: the scale parameter λ in (5), any parameters of the error precision matrix Ψ, and potentially parameters of the kernel h. In the application Section 6, we used an empirical Bayes approach whereby these hyperparameters are estimated by their maximum likelihood estimators. Although the philosophical aspects of empirical Bayes have not been fully resolved (see Efron, 2019, with discussion), estimating hyperparameters from the data is in line with the usual approach in the random effects literature and the regularization literature. Naturally, a fully Bayes approach could be used as well by assigning priors to the hyperparameters, but we did not pursue this avenue here as we expect no major differences in outcomes for the examples considered. Furthermore, it is not clear to us what could be a reasonable hyperprior for the scale parameter λ.

Though the I-prior is new, the idea of using the Fisher information to define an objective prior is not. The Jeffreys prior and Zellner's g-prior (Zellner, 1986)

are also based on the Fisher information, but in a different way. In particular, for a multiple regression model, the g-prior covariance matrix for the vector of regression coefficients is proportional to its *inverse* Fisher information matrix, in contrast to the actual Fisher information matrix for the I-prior. However, we show in Section 3.1 that the *standard g-prior* can be interpreted as an I-prior, if the covariate space is equipped with the Mahalanobis distance.

Jamil (2018) builds on an earlier version of the present paper and another unpublished manuscript. He 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).

An overview of the paper is as follows. In Section 2, we give an expression for the Fisher information on the regression function and define the I-prior, illustrating with basic examples of regression with linear regression functions and one-dimensional smoothing. In Section 3, we compare the methodology with existing methods, including Zellner's g-priors (which can be viewed as a special case of I-priors), cubic spline smoothing and more generally Tikhonov regularization, and Jeffreys priors. In Section 4, the posterior distribution of the regression function under the I-prior is given. Section 5 introduces the RKHSs used in the real data analyses of Section 6. In particular the *canonical RKHS* of linear functions is briefly described and a more detailed description is given of the family of fractional Brownian motion (FBM) RKHSs over a Hilbert space. Smoothness properties of the functions in the FBM RKHS are given, as well as of corresponding I-prior paths. In Section 6, we apply the I-prior methodology to a number of data sets and compare predictive performance with a number of published results for the same data sets, showing the I-prior methodology compares well. In Section 7, a simulation study is done for one-dimensional smoothing, in order to compare with Tikhonov regularization and GPR with a squared exponential prior. The concluding Section 8 briefly summarizes the paper and gives some directions for future work. Appendix A puts the proposed methodology in a broader setting and may be of interest in its own right.

The main new contribution of this paper is Section 2. The results in Section 5 are largely well-known among experts, but may be difficult to find in the literature. Section 6.1 gives some numerical insights that may also be useful for Gaussian process regression. The developments in Appendix A are new except when indicated otherwise.

2. I-priors

A definition of RKHSs is recalled in Section 2.1. The Fisher information on the regression function is derived in Section 2.2. Being positive definite, the Fisher information induces a new RKHS, which is described in Section 2.3. The I-prior is defined in Section 2.4. We give a justification of the data dependence of the I-prior in Section 2.5, and some relatively straightforward applications, to regression with linear functions and one-dimensional smoothing, are given in Section 2.6. Except for the first subsection, all developments here are new unless indicated otherwise.

2.1. Definition of RKHSs and of tensor products of RKHSs

Recall that a Hilbert space is a complete inner product space with a positive definite inner product. Suppose $\mathcal{F}$ is a Hilbert space of functions over a set $\mathcal{X}$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. 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 of functions is called a *reproducing kernel Hilbert space* (RKHS) if it possesses a reproducing kernel. If $\mathcal{X}$ is a set, 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 $n = 1, 2, \ldots$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and $x_1, \ldots, x_n \in \mathcal{X}$. (Note that standard terminology is slightly different for kernels than for matrices and a positive definite kernel is, in fact, the generalization of a positive semi-definite matrix; see Sriperumbudur et al. (2011) for restricted versions of positive definiteness for kernels.) By (a) and (b) above, a reproducing kernel h satisfies $h(x, x') = \langle h(x, \cdot), h(x', \cdot) \rangle_{\mathcal{F}}$, and is hence positive definite. The Moore-Aronszajn theorem states that every symmetric positive definite kernel defines a unique RKHS.

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

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

A Hilbert space $\mathcal{B}$ over $\mathcal{X}$ is called a *feature space* of $\mathcal{F}$ with feature $\phi : \mathcal{X} \to \mathcal{B}$ if $f(x) = \langle \phi(x), f \rangle_{\mathcal{B}}$ for all $x \in \mathcal{X}$. $\mathcal{F}$ is called the *canonical feature space*, and has feature $h(x, \cdot)$.

A function $e_x : \mathcal{F} \to \mathbb{R}$ is called a *point evaluator* at x if $e_x(f) = f(x)$. It can be shown that a Hilbert space of functions is an RKHS if and only if the point evaluators are continuous.

Example 1. *Let $\mathcal{F}$ be the RKHS over $\mathcal{X} = \mathbb{R}^p$ with reproducing kernel $h(x, x') = x^\top x'$. Then $\mathcal{F}$ consists of functions of the form $f(x) = x^\top \beta$ with norm $\|f\|_\mathcal{F} = \|\beta\|_{\mathbb{R}^p}$. $\mathcal{F}$ is also called the dual space of $\mathbb{R}^p$.*

Example 2. *Let $\mathcal{F}$ be the RKHS over $\mathbb{R}$ with reproducing kernel*

$$h(x, x') = -\frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n \left(|x - x'| - |x - x_i| - |x' - x_j| + |x_i - x_j| \right), \qquad (6)$$

for real numbers $x_1, \ldots, x_n$. Then $\mathcal{F}$ is called a centered Brownian motion RKHS (see Section 5.1 for a generalization). It follows from van der Vaart and van Zanten (2008, Section 10) that $\mathcal{F}$ consists of functions $f : \mathbb{R} \to \mathbb{R}$ possessing a square integrable derivative, satisfying $\sum f(x_i) = 0$, and with norm

$$\|f\|_\mathcal{F}^2 = \int \dot{f}(x)^2 dx,$$

where $\dot{f}$ denotes the derivative of f.

2.2. The Fisher information on the regression function

The log-likelihood of parameter f in (1) subject to (2) is given by

$$L(f|y) = C - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \psi_{ij}(y_i - f(x_i))(y_j - f(x_j))$$

for a constant C. The next lemma gives the Fisher information $I[f] = -E\nabla^2 L(f|y)$ for f.

Lemma 1. *Suppose (1) subject to (2) and (3) holds. Then the Fisher information $I[f] \in \mathcal{F} \otimes \mathcal{F}$ for f is given by*

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

More generally, if $\mathcal{F}$ has feature space $\mathcal{B}$ with feature $\phi : \mathcal{X} \to \mathcal{B}$, then if $f(x) = \langle \phi(x), \beta \rangle_\mathcal{B}$ the Fisher information $I[\beta] \in \mathcal{B} \otimes \mathcal{B}$ for β is

$$I[\beta] = \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} \phi(x_i) \otimes \phi(x_j).$$

For any fixed $g \in \mathcal{F}$, the Fisher information on $f_g = \langle f, g \rangle_\mathcal{F}$ is

$$I[f_g] = \sum_{i,j=1}^n \psi_{ij} g(x_i) g(x_j).$$

Remark 1. *If $\mathcal{F}$ is a Hilbert space of functions but not an RKHS, then there is an $x \in \mathcal{X}$ such that the point evaluator $e_x(f) := f(x)$ is discontinuous (Aronszajn, 1950). Thus, if there is an x_i in the sample such that the point evaluator at x_i is discontinuous, the Fisher information on f does not exist because the gradient of the likelihood does not exist.*

Proof of Lemma 1. For $x \in \mathcal{X}$, let $e_x : \mathcal{B} \to \mathbb{R}$ be defined by $e_x(\beta) = \langle \phi(x), \beta \rangle_{\mathcal{B}}$. Clearly, e_x is linear and continuous. Hence, the directional derivative of $e_x(\beta)$ in the direction $\gamma \in \mathcal{B}$ is

$$\nabla_\gamma e_x(\beta) = \lim_{\delta \to 0} \frac{e_x(\beta + \delta\gamma) - e_x(\beta)}{\delta} = e_x(\gamma) = \langle \phi(x), \gamma \rangle_{\mathcal{B}}.$$

Hence by definition of the gradient (see Appendix C)

$$\nabla e_x(\beta) = \phi(x). \tag{7}$$

The log-likelihood of β is given by

$$L(\beta | \mathbf{y}, \Psi) = C - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij}(y_i - e_{x_i}(\beta))(y_j - e_{x_j}(\beta)),$$

for some constant C. Then after standard calculations and using (7),

$$I[\beta] = -E\left[\nabla^2 L(\beta | \mathbf{y}, \Psi)\right] = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} + \nabla e_{x_i}(\beta) \otimes \nabla e_{x_j}(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij}\, \phi(x_i) \otimes \phi(x_j).$$

Taking the canonical feature $\phi(x) = h(x, \cdot)$, the formula for $I[f]$ follows.

For any fixed $g \in \mathcal{F}$, the Fisher information on $\langle f, g \rangle_{\mathcal{F}}$ is

$$I[\langle f, g \rangle_{\mathcal{F}}] = \langle I[f], g \otimes g \rangle_{\mathcal{F} \otimes \mathcal{F}} = \sum_{i,j=1}^{n} \psi_{ij} \langle h(\cdot, x_i) \otimes h(\cdot, x_j), g \otimes g \rangle_{\mathcal{F} \otimes \mathcal{F}}$$

$$= \sum_{i,j=1}^{n} \psi_{ij} \langle h(\cdot, x_i), g \rangle_{\mathcal{F}} \langle h(\cdot, x_j), g \rangle_{\mathcal{F}} = \sum_{i,j=1}^{n} \psi_{ij} g(x_i) g(x_j).$$

$\square$

2.3. The RKHS induced by the Fisher information

The Fisher information, being positive definite, induces a new RKHS over a subspace of $\mathcal{F}$. This RKHS is important because it describes the available information on f in the sense explained below. We describe this RKHS next.

Define

$$\mathcal{F}_n = \left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, f(x) = \sum_{i=1}^{n} h(x, x_i) w_i \text{ for some } w_1, \ldots, w_n \in \mathbb{R} \right\} \tag{8}$$

and let h_n be the kernel over $\mathcal{X}$ defined by $h_n(x, x') = I[f](x, x')$, i.e.,

$$h_n(x, x') = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} h(x, x_i) h(x', x_j). \tag{9}$$

Note that, since $h(\cdot, x_i) \in \mathcal{F}$, $\mathcal{F}_n$ is a subspace of $\mathcal{F}$. The next lemma describes the RKHS induced by the Fisher information.

Lemma 2. *Let $\mathcal{F}_n$ be equipped with the inner product*

$$\langle f_w, f_{w'} \rangle_{\mathcal{F}_n}^2 = w^\top \Psi^{-1} w' \tag{10}$$

where $w = (w_1, \ldots, w_n)^\top$ and $f_w(x) = \sum h(x, x_i) w_i$. Then h_n defined by (9) is a reproducing kernel of $\mathcal{F}_n$.

Proof. Denote by ψ_{ij}^- the (i, j)th element of Ψ^{-1}. By (10) we have $\langle h(\cdot, x_i), h(\cdot, x_j) \rangle_{\mathcal{F}_n} = \psi_{ij}^-$ so

$$\langle f_w, h_n(x, \cdot) \rangle_{\mathcal{F}_n} = \left\langle \sum_{i=1}^{n} h(\cdot, x_i) w_i, \sum_{j=1}^{n} \sum_{k=1}^{n} \psi_{jk} h(x, x_j) h(\cdot, x_k) \right\rangle_{\mathcal{F}_n}$$

$$= \sum_{i=1}^{n} w_i \sum_{j=1}^{n} \sum_{k=1}^{n} h(x, x_j) \psi_{jk} \langle h(\cdot, x_i), h(\cdot, x_k) \rangle_{\mathcal{F}_n} = \sum_{i=1}^{n} w_i \sum_{j=1}^{n} \sum_{k=1}^{n} h(x, x_j) \psi_{jk} \psi_{ik}^-$$

$$= \sum_{i=1}^{n} w_i \sum_{j=1}^{n} h(x, x_j) \delta_{ij} = \sum_{i=1}^{n} w_i h(x, x_i) = f_w(x).$$

Hence, h_n is a reproducing kernel for $\mathcal{F}_n$. $\qquad\square$

We immediately obtain the following interpretation of $\|\cdot\|_{\mathcal{F}_n}^2$. Let $\hat{f}$ be an unbiased estimator of f in model (1) subject to (2) and (3). Then for $g \in \mathcal{F}_n$, $\|g\|_{\mathcal{F}_n}^2$ is the Cramér-Rao lower bound for the variance of $\mathrm{var}(\langle g, \hat{f} \rangle_{\mathcal{F}})$, i.e.,

$$\mathrm{var}(\langle g, \hat{f} \rangle_{\mathcal{F}}) \geq \|g\|_{\mathcal{F}_n}^2. \tag{11}$$

It follows from the theory of weighted least squares that the lower bound is achieved if $\hat{f}$ is a maximum likelihood estimator of f.

