Bayesian Analysis (2017) **12**, Number 1, pp. 219–238

Adaptive Empirical Bayesian Smoothing Splines

Paulo Serra[*‡] and Tatyana Krivobokova[†]

Abstract. In this paper we develop and study adaptive empirical Bayesian smoothing splines. These are smoothing splines with both smoothing parameter and penalty order determined via the empirical Bayes method from the marginal likelihood of the model. The selected order and smoothing parameter are used to construct adaptive credible sets with good frequentist coverage for the underlying regression function. We use these credible sets as a proxy to show the superior performance of adaptive empirical Bayesian smoothing splines compared to frequentist smoothing splines.

Keywords: adaptive estimation, unbiased risk minimiser, maximum likelihood, oracle parameters.

1 Introduction

Consider n observations from the non-parametric regression model

$$Y_i = f(x_i) + \sigma\epsilon_i, \quad i = 1, \dots, n. \tag{1}$$

The function f is assumed to belong to a Sobolev class $\mathcal{W}_\beta(M)$, a collection of continuous functions $f \in L_2$ such that $f^{(\beta-1)}$ is absolutely continuous and $\|f^{(\beta)}\|^2 < M^2$, where $\|\cdot\|$ is the ℓ_2-norm. The design points $\boldsymbol{x} = (x_1, \dots, x_n) \in [0,1]^n$ are $x_i = (2i-1)/(2n)$, the observation errors $\epsilon_1, \dots, \epsilon_n$ are assumed to be i.i.d. standard Gaussian random variables and $\sigma^2 > 0$. Parameters f, β, and σ^2 are unknown and of interest.

In this paper we study a smoothing spline estimator for f, which is the unique minimiser in $\mathcal{W}_q$ of the penalised least squares criterion

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - f(x_i)\right\}^2 + \lambda\int_0^1 \left\{f^{(q)}(t)\right\}^2 dt, \quad \lambda > 0, \quad q \in \mathbb{N} \tag{2}$$

and is well known to be a natural polynomial smoothing spline of degree $2q-1$ with knots at the observation points; see Wahba (1990).

The performance of smoothing splines as data-smoothers crucially depends on the choice of the smoothing parameter λ, which balances fidelity to the data and smoothness of the estimator. Criteria to select such a smoothing parameter can be obtained

*Korteweg–de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, the Netherlands, p.serra@uva.nl

†Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Göttingen, Germany, tkrivob@gwdg.de

‡The research took place while the author was a postdoctoral researcher at Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Göttingen, Germany.

 DOI: 10.1214/16-BA997

under two paradigms, which correspond to making different assumptions on the data-generating mechanism. One possibility is to assume that the regression function f is some fixed function from a certain class (frequentist model). In this case λ is estimated by minimising an unbiased estimator of the mean integrated squared error (unbiased risk estimator). Generalised cross validation (GCV), Mallow's C_p and Akaike's information criterion are all asymptotically equivalent criteria of this type. In the following $\hat{\lambda}_f$ denotes a minimiser of one of these criteria.

Another possibility is to assume that the regression function f is a realisation of some stochastic process (Bayesian model). Here a conjugate prior is put on the regression function, such that the resulting posterior mean coincides with the smoothing spline estimator. The smoothing parameter λ is then a so-called *hyper-parameter* of the prior. Its estimator is set to a maximiser of the resulting marginal likelihood (empirical Bayes method) and will be denoted by $\hat{\lambda}_q$.

The prior we use is a Gaussian prior. For these, conjugacy properties are often explored to directly study the posterior and specific Bayes estimators. Properties of Gaussian process priors (not necessarily conjugate) can be found in van der Vaart and van Zanten (2008); modifications to obtain adaptive priors were proposed in van der Vaart and van Zanten (2009). A (by no means extensive) list of results on adaptation using Gaussian priors in regression and the closely related Gaussian white noise model include Belitser and Ghosal (2003); Babenko and Belitser (2010); Knapik et al. (2011, 2013); de Jonge and van Zanten (2010, 2012); Szabó et al. (2014).

Bayesian smoothing splines with Gaussian priors have been first considered in Kimeldorf and Wahba (1970). Extensions and modifications of these splines have been discussed e.g. in Kohn and Ansley (1987), Speckman and Sun (2003) or Yue et al. (2014).

The asymptotic distributions of $\hat{\lambda}_f$ and $\hat{\lambda}_q$ can be computed under the assumption that the data come from the frequentist model with f as a fixed, "true" regression function of interest. This allows a direct comparison of these two estimators obtained under different paradigms. Krivobokova (2013) shows that $\hat{\lambda}_f$ and $\hat{\lambda}_q$ are consistent for certain oracles and that the asymptotic variance of $\hat{\lambda}_f$ can be several times larger than that of $\hat{\lambda}_q$.

The literature on adaptive Bayesian non-parametric estimation in non-parametric regression, and their frequentist performance is already quite extensive. For general priors, sufficient conditions for so called posterior contraction were proposed in Ghosal et al. (2000); Shen and Wasserman (2001); Ghosal and van der Vaart (2007) – posterior contraction at a given rate ensures the existence of frequentist estimators with the same rate. Adaptation is usually achieved by considering a family of priors indexed by a hyper-parameter (like λ and q above). If the regression function f belongs to a given smoothness class and there is some value of the hyper-parameter such that the resulting posterior contracts at the minimax rate for f in that class, then either endowing the hyper-parameter with a prior (hierarchical Bayes), or picking the hyper-parameter in a data-driven way (empirical Bayes), can lead to posteriors that contract adaptively. This approach has be used in Belitser and Levit (2003); Zhang (2005); Johnstone and

Silverman (2005); McAuliffe et al. (2006); Ghosal et al. (2008); Knapik et al. (2013); Shen and Ghosal (2015), among others. Here we consider the empirical Bayes approach.

Smoothing parameters that minimise an unbiased risk estimator (e.g., GCV) are predominant in practice and known to have good theoretical properties. In particular, if there is a mismatch between the order of the spline $(2q-1)$ and the smoothness of the regression function (f admits more than just q square integrable derivatives), then $\hat{\lambda}_f$ adapts to this extra smoothness, but only up to $2q$. In contrast, $\hat{\lambda}_q$ does not adapt and its rate is determined by q only. The main question that we address in this paper is whether one can obtain a *selector* $\hat{q}$, such that the resulting $\hat{\lambda}_{\hat{q}}$ not only adapts to the underlying smoothness of f, but also outperforms $\hat{\lambda}_f$ due to a much smaller variance.

To derive such a selection criterion for q we use the fact that the prior distribution depends on q, albeit in an implicit way, and apply the empirical Bayes approach. Contrary to the selection of the smoothing parameter λ, the selection of the order of smoothing splines has received very little attention in the literature.

Since $\hat{\lambda}_f$ and $\hat{\lambda}_{\hat{q}}$ are associated with splines of different order, direct comparison between the two smoothing parameters is not adequate. Instead, we construct credible ℓ_2 balls with good frequentist coverage, obtained from a high probability region of the posterior corresponding to hyper-parameters $\hat{q}$ and $\hat{\lambda}_{\hat{q}}$ selected via empirical Bayes. Subsequently, we show that if the centre of this ball is replaced by a smoothing spline with the smoothing parameter $\hat{\lambda}_f$, then the coverage property is lost, proving superiority of adaptive empirical Bayesian smoothing splines.

This paper is structured as follows. In Section 2 we describe the empirical Bayes approach and define some estimators. The asymptotic behaviour of the estimators is described in Section 3. In Section 4 we establish frequentist properties of a specific type of Bayesian credible set. In Section 5 we compare our approach for the selection of the smoothing parameter with generalised cross validation. Some numerical experiments can be found in Section 6. Section 7 contains some conclusions. We refer the reader to the supplementary materials (Serra and Krivobokova, 2016) for technical details and proofs.

2 Empirical Bayesian smoothing splines

Let us denote the minimiser of (2) by $\hat{f}_{\lambda,q}$, which is a natural smoothing spline of degree $2q-1$ with knots at the observation points. This smoothing spline estimator is a linear estimator that satisfies for each λ and q

$$\hat{\boldsymbol{f}}_{\lambda,q} = \hat{f}_{\lambda,q}(\boldsymbol{x}) = \boldsymbol{S}_{\lambda,q}\boldsymbol{Y}. \tag{3}$$

The positive-definite smoother matrix $\boldsymbol{S}_{\lambda,q} \in \mathbb{R}^{n\times n}$ is known explicitly and the vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ comprises of the observations from model (1).