The next lemma implies that the data do not contain any Fisher information to distinguish between two functions f and f' if $f(x_i) = f'(x_i)$ for $i = 1, \ldots, n$.

8

Lemma 3. *The orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$ is*

$$\mathcal{F}_n^\perp = \left\{ f \in \mathcal{F} \middle| f(x_1) = \ldots = f(x_n) = 0 \right\}. \tag{12}$$

We can hence uniquely decompose $f \in \mathcal{F}$ as

$$f(x) = f_n(x) + r_n(x) \quad f_n \in \mathcal{F}_n, r_n \in \mathcal{F}_n^\perp. \tag{13}$$

Then $I[f_n] = I[f]$ and $I[r_n] = 0$. Furthermore, the Fisher information on any nonzero linear functional of f_n is strictly positive.

Proof of Lemma 3. Let $f_n = \sum_{i=1}^n h(\cdot, x_i)w_i \in \mathcal{F}_n$ and let $g \in \mathcal{F}$. Then by the reproducing property of h, $\langle f_n, g \rangle_\mathcal{F} = \sum_{i=1}^n w_i \langle h(x_i, \cdot), g \rangle_\mathcal{F} = \sum_{i=1}^n w_i g(x_i)$. But this vanishes for any $w_1, \ldots, w_n$ iff $g(x_1) = \ldots g(x_n) = 0$, proving (12).

Denote the log-likelihood of a parameter by $L(\cdot|y)$. Since $L(f|y) = L(f_n|y)$ and $f_n \perp_\mathcal{F} r_n$, the definition of Fisher information immediately implies $I[f_n] = I[f]$ and $I[r_n] = 0$. For $g \in \mathcal{F}$, $\langle f_n, g \rangle_\mathcal{F} \neq 0$ iff $g \notin \mathcal{F}_n^\perp$. But then, if $g \in \mathcal{F} \setminus \mathcal{F}_n^\perp$, Lemma 1 implies that $I[\langle f_n, g \rangle_\mathcal{F}] = \sum \psi_{ij} g(x_i) g(x_j) > 0$. $\square$

Remark 2. *Another way to obtain the conclusion that the data contain no Fisher information to distinguish between two functions which have the same values at $x_1, \ldots, x_n$ is as follows. The* Fisher information metric *over $\mathcal{F}$ is the distance induced by semi-norm over $\mathcal{F}$ given by*

$$\|f\|_I^2 = \left\langle I[f], f \otimes f \right\rangle_{\mathcal{F} \otimes \mathcal{F}} = \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} f(x_i) f(x_j).$$

The quantity $\|f - f'\|_I$ can be thought of as the 'amount of information between f and f''. We see that $\|f - f'\|_I = 0$ if and only if $f(x_i) = f'(x_i)$ for $i = 1, \ldots, n$.

Example 3. *Suppose $f(x) = \sum_{i=1}^n h(x, x_i)w_i \in \mathcal{F}_n$ and the errors in (1) are autoregressive, in particular, $\varepsilon_1 = \eta_1$, $\varepsilon_{i+1} = \alpha\varepsilon_i + \eta_i$ $(i = 2, \ldots, n)$, with the η_i i.i.d. $N(0, \sigma^2)$ and $-1 \leq \alpha \leq 1$. Then from Lemma 2 and Appendix D,*

$$\|f\|_{\mathcal{F}_n}^2 = w^\top \Psi^{-1} w = \frac{1}{\sigma^2} \sum_{i=1}^n \left(\sum_{j=i}^n \alpha^{j-i} w_j \right)^2,$$

where $0^0 := 1$. We have the special cases

$$\|f\|_{\mathcal{F}_n}^2 = \frac{1}{\sigma^2} \times \begin{cases} \sum_{i=1}^{i-1}(w_{i+1} - w_i)^2 & \alpha = -1 \\ \sum_{i=1}^n w_i^2 & \alpha = 0 \\ \sum_{i=1}^n \left(\sum_{j=i}^n w_j \right)^2 & \alpha = 1 \end{cases}.$$

Example 4. *(Continuation of Example 2.) With $\mathcal{F}$ the centered Brownian motion RKHS over $\mathbb{R}$, it can be seen that $\mathcal{F}_n$ is the set of functions which integrate to zero and are piecewise linear with knots at $x_1, \ldots, x_n$. Hence for $f \in \mathcal{F}_n$,*

$$\|f\|_{\mathcal{F}}^2 = \int \dot{f}(x)^2 dx = \sum_{i=1}^{n-1} \frac{(f(x_{i+1}) - f(x_i))^2}{x_{i+1} - x_i}.$$

Furthermore, if $f \in \mathcal{F}_n$ satisfies $f(x) = \sum h(x, x_i) w_i$ and assuming $x_1 \leq x_2, \ldots \leq x_n$, then it is straightforward to check (by substituting $f(x_k) = \sum h(x_k, x_i) w_i$ into the right hand sides) that

$$w_1 = \frac{f(x_2) - f(x_1)}{x_2 - x_1}, \qquad w_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

and for $i = 2, \ldots, n-1$,

$$w_i = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} - \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}.$$

It follows that $f \in \mathcal{F}_n$ can be represented as

$$f(x) = \int_{-\infty}^{x} \beta(t) dt, \tag{14}$$

where

$$\beta(t) = \sum_{i: x_i \leq t} w_i = \frac{f(x_{i_t+1}) - f(x_{i_t})}{x_{i_t+1} - x_{i_t}}, \tag{15}$$

with $i_t = \max_{x_i \leq t} i$. Note that $\sum w_i = 0$ and hence $\lim_{t \to \pm\infty} \beta(t) = 0$.

By Lemma 2, $\|f\|_{\mathcal{F}_n}^2 = w^\top \Psi^{-1} w$. For i.i.d. errors, the above expressions for the w_i show this is proportional to

$$\left(\frac{f(x_2) - f(x_1)}{x_2 - x_1}\right)^2 + \sum_{i=2}^{n-1} \left(\frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} - \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}\right)^2 + \left(\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}\right)^2. \tag{16}$$

2.4. Definition of I-priors

By Lemma 3, the set $\mathcal{F}$ is too big for the purpose of estimating f, in the sense that, for pairs of functions in $\mathcal{F}$ with the same values at $x_1, \ldots, x_n$, the data do not contain information on whether one is closer to the truth than the other. An objective prior for f therefore need not have support $\mathcal{F}$, instead it is sufficient to consider priors with support $f_0 + \mathcal{F}_n$, where $f_0 \in \mathcal{F}$ is fixed and chosen a priori

as a 'best guess' of f. Lemma 3 implies the data contain information to allow a comparison between any pair of functions in $f_0 + \mathcal{F}_n$.

We follow Jaynes (1957a,b, 2003) and define an objective prior using the maximum entropy principle. The entropy of a prior π over $f_0 + \mathcal{F}_n$ relative to a measure ν is defined as

$$\mathcal{E}(\pi) = -\int_{f_0 + \mathcal{F}_n} \pi(f) \log \pi(f) \nu(\mathrm{d}f).$$

We take ν to be volume measure induced by $\|\cdot - f_0\|_{\mathcal{F}_n}$, which is flat. An I-prior for f is now defined as a prior maximizing entropy subject to a constraint of the form

$$E_\pi \|f - f_0\|^2_{\mathcal{F}_n} = \text{constant.}$$

Variational calculus shows that I-priors for f are the Gaussian variables with mean f_0 and covariance kernel proportional to h_n given by (9), i.e.,

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

for some $\lambda > 0$. Thus, if f has an I-prior distribution, we can use the convenient representation (5).

The posterior distribution of f and the marginal likelihood of (λ, Ψ) are given in Section 4.

2.5. Data dependence of I-prior

It can be seen that the I-prior depends on the data $x_1, \ldots, x_1$. An argument can be made that any objective prior must in fact be data dependent, and representable in the form

$$f(x) = f_0(x) + \sum_{i=1}^{n} h(x, x_i) \alpha_i \text{ for random } \alpha_i. \tag{17}$$

The argument is as follows (details on the assertions are given in Section 2.2). Any $f \in \mathcal{F}$ can be uniquely decomposed as $f(x) = f_n(x) + r_n(x)$, where $f_n(x) = \sum_{i=1}^{n} h(x, x_i) w_i$ for some $w_1, \ldots, w_n$ and $r_n(x_i) = 0$ for $i = 1, \ldots, n$. Since the likelihood for f does not depend on r_n, and f_n and r_n are orthogonal in $\mathcal{F}$, the data contain no (Fisher) information on r_n. Therefore, unless we have actual prior information about the relation between r_n and f_n, it is not possible to do statistical inference on r_n using the data at hand. If the prior for f is representable as in (17), this implies the prior for r_n is a point mass at our prior guess of it, and so is the posterior for r_n (note that our prior guess for r_n is the orthogonal projection of f_0 onto the subspace of $\mathcal{F}$ consisting of functions r for which $r(x_1) = \ldots = r(x_n) = 0$). In summary, only with a prior representable as in (17), all the 'information'

11

$\mathcal{X}$	kernel $h(x, x')$	$f(x)$	I-prior
$\mathbb{R}^p$ (dot product)	$x^\top x'$	$x^\top \beta$	$\beta \sim \text{MVN}(\beta_0, \lambda X^\top X)$
$\mathbb{R}^p$ (Mahal. metric)	$x^\top (X^\top X)^{-1} x'$	$x^\top \beta$	$\beta \sim \text{MVN}(\beta_0, g(X^\top X)^{-1})$
$\mathbb{R}$	centered BM	$\int_{-\infty}^{x} \beta(t) dt$	$\beta \sim$ Brownian bridge
$\mathbb{R}^p$	centered BM	f is Hölder $\geq 1/2$	f is Hölder 1 (a.s.)

Table 1: I-priors for the illustrative examples in Section 2.6 (first three rows), assuming model (1) with i.i.d. normal errors. The model in the last row is discussed in Section 5. The I-prior when $\mathbb{R}^p$ is equipped with the Mahalanobis metric is also known as the g-prior. BM stands for Brownian motion (the same as FBM-1/2). See the text for further details.

available about r_n is our prior guess of it, and it remains nothing more than a mere prior guess even after observing the data. Our maximum entropy argument in Section 2.4 and Appendix A then leads to the I-prior represented in (5).

2.6. Applications

We give the I-priors for linear regression functions and for one-dimensional smoothing, summarized in Table 1

2.6.1. Linear regression functions

Consider the model

$$y_i = x_i^\top \beta + \varepsilon_i, \qquad x_i \in \mathbb{R}^p, \ i = 1, \dots, n$$

subject to (2). With X the $n \times p$ matrix whose ith row is $x_i^\top$, the Fisher information on β is

$$I[\beta] = X^\top \Psi X.$$

Hence, the I-prior for β with prior mean 0 is the multivariate normal distribution with covariance matrix $\lambda X^\top \Psi X$, i.e., under the I-prior

$$\beta \sim \text{MVN}(\beta_0, \lambda X^\top \Psi X) \tag{18}$$

for a scale parameter $\lambda > 0$ and a prior mean $\beta_0 \in \mathbb{R}^p$. This prior is suitable whether p is small or large (see Section 6 where the prior is used for potentially large p).

In the above, we assumed $\mathcal{F}$ is the dual space of $\mathbb{R}^p$, i.e., if $f(x) = x^\top \beta$, $\|f\|_{\mathcal{F}} = \|\beta\|_{\mathbb{R}^p}$. This is a suitable space if, say, a vector x is a series of repeated measurements on the same scale, but is not suitable if x consists of measurements on different scales, such as height in metres and weight in kilograms. In that case, it is better to instead equip $\mathbb{R}^p$ with the Mahalanobis distance, i.e.,

$$\|\beta\|_{\text{Mah}}^2 = \beta^\top (X^\top \Psi X)^{-1} \beta.$$

With this metric, the Fisher information on β is $(X^\top \Psi X)^{-1}$ (rather than $X^\top \Psi X$ in the standard Euclidean metric), and the I-prior becomes

$$\beta \sim \text{MVN}(\beta_0, \lambda\, (X^\top \Psi X)^{-1}). \tag{19}$$

This is the usual g-prior with $g = \lambda$ (Zellner, 1986).