The L_2-risk of the smoothing spline estimator (3) of $f \in \mathcal{W}_\beta$ is a function of λ and q known asymptotically; see Wahba (1990). In particular, with a suitable λ that

minimises that risk, it holds for each q

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\lambda,q} - \boldsymbol{f}\|^2 \asymp n^{-\frac{2\min(\beta,q)}{2\min(\beta,q)+1}}. \tag{4}$$

This is of the order of the minimax risk for estimating $f \in \mathcal{W}_{\min(\beta,q)}$ in model (1). Apparently, to minimise this risk q should be larger than the maximum smoothness of f. This fact is known in an wider context of Tikhonov regularisation, see Lukas (1998) and references therein. However, we are not aware of any practical methods to select the optimal penalisation order q. In this work, estimates for both λ and q are obtained via the empirical Bayes method. We start by specifying a prior on the regression function f and on σ^2.

Given $(\boldsymbol{x}, \lambda, q)$ we endow σ^2 with an inverse-gamma prior with shape parameter a, and scale parameter b (both left unspecified for now), and given $(\sigma^2, \boldsymbol{x}, \lambda, q)$ we endow $\boldsymbol{f}$ with a so called partially informative Gaussian prior with mean vector $\mathbf{0}$ and precision matrix $(\boldsymbol{S}_{\lambda,q}^{-1} - \boldsymbol{I}_n)/\sigma^2$, which we denote $PN\{\mathbf{0}, (\boldsymbol{S}_{\lambda,q}^{-1} - \boldsymbol{I}_n)/\sigma^2\}$ and admits a density which is proportional to

$$\Big|\frac{\boldsymbol{S}_{\lambda,q}^{-1} - \boldsymbol{I}_n}{\sigma^2}\Big|_+^{1/2} \exp\Big\{-\frac{1}{2\sigma^2}\boldsymbol{f}^T(\boldsymbol{S}_{\lambda,q}^{-1} - \boldsymbol{I}_n)\boldsymbol{f}\Big\}, \tag{5}$$

where $|\cdot|_+$ represents the product of the non-zero eigenvalues (the smoother matrix $\boldsymbol{S}_{\lambda,q}$ has exactly q eigenvalues equal to 1; cf. Speckman, 1985 and (31) in the supplementary materials). This prior has two parts, a constant, non-informative prior on the null space of $\boldsymbol{S}_{\lambda,q}^{-1} - \boldsymbol{I}_n$, and a proper, degenerate Gaussian prior on the range of $\boldsymbol{S}_{\lambda,q}^{-1} - \boldsymbol{I}_n$; cf. Speckman and Sun, 2003.

We say that the prior on $(\boldsymbol{f}, \sigma^2)$ is a partially-informative-Gaussian-inverse-gamma distribution, and denote it by

$$\Pi_{\lambda,q}(\,\cdot \mid \boldsymbol{x}) = PNIG\Big\{\mathbf{0},\ \boldsymbol{S}_{\lambda,q}^{-1} - \boldsymbol{I}_n,\ a,\ b\Big\}. \tag{6}$$

This prior is conjugate for model (1) and the corresponding posterior distribution is

$$\Pi_{\lambda,q}(\,\cdot \mid \boldsymbol{Y}, \boldsymbol{x}) = PNIG\Big\{\boldsymbol{S}_{\lambda,q}\boldsymbol{Y},\ \boldsymbol{S}_{\lambda,q}^{-1},\ a + \frac{n}{2},\ b + \frac{1}{2}\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q})\boldsymbol{Y}\Big\}. \tag{7}$$

The posterior mean of (7) is $\hat{\boldsymbol{f}}_{\lambda,q}$, and the mean of the predictive posterior distribution can be shown to be the smoothing spline $\hat{f}_{\lambda,q}$. The prior (6) is improper but the corresponding posterior (7) is proper. By definition, the marginal posterior for σ^2 is an inverse-gamma distribution

$$\Pi_{\lambda,q}^{\sigma^2}(\,\cdot\,|\boldsymbol{Y}, \boldsymbol{x}) = \int \Pi_{\lambda,q}(d\boldsymbol{f},\,\cdot\,|\boldsymbol{Y}, \boldsymbol{x}) = IG\Big\{a + \frac{n}{2},\ b + \frac{1}{2}\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q})\boldsymbol{Y}\Big\}, \tag{8}$$

and the marginal posterior for $\boldsymbol{f}$ is a non-central, n-variate t-distribution (cf. Kotz and Nadarajah, 2004)

$$\Pi_{\lambda,q}^{\boldsymbol{f}}(\,\cdot\,|\boldsymbol{Y}, \boldsymbol{x}) = \int \Pi_{\lambda,q}(\,\cdot\,, d\sigma^2|\boldsymbol{Y}, \boldsymbol{x}) = t_{2a+n-q}\Big\{\boldsymbol{S}_{\lambda,q}\boldsymbol{Y},\ \frac{2b + \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q})\boldsymbol{Y}}{2a + n - q}\boldsymbol{S}_{\lambda,q}\Big\}. \tag{9}$$

The posterior distribution $\Pi_{\lambda,q}(\cdot \mid \boldsymbol{Y}, \boldsymbol{x})$ depends on λ and q, and on the (hyper-) parameters a and b. We select the unknown parameters λ and q in a data-driven way via estimators $\hat{\lambda}$ and $\hat{q}$ and plug them into the posterior (7) resulting in a new random measure, the empirical Bayes posterior, which is defined as

$$\Pi_{\hat{\lambda},\hat{q}}\big(\cdot \mid \boldsymbol{Y}, \boldsymbol{x}\big) = \Pi_{\lambda,q}\big(\cdot \mid \boldsymbol{Y}, \boldsymbol{x}\big)\big|_{(\lambda,q)=(\hat{\lambda},\hat{q})}. \tag{10}$$

Following Robbins (1955), the empirical Bayes method consists of setting λ and q to maximisers of the marginal likelihood of model (1) under the Bayesian paradigm. We designate the mean of the empirical Bayes marginal posterior (10) as the *adaptive empirical Bayesian smoothing spline*.

Since the data $\boldsymbol{Y}|(f, \sigma^2, \boldsymbol{x})$ are distributed like a $N(\boldsymbol{f},\ \sigma^2\boldsymbol{I}_n)$ random vector, and we endow $\boldsymbol{f}|(\sigma^2, \boldsymbol{x}, \lambda, q)$ with a $PN\{\boldsymbol{0},\ (\boldsymbol{S}_{\lambda,q}^{-1} - \boldsymbol{I}_n)/\sigma^2\}$ prior, then $\boldsymbol{Y}|(\sigma^2, \boldsymbol{x}, \lambda, q)$ is distributed like a $PN\{\boldsymbol{0},\ (\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q})/\sigma^2\}$ random vector. The variance $\sigma^2|(\boldsymbol{x}, \lambda, q)$ is endowed with an $IG(a, b)$ prior, such that $(\boldsymbol{Y}, \sigma^2)|(\boldsymbol{x}, \lambda, q)$ is by definition jointly distributed like a $PNIG\{\boldsymbol{0},\ \boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q},\ a,\ b\}$ random vector. By integrating out σ^2, $\boldsymbol{Y}|(\boldsymbol{x}, \lambda, q)$ is distributed like a $t_{2a-q}\{\boldsymbol{0},\ 2b(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q})^{-}/(2a-q)\}$ random vector, where the superscript "–" indicates the pseudo-inverse. It can be shown that this distribution admits a density with respect to an appropriate dominating measure, resulting, up to some constant $h_n(a, b)$, in the following marginal log-likelihood for (λ, q),

$$\ell_n(\lambda, q \mid a, b) = -\Big(a + \frac{n-q}{2}\Big) \log\Big\{\boldsymbol{Y}^T\big(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q}\big)\boldsymbol{Y} + 2b\Big\} + \frac{1}{2}\log|\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q}|_{+}.$$

The hyper-parameters a and b do not play an important role in our approach so we set $a = q/2$, $b = 0$ with the convention that $0^0 = 1$ (this corresponds to placing an improper prior on σ^2). This does simplify the expressions that follow, but a and b can be set to any non-negative value that is $o(n)$ and does not depend on λ or q, without affecting our results. We obtain, up to a constant, the marginal log-likelihood

$$\ell_n(\lambda, q) = \ell_n(\lambda, q \mid q/2, 0) = -\frac{n}{2}\log\Big\{\boldsymbol{Y}^T\big(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q}\big)\boldsymbol{Y}\Big\} + \frac{1}{2}\log|\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q}|_{+}, \tag{11}$$

where (λ, q) lives on $(0, \infty) \times \mathbb{N}$; note that $h_n(q/2, 0)$ does not depend on λ or q.