In contrast to (18), the prior (19) is scale invariant and hence suitable for covariates measured on different scales. However, it has the drawback that it is only suitable if $p \ll n$, whereas (18) can be used even if $p > n$. In Appendix A a different derivation of the g-prior is given as well as a generalization.

2.6.2. One-dimensional smoothing with I-priors and connection with cubic spline smoothing

We continue Examples 2 and 4, where we assumed that the regression function lies in the centered Brownian motion RKHS over $\mathbb{R}$.

With i.i.d. errors, under the I-prior the w_i in (15) are i.i.d. zero mean normals, so that β defined there is an ordinary Brownian bridge with respect to the empirical distribution function $P_n(x) = \sum_{i=1}^n I(x_i < x)$. It is straightforward to verify that β then has covariance kernel

$$\text{cov}(\beta(x), \beta(x')) = n^2 \text{cov}(P_n(x), P_n(x')) = n\big[\min(P_n(x), P_n(x')) - P_n(x)P_n(x')\big].$$

From (14), the prior process for f is thus an integrated Brownian bridge. This shows a close relation with cubic spline smoothers, which can be interpreted as the posterior mean when the prior is an integrated Brownian motion (Wahba, 1978, 1990b; Green and Silverman, 1994, Section 3.8.3). Under the I-prior, we have $\text{var}(\beta(x)) = P_n(X < x)(1 - P_n(X < x))$, which shows an automatic boundary correction: close to the boundary there is little Fisher information on the derivative of the regression function, so the prior variance is small. This will then lead to more shrinkage of the posterior derivative of f towards the derivative of the prior mean.

Note that the problem of finding the posterior mean of f under the I-prior can be formulated as a penalized generalized least squares problem with penalty proportional to $\|f\|_{\mathcal{F}_n}^2$ which is proportional to (16).

The natural cubic spline smoother and I-prior estimator under the Brownian motion RKHS are hence similar, but have the following main differences (we assume for simplicity that the prior mean is zero). In the range of the observed x-values, the former is piecewise cubic and the latter is piecewise linear; outside this range, they are linear and constant, respectively. However, the two methods are based on different models: due to the penalty $\int \ddot{f}(x)^2 dx$, the cubic spline smoother assumes two derivatives, whereas the I-prior estimator only assumes one, i.e., at least from a theoretical perspective the I-prior has broader applicability.

In the present setting, the smoother the errors (e.g., the more positively autocorrelated the errors are), the more difficult it is to estimate the regression function. This is because smoother errors are more like a function in the RKHS than rougher errors. The I-prior accommodates for this fact by roughening the prior, so that rough functions in the RKHS can still be estimated reasonably even if the errors are relatively smooth. Let us consider AR(1) or MA(1) errors. If the errors are dependent, β is a generalized Brownian bridge because, whilst being tied to zero outside the range of the x_is, the increments which are summed over are dependent. Note that β is piecewise constant with jumps at the x_i, so the I-prior for f is piecewise linear with knots at the x_i, and the same holds true for the posterior mean. As follows from Lemma 11 in Appendix D, if the errors are an AR(1) process with parameter α and error variance σ^2, the w_i form an MA(1) process with parameter $-\alpha$ and error variance σ^{-2}, and if the errors are an MA(1) process with parameter α and error variance σ^2, the w_i form an AR(1) process with parameter $-\alpha$ and error variance σ^{-2}. It follows that if the errors are a random walk (i.e., and AR(1) process with parameter $\alpha = 1$) then it can be checked that the I-prior is also a random walk and the model is not identified, i.e., the I-prior has (essentially) the same distribution as the errors and the regression curve cannot be separated from the errors. Thus, if the errors form a random walk, and all we know about the regression curve is that it is weakly differentiable, there is no way of determining what part of the variation in the y_is is due to the regression curve or due to the errors. To estimate f, a stronger assumption has to be made, e.g., that it is twice weakly differentiable.

3. Comparison with other methods

We now give a brief overview of some other existing methods for estimating the regression function in (1). In Section 1 we compared the I-prior methodology with Gaussian process regression.

3.1. Zellner's g-priors

Zellner's g-prior (Zellner, 1986) is a scale invariant prior for linear regression functions, and is suitable when covariates are measured on different scales (such as height in metres and weight in kilograms). In Section 2.6.1 we showed that it can be viewed as a special case of the I-prior, when the covariate space is equipped with Mahalanobis distance. Alternatively, in Appendix A.2 a different construction (and a generalization) of the g-prior is given, namely as a maximum entropy prior subject to a constraint involving the Fisher information metric; this construction is essentially the one originally given by Zellner. A drawback of the g-prior is that it cannot be used if the number of covariates p is large compared to the sample size n.

3.2. Cubic spline smoothing

If $\mathcal{F}$ is the aforementioned centered Brownian motion RKHS over $\mathbb{R}$ (see Section 2.6.2 for more details) and the errors are i.i.d., I-prior estimation is similar to cubic spline smoothing. Whereas the cubic spline smoother minimizes (4) with $\|f\|_{\mathcal{F}}^2 = \int \ddot{f}(x)^2 dx$ and $\psi_{ij} = I(i = j)$ (I is the indicator function), the I-prior estimator minimizes a similar expression with $\int \ddot{f}(x)^2 dx$ replaced by (16). Note that if the x_i are equally spaced, then (16) multiplied by n^2 is a discrete approximation of $\int \ddot{f}(x)^2 dx$.

Although the two methods will therefore tend to yield similar smoothers, there is a big theoretical difference between the two methods, in that I-prior estimation only assumes f has one derivative, while cubic spline smoothing assumes two derivatives.

3.3. Tikhonov regularization

The Tikhonov regularizer of a regression function f is the function minimizing (4). It is well-known to have a Bayesian interpretation, namely as the posterior mean of f when the prior is a Gaussian with mean f_0 and covariance kernel λh. Although at first sight Tikhonov regularization seems intuitively reasonable, it has been shown to be inadmissible with respect to squared error loss (Chakraborty and Panaretos, 2019). In particular, it may undersmooth every true regression function in $\mathcal{F}$ in the sense we explain now. In the Bayesian interpretation of regularization, for infinite dimensional $\mathcal{F}$, the prior probability of $\mathcal{F}$ is well known to be zero (e.g. Lifshits, 2012, Section 4.1). Undersmoothing can be said to occur if the prior function paths are, with probability one, rougher than those in $\mathcal{F}$. This happens, for example, if $\mathcal{F}$ is a (centered) Brownian motion RKHS, in which case the prior sample paths have regularity 0.5, while the functions in the RKHS have regularity at least 1 (see Figure 1 for an illustration and Section 5.4 for more details). Although optimal asymptotic convergence rates can often still be obtained with undersmoothing (van der Vaart and van Zanten, 2007), the simulations in Section 7 show that for finite samples undersmoothing can have significant adverse effects on the estimation of functions in $\mathcal{F}$.

3.4. Jeffreys priors

Like the I-prior, the Jeffreys prior is based on the Fisher information, in particular, the Jeffreys prior is proportional to the square root of the determinant of the Fisher information. Hence, it is suitable only for low-dimensional problems. An interesting property of the Jeffreys prior is that it is invariant to parameterization, which the I-prior is not (this is easy to see as the I-prior depends on the reproducing kernel of the RKHS). For model (5) subject to (2) and (3), the Fisher information on the regression function is given by Lemma 1 and can be seen not to depend on f, so the Jeffreys prior is flat, and for the purposes of this paper
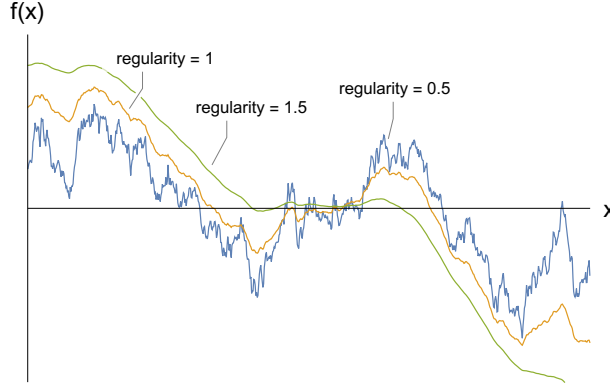
Figure 1: Randomly generated paths of different regularity. The path with regularity 0.5 is a centered Brownian motion path. Functions in the corresponding centered Brownian motion RKHS (also called FBM-1/2 RKHS) have regularity greater than 1, and can be seen to be significantly smoother than the corresponding process paths.

not very useful except potentially in low-dimensional regression (e.g. Ibrahim and Laud, 1991). A more extensive discussion is given in Appendix A.

3.5. Reference priors

Bernardo introduced *reference priors* (Bernardo, 1979, 2005; Berger et al., 2009), which for one-dimensional parameters coincide with Jeffreys priors. Consider a family of probability distributions $P(x|\theta)$, $\theta \in \Theta$. A reference prior π for θ maximizes expected Kullback-Leibler divergence of the prior from the posterior $\pi(\theta|x)$, that is, it maximizes

$$\pi \mapsto E_P\big\{\mathrm{KL}(\pi(\cdot|X)\,|\,\pi)\big\} = E_P \int_\Theta \pi(\theta|x) \log \frac{\pi(\theta|x)}{\pi(\theta)} d\theta.$$

Like the Jeffreys prior, reference priors are parameterization invariant, and unlike the I-prior is only suitable for low-dimensional parameters.

3.6. Fisher kernels

Jaakkola and Haussler (1998) introduced the *Fisher kernel*, defined for a broad range of models, which can be used with kernel methods, for example in support vector machines or as a covariance kernel in Gaussian process regression. Like I-prior, it is a method based on the Fisher information, but that is the only connection. Suppose $P(x|\theta)$ is a probability function depending on a parameter $\theta \in \mathbb{R}^p$. With $s_x(\theta) = \nabla_\theta \log P(x|\theta)$ the score vector for θ and $I[\theta]$ the Fisher information on θ, the Fisher kernel is defined as

$$K(x, x') = \langle s_x, s_{x'} \rangle_{\mathcal{F}_n} = s_x(\theta)^\top I[\theta]^{-1} s_{x'}(\theta).$$

16

4. Posterior distribution of the regression function under the I-prior

This section contains no new results, but is provided for convenience as the notation is different than for standard Gaussian process regression.

Denote $y = (y_1, \ldots, y_n)^\top$, $\mathbf{f} = (f(x_1), \ldots, f(x_n))^\top$, $\mathbf{f}_0 = (f_0(x_1), \ldots, f_0(x_n))^\top$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, $w = (w_1, \ldots, w_n)^\top$ and let H be the $n \times n$ matrix with (i,j)th coordinate $h(x_i, x_j)$. Then (1) implies $y = \mathbf{f} + \varepsilon$. Under the I-prior,

$$\mathbf{f} \sim \mathrm{MVN}(\mathbf{f}_0, \lambda^2 H \Psi H).$$

The marginal distribution of y then is

$$y \sim \mathrm{MVN}(\mathbf{f}_0, V_y), \tag{20}$$

where the marginal covariance is given as

$$V_y = \lambda^2 H \Psi H + \Psi^{-1}.$$

Thus, the marginal log likelihood of (λ, Ψ) is

$$L(\lambda, \Psi | y) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |V_y| - \frac{1}{2}(y - \mathbf{f}_0)^\top V_y^{-1}(y - \mathbf{f}_0). \tag{21}$$

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

$$u_{ij} = \frac{1}{2} \mathrm{tr}\left(V_y^{-1} \frac{\partial V_y}{\partial \theta_i} V_y^{-1} \frac{\partial V_y}{\partial \theta_j}\right),$$

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

The next lemma gives the posterior distribution of f under the I-prior.

Lemma 4. *The posterior distribution of f in (1) subject to (2) given $(y_1, \ldots, y_n)$ under the I-prior π is Gaussian with mean given by*

$$E_\pi\big[f(x)|y_1, \ldots, y_n\big] = f_0(x) + \lambda \sum_{i=1}^n h(x, x_i)\hat{w}_i$$

where $\hat{w} = \lambda \Psi H^\top V_y^{-1}(y - \mathbf{f}_0)$, and covariance kernel given by

$$\mathrm{cov}_\pi\big(f(x), f(x')|y_1, \ldots, y_n\big) = \lambda^2 \sum_{i=1}^n \sum_{j=1}^n h(x, x_i)h(x', x_j)(V_y^{-1})_{ij}$$

.

Proof of Lemma 4. Under the I-prior π, (5) holds and the joint distribution of (w, y) is given by

$$\begin{pmatrix} w \\ y \end{pmatrix} \sim \text{MVN} \left[\begin{pmatrix} 0 \\ \mathbf{f}_0 \end{pmatrix}, \begin{pmatrix} \Psi & \lambda \Psi H^\top \\ \lambda H \Psi & V_y \end{pmatrix} \right].$$

From this standard results give the posterior distribution of w given y, i.e, the conditional distribution of w given y, which is multivariate normal with mean $\hat{w}$ and covariance matrix

$$\tilde{V}_w = \Psi - \lambda^2 \Psi H^\top V_y^{-1} H \Psi = V_y^{-1}, \tag{22}$$

where the last equality is the Woodbury matrix identity. The posterior mean of f is now obtained by substituting each w_i in (5) by $\tilde{w}_i$, and the posterior covariance matrix is as in the lemma. $\qquad\square$

It follows from the lemma that given the I-prior, the posterior of f can be represented by the left part of (5) where $(w_1, \ldots, w_n)^\top$ is multivariate normal with mean $\tilde{w}$ and covariance matrix V_y^{-1}. The computational complexity of computing the posterior distribution is $O(n^3)$, the same as in Gaussian process regression. This can be reduced in very specific cases, such as for parametric (i.e., finite dimensional) models or for one dimensional smoothing via the Reinsch algorithm (Green and Silverman, 1994, Section 2.3.3). A number of approximation methods to overcome this computational problem is listed in Chapter 8 of Rasmussen and Williams (2006).

5. I-priors and FBM RKHSs

The main aim of this section is to describe smoothness properties of functions in the FBM RKHS and those of I-prior paths when the regression function is assumed to be in the FBM RKHS. The results on smoothness of Gaussian process paths and functions in the associated RKHSs given in this section are well-known, but explicit references may be difficult to find. However, very general related results can be found in Steinwart (2018).

In Section 5.1, the FBM RKHS with Hurst coefficient γ is defined, and in Section 5.2 centering of an RKHS is defined. Section 5.3 concerns Hölder smoothness, and the main results are that functions in the FBM RKHS with Hurst coefficient γ are Hölder of order γ, while I-prior paths are Hölder of order 2γ. This can be compared with FMB-γ process paths, which are Hölder of any order less than γ. Section 5.4 concerns a different concept of smoothness, called regularity, which is based on the rate of decay of Karhunen-Loeve coefficients. This concept is perhaps most useful for one-dimensional functions. Differently from Hölder smoothness, regularity shows a *gap* in smoothness between FBM process paths and functions

Type of function	Hölder degree	Regularity (one dimensional case)
FBM-γ process paths	Any $< \gamma$	γ
FBM-γ RKHS functions	$\geq \gamma$	$> \gamma + 1/2$
FBM-γ I-prior paths	2γ	$2\gamma + 1$ (asymptotically if errors i.i.d.)

Table 2: Smoothness of functions related to the FBM-γ kernel. It is seen that all functions in an FBM RKHS are smoother than the corresponding FBM process paths, while the RKHS contains both rougher and smoother functions than I-prior paths. Note that, with probability 1, the FBM RKHS does not contain an FBM path but does contain an I-prior path.

in the FBM RKHS (see also Figure 1). The results of Sections 5.3 and 5.4 are summarized in Table 2. In Section 2.6.2, we look in some detail at smoothing with a one-dimensional Brownian motion RKHS. In this case, the I-prior methodology with i.i.d. errors gives similar results as cubic spline smoothing, but with a different theoretical justification.

5.1. Canonical and Fractional Brownian motion RKHS

In this paper we consider two (families of) RKHSs of functions over a Hilbert space $\mathcal{X}$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{X}}$.

Firstly, the *canonical RKHS* is the dual space of $\mathcal{X}$ and is defined by the canonical kernel

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

Being the dual space, it consists of all linear functions over $\mathcal{X}$. The Riesz representation theorem implies that for any linear function f over $\mathcal{X}$ there exists a $\beta \in \mathcal{X}$ such that $f(x) = \langle x, \beta \rangle_{\mathcal{X}}$. In that case, $\|f\|_{\mathcal{F}} = \|\beta\|_{\mathcal{X}}$.

Secondly, we consider the Fractional Brownian Motion (FBM) RKHS. Schoenberg (1937) has shown that, for $0 < \gamma < 1$, there exists a Hilbert space $\mathcal{B}$ and a function $\phi_\gamma : \mathcal{X} \to \mathcal{B}$ such that

$$\|\phi_\gamma(x) - \phi_\gamma(x')\|_{\mathcal{B}} = \|x - x'\|_{\mathcal{X}}^{\gamma} \qquad \forall x, x' \in \mathcal{X}. \tag{23}$$

Using the polarization identity, we obtain

$$h_\gamma(x, x') = \langle \phi_\gamma(x), \phi_\gamma(x') \rangle_{\mathcal{B}} = -\frac{1}{2}\Big(\|x - x'\|_{\mathcal{X}}^{2\gamma} - \|x\|_{\mathcal{X}}^{2\gamma} - \|x'\|_{\mathcal{X}}^{2\gamma} \Big). \tag{24}$$

From its construction, it is clear that h_γ is positive definite. It is in fact the covariance kernel of fractional Brownian motion (FBM) on $\mathcal{X}$ with Hurst coefficient γ (Kolmogorov, 1940; Mandelbrot and Ness, 1968). Note that if $\gamma = 1$ then $h_\gamma(x, x')$ is the canonical kernel $\langle x, x' \rangle_{\mathcal{X}}$. Following Cohen (2002), we call the RKHS with kernel h_γ the FBM RKHS of order γ, and we denote it $\mathcal{F}_\gamma$. An alternative name is the Cameron-Martin space of FBM (see, e.g. Picard, 2011).

5.2. Centering of an RKHS

The functions in an RKHS may be arbitrarily positioned, for example, if f is in a canonical or in an FBM RKHS, then $f(0) = 0$, which is undesirable for the purposes of this paper. To remedy this, an RKHS may be centered. The functions in a centered RKHS have zero mean.

If P is a probability distribution over $\mathcal{X}$ and $X, X' \sim P$ are independent, a kernel h over $\mathcal{X}$ may be centered as follows:

$$h_{\text{cent}}(x, x') = E_P(h(x, x') - h(x, X) - h(x', X') + h(X, X')).$$

The RKHS with kernel h_{cent} is then centered in the sense that $E_P(f(X)) = 0$ for all functions f in the RKHS.

In the present paper we center with respect to the empirical distribution of data $x_1, \ldots, x_n$, so that in the centered RKHS, $\sum_{i=1}^n f(x_i) = 0$. The centered canonical kernel then becomes $h_{\text{cent}}(x, x') = \langle x - \bar{x}, x - \bar{x} \rangle_{\mathcal{X}}$ where $\bar{x} = n^{-1} \sum_{i=1}^n x_i$. The centered FBM kernel becomes

$$h_{\gamma, P}(x, x') = -\frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n \left(\|x - x'\|_{\mathcal{X}}^{2\gamma} - \|x - x_i\|_{\mathcal{X}}^{2\gamma} - \|x' - x_j\|_{\mathcal{X}}^{2\gamma} + \|x_i - x_j\|_{\mathcal{X}}^{2\gamma} \right). \quad (25)$$

5.3. Hölder smoothness

A function f over a set $\mathcal{X}$ with norm $\|\cdot\|_{\mathcal{X}}$ is Hölder of order $0 < \gamma \leq 1$ if there exists a $C > 0$ such that

$$|f(x) - f(x')| < C\|x - x'\|_{\mathcal{X}}^{\gamma} \quad \forall x, x' \in \mathcal{X} \tag{26}$$

and f is Hölder of order $1 < \gamma \leq 2$ if

$$|f(x + t) - 2f(x) + f(x - t)| \leq K\|t\|_{\mathcal{X}}^{\gamma} \quad \forall x, t \in \mathcal{X}, \tag{27}$$

for some $K > 0$ (see Gilbarg and Trudinger, 1998, Chapter 4, or Stein, 1970, Section 4.3).

It is well-known that realizations of an FBM-γ process are a.s. Hölder continuous of any order less than γ (e.g., Theorem 4.1.1 in Embrechts and Maejima, 2002). The next lemma shows that functions in the FBM RKHS are strictly smoother than FBM realizations.

Lemma 5. *The functions in the FBM-γ RKHS are Hölder of order γ.*

Proof of Lemma 5. Let $f \in \mathcal{F}_{\gamma}$. By the reproducing property of h_{γ}, $f(x) = \langle h_{\gamma}(x, \cdot), f \rangle_{\mathcal{F}_{\gamma}}$ for $f \in \mathcal{F}_{\gamma}$. From this, the Cauchy-Schwarz inequality and (23) with $\phi_{\gamma}(x) = h_{\gamma}(x, \cdot)$,

$$\begin{aligned}
|f(x) - f(x')| &= |\langle h_{\gamma}(x, \cdot) - h_{\gamma}(x', \cdot), f \rangle_{\mathcal{F}_{\gamma}}| \\
&\leq \|h_{\gamma}(x, \cdot) - h_{\gamma}(x', \cdot)\|_{\mathcal{F}_{\gamma}} \|f\|_{\mathcal{F}_{\gamma}} = \|x - x'\|_{\mathcal{X}}^{\gamma} \|f\|_{\mathcal{F}_{\gamma}},
\end{aligned} \tag{28}$$

proving the lemma. $\qquad \square$

The next example illustrates Lemma 5 is sharp.

Example 5. If $\mathcal{X} = [0,1]$, the FBM-1/2 RKHS consists of all functions f that are absolutely continuous possessing a weak derivative $\dot{f}$ and satisfying $f(0) = 0$. The norm is given by

$$\|f\|_{\mathcal{F}}^2 = \|\dot{f}\|_{L^2([0,1])}^2 = \int_{[0,1]} \dot{f}(x)^2 dx \quad \forall f \in \mathcal{F}.$$

See van der Vaart and van Zanten (2008), Section 10, for a proof. It is now straightforward to verify that the function f over $[0,1]$ defined by $f(x) = \sqrt{1-x^2}$ is in $\mathcal{F}_{1/2}$. This function is Hölder of order $1/2$ but not more.

By (5), I-prior realizations are finite combinations of basis functions $h_\gamma(x_i, \cdot)$. Hence Lemma 10 in Appendix B directly implies the following:

Lemma 6. *If $\mathcal{F}$ in (1) subject to (2) is the FBM-γ RKHS, the I-prior has realizations that are Hölder of order 2γ.*

5.4. Regularity

As seen above, in terms of Hölder smoothness, FBM paths and functions in the corresponding FBM RKHS may differ by an infinitesimally small amount. If we look at a different concept of smoothness, it turns out there is a gap between the two.

Let $\mathcal{F}$ be a Hilbert space of functions over a set $\mathcal{X}$ with orthonormal basis $\{g_i\}$. For $\beta > 0$ and $f \in \mathcal{F}$, consider the squared norm

$$\|f\|_\beta^2 = \frac{1}{2} \sum_{i=1}^{\infty} f_i^2 i^{2\beta}, \tag{29}$$

where $f_i = \langle f, g_i \rangle_{\mathcal{F}}$. A function f for which $\|f\|_\beta < \infty$ is said to have *regularity* β relative to the basis $\{g_i\}$. We have:

Lemma 7. *The FBM-γ RKHS over $[0,1]$ is regular of order $1/2 + \gamma$ relative to the Karhunen-Loeve basis for the FBM-γ process.*

Proof of Lemma 7. It follows from results in Bronski (2003) that the FBM-γ kernel over $[0,1]$ has eigenvalues $\lambda_i \sim i^{-1-2\gamma}$. Now $f \in \mathcal{F}_\gamma$ if and only if $\sum f_i^2/\lambda_i < \infty$ (e.g., Lemma 1.1.1 in Wahba (1990b)). But then $f_i = o(i^{-1-\gamma})$ as $i \to \infty$ such that the sum converges. It follows that $\|f\|_\beta < \infty$ if and only if $\beta \le 1/2 + \gamma$. $\square$

I-prior paths are finite dimensional and hence have infinite regularity. More interestingly, we can may consider asymptotic regularity. A series of functions $a_1, a_2, \ldots$ can be defined to have asymptotic regularity β if $\lim_{n \to \infty} \|a_n(x)\|_\beta < \infty$. From Bronski (2003), with $\mathcal{X} = [0,1]$, h_γ has a Mercer expansion $h_\gamma(x, x') =$
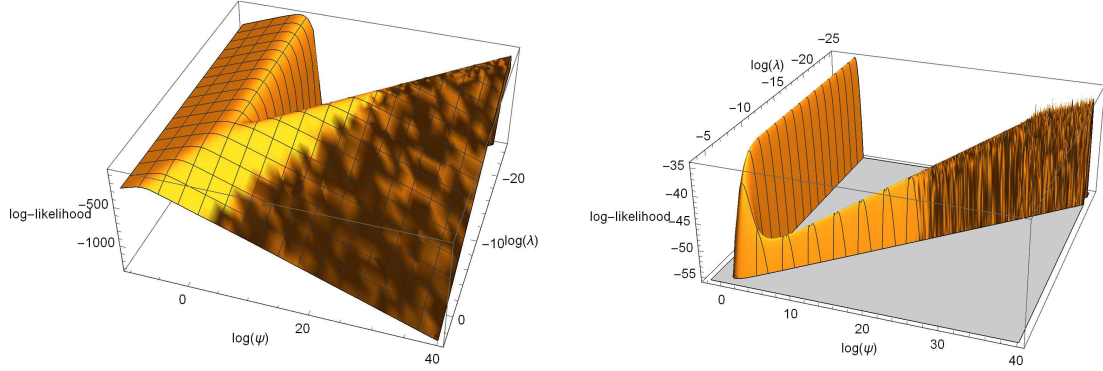
Figure 2: Marginal log likelihood for Eye State data (global view on left and zoom of very top of graph on right). Darker areas reflect a nonsmooth surface due to numerical errors in evaluating the log-likelihood. There are local maxima on each of the two ridges, with the maximum on the diagonal ridge leading to the best predictive performance. It can be seen that the local maximum on the diagonal ridge is numerically hard to find, in fact, it is hard to establish what the global maximum is. Fortunately, predictive performance is near-identical anywhere on the diagonal ridge with say $\log(\psi) > 6$, so for the purpose of prediction there is no need to find the actual global maximum.