The dependence of (11) on q is rather implicit, so that it is convenient to represent the smoother matrix as $\boldsymbol{S}_{\lambda,q} = \boldsymbol{\Phi}\{\boldsymbol{I}_n + \lambda n\,\mathrm{diag}(\boldsymbol{\eta}_q)\}^{-1}\boldsymbol{\Phi}^T$. Here $\boldsymbol{\Phi}$ is the Demmler–Reinsch basis matrix, such that $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T = \boldsymbol{I}_n$, and $\boldsymbol{\eta}_q = (\eta_{q,1}, \ldots, \eta_{q,n})^T$, see Section 8.1 in the supplementary materials for details. In particular, (31) in Section 8 gives an approximation of $\eta_{q,i}$ as a function of q. With this, we can re-express (11) as

$$\ell_n(\lambda, q) = -\frac{n}{2}\log\Bigg(\sum_{i=q+1}^{n}\frac{X_i^2\lambda n\eta_{q,i}}{1+\lambda n\eta_{q,i}}\Bigg) + \frac{1}{2}\sum_{i=q+1}^{n}\log\frac{\lambda n\eta_{q,i}}{1+\lambda n\eta_{q,i}}, \tag{12}$$

where $\boldsymbol{X} = (X_1, \ldots, X_n) = \boldsymbol{\Phi}^T\boldsymbol{Y}$. Further, based on the approximations from (31), $\ell_n(\lambda, q)$ is continuously differentiable.

A maximiser $(\hat{\lambda}, \hat{q})$ of $\ell_n(\lambda, q)$ is found by means of estimating equations, as zeroes of appropriately rescaled partial derivatives of $\ell_n(\lambda, q)$. Our estimating equations for λ and q are respectively

$$\begin{aligned} T_\lambda(\lambda, q) &= -\frac{2\lambda}{n^2}\Big\{\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q})\boldsymbol{Y}\Big\}\frac{\partial \ell_n(q,\lambda)}{\partial \lambda} \\ &= \frac{1}{n}\sum_{i=q+1}^{n}\frac{X_i^2 \lambda n \eta_{q,i}}{(1+\lambda n \eta_{q,i})^2} - \frac{1}{n^2}\sum_{i=q+1}^{n}\frac{X_i^2 \lambda n \eta_{q,i}}{1+\lambda n \eta_{q,i}}\sum_{i=q+1}^{n}\frac{1}{1+\lambda n \eta_{q,i}}, \quad \text{and} \\ T_q(\lambda, q) &= -\frac{2q}{n^2}\Big\{\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q})\boldsymbol{Y}\Big\}\frac{\partial \ell_n(q,\lambda)}{\partial q} \\ &= \frac{1}{n}\sum_{i=q+1}^{n}\frac{X_i^2 \lambda n \eta_{q,i}\log(n\eta_{q,i})}{(1+\lambda n \eta_{q,i})^2} - \frac{1}{n^2}\sum_{i=q+1}^{n}\frac{X_i^2 \lambda n \eta_{q,i}}{1+\lambda n \eta_{q,i}}\sum_{i=q+1}^{n}\frac{\log(n\eta_{q,i})}{(1+\lambda n \eta_{q,i})} \\ &\quad + \frac{1}{n}\sum_{i=q+1}^{n}\frac{\partial X_i^2}{\partial q}\frac{\lambda n \eta_i}{1+\lambda n \eta_i}, \end{aligned} \tag{13}$$

See Section 8.1 in the supplementary materials for details on the derivation of these expressions. In particular, it is shown that the contribution of the last term in $T_q(\lambda, q)$ with $\partial X_i^2/\partial q$ is negligible. Note that since $\ell_n(\lambda, q)$ is continuously differentiable, if $\hat{\lambda}_q$ solves $T_\lambda(\lambda, q) = 0$ for each q and $\hat{q}$ solves $T_q(\hat{\lambda}_q, q) = 0$, then $(\hat{\lambda}, \hat{q}) = (\hat{\lambda}_{\hat{q}}, \hat{q})$. Note that for each q, $\hat{\lambda}_q$ is essentially the generalised maximum likelihood estimator from Wahba (1985); cf. (1.5) in Wahba (1985) and the criterion $T_\lambda(\lambda, q)$ above.

We briefly address some practical issues involving the optimisation of the criteria in (13). The estimate $\hat{\lambda}$ is taken on $[1/n, 1]$. Parameter q enters the eigenvalues $\eta_{q,i}$ according to (31) from the supplementary materials and X_i as the degree $2q-1$ of basis $\boldsymbol{\Phi}$. While values $\eta_{q,i}$ are defined for each q, the degree of a spline is, in practice, typically an integer. One practical way to proceed in minimisation of (13) would be to restrict $q \in \mathbb{N}$. Alternatively, one could generalise splines to a fractional order (cf. Unser and Blu, 2000, for a representation of fractional splines in terms of fractional B-splines), which we do not pursue. Instead, we can use the fact that the contribution of the term with $\partial X_i^2/\partial q$ is negligible (see Section 8.1 in the supplementary materials for details). Therefore, we suggest to relax $q \in (1/2, \log(n)]$ to be real-valued and in practice set $X_{q,i} = X_{\lfloor q \rfloor, i}$, which allows to estimate non-integer q's. Additionally, we also have to define $\mathcal{W}_\beta(M)$ for real-valued $\beta > 1/2$, which we do in Section 8 equation (29) in the supplementary materials. Hence, throughout the paper both q and β are understood as real-valued numbers. In particular, all results and proofs hold also for $q, \beta \in \mathbb{N}$.

In practice, finding $(\hat{\lambda}, \hat{q})$ that optimises the criteria in (13) consists of finding $\hat{\lambda}_q$ that solves, for each q in some fine grid $\mathbb{Q}_n$, the criterion $T_\lambda(\lambda, q)$ up to an $o(1/n)$ factor, then finding $\hat{q} \in \mathbb{Q}_n$ that solves $T_q(\hat{\lambda}_q, q)$ and setting $(\hat{\lambda}, \hat{q}) = (\hat{\lambda}_{\hat{q}}, \hat{q})$. The grid $\mathbb{Q}_n = \{q_1, \cdots, q_{N_n}\} \in (1/2, \log(n)]^{N_n}$ must be such that $|q_{i-1} - q_i| = o\{1/\log(n)\}$, $i = 1, \ldots, N_n$, with $q_0 = 1/2$. This ensures that

$$n^{-\frac{2q_{i-1}}{2q_{i-1}+1}} = n^{-\frac{2q_i}{2q_i+1}}\{1 + o(1)\}, \quad i = 1, \ldots, N_n,$$

which means that the discretisation is sufficiently fine.

3 Asymptotics of the solutions of the estimating equations

Fix some continuous regression function $f \in L_2$, and denote $\boldsymbol{B} = (B_1, \ldots, B_n)^T = \boldsymbol{\Phi}^T \boldsymbol{f}$ such that $\mathbb{E}\boldsymbol{X} = \mathbb{E}\boldsymbol{\Phi}^T\boldsymbol{Y} = \boldsymbol{B}$. The oracle smoothing parameters will be defined as a solution to the system of equations

$$0 = \mathbb{E}T_\lambda(\lambda, q) = \frac{1}{n}\Big\{ \sum_{i=q+1}^{n} \frac{B_i^2 \lambda n \eta_{q,i}}{(1+\lambda n \eta_{q,i})^2} - \sum_{i=q+1}^{n} \frac{\sigma^2}{(1+\lambda n \eta_{q,i})^2} + o(1) \Big\}, \tag{14}$$

$$0 = \mathbb{E}T_q(\lambda, q) = \frac{1}{n} \sum_{i=q+1}^{n} \frac{B_i^2 \lambda n \eta_{q,i} \log(\lambda n \eta_{q,i})}{(1+\lambda n \eta_{q,i})^2} + \log(1/\lambda)\mathbb{E}T_\lambda(\lambda, q), \tag{15}$$

where the expectation is taken under model (1). These expressions follow by several applications of Lemmas 1 and 2 from the supplementary materials, similar to derivations in Krivobokova (2013). To solve this system assume that for each $q > 1/2$, equation (14) has a unique solution λ_q. Then equation (15) at $\lambda = \lambda_q$ becomes

$$0 = \mathbb{E}T_q(\lambda_q, q) = \frac{1}{n} \sum_{i=q+1}^{n} \frac{B_i^2 \lambda_q n \eta_{q,i} \log(\lambda_q n \eta_{q,i})}{(1+\lambda_q n \eta_{q,i})^2}. \tag{16}$$

If this equation has a unique solution $\bar{\beta}$, then the solution to the system (14), (15) on $[1/n, 1] \times (1/2, \log(n)]$ will be called the oracle parameter $(\lambda_{\bar{\beta}}, \bar{\beta})$.