$\sum_{i=1}^{\infty} \lambda_i g_i(x) g_i(x')$ where $\lambda_i \sim i^{-1-2\gamma}$ and the g_i form the Karhunen-Loeve basis for FBM-γ. Hence I-prior paths can be written as

$$f_n(x) = \sum_{j=1}^{n} h(x, x_j) w_j = \sum_{i=1}^{\infty} \lambda_i u_i g_i(x)$$

where $u_i = \sum_{j=1}^{n} g_i(x_j) w_j$. If the errors in (1) are i.i.d. normal, the w_i are i.i.d. normal under the I-prior. Under some conditions on the x_i, and assuming the g_i have a common bound (unfortunately we have no proof of this but experimental results support the assertion), we can verify that $\lim_{n \to \infty} \| f_n(x)/n \|_\beta < \infty$ a.s. for $\beta \leq 1 + 2\gamma$, i.e., I-prior paths multiplied by $1/n$ are a.s. asymptotically regular of order $1 + 2\gamma$.

6. Application to real data

In this section we apply the I-prior methodology to nine real data sets which have been extensively analyzed in the literature, and compare performance to published methods as a well as to Tikhonov regularization. We analyze one functional regression data set and eight classification problems, and obtain competitive performance of the I-prior methodology.

6.1. Model assumptions and estimation of hyperparameters

For the real data examples below, we assumed model (1) subject to i.i.d. $N(0, \psi^{-1})$ errors, and for the set of regression functions $\mathcal{F}$ we used the canonical

RKHS of linear functions and the FBM-1/2 RKHS. We also made some limited use of the FBM-γ RKHS where the Hurst coefficient γ was estimated, and the squared exponential RKHS (30) with $\xi = 1$ where σ was estimated. For Tikhonov regularization, we only used the canonical and FBM-1/2 RKHSs, which leave just the scale parameter parameter λ in (4) to be estimated, which we did using generalized cross-validation.

For the I-prior methodology, λ, ψ, and possibly γ or σ were to be estimated. We first discuss estimation of λ and ψ, which we did by maximizing the marginal likelihood (21) (we also tried minimizing various cross-validation criteria, but this gave worse performance). Maximum likelihood estimation was not straightforward for two reasons: the possible occurrence of multiple local maxima, and numerical difficulties in evaluating the likelihood. A typical situation is as pictured in Figure 2: the likelihood has two ridges, one parallel to the $\log(\lambda)$ axis, and one running diagonally across the graph. We found empirically that each ridge may have a local maximum, and there may be a local maximum on or near the cusp as well. Usually if there was a local maximum on the diagonal ridge, we were not able to find it because it was too difficult to numerically evaluate the likelihood. In some cases, particularly for the canonical kernel (e.g., for the Hill-Valley data below), we could not get a decent estimate of any part of the diagonal ridge. Fortunately, in most cases, it was only necessary to be able to estimate the part of the diagonal ridge near the cusp, as predictive performance did not noticeably change moving up the ridge. Some more details are given in the caption of Figure 2. We selected the local maximum (or near-local-maximum on the diagonal ridge) that gave the smallest cross-validation error.

As a result, in most cases we could not determine the value of the maximum marginal likelihood, and in these cases it was impossible to find the maximum likelihood estimator of the Hurst coefficient γ. Instead, for different values of γ we estimated λ and ψ using maximum likelihood, and selected the value of γ which minimized cross-validation error. As this was quite time consuming, particularly due to the difficulty of finding the maximum likelihood estimators of λ and ψ for different values of γ, we omitted estimation of γ from the simulations below.

Notwithstanding some philosophical advantages of a fully Bayes approach compared to an empirical Bayes one, it does not appear to be the case that a fully Bayes approach would alleviate the aforementioned problems. In a fully Bayes setting, we may expect the posterior to be multimodal, so the posterior mean would be an inadequate summary measure, and the posterior mode may be more appropriate. However, finding it may be computationally difficult using Monte Carlo methods. Furthermore, unlike the marginal likelihood, the posterior of f cannot easily be visualized, and it may be difficult to find the true mode among multiple local modes. The numerical difficulties we encountered in evaluating the likelihood and the posterior (with hyperparameters substituted by their max-

imum likelihood estimators) will likely be replicated in the posterior in a fully Bayes approach.

6.2. Motivation for use of FBM RKHS

As explained in more detail in Section 5, 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* are (weakly) differentiable if the Hurst coefficient is at least $1/2$ and have minimum Hölder smoothness ranging from 0 to 2. This wide range of smoothnesses make it an attractive general purpose function space for non-parametric 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

$$r(x, x') = \exp\Big(- \frac{\|x - x'\|^{2\xi}}{2\sigma^2} \Big), \tag{30}$$

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.

Besides smoothing with FBM kernels, the only other smoothing method that we are aware of that requires only a single hyperparameter to be estimated is thin plate spline smoothing. However, for larger dimensions, thin plate splines seem to be harder to interpret and implement.

6.3. 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 3). 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
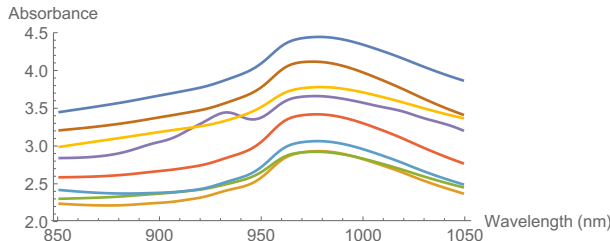
Figure 3: Sample of spectrometric curves used to predict fat content of meat

remaining 43 observations are used as a test sample (following Thodberg's original recommendation).

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 et al. (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 3 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. To be able to estimate a linear or smooth effect using the canonical or FBM RKHSs, an appropriate inner product for the x_i needs to be found. From Figure 3 it appears the curves are differentiable, and it seems reasonable to assume the x_i lie in a Sobolev-Hilbert space $\mathcal{X}$ with inner product

$$\langle x, x' \rangle^2_{\mathcal{X}} = \int \dot{x}(t)\dot{x}'(t)dt.$$

A linear effect of the spectral curve on fat content can be modelled using the canonical RKHS over $\mathcal{X}$. We see in Table 3 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. 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

Method	RMSE	
	Training	Test
Global constant model	12.50	13.3
Neural network (Vila et al., 2000)		0.34
Kernel smoothing (Ferraty and Vieu, 2006, Section 7.2)		1.85
Double index model (Chen et al., 2011)		1.58
Single index model (Goia and Vieu, 2014)		1.18
Sliced inverse regression (Lian and Li, 2014)		0.90
MARS (Zhu et al., 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 3: RMSEs for predicting fat content from spectrometric functional covariate (see Figure 3): previously published results, Tikhonov regularization, and I-prior methodology.

exponential RKHS we did manage to find the ML estimator $\hat{\sigma}$ of σ, and it is given in Table 3.

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.

6.4. Classification

We now apply the I-prior methodology to classification problems, assuming model (1) with $y_i \in \{0, 1\}$ denoting the class label of observation i, and $x_i \in \mathbb{R}^p$ a p-dimensional covariate. A newly observed unit $n + 1$ with covariate value x_{n+1} is classified into class 0 if $\hat{f}(x_{n+1}) < 0.5$ and into class 1 if $\hat{f}(x_{n+1}) > 0.5$.

An extensive analysis of eight data sets with sixteen different methods has recently been done by Cannings and Samworth (2017). The methods are the following: linear and quadratic discriminant analysis (LDA and QDA), k nearest neighbours ($k$nn), Cannings and Samworth's random projection version of these methods (RP-LDA, RP-QDA and RP-$k$nn), a single projection version of LDA and $k$nn, random forests (RF, Breiman, citeyearbreiman01), support vector machines (SVMs) with linear and radial kernels, Gaussian process regression with a radial kernel, penalized LDA (Witten and Tibshirani, 2011), nearest shrunken

centroids (Tibshirani et al., 2003), L_1 penalized logistic regression (Goeman et al., 2015), optimal tree ensembles (Khan et al., 2016), and an ensemble of a subset of $k$nn classifiers (Gul et al., 2016).

For each data set, random subsamples of sizes between 50 and 1000 were taken, and for each subsample the model was fitted and model based predicted class labels of the remaining data were computed. The number of random random subsamples ranged between 40 and 1000, and the average misclassification percentage for the predictions was computed as well as corresponding standard errors.

In Table 4, for each data set the best results provided in Cannings and Samworth (2017) are reproduced, along with results for the I-prior methodology and Tikhonov regularization based on the canonical and FBM-1/2 RKHSs. For the Hill-Valley and Mice data, we also included results for the FBM-0.9 RKHS, which dramatically reduced misclassification rates. For most subsamples, $\hat{\gamma} = 0.9$ approximately minimized cross-validation error (it is coincidental that the number is the same for both data sets). For all datasets, initial analyses indicated further improvements of results could be obtained by estimating γ rather than using $\gamma = 1/2$ or $\gamma = 0.9$, but this was too time consuming to carry out. For Tikhonov regularization the scale parameter was estimated using generalized cross-validation, and for the I-prior methodology hyperparameters were estimated using a modified maximum likelihood approach (see Section 6.1). For five out of eight data sets (Eye state, Mice, Hill-Valley, Musk, and Activity recognition), the I-prior methodology gives better results than the best method reported by Cannings and Samworth, and for the Gisette data there is a tie for first place with Linear SVMs. For six out of the eight data sets, the I-prior methodology improved on the random projection ensemble results of Cannings and Samworth (2017), while only for the Ionosphere data did an ensemble method perform better; however, it is possible that random projection ensembles can further improve the I-prior methodology. Furthermore, in most instances the I-prior methodology gives better results then Tikhonov regularization, often by a large margin. Part of the reason for this may be that for Tikhonov regularization, only a single parameter is estimated (the scale parameter λ), while for the I-prior methodology two are estimated (λ and ψ), giving more flexibility to adapt to the data.

We did not show results for GP regression with the FBM-1/2 and canonical kernels, which gave results comparable to the I-prior methodology, sometimes better, sometimes worse.