Apparently, the risk (4) depends on the relationship between q and β, whereby q should be chosen, while β is unknown. Therefore, we analyse both oracle parameters under two scenarios: a *low order penalty* scenario where $q \leq \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$, and a *high order penalty* scenario where $q > \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$. Here $\max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$ is considered, since $f \in \mathcal{W}_\beta(M)$ does not preclude $f \in \mathcal{W}_{\beta'}(M)$ for some $\beta' > \beta$. Additionally to $f \in \mathcal{W}_\beta(M)$, we also discuss the case when f is an analytic signal. $\mathcal{P}_\infty$ will denote the space of all analytic functions on $[0, 1]$ such that $\mathcal{P}_\infty \subset \mathcal{W}_\infty(M)$, while the space of all polynomials of degree $d-1$ is denoted by $\mathcal{P}_d$, $d \in \mathbb{N}$.

3.1 Empirical Bayes estimate for λ

First we consider the solution to (14) for each $q > 1/2$.

Theorem 1. *Let $f \in \mathcal{W}_\beta(M)$, and assume that $\|f^{(\beta)}\|^2 > 0$.*

If $1/2 < q \leq \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$, then

$$\lambda_q = \left[n \frac{\|f^{(q)}\|^2}{\sigma^2 \kappa_q(0,2)} \{1 + o(1)\} \right]^{-2q/(2q+1)}, \tag{17}$$

where the constants $\kappa_q(m, l)$ are defined in Section 8 in the supplementary materials.

If $q > \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$, *then*

$$\lambda_q \geq \left[n \frac{\|f^{(\beta)}\|^2}{\sigma^2 \kappa_q(0,2)} \{1 + o(1)\} \right]^{-2q/(2\beta+1)}. \tag{18}$$

Moreover, for any $q > 1/2$, $\hat{\lambda}_q$ *is consistent for* λ_q *and*

$$\lambda_q^{-1/(4q)} \left(\hat{\lambda}_q / \lambda_q - 1 \right) \xrightarrow{d} \mathcal{N}\left[0, \frac{2\kappa_q(2,2)}{\left\{ 3\kappa_q(0,2) - 2\kappa_q(0,3) \right\}^2} \right], \quad \text{as } n \to \infty.$$

Proof of equations (17) and (18) follows from Lemmas 1 and 3 from the supplementary materials. The consistency of $\hat{\lambda}_q$ and its asymptotic distribution in the case of $f \in \mathcal{W}_q(M)$ has been studied in Krivobokova (2013). Inspection of the proofs in Krivobokova (2013) shows that they hold with no changes for the case $q > \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$.

Note that if $f \in \mathcal{P}_q$ such that $\|f^{(q)}\| = 0$, then $\lambda_q = \infty$.

3.2 Empirical Bayes estimate for q

First, consider the low penalty scenario where $1/2 < q \leq \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$ holds, so that in particular $f \in \mathcal{W}_q(M)$. By Lemma 3 (cf. the supplementary materials),

$$\mathbb{E}T_q(\lambda_q, q) = -\lambda_q \log(1/\lambda_q) \|f^{(q)}\|^2 \{1 + o(1)\}. \tag{19}$$

Hence, for all $1/2 < q \leq \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$ the estimating equation $\mathbb{E}T_q(\lambda_q, q)$ remains strictly negative for $f \in \mathcal{W}_q(M) \backslash \mathcal{P}_q$ (that is as long as $\|f^{(q)}\| \neq 0$). If $f \in \mathcal{P}_q$, then $\mathbb{E}T_q(\lambda_q, q) = 0$.

Consider now the high order penalty scenario where $q > \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$. Contrary to the low penalty scenario, the sign of (16) is not characterised by just the assumption $f \in \mathcal{W}_\beta(M)$, which implies $f \notin \mathcal{W}_q(M)$. It turns out, that not every signal f that belongs to $\mathcal{W}_\beta(M)$ but not to $\mathcal{W}_{\beta+\delta}(M)$ for any $\delta > 0$ will be such that (16) is positive for $q > \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$. Such a mismatch between smoothness as "measured" by $\max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$, and smoothness as "measured" by a change in the sign of the sum in (16) (which we can estimate), seems unavoidable (cf. Belitser and Enikeeva, 2008 for a similar issue in the context of hypothesis testing for smoothness in the Gaussian white noise model, and Giné and Nickl, 2010 in the context of Hölder smoothness in the construction of adaptive L_∞ credible bands in density estimation). From such issues stems, for example, the inability to construct adaptive credible sets in certain models with good coverage probability for $f \in \bigcup_{\beta \in B} \mathcal{W}_\beta$ if the range of smoothnesses B is large (cf. Low, 1997 and Section 4).

To "estimate" $\max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$ for as large a family of models as possible, it is customary to remove from each model the functions for which the mismatch occurs, and thus consider the estimation problem over a smaller family. Possible functions to remove are those which are not *self-similar* (cf. Picard and Tribouley, 2000;

Giné and Nickl, 2010), or that do not satisfy a *polished-tail* condition (cf. Szabó et al., 2014).

In our context, the set of functions with polished tails, call it $\mathcal{M} = \mathcal{M}(L, N, \rho)$, $L > 0$, $N \in \mathbb{N}$, $\rho \geq 2$, corresponds the set of all square integrable sequences such that

$$\frac{1}{n}\sum_{i=j}^{n} B_i^2 \leq \frac{L}{n}\sum_{i=j}^{\rho j} B_i^2, \quad N \leq j \leq n/\rho. \tag{20}$$

Such a condition has the role of excluding "irregular" signals, and unlike self-similarity conditions, it is not associated with any specific smoothness class. The condition ensures that the energy contained in the blocks $(B_j, \ldots, B_{\rho j})$ does not surge over contributions of earlier blocks – the signal can only get more "polished" as one runs along the sequence B_i. This effectively excludes irregular signals that contain artefacts in their tails that influence the smoothness of the signal but are not detectable due to the noise.

For signals with polished tails, the criterium T_q can actually pick up on the smoothness of the signal. For fixed parameters L, N, ρ, there is a well defined notion of smoothness which we denote

$$\bar{\beta} = \max\big\{\beta > 1/2 : f \in \mathcal{W}_\beta(M) \cap \mathcal{M}\big\}, \tag{21}$$

If f satisfies the polished tail conditions and is not in $\mathcal{W}_\beta(M)$, then by Lemma 3 from the supplementary materials, for some $c > 0$,

$$\mathbb{E}T_q(\lambda_q, q) \geq c\lambda_q^{\beta/q} > 0, \tag{22}$$

for all large enough n. The conclusion is that for all sufficiently large n, $\mathbb{E}T_q(\lambda_q, q)$ changes signs at $\bar{\beta}$.

We conclude that the behaviour of $\mathbb{E}T_q(\lambda_q, q)$ can be described as follows. If for some $\beta > 1/2$, $f \in \mathcal{W}_\beta(M) \cap \mathcal{M}$ then the (continuous) criterion $\mathbb{E}T_q(\lambda_q, q)$ has a zero at $\min\{q \in \mathbb{Q}_n : q > \bar{\beta}\}$ since it is negative for $q \leq \bar{\beta}$, and positive for $q > \bar{\beta}$. If $f \in \mathcal{P}_\infty$ then $\mathbb{E}T_q(\lambda_q, q) \leq 0$, $q \in \mathbb{Q}_n$; if for some $d \in \mathbb{N}$, $f \in \mathcal{P}_d \backslash \mathcal{P}_{d-1}$ (such that f is a polynomial of degree exactly $d-1$), then $\mathbb{E}T_q(\lambda_q, q)$ is negative for $1/2 < q \leq d-1$ and is zero for $q \geq d$.