7. Simulation study

The I-prior methodology is generally applicable, and in this section we attempt to gain some insight into its performance by considering the special case of smoothing over $[0, 1]$. We compare the I-prior with Tikhonov regularization and

Method	Eye state data			Ionosphere data		
	$n = 50$	$n = 200$	$n = 1000$	$n = 50$	$n = 100$	$n = 200$
Best previous result	$39.0_{0.4}$	$26.9_{0.3}$	$13.5_{0.2}$	$8.1_{0.4}$	$6.2_{0.2}$	$5.2_{0.2}$
— using method	RP-QDA	RP-$k$nn	RP-$k$nn	RP-QDA	RP-QDA	RP-QDA
Tikh. reg. (linear)	$46.1_{0.2}$	$42.7_{0.2}$	$37.8_{0.4}$	$19.0_{0.2}$	$15.2_{0.2}$	$13.7_{0.1}$
Tikh. reg. (FBM-1/2)	$46.3_{0.2}$	$42.2_{0.3}$	$25.8_{0.5}$	$19.4_{0.2}$	$11.4_{0.1}$	$8.4_{0.1}$
I-prior (linear)	$46.2_{0.4}$	$40.0_{0.2}$	$38.1_{0.2}$	$17.3_{0.2}$	$14.6_{0.1}$	$13.6_{0.2}$
I-prior (FBM-1/2)	$37.0_{0.1}$	$24.0_{0.1}$	$10.3_{0.1}$	$11.3_{0.1}$	$7.7_{0.1}$	$6.1_{0.1}$

Method	Mice data			Hill-valley data		
	$n = 200$	$n = 500$	$n = 1000$	$n = 100$	$n = 200$	$n = 500$
Best previous	$6.4_{0.1}$	$2.2_{0.1}$	$0.6_{0.1}$	$36.8_{0.8}$	$36.5_{0.9}$	$32.6_{1.1}$
— using method	LDA	RP-$k$nn	RP-$k$nn	RP-LDA	RP-LDA	RP-LDA
Tikh. reg. (linear)	$9.5_{0.1}$	$4.3_{0.1}$	$3.6_{0.2}$	$50.2_{0.0}$	$49.8_{0.1}$	$35.5_{0.2}$
Tikh. reg. (FBM-1/2)	$25.2_{0.2}$	$12.2_{0.2}$	$5.4_{0.3}$	$50.3_{0.0}$	$50.5_{0.0}$	$50.8_{0.1}$
I-prior (linear)	$6.3_{0.1}$	$4.3_{0.1}$	$3.8_{0.2}$	–	–	–
I-prior (FBM-1/2)	$6.8_{0.0}$	$1.1_{0.0}$	$0.1_{0.0}$	$47.4_{0.1}$	$43.4_{0.2}$	$32.9_{0.2}$
I-prior (FBM-0.9)	$3.7_{0.1}$	$0.6_{0.0}$	$0.1_{0.0}$	$32.5_{0.1}$	$25.4_{0.1}$	$18.9_{0.2}$

Method	Musk data			Arrhythmia data		
	$n = 100$	$n = 200$	$n = 500$	$n = 50$	$n = 100$	$n = 200$
Best previous	$11.8_{0.3}$	$9.7_{0.2}$	$7.4_{0.1}$	$30.5_{0.3}$	$26.7_{0.3}$	$22.4_{0.3}$
— using method	RP-$k$nn	RP-$k$nn	Linear-SVM	RP-QDA	RF	RF
Tikh. reg. (linear)	$13.6_{0.1}$	$10.6_{0.1}$	$7.7_{0.0}$	$44.8_{0.2}$	$36.9_{0.3}$	$29.1_{0.1}$
Tikh. reg. (FBM-1/2)	$15.1_{0.1}$	$11.9_{0.1}$	$7.8_{0.1}$	$46.3_{0.2}$	$39.6_{0.4}$	$29.6_{0.1}$
I-prior (linear)	$15.1_{0.1}$	$11.5_{0.2}$	$9.1_{0.1}$	$39.1_{0.2}$	$33.1_{0.3}$	$27.5_{0.2}$
I-prior (FBM-1/2)	$9.5_{0.1}$	$7.0_{0.1}$	$5.0_{0.1}$	$31.8_{0.1}$	$28.1_{0.1}$	$25.5_{0.1}$

Method	Activity recognition data			Gisette data		
	$n = 50$	$n = 200$	$n = 1000$	$n = 50$	$n = 200$	$n = 1000$
Best previous	$0.11_{0.02}$	$0.04_{0.01}$	$0.00_{0.00}$	$11.9_{0.3}$	$6.8_{0.1}$	$4.5_{0.1}$
— using method	Pen-LDA	Pen-LDA	Pen-LDA	Linear-SVM	Linear-SVM	Linear-SVM
Tikh. reg. (linear)	$0.28_{0.00}$	$0.23_{0.01}$	$0.10_{0.01}$	$41.2_{0.5}$	$11.0_{0.1}$	$6.9_{0.1}$
Tikh. reg. (FBM-1/2)	$0.28_{0.01}$	$0.25_{0.01}$	$0.19_{0.01}$	$46.0_{0.4}$	$15.4_{0.4}$	$7.6_{0.1}$
I-prior (linear)	$0.04_{0.00}$	$0.00_{0.00}$	0_0	$12.3_{0.1}$	$6.9_{0.1}$	$4.5_{0.1}$
I-prior (FBM-1/2)	$0.16_{0.00}$	$0.03_{0.00}$	$0.00_{0.00}$	$14.1_{0.1}$	$7.1_{0.1}$	$4.2_{0.1}$

Table 4: Average percentage test-set misclassification for eight data sets with standard errors in the subscript. The best previous results are taken from Cannings and Samworth (2017), and 'RP' refers to their random projection ensemble method. It can be seen that the Tikhonov regularization method performs poorly. The best performer is colored red, which for five out of eight datasets is the I-prior, with a tie for the Gisette data set. Dashes indicate the model could not be fitted due to numerical problems.

with GPR based on the squared exponential process prior. The main result is that the I-prior estimator has better small sample performance than the Tikhonov regularizer, even in cases most favourable to the latter. Furthermore, compared to the other two methods, the squared exponential prior gives very poor performance for the roughest functions in $\mathcal{F}$.

The assumed model is given by (1) where $\mathcal{F}$ is the centered Brownian motion RKHS over $[0, 1]$ given by (6), with norm $\|f\|_{\mathcal{F}} = \left(\int_0^1 \dot{f}(x)^2 dx \right)^{1/2}$, and the errors are i.i.d. $N(0, 1)$. We consider the following three estimators of f:

- The posterior mean under the I-prior

- The Tikhonov regularizer, i.e., the minimizer of

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 \dot{f}(x)^2 dx.$$

- The posterior mean under the squared exponential Gaussian process prior (subsequently referred to as SE estimator), with covariance kernel given by (30) with $\xi = 1$.

In this case, the Tikhonov regularizer is the posterior mean of the regression function under a centered Brownian motion prior. In all cases, we estimate the smoothing parameter by maximum marginal likelihood or the implied marginal likelihood for Tikhonov regularization.

We included the SE estimator as it is commonly used, and, if the scaling parameter λ is suitably chosen, it has optimal asymptotic convergence rate for all functions in $\mathcal{F}$ (van der Vaart and van Zanten, 2007). As mentioned above, in the present case, the regularizer of f is its posterior mean under a Brownian motion prior. Brownian motion paths have regularity $1/2$, while functions in $\mathcal{F}$ have regularity greater than 1. Hence the sample paths of Brownian motion are 'too rough', and the posterior mean (i.e., the Tikhonov regularizer) is expected to undersmooth. As shown in Section 5, the I-prior for f is an integrated Brownian bridge which has regularity 1.5, so should perform well for functions of intermediate smoothness, but not necessarily for very rough or very smooth functions in $\mathcal{F}$. Similarly, the SE estimator would not necessarily be expected to perform well for non-analytic functions.

The functions in $\mathcal{F}$ have a wide range of smoothness, ranging from functions which merely have one derivative to analytic functions. Hence, no estimator can be expected to perform well for all functions in $\mathcal{F}$, but a desirable estimator would perform reasonably across a wide range of smoothnesses. Normally, we would probably desire good performance for the rougher functions in $\mathcal{F}$.

To assess performance, we simulated regression functions with regularities 1, 1.5, and ∞ (see Figure 4). Note that for the simulations it is only necessary to evaluate the simulated functions at $x_1, \ldots, x_n$. With h the covariance kernel of centered Brownian motion given by (6), H the matrix with (i, j)th element $h(x_i, x_j)$ and $w = (w_1, \ldots, w_n)$ a vector of i.i.d. normals, we simulated the following:

(a) 'Rough' functions, generated as $\mathbf{f} = H^{3/4}w$. (In the limit as $n \to \infty$, these can be shown to have regularity 1 and hence are slightly rougher than the roughest functions in $\mathcal{F}$.) Due to their roughness, these the functions should most favour Tikhonov regularization.

(b) Functions of regularity 1.5, generated as $\mathbf{f} = Hw$. These are sample paths of the I-prior and this scenario should hence favour the I-prior.

(c) Analytic functions (regularity=∞) generated as sample paths of the squared exponential process with $\sigma = 0.02$. Clearly, this scenario is expected to favour the squared exponential prior.

We standardized simulated sample paths so that their RKHS norm equals 1, see Figure 4 for examples of sample paths. The centering of the paths means no intercept needs to be estimated, simplifying the simulations.

We measured quality of estimation by the L_2 median absolute error (MAE):

$$\mathrm{MAE}(L_2) := \mathrm{median}(\|\hat{f} - f\|_{L_2}).$$

The reason we took median rather than root mean square error was related to robustness. In particular, in a small number of cases, where we did not manage to obtain good convergence. As a result, we worried there might be some bias in the estimation of the mean squared error due to outliers, and took median squared error instead.

In a supplementary online document we also report the MAE based on two other norms, namely $\mathrm{MAE}(\mathcal{F}_n) := \mathrm{median}(\|\hat{f} - f\|_{\mathcal{F}_n})$ and $\mathrm{MAE}(\mathcal{F}) := \mathrm{median}(\|\hat{f} - f\|_{\mathcal{F}})$. Essentially the same results are seen, but more strongly.

Further simulation details are as follows. We took a sample size $n = 50$ and the x_i equally spaced over $[0, 1]$. This sample size makes the computations tractable, and our explorations with other sample sizes showed no essential differences in conclusions. Hyperparameters were estimated using maximum marginal likelihood. For regularization and the I-prior method, only one hyperparameter needs to be estimated, namely the scale parameter (denoted λ in the paper). For the SE estimator, an additional hyperparameter needs to be estimated, namely the parameter σ in the formula above. The latter makes the SE estimator significantly more difficult to compute for two reasons: (i) it takes more time to search for a local maximum of the marginal likelihood, and (ii) it is often more difficult to find
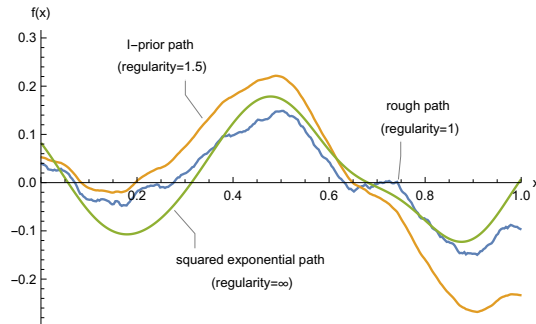
Figure 4: Example sample paths of three different regularities used in simulation. The true regression function is in the centered Brownian motion RKHS, consisting of functions of regularity greater than 1. Paths are centered to integrate to zero and have unit RKHS length.

the global maximum because more starting values need to be tried. As can be seen in Figure 5, estimation of the SE estimator broke down for very small error standard deviations and rough truths. For all estimators multiple local maxima of the marginal likelihood were sometimes encountered so we used several starting values, so that most of the time we could find the global maximum. However, in particular in some extreme cases (such as very small error standard deviations) we found for some data it could be very difficult to find the global maximum, especially for the SE estimator. As mentioned, we computed the *median* absolute error (MAE) rather than the mean squared error for robustness purposes.

The simulation results are displayed in Figure 5 using log-log plots of the MAE as a function of the error standard deviation. It is seen that the I-prior method always outperforms regularization, though the advantage of the former is small for the roughest functions in the RKHS (see the subfigures (a)). For rougher true regression functions in the RKHS, the I-prior estimator outperforms the SE estimator, which breaks down numerically for small errors. For analytic truths, the SE estimator outperforms the I-prior, as was to be expected. Overall, because the I-prior method can estimate very smooth functions quite well, but the SE estimator cannot estimate rough functions well, the SE estimator does not seem satisfactory for use in the present case. Furthermore, as mentioned before, the SE estimator is numerically more difficult to find.

8. Conclusion

We introduced the I-prior methodology for estimating a parametric or non-parametric regression function in a likelihood setting. One advantage of the methodology is that, because the I-prior is proper, the posterior mean is an admissible estimator. This is unlike the Tikhonov regularizer, which is the main alternative methodology that can be used in the same setting. Simulations and real

31
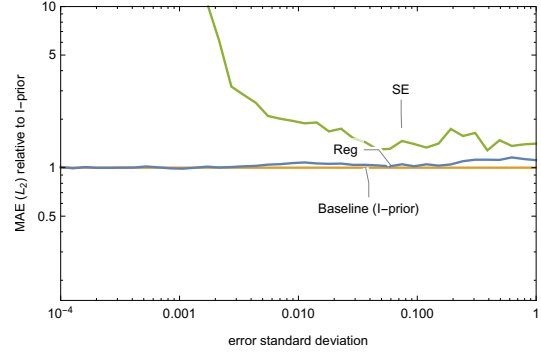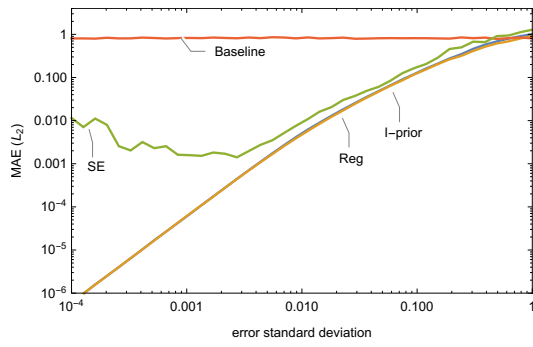
(a) True regression function has regularity 1
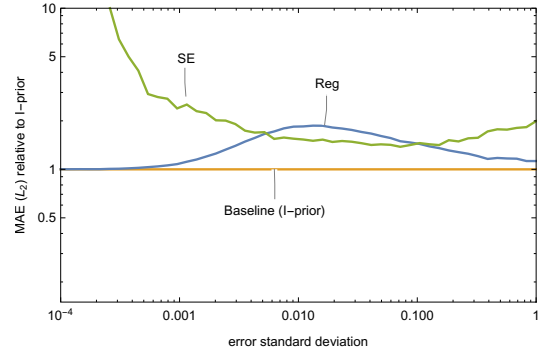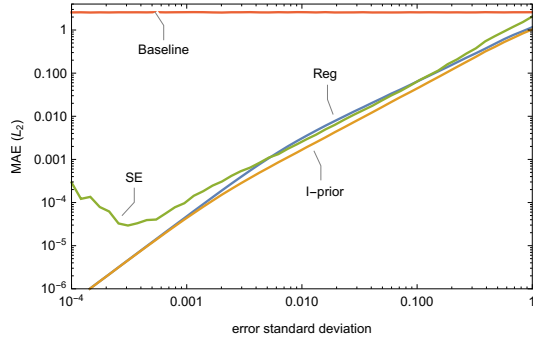
(b) True regression function has regularity 1.5
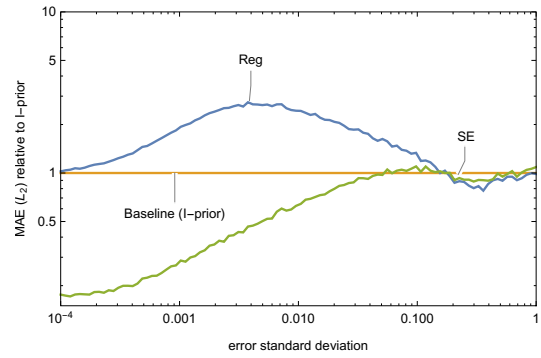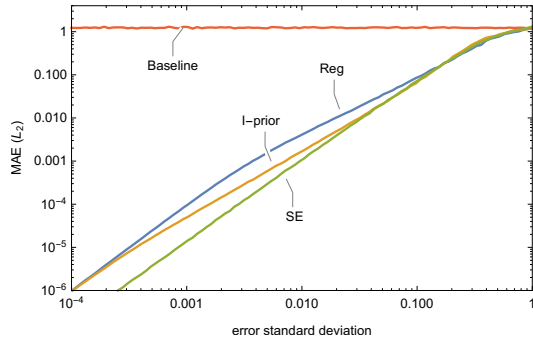
(c) True regression function is a squared exponential Gaussian process path

Figure 5: Panels on left: simulated $\text{MAE}(L_2)$ for Tikhonov regularizer ('Reg'), I-prior estimator ('I-prior'), and SE estimator ('SE'). The baseline level is the MAE if the zero function is fitted. Panels on right: ratio of $\text{MAE}(L_2)$ for regularizer and SE estimator compared to I-prior. Model (1) is assumed with $\mathcal{F}$ the FBM-1/2 RKHS and i.i.d. normal errors.

data analyses also show better performance of the I-prior methodology compared to Tikhonov regularization. Furthermore, the I-prior methodology is as automatic as would seem possible. The user *does* need to specify a distribution for the errors and a reproducing kernel for the space of regression functions, but it is hard to envisage how one could proceed without this effort. In practice, suitable choices are often i.i.d. or autoregressive errors, and a canonical or FBM-1/2 RKHS, for linear and smooth regression functions, respectively. In the real data examples, these were shown to give good predictive performance. The I-prior methodology is general in the following sense: if a regression function f is assumed to lie in a Hilbert space which is not an RKHS, then the Fisher information on f need not exist, in which case it is difficult to see how f could be estimated in a pointwise consistent way.

Further work consists of generalizing the I-prior methodology to multiple, possibly multidimensional covariates. Each covariate then requires a separate reproducing kernel which can be combined using the ANOVA construction (Wahba, 1990a; Gu and Wahba, 1993). Each kernel requires a scale parameter which can take values on the real line, and if some are negatively valued, the resulting kernel becomes indefinite generating a *reproducing kernel Krein space*. As the Fisher information remains positive definite, the I-prior methodology straightforwardly extends to this case. In this setting, variable selection can be performed.

Appendix A. Fisher information, associated distances, and maximum entropy priors

This section puts the I-prior methodology in a somewhat broader context and shows how it can be generalized. We also give a generalization of Zellner's g-prior, which we call Rao-Jeffreys prior, being based on the Rao metric and Jeffreys measure. The I-prior and Rao-Jeffreys priors are based on two different Riemannian metrics derived from the Fisher information.

Appendix A.1. Fisher information and distances between probability distributions
We first define the Fisher information, and then describe two distances between probability distributions based on it, namely (i) the Rao distance, which is the length of the shortest geodesic in the Riemannian metric induced by the Fisher information matrix, and (ii) the length of the shortest geodesic in the Riemannian metric induced by the *inverse* Fisher information matrix. The former is parameterization invariant and measures the amount of information between two parameter values. If the Fisher information does not depend on the parameter of interest, the latter can be easily related to the Cramér-Rao bound (see (11)). Finally, we give the volume measures associated with the distances. The volume measure associated with the Rao distance is well-known to be the Jeffreys measure.

Let Θ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_\Theta$, and let X be a random variable with density in the parametric family $\{P(\cdot|\theta)|\theta \in \Theta\}$. If $P(X|\theta) > 0$, the log-likelihood function of θ is denoted $L(\theta|X) = \log P(X|\theta)$. Assuming existence, the score is defined as the gradient $\nabla L(\theta|X)$ (see Appendix C for the definition of the gradient), and the Fisher information $I[\theta] \in \Theta \otimes \Theta$ for θ as

$$I[\theta] = -E\big[\nabla^2 L(\theta|X)\,\big|\,\theta\big].$$

For $b \in \Theta$, denote $\theta_b = \langle \theta, b \rangle_\Theta$. We define the Fisher information on θ_b as

$$I[\theta_b] = \big\langle I[\theta], b \otimes b \big\rangle_{\Theta \otimes \Theta},$$

and the Fisher information between θ_b and $\theta_{b'}$ as

$$I[\theta_b, \theta_{b'}] = \big\langle I[\theta], b \otimes b' \big\rangle_{\Theta \otimes \Theta},$$

where $\langle \cdot, \cdot \rangle_{\Theta \otimes \Theta}$ is the usual inner product on the tensor product space $\Theta \otimes \Theta$.

We now consider two distances between probability distributions $P(\cdot|\theta)$, $\theta \in \Theta$, based on the Fisher information on θ. We assume Θ possesses a finite dimensional parameterization such that the Fisher information $I[\theta]$ for θ is nonsingular.

The first is the well-known Rao distance D_{Rao}, defined as the length of the shortest geodesic on the Riemannian manifold whose metric tensor is the Fisher information (Rao, 1945; Atkinson and Mitchell, 1981; Amari, 1985). As an example, consider the family of multivariate normal distributions with unknown mean $\mu \in \mathbb{R}^p$ and known covariance matrix Σ. The Fisher information on μ is Σ^{-1}, which does not depend on μ so the metric is flat, and the Rao distance between distributions indexed by their mean is the Mahalanobis distance and is given by

$$D_{\mathrm{Rao}}(\mu, \mu')^2 = (\mu - \mu')^\top \Sigma^{-1} (\mu - \mu').$$

Methods for computing the Rao distance are given by Atkinson and Mitchell (1981), and a list of further examples is given by Rao (1987). The Rao distance is invariant to reparameterization, which is an advantage if the parameterization of the model is arbitrary, but may be a disadvantage if the parameterization is not arbitrary, because scale information is lost.

This paper introduces a second distance that depends on the Fisher information, namely the distance D_I defined as the length of the shortest geodesic on the Riemannian manifold with metric tensor the *inverse* Fisher information.

For a distance D, define ν_D to be the associated volume measure. For Euclidean θ, the densities relative to Lebesgue measure are

$$\nu_{D_{\mathrm{Rao}}}(\theta) = \sqrt{\big|I[\theta]\big|} \tag{A.1}$$

and

$$\nu_{D_I}(\theta) = \sqrt{|I[\theta]^{-1}|}. \tag{A.2}$$

The measure $\nu_{D_{\mathrm{Rao}}}$ is the well-known Jeffreys measure or 'Jeffreys prior' (Jeffreys, 1946).

Appendix A.2. Maximum entropy distributions

In a class of distributions, we may consider the one maximizing entropy. Such maximum entropy distributions can be thought of as the 'least informative' concerning a parameter of interest, and may hence by useful as so-called noninformative prior distributions in Bayesian inference.

Let (Θ, D) be a metric space and let $\nu = \nu_D$ be a volume measure over Θ induced by D (e.g., Hausdorff measure). Denote by π a density on Θ relative to ν, i.e., if θ is a random variable with density π, then for any measurable subset $A \subset \Theta$, $\Pr(\theta \in A) = \int_A \pi(t)\nu(dt)$. With $\theta_0 \in \Theta$, let Π_D be the class of distributions π such that

$$E_\pi D(\theta, \theta_0)^2 = \text{constant}.$$

The entropy of π relative to ν is

$$\mathcal{E}(\pi) = \int_\Theta \pi(t) \log \pi(t)\nu(dt).$$

Standard variational calculus shows that, if it exists, the density maximizing $\mathcal{E}(\pi)$ subject to the constraint that $\pi \in \Pi_D$ is given by

$$\pi_D(t) = \frac{e^{-\frac{1}{2\lambda}D(t,\theta_0)^2}}{\int_\Theta e^{-\frac{1}{2\lambda}D(s,\theta_0)^2}\nu(ds)} \propto e^{-\frac{1}{2\lambda}D(t,\theta_0)^2},$$

where λ is a function of the above constant. This distribution can be thought of as maximizing 'uncertainty' subject to the constraint that the expected squared distance of the random variable θ from some 'best guess' θ_0 is fixed. If (Θ, D) is a Euclidean space, ν is a flat (Lebesgue) measure and π_D is a multivariate normal density.

We now give three such maximum entropy priors, including two based on the two distances defined in the previous section and derived from the Fisher information. Suppose Θ is a finite dimensional affine subspace of a Hilbert space with norm $\|\cdot\|_\Theta$. Set $D_\Theta(\theta, \theta') = \|\theta - \theta'\|_\Theta$. Defining $\pi_{\mathrm{reg}} := \pi_{D_\Theta}$ we obtain

$$\pi_{\mathrm{reg}}(\theta) \propto e^{-\frac{1}{2\lambda}\|\theta - \theta_0\|^2},$$

35

which is a Gaussian density as the volume measure ν_{D_Θ} is flat. The subscript 'reg' refers to regularization, because the posterior mode based on π_{reg} is the usual regularizer of θ, based on minimizing

$$\theta \mapsto -\log \Pr(X|\theta) + \frac{1}{2\lambda}\|\theta - \theta_0\|^2.$$

Alternatively, we can define a prior based on the aforementioned distances D_{Rao} and D_I, which are based on the Fisher information. We denote $\pi_{RJ} := \pi_{D_{\text{Rao}}}$ and refer to this as the *Rao-Jeffreys prior*, being based on the Rao distance and the Jeffreys measure (A.1). We denote $\pi_I := \pi_{D_I}$ and refer to this as the *I-prior*. The prior densities relative to Lebesgue measure are given by

$$\pi_{\text{RJ}}(f_w) \propto \sqrt{|I[\theta]|}\, e^{-\frac{1}{2\lambda}D_{\text{Rao}}(\theta,\theta_0)^2} \quad \text{and} \quad \pi_I(f_w) \propto \sqrt{|I[\theta]^{-1}|}\, e^{-\frac{1}{2\lambda}D_I(\theta,\theta_0)^2}.$$

The I-prior can be generalized to infinite dimensional spaces, as done in this paper, but the Rao-Jeffreys prior cannot.

Appendix B. Hölder smoothness of FBM RKHS basis functions

For Lemma 10 below we need the following two lemmas on geometric inequalities.