The proof of the following theorem is in Section 8.3 of the supplementary materials.

Theorem 2. *Assume that for some $\beta > 1/2$, $f \in \mathcal{W}_\beta(M)$, then*

$$\mathbb{P}\big\{\beta < \hat{q} \leq \log(n)\big\} \to 1, \quad n \to \infty.$$

If furthermore $\beta = \bar{\beta}$ as defined in (21), and $f \in \mathcal{M}$, then

$$\hat{q} \xrightarrow{P} \bar{\beta}, \quad n \to \infty,$$

If for some $d \in \mathbb{N}$, $f \in \mathcal{P}_d$ then

$$\mathbb{P}\big\{d \leq \hat{q} \leq \log(n)\big\} \to 1, \quad n \to \infty.$$

As a side-note, the oracle λ_q can also be made more explicit for $f \in \mathcal{W}_\beta(M) \cap \mathcal{M}$. By (18) and (20), one finds that for each $f \in \mathcal{M}_{\bar{\beta}}(M) \cap \mathcal{M}$,

$$\lambda_q(f) = \left[n \frac{c_f M^2}{\sigma^2 \kappa_q(0,2)} \{1 + o(1)\} \right]^{-2q/(2\bar{\beta}+1)}, \tag{23}$$

for some constant $0 < c_f \leq 1$, which depends on f and is bounded away from zero uniformly over $f \in \mathcal{W}_\beta(M) \cap \mathcal{M}$.

The exclusion of certain signals (so that the behaviour above holds) can be argued to be innocuous since the set of all signals that do not satisfy the polished-tail condition (or that are not self-similar) is "small". This can be justified following several arguments: removing such signals leaves the minimax rate (almost) unchanged so that the statistical problem does not become simpler; the probability that a function sampled from the prior does not satisfy such conditions is zero; there are also topological arguments for this. For a more extensive discussion cf. Giné and Nickl, 2010; Szabó et al., 2014 and the references therein. However, from the practical perspective, exactly which signals are removed is not relevant since one cannot check if the data come from a regression function satisfying a polished tail condition or not. In this sense one might as well implicitly exclude all signals for which (22) or (23) do not hold.

4 Bayesian credible sets as adaptive confidence sets

In Section 2 we propose a method for selecting the penalty order q of smoothing splines and the corresponding smoothing parameter λ_q. An immediate application of the consistency results in Section 3 is that $\hat{q}$ and $\hat{\lambda}_{\hat{q}}$ can be directly plugged into the smoothing spline $\hat{f}_{\lambda,q}$ to obtain adaptive estimates for any continuous regression function $f \in L_2$. (This follows immediately from the consistency of the parameters, and standard arguments for smoothing spline estimators; cf. Wahba, 1990.) In this section we present another application: the construction of rate adaptive confidence sets based on the empirical Bayes posterior (10).

One of the often mentioned advantages of the Bayesian approach is that a posterior distribution provides statisticians with more than just point estimates. For appropriate priors, if the data are distributed according to a fixed distribution in the model, then with probability going to one, the posterior concentrates around this distribution. If this is the case, then for appropriate q and λ a small ℓ^2-ball centred at the posterior mean $\hat{\boldsymbol{f}}_{\lambda,q}$, can capture most of the mass of the marginal posterior for $\boldsymbol{f}$. Since for each λ and q the posterior is known explicitly, simulating a high-probability region of the posterior (a credible set) and using it as a *frequentist* confidence set is of great appeal. However, it is known that such credible sets do not always have good frequentist coverage properties. This is more so the case when dealing with posteriors that adapt to the smoothness of the underlying signal to be estimated. In this section we adapt a technique developed by Szabó et al. (2014) for the Gaussian white noise model, to study the behaviour of a specific Bayesian credible set for our regression model (1). Complicating factors in our setup are that the variance of the noise is not assumed to be known, and that we

work with two empirically chosen parameters (λ and q) simultaneously. We outline the technique in some detail since it is of independent interest.

We remind that the marginal posterior $\Pi^{\boldsymbol{f}}_{\lambda,q}(\,\cdot\,|\boldsymbol{Y},\boldsymbol{x})$ equals

$$t_n\Big(\hat{\boldsymbol{f}}_{\lambda,q},\,\hat{\sigma}^2_{\lambda,q}\boldsymbol{S}_{\lambda,q}\Big),\quad \text{where}\quad \hat{\sigma}^2_{\lambda,q}=\frac{1}{n}\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{S}_{\lambda,q})\boldsymbol{Y}.$$

Representation properties of the multivariate t-distribution state that if $\boldsymbol{f}$ is distributed like the marginal posterior above, then $\|\boldsymbol{f}-\hat{\boldsymbol{f}}_{\lambda,q}\|^2/\hat{\sigma}^2_{\lambda,q}$ is distributed like $\boldsymbol{Z}^T\boldsymbol{S}_{\lambda,q}\boldsymbol{Z}/N$, where $\boldsymbol{Z}\sim N(\boldsymbol{0},\boldsymbol{I}_n)$, $N\sim\mathcal{X}^2_n$, and N is independent of $\boldsymbol{Z}$. Conclude that for any $\alpha\in(0,1)$ there exists a (known, deterministic) sequence $r_n(\lambda,q)$ such that for every n,λ,q,

$$\Pi^{\boldsymbol{f}}_{\lambda,q}\Big(\|\boldsymbol{f}-\hat{\boldsymbol{f}}_{\lambda,q}\|\le\hat{\sigma}_{\lambda,q}\,r_n(\lambda,q)\Big|\boldsymbol{Y},\boldsymbol{x}\Big)=1-\alpha.$$

The level α is fixed for the remainder of this section. It is therefore natural to consider, for any $L\ge 1$, the empirical credible ball

$$\hat{\mathcal{C}}_n(L)=\Big\{\boldsymbol{f}\in\ell^2:\|\boldsymbol{f}-\hat{\boldsymbol{f}}_{\hat{\lambda}_{\hat{q}},\hat{q}}\|\le\hat{\sigma}_{\hat{\lambda}_{\hat{q}},\hat{q}}\,L\,r_n(\hat{\lambda}_{\hat{q}},\hat{q})\Big\}. \tag{24}$$

By definition of the sequence $r_n(\lambda,q)$, for any $L\ge 1$,

$$\Pi^{\boldsymbol{f}}_{\hat{\lambda}_{\hat{q}},\hat{q}}\big(\hat{\mathcal{C}}_n(L)\,\big|\boldsymbol{Y},\boldsymbol{x}\big)\ge 1-\alpha,$$

such that we can sample functions in $\hat{\mathcal{C}}_n(L)$ by sampling functions from the posterior and then keeping those that satisfy the inequality in (24). Such functions give a visual impression of the uncertainty in the point estimate $\hat{\boldsymbol{f}}_{\hat{\lambda}_{\hat{q}},\hat{q}}$ – the adaptive empirical Bayesian smoothing spline. Note that since for each λ and q the posterior is known explicitly, simulating $\hat{\mathcal{C}}_n(L)$ is straightforward.

The following theorem is proved in Section 8.4 of the supplementary materials.

Theorem 3. *Consider an interval $B=[\underline{b},\bar{b}]$, where $1/2<\underline{b}\le\bar{b}<\infty$, and define $\mathcal{W}=\bigcup_{\beta\in B}\mathcal{W}_\beta(M)\cap\mathcal{M}(L,N,\rho)$. Then, for all large enough L,*

$$\inf_{\boldsymbol{f}\in\mathcal{W}}\mathbb{P}_f\big\{\boldsymbol{f}\in\hat{\mathcal{C}}_n(L)\big\}=1+o(1),\quad \textit{and} \tag{25}$$

$$\inf_{\boldsymbol{f}\in\mathcal{W}_\beta(M)}\mathbb{P}_f\big\{r_n(\hat{\lambda}_{\hat{q}},\hat{q})\le Kn^{-\beta/(2\beta+1)}\big\}=1+o(1),\quad \beta\in B, \tag{26}$$

for some large enough constant $K>0$, depending on L, M, ρ, σ^2, $\underline{b}$, and $\bar{b}$.

Statement (25) is usually referred to as *honest coverage*, while (26) means that the credible ball $\hat{\mathcal{C}}_n(L)$ has a radius of the optimal order.

Ideally one would like to take $\mathcal{W}=\ell^2$. However, as mentioned in Section 3.2, it is known (cf. Low, 1997) that it is in general not possible to fulfil conditions (25) and (26) simultaneously if $\bar{b}/\underline{b}>2$ and $\mathcal{W}=\ell^2$. To allow (25) and (26) to hold simultaneously

for a wide (but bounded) range of smoothnesses one usually identifies "problematic" functions which are either removed from the model (cf. Giné and Nickl, 2010; Szabó et al., 2014) or replaced with a collection of so called surrogates, "non-problematic" replacement functions that retain the main features of interest of the functions that were removed from the model (cf. Genovese and Wasserman, 2008). Imposing an upper bound $\bar{b}$ on the smoothness is also necessary if we are to establish (25) and (26). Such a bound can also be justified from a computational standpoint.

The constant L, which is the multiplicative factor for the radius of the credible set $\hat{\mathcal{C}}_n$, must be taken appropriately large for (25) and (26) to hold. It is possible to provide an explicit lower-bound L by inspecting the constants in the proof of Theorem 3: for all sufficiently large n we may take

$$L \geq 1 + \left\{\kappa_q(0,2)/\kappa_q(0,1)\right\}^{1/2} = 1 + \left\{(2q-1)/(2q)\right\}^{1/2},$$

so that uniformly over q, $L \geq 2$. Inspection of the computations in Section 8 from the supplementary materials shows that the level α appears associated with lower order terms so that L does not depend on α, even if we were to allow L to depend on q. Because of this one does not get exactly coverage $1-\alpha$ and the credible sets $\hat{\mathcal{C}}_n$ are always conservative (in that the asymptotic probability of coverage is $1 > 1-\alpha$).

It follows, in fact, from the inequalities established in the proof of Theorem 3 that the posterior distribution contracts (with respect to $\|\cdot\|$) around the true regression function at the optimal rate for $f \in \mathcal{W}_\beta$, when f is indeed in this space. To establish this, condition (20) is not needed.

5 Comparison with frequentist smoothing splines

In the frequentist framework there are several competing ways of selecting the smoothing parameter λ (for a fixed q). Typically, λ is selected as a minimiser of some asymptotically unbiased estimator of the risk $\mathbb{E}\|\hat{\boldsymbol{f}}_{\lambda,q} - \boldsymbol{f}\|^2$. Generalised cross validation, Mallow's C_p and Akaike's information criterion are particular examples of such estimates; let $\hat{\lambda}_f$ denote a minimiser of one of such criteria. If the regression function f belongs to $\mathcal{W}_\beta(M)$, and we set $q \in [\beta/2, \beta]$ such that $\beta/q \in [1,2]$, then $\hat{\lambda}_f$ adapts. This means that generalised cross-validated smoothing splines adapt to the unknown smoothness β in the sense that the estimator $\hat{\lambda}_f$ is consistent for the oracle

$$\lambda_f \geq \left[n \frac{\|f^{(\beta)}\|^2}{\sigma^2 \kappa_q(1,2)} \{1 + o(1)\} \right]^{-2q/(2\beta+1)}. \tag{27}$$

Furthermore, Theorem 3 of Krivobokova (2013) states that

$$\lambda_f^{-1/(4q)} \left(\hat{\lambda}_f/\lambda_f - 1\right) \xrightarrow{d} \mathcal{N}\left[0, \frac{2\kappa_q(4,2)}{\left\{4\kappa_q(1,2) - 3\kappa_q(1,3)\right\}^2}\right], \quad \text{as } n \to \infty,$$

so that the asymptotic variance above can be much larger than that associated with the empirical Bayes estimate $\hat{\lambda}_\beta$ for the range of values of q for which the GCV $\hat{\lambda}_f$ adapts, see Krivobokova (2013) for more discussion and simulations.

In this section we investigate how the asymptotic variance of $\hat{\lambda}_f$ compares to that of $\hat{\lambda}_{\hat{q}}$; or, more specifically, we compare the distances of $\hat{f}_{\hat{\lambda}_f,q}$ and of $\hat{f}_{\hat{\lambda}_{\hat{q}},\hat{q}}$ to the true regression function. We use the credible sets from the previous section as a proxy for this comparison: we bound the probability that the regression function belongs to a ball centred at $\hat{f}_{\hat{\lambda}_f,q}$ with a radius that assures coverage for the Bayesian credible set $\hat{\mathcal{C}}_n(L)$ – by construction, that is the radius of $\hat{\mathcal{C}}_n(1)$, since we are only interested here in coverage, and not honest coverage.

The proof of the following theorem is in Section 8.5 of the supplementary materials.

Theorem 4. *Assume that the regression function f belongs to $\mathcal{W}_\beta(M)$, $\beta > 1/2$, such that with probability going to 1 the radius of the credible set $\hat{\mathcal{C}}_n(1)$ is $\sigma\, r_n(\lambda_\beta, \beta)$. Define*

$$\hat{\mathcal{D}}_n = \hat{\mathcal{D}}_n\big(\hat{\boldsymbol{f}}_{\hat{\lambda}_f,q}\big) = \Big\{\boldsymbol{f} \in \ell^2 : \|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\hat{\lambda}_f,q}\| \le \sigma\, r_n(\lambda_\beta, \beta)\Big\}.$$

Then, for any q such that $\hat{\lambda}_f$ adapts to the smoothness β, i.e., for any $q \in [\beta/2, \beta]$,

$$\mathbb{P}_f\Big\{\boldsymbol{f} \in \hat{\mathcal{D}}_n\big(\hat{\boldsymbol{f}}_{\hat{\lambda}_f,q}\big)\Big\} = o(1), \quad n \to \infty.$$

The conclusion is that $\hat{\lambda}_f$ can somewhat adapt to the smoothness of the regression function, but at the cost of a high asymptotic variance. Using the fact that for each fixed q, the empirical Bayes selected $\hat{\lambda}_q$ has much lower asymptotic variance, we show that the smoothing parameter $\hat{\lambda}_{\hat{q}}$ outperforms $\hat{\lambda}_f$: if the centre of the empirical Bayes credible ball (the adaptive empirical Bayesian smoothing spline) is replaced by a frequentist smoothing spline with the smoothing parameter $\hat{\lambda}_f$, then the coverage property is lost. Note that even if this were not the case, the empirical Bayes smoothing spline would still adapt to a wider range of smoothnesses than the risk-based smoothing spline (which adapts only within a $[q, 2q]$ range).

6 Numerical simulations

The following simulation study aims to verify our theoretical findings in finite samples and compare frequentist and proposed adaptive empirical Bayesian smoothing splines in terms of the average mean squared error.

In all settings the Monte-Carlo sample $M = 1000$, the sample size is $n = 1000$, the design points are fixed and equidistant $x = i/n$, $i = 1, \dots, n$ and $\sigma = 0.1^2$. The results for other sample sizes and higher σs were found conceptually similar and are not reported. We consider two mean functions f_j, $j = 1, 2$ that are scaled by its range. Function f_1, shown on the top left plot of Figure 1, has a known decay of its Demmler–Reinsch coefficients, namely we set

$$f_1(x) = \sum_{i=q+1}^{n} \psi_{\beta,i}(x)(i+1)^{-\beta}\cos(2i), \quad \beta = 3,$$

where $\psi_{\beta,i}$ is the ith basis function of the Demmler–Reinsch basis of degree $\beta = 3$,

defined in (28). The second function is analytical $f_2(x) = \cos(5\pi x)$ (Figure 2) with an exponential decay of its Demmler–Reinsch coefficients.

We estimated both functions by smoothing splines with the empirical Bayesian smoothing parameter $\hat{\lambda}_{\hat{q}}$ and with the smoothing parameter $\hat{\lambda}_f$. To get $\hat{\lambda}_{\hat{q}}$, first $\hat{\lambda}_q$ is obtained for $q = 1, 2, \ldots, 6$ as a solution to $T_\lambda(\lambda, q) = 0$ and then $\hat{q}$ is set to the nearest integer to q^* such that $T_q(\hat{\lambda}_q, q^*) = 0$. The smoothing parameter $\hat{\lambda}_f$ has been calculated by generalised cross-validation for different values of $q = 2, \ldots, 6$.