Lemma 8. *[A triangle inequality] Let $0 \le \gamma \le 1$ and let d_{ij} denote pairwise distances between points i and j in a metric space. Then*

$$d_{12}^\gamma \le d_{13}^\gamma + d_{23}^\gamma. \tag{B.1}$$

Proof. Let $r(z) = (1 + z^\gamma) - (1 + z)^\gamma$. Then $\dot{r}(z) = \gamma/z^{1-\gamma} - \gamma/(1+z)^{1-\gamma} > 0$ for $z > 0$. Hence, r is increasing and since $\lim_{z\to 0} r(z) = 0$, $r(z) > 0$ for $z > 0$. Thus, $(1 + z)^\gamma \le 1 + z^\gamma$ for $z > 0$. By the triangle inequality and this result,

$$d_{12}^\gamma \le (d_{13} + d_{23})^\gamma = d_{13}^\gamma(1 + d_{23}/d_{13})^\gamma \le d_{13}^\gamma(1 + (d_{23}/d_{13})^\gamma) = d_{13}^\gamma + d_{23}^\gamma.$$

$\square$

Lemma 9. *[A parallelogram inequality] Let x and x' be points in a Hilbert space $(\mathcal{X}, \|\cdot\|)$. Then for $0 \le \gamma \le 1$,*

$$\|x + x'\|^{2\gamma} + \|x - x'\|^{2\gamma} \le 2\|x\|^{2\gamma} + 2\|x'\|^{2\gamma}, \tag{B.2}$$

with equality if $\gamma = 1$.

Proof. With ϕ a metric embedding into a Hilbert space satisfying (23), let $A = \phi_\gamma(0)$, $B = \phi_\gamma(x)$, $C = \phi_\gamma(x + x')$ and $D = \phi_\gamma(x')$. Denoting the length of the line segment between A and B by AB, and so on, we have $AB = CD = \|x\|^\gamma$, $AD = BC = \|x'\|^\gamma$, and $BD = \|x - x'\|^\gamma$. With $E = B + D - A$, the points A, B, E, D form a parallelogram, and the parallelogram law gives

$$AE^2 + BD^2 = 2AB^2 + 2AD^2 \tag{B.3}$$

Let $M = (A + E)/2 = (B + D)/2$ be the midpoint of the parallelogram. By a symmetry argument, $AM = CM$, and the triangle inequality gives $AC \leq AM + CM = AE$. Hence, using (B.3), $AC^2 + BD^2 \leq 2AB^2 + 2AD^2$, which is equivalent to (B.2) and completes the proof. $\qquad\square$

For $\gamma = 1$, (B.1) is the usual triangle inequality and (B.2) is the parallelogram law.

Lemma 10. *For $0 < \gamma < 1$ and $x_0 \in \mathcal{X}$, the function $h_\gamma(x_0, \cdot)$ over $\mathcal{X}$ is Hölder of order 2γ.*

Proof. For $0 < \gamma \leq 1/2$ we have

$$
\begin{aligned}
|h_\gamma(x_0, x) - h_\gamma(x_0, x')| &= \frac{1}{2}\Big|\|x_0 - x\|_{\mathcal{X}}^{2\gamma} - \|x_0 - x'\|_{\mathcal{X}}^{2\gamma} - \|x\|_{\mathcal{X}}^{2\gamma} + \|x'\|_{\mathcal{X}}^{2\gamma}\Big| \\
&\leq \frac{1}{2}\Big|\|x_0 - x\|_{\mathcal{X}}^{2\gamma} - \|x_0 - x'\|_{\mathcal{X}}^{2\gamma}\Big| + \frac{1}{2}\Big|\|x\|_{\mathcal{X}}^{2\gamma} - \|x'\|_{\mathcal{X}}^{2\gamma}\Big| \\
&\leq \|x - x'\|_{\mathcal{X}}^{2\gamma},
\end{aligned}
$$

where the last inequality is due to the triangle inequality given in Lemma 8. For $1/2 < \gamma < 1$,

$$
\begin{aligned}
&|h_\gamma(x_0, x - t) - 2h_\gamma(x_0, x) + h_\gamma(x_0, x + t)| \\
&= \frac{1}{2}\Big|\|x_0 - x + t\|_{\mathcal{X}}^{2\gamma} - 2\|x_0 - x\|_{\mathcal{X}}^{2\gamma} + \|x_0 - x - t\|_{\mathcal{X}}^{2\gamma} - \|x - t\|_{\mathcal{X}}^{2\gamma} + 2\|x\|_{\mathcal{X}}^{2\gamma} - \|x + t\|_{\mathcal{X}}^{2\gamma}\Big| \\
&\leq \frac{1}{2}\Big|\|x_0 - x + t\|_{\mathcal{X}}^{2\gamma} - 2\|x_0 - x\|_{\mathcal{X}}^{2\gamma} + \|x_0 - x - t\|_{\mathcal{X}}^{2\gamma}\Big| + \frac{1}{2}\Big|\|x - t\|_{\mathcal{X}}^{2\gamma} - 2\|x\|_{\mathcal{X}}^{2\gamma} + \|x + t\|_{\mathcal{X}}^{2\gamma}\Big| \\
&\leq \|x - x'\|_{\mathcal{X}}^{2\gamma},
\end{aligned}
$$

where the last inequality is due to the parallelogram inequality given in Lemma 9. $\qquad\square$

Appendix C. The gradient

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be an inner product space and consider a function $g : \mathcal{H} \to \mathbb{R}$. Denote the directional derivative of g in the direction $s \in \mathcal{H}$ by $\nabla_s g$, that is,

$$\nabla_s g(x) = \lim_{\delta \to 0} \frac{g(x + \delta s) - g(x)}{\delta}. \tag{C.1}$$

37

The *gradient* of g, denoted by ∇g, is the unique vector field satisfying

$$\langle \nabla g(x), s \rangle = \nabla_s g(x) \quad \forall x, s \in \mathcal{H}.$$

Appendix D. Duality between AR(1) and MA(1) processes

Let α be a real number. Let $u = (u_1, \ldots, u_n)$ be the AR(1) process with parameter α defined by

$$u_1 = \epsilon_1 \qquad u_i = \alpha u_{i-1} + \epsilon_i \ (i = 2, \ldots, n),$$

where the ϵ_i are i.i.d. $N(0, \sigma^2)$. Let $v = (v_1, \ldots, v_n)$ be the MA(1) process with parameter $-\alpha$ defined by

$$v_i = \zeta_i - \alpha \zeta_{i+1} \ (i = 1, \ldots, n-1) \qquad v_n = \zeta_n,$$

where the ζ_i are i.i.d. $N(0, \sigma^{-2})$. Denote the covariance matrices of u and v by V_u and V_v.

Lemma 11. $V_v = V_u^{-1}$.

Proof. Write $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ and $\zeta = (\zeta_1, \ldots, \zeta_n)^\top$. Then $u = A\epsilon$ and $v = B\zeta$ where A and B have elements

$$a_{ij} = \begin{cases} 0 & i < j \\ \alpha^{i-j} & i \geq j \end{cases} \qquad b_{ij} = \begin{cases} 1 & i = j \\ -\alpha & i = j - 1 \\ 0 & \text{otherwise} \end{cases}.$$

Direct multiplication shows that $AB^\top = A^\top B = I$. Now $V_u = \sigma^2 AA^\top$ and $V_v = \sigma^{-2} BB^\top$, so $V_u V_v = I$, which is the desired result. $\qquad \square$

References

Amari, S., 1985. Differential-geometrical methods in statistics. volume 28. Springer.

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

Atkinson, C., Mitchell, A.F., 1981. Rao's distance measure. Sankhyā: The Indian Journal of Statistics, Series A , 345–365.

Berger, J.O., Bernardo, J.M., Sun, D., 2009. The formal definition of reference priors. The Annals of Statistics , 905–938.

Bernardo, J.M., 1979. Reference posterior distributions for Bayesian inference. J. Roy. Statist. Soc. Ser. B , 113–147.

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

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Bronski, J.C., 2003. Small ball constants and tight eigenvalue asymptotics for fractional brownian motions. Journal of Theoretical Probability 16, 87–100.

Cannings, T.I., Samworth, R.J., 2017. Random-projection ensemble classification. J. Roy. Stat. Soc. .

Chakraborty, A., Panaretos, V.M., 2019. Hybrid regularisation and the (in)admissibility of ridge regression in infinite dimensional hilbert spaces. Bernoulli 25, 1939–1976. URL: 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, 1720–1747.

Cohen, S., 2002. Champs localement auto-similaires, in: Abry, P., Goncalvès, P., Véhel, J.L. (Eds.), Lois d'échelle, fractales et ondelettes.

Efron, B., 2019. Bayes, oracle bayes and empirical bayes (with discussion). Statistical Science 34, 177–201.

Embrechts, P., Maejima, M., 2002. Selfsimilar processes. Princeton University Press.

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

Gilbarg, D., Trudinger, N.S., 1998. Elliptic partial differential equations of second order. Springer.

Goeman, J., Meijer, R., Chaturvedi, N., 2015. penalized: L1 (lasso and fused lasso) and l2 (ridge) penalized estimation in glms and in the cox model. URL http://cran. r-project. org/web/packages/penalized/index. html .

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

Green, P.J., Silverman, B.W., 1994. Nonparametric regression and generalized linear models: a roughness penalty approach. Monographs on Statistics and Applied Probability, Chapman & Hall, London.

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.

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O., Miftahuddin, M., Adler, W., Lausen, B., 2016. Ensemble of a subset of knn classifiers. Advances in Data Analysis and Classification , 1–14.

Ibrahim, J.G., Laud, P.W., 1991. On Bayesian analysis of generalized linear models using Jeffreys's prior. Journal of the American Statistical Association 86, 981–986.

Jaakkola, T., Haussler, D., 1998. Exploiting generative models in discriminative classifiers. Advances in neural information processing systems 11 , 487–493.

Jamil, H., 2019. iprior: Regression Modelling using I-Priors. URL: `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://cran.r-project.org/web/packages/iprior/vignettes/ .

Jamil, H.M., 2018. Regression modelling using priors depending on Fisher information covariance kernels (I-priors). Ph.D. thesis. The London School of Economics and Political Science (LSE).

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

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

Jaynes, E.T., 2003. Probability theory: the logic of science. Cambridge university press.

Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London (Ser. A) 186, 453–461.

Khan, Z., Gul, A., Mahmoud, O., Miftahuddin, M., Perperoglou, A., Adler, W., Lausen, B., 2016. An ensemble of optimal trees for class membership probability estimation, in: Analysis of Large and Complex Data. Springer, pp. 395–409.

Kolmogorov, A.N., 1940. The Wiener helix and other interesting curves in Hilbert space (in Russian). Dokl. Akad. Nauk SSSR 26, 115–118.

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

Lifshits, M., 2012. Lectures on Gaussian processes. Springer.

Mandelbrot, B.B., Ness, J.W.V., 1968. Fractional Brownian motions, fractional noises and applications. SIAM Review 10, pp. 422–437.

Picard, J., 2011. Representation formulae for the fractional Brownian motion, in: Séminaire de Probabilités XLIII. Springer, pp. 3–70.

Pillai, N.S., Wu, Q., Liang, F., Mukherjee, S., Wolpert, R.L., 2007. Characterizing the function space for bayesian kernel models. Journal of Machine Learning Research 8, 1769–1797.

Rao, C.R., 1945. Information and accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society 37, 81–91.

Rao, C.R., 1987. Differential metrics in probability spaces. Differential geometry in statistical inference 10, 217–240.

Rasmussen, C., Williams, C., 2006. Gaussian processes for machine learning. MIT press Cambridge, MA.

Schoenberg, I.J., 1937. On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space. The Annals of Mathematics 38, pp. 787–793.

Sriperumbudur, B., Fukumizu, K., Lanckriet, G., 2011. Universality, characteristic kernels and RKHS embedding of measures. Journal of Machine Learning Research 12, 2389–2410.

Stein, E.M., 1970. Singular integrals and differentiability properties of functions. volume 2. Princeton university press.

Steinwart, I., 2018. Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties. Potential Analysis , 1–35.

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

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2003. Class prediction by nearest shrunken centroids, with applications to dna microarrays. Statistical Science , 104–117.

van der Vaart, A., van Zanten, H., 2007. Bayesian inference with rescaled Gaussian process priors. Electronic Journal of Statistics 1, 433–448.

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

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, 265–278.

Wahba, G., 1978. Improper priors, spline smoothing and the problem of guarding against model errors in regression. J. Roy. Statist. Soc. B 40, pp. 364–372.

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

Wahba, G., 1990b. Spline models for observational data. volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

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

Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti 6, 233–243.

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