We compare the resulting smoothing parameter estimators $\hat{\lambda}_{\hat{q}}$ and $\hat{\lambda}_f$ for $q = 2, \ldots, 6$ in terms of the sample mean and sample variance in the Monte Carlo sample (see mean$(\hat{\lambda})$ and var$(\hat{\lambda})$ in Table 1). Further, we compare the empirical average mean squared error (AMSE) of the resulting estimators, that is

$$A(\hat{\lambda}) = \frac{1}{Mn} \sum_{i=1}^{M} \sum_{j=1}^{n} \left\{ \hat{f}_i(x_j, \hat{\lambda}) - f(x_j) \right\}^2,$$

where $\hat{f}_i$ denotes a smoothing spline estimator in the ith Monte Carlo simulation. The values of $A(\hat{\lambda})$ are given in Table 1 together with the AMSE ratio, which is defined as $R = A(\hat{\lambda}_f)/A(\hat{\lambda}_{\hat{q}})$, so that values of $R > 1$ imply superiority of the adaptive empirical Bayesian estimator.

Let us first consider the simulation results for f_1. According to the theoretical results on $\hat{\lambda}_f$ discussed in Section 5, as long as $q \geq \beta/2$, its oracle $\lambda_f \geq \text{const } n^{-2q/(2\beta+1)}$, leading to $\mathbb{E}\|\hat{\boldsymbol{f}}(\hat{\lambda}_f) - \boldsymbol{f}\|^2 \asymp n^{-2\beta/(2\beta+1)}$. For $\beta = 3$ we expect to observe that λ_f for $q = 2$ is larger than for $q = 3$ and goes very fast to zero for $q > \beta = 3$. Moreover, $\hat{f}(\hat{\lambda})$ for $\hat{\lambda}_{\hat{q}}$ and $\hat{\lambda}_f$ for $q = 2, \ldots, 6$ should all have the same convergence rates and differ only in constants. Results of the simulations given in Table 1 confirm these theoretical findings. We observe also that the means of $\hat{\lambda}_{\hat{q}}$ and $\hat{\lambda}_f$ for $q = 3$ are of the same rate, but the variance of $\hat{\lambda}_f$ is much higher, which is also visible in the boxplots, given on the

	EB	GCV $q = 2$	$q = 3$	$q = 4$	$q = 5$	$q = 6$
f_1						
mean$(\hat{\lambda})$	$5.7 \cdot 10^{-12}$	$1.8 \cdot 10^{-08}$	$7.1 \cdot 10^{-12}$	$1.8 \cdot 10^{-15}$	$4.5 \cdot 10^{-19}$	$1.2 \cdot 10^{-22}$
var$(\hat{\lambda})$	$5.0 \cdot 10^{-25}$	$2.3 \cdot 10^{-17}$	$1.2 \cdot 10^{-23}$	$1.7 \cdot 10^{-30}$	$2.0 \cdot 10^{-37}$	$2.4 \cdot 10^{-44}$
$A(\hat{\lambda})$	$2.5 \cdot 10^{-06}$	$2.9 \cdot 10^{-06}$	$2.6 \cdot 10^{-06}$	$2.7 \cdot 10^{-06}$	$2.9 \cdot 10^{-06}$	$2.9 \cdot 10^{-06}$
R	–	1.147912	1.032582	1.075522	1.119951	1.139055
f_2						
mean$(\hat{\lambda})$	$1.7 \cdot 10^{-19}$	$5.9 \cdot 10^{-09}$	$1.6 \cdot 10^{-11}$	$3.4 \cdot 10^{-14}$	$3.6 \cdot 10^{-17}$	$2.8 \cdot 10^{-19}$
var$(\hat{\lambda})$	$1.1 \cdot 10^{-36}$	$2.0 \cdot 10^{-18}$	$2.4 \cdot 10^{-23}$	$4.8 \cdot 10^{-28}$	$4.0 \cdot 10^{-34}$	$2.4 \cdot 10^{-38}$
$A(\hat{\lambda})$	$1.5 \cdot 10^{-06}$	$3.8 \cdot 10^{-06}$	$2.1 \cdot 10^{-06}$	$1.9 \cdot 10^{-06}$	$1.8 \cdot 10^{-06}$	$1.6 \cdot 10^{-06}$
R	–	2.483676	1.406107	1.266803	1.164454	1.040263

Table 1: Simulation results for functions f_1 and f_2.

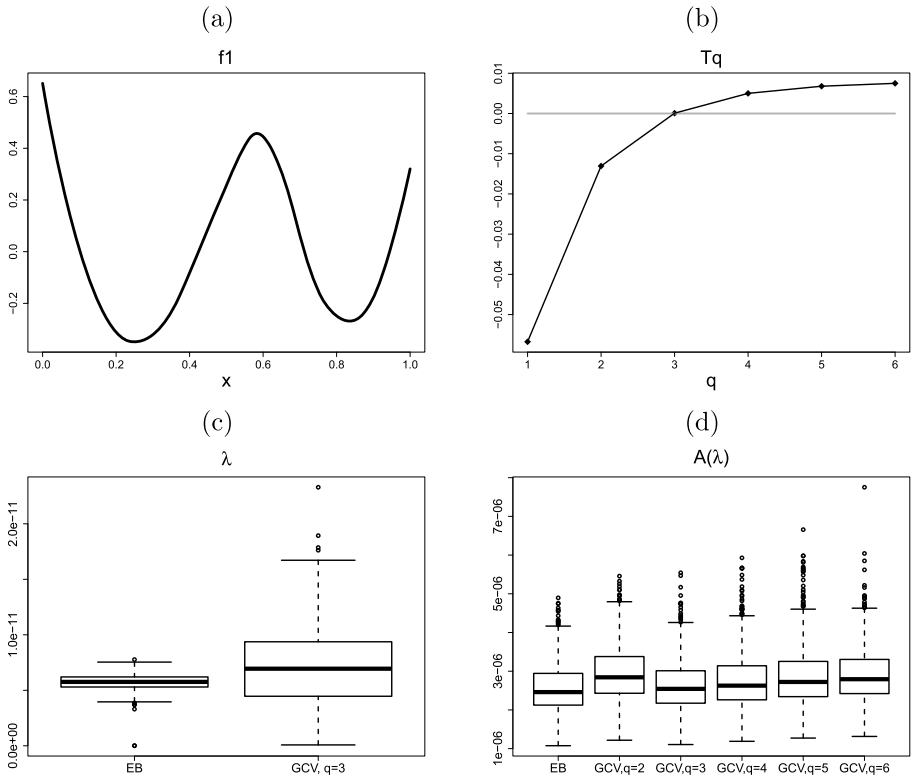

Figure 1: (a) Function f_1, (b) Criterion $T_q(\hat{\lambda}_q, q)$ for f_1, (c) Boxplots of $\hat{\lambda}_{\hat{q}}$ and $\hat{\lambda}_f$ obtained with GCV and $q = \beta = 3$, (d) Boxplots of $A(\hat{\lambda}_{\hat{q}})$ (EB) and $A(\hat{\lambda}_f)$.

bottom left plot of Figure 1. The differences between $A(\hat{\lambda})$, shown on the bottom right plot of Figure 1, are less pronounced, but from the AMSE ratio given in Table 1 we find that the empirical Bayesian smoothing spline estimator outperforms the frequentist smoothing spline uniformly in q. The smallest difference is observed between $A(\hat{\lambda}_{\hat{q}})$ and $A(\hat{\lambda}_f)$ for the true $q = \beta = 3$, which is, of course, unknown in practice and is not estimated in the frequentist framework. Finally, we remark that the empirical Bayesian estimator of q appeared to be very reliable for this value of β: out of $M = 1000$ samples in 998 cases $\hat{q} = 3$ has been obtained and in two cases $\hat{q} = 4$. The estimating equation $T_q(\hat{\lambda}_q, q)$ is shown on the top right plot of Figure 1.

Now we consider the simulation results for the analytical function $f_2(x) = \cos(5\pi x)$ which is known to have exponentially decaying Demmler–Reinsch coefficients. In this setting, criterion $T_q(\hat{\lambda}_q, q)$ should remain negative, asymptotically approaching zero from below, which is visible in the top right plot of Figure 2. We estimated $\hat{\lambda}_{\hat{q}}$ setting $\hat{q} = 6$ if $T_q(\hat{\lambda}_q, q) < 0$ for $q = 6$. For function f_2 estimator $\hat{q}$ is slightly more variable, resulting in $\hat{q} = 6$ in 986 cases and in $\hat{q} = 5$ in 14 cases. This can be due to

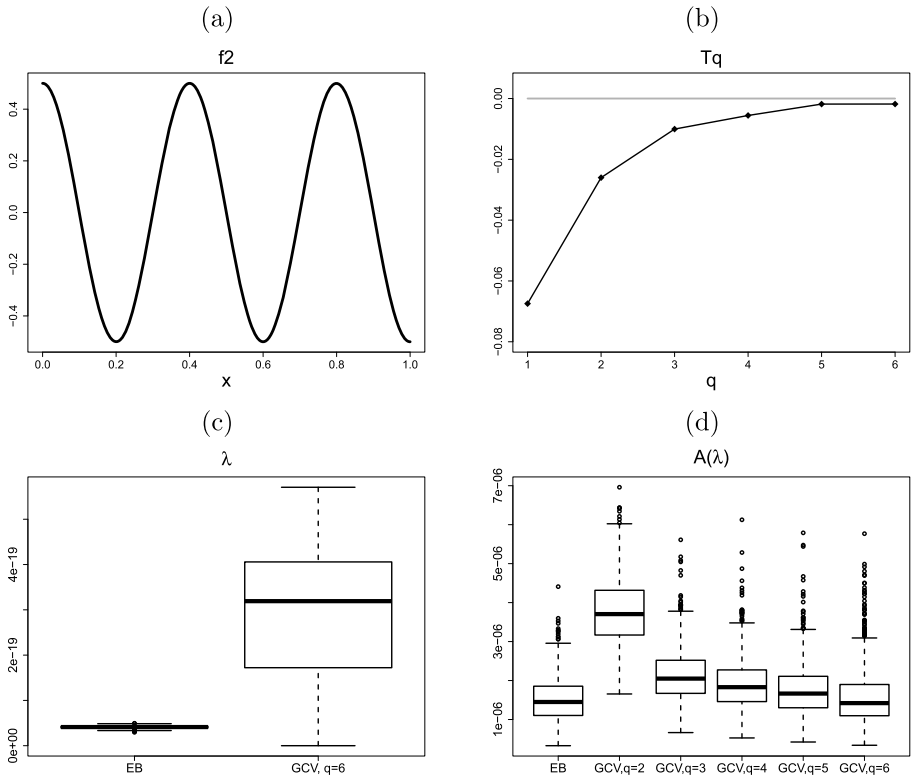

Figure 2: (a) Function f_2, (b) Criterion $T_q(\hat{\lambda}_q, q)$ for f_2, (c) Boxplots of $\hat{\lambda}_{\hat{q}}$ and $\hat{\lambda}_f$ obtained with GCV and $q = 6$, (d) Boxplots of $A(\hat{\lambda}_{\hat{q}})$ (EB) and $A(\hat{\lambda}_f)$.

the fact that in small samples it is difficult to distinguish between an exponential decay $\exp(-\pi i)$ and a decay with $(\pi i)^{-q}$, $q > 5$, $i = 1, \ldots, n$. Also, $\hat{\lambda}_f$ for $q = 2, \ldots, 6$ with generalised cross-validation have been calculated. It appears that $\hat{\lambda}_{\hat{q}}$ and $\hat{\lambda}_f$ with $q = 6$ have the same rate, but again, $\hat{\lambda}_f$ is much more variable. For these smoothing parameters we also observe, that the corresponding AMSE ratio is closest to one. In general, for this function f_2 the adaptive empirical Bayesian smoothing spline estimator again outperforms frequentist splines uniformly in q with the largest AMSE ratio of about 2.5 for $q = 2$.

Finally, we remark on the implementation of the procedure. It is well-known that the spline based basis with knots at observations becomes numerically unstable for higher qs. In fact, it seems impossible to get numerically stable Demmler–Reinsch basis for the natural spline space for $q > 3$ with usual approaches. Instead, we relied on an approximation based on the Demmler–Reinsch basis (28) for $\mathcal{W}_q$. The details of this approach will be reported elsewhere, but the implementation in R is available from the authors on request.

7 Conclusions

The selection of the order of smoothing splines in non-parametric regression is a topic mostly absent from the literature. The empirical Bayes method is shown to provide an adequate framework to produce data driven choices for this parameter. Although the dependence of the prior on the parameter q – which controls the order of the smoothing spline – is rather implicit, if the regression function has a well defined smoothness (determined with the help of so the so called polished-tail condition), then $\hat{q}$ is consistent and we identify the smallest Sobolev space containing the regression function. Hence, our adaptive empirical Bayesian smoothing spline estimator (which is the mean of the empirical Bayes posterior) adapts to the underlying smoothness of the signal.

High probability regions of the empirical posterior are shown to have good frequentist coverage properties. For a large class of functions the size of these regions is shown to adapt to the underlying smoothness of the signal, effectively quantifying the amount of uncertainty of the empirical Bayesian smoothing spline estimator. These results are used to show that frequentist smoothing splines are outperformed by empirical Bayesian smoothing splines.

Supplementary Material

Supplementary materials: Adaptive empirical Bayesian smoothing splines (DOI: 10.1214/16-BA997SUPP; .pdf).

References

Babenko, A. and Belitser, E. (2010). "Oracle convergence rate of posterior under projection prior and Bayesian model selection." *Mathematical Methods of Statistics*, 219–245. MR2742927. doi: http://dx.doi.org/10.3103/S1066530710030026. 220

Belitser, E. and Enikeeva, F. (2008). "Empirical Bayesian test of the smoothness." *Mathematical Methods of Statistics*, 17(1): 1–18. MR2400361. doi: http://dx.doi.org/10.3103/S1066530708010018. 226

Belitser, E. and Ghosal, S. (2003). "Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution." *Annals of Statistics*, 31(2): 536–559. MR1983541. doi: http://dx.doi.org/10.1214/aos/1051027880. 220

Belitser, E. and Levit, B. (2003). "On the empirical Bayes approach to adaptive filtering." *Mathematical Methods of Statistics*, 12(2): 131–154. MR2025355. 220

de Jonge, R. and van Zanten, J. (2010). "Adaptive nonparametric Bayesian inference using location-scale mixture priors." *Annals of Statistics*, 3300–3320. MR2766853. doi: http://dx.doi.org/10.1214/10-AOS811. 220

de Jonge, R. and van Zanten, J. H. (2012). "Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors." *Electronic Journal of Statistics*, 6: 1984–2001. MR3020254. doi: http://dx.doi.org/10.1214/12-EJS735. 220

Genovese, C. and Wasserman, L. (2008). "Adaptive confidence bands." *Annals of Statistics*, 36(2): 875–905. MR2396818. doi: http://dx.doi.org/10.1214/07-AOS500. 230

Ghosal, S., Ghosh, J., and van der Vaart, A. (2000). "Convergence rates of posterior distributions." *Annals of Statistics*, 28: 500–531. MR1790007. doi: http://dx.doi.org/10.1214/aos/1016218228. 220

Ghosal, S., Lember, J., and van der Vaart, A. (2008). "Nonparametric Bayesian model selection and averaging." *Electronic Journal of Statistics*, 2: 63–89. MR2386086. doi: http://dx.doi.org/10.1214/07-EJS090. 220

Ghosal, S. and van der Vaart, A. (2007). "Convergence rates of posterior distributions for non-i.i.d. observations." *Annals of Statistics*, 35(1): 192–223. MR2332274. doi: http://dx.doi.org/10.1214/009053606000001172. 220

Giné, E. and Nickl, R. (2010). "Confidence bands in density estimation." *Annals of Statistics*, 38(2): 1122–1170. MR2604707. doi: http://dx.doi.org/10.1214/09-AOS738. 226, 228, 230

Johnstone, I. and Silverman, B. (2005). "Empirical Bayes selection of wavelet thresholds." *Annals of Statistics*, 1700–1752. MR2166560. doi: http://dx.doi.org/10.1214/009053605000000345. 220

Kimeldorf, G. and Wahba, G. (1970). "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines." *Annals of Mathematical Statistics*, 41: 495–502. MR0254999. 220

Knapik, B., van der Vaart, A., and van Zanten, J. (2011). "Bayesian inverse problems with Gaussian priors." *Annals of Statistics*, 2626–2657. MR2906881. doi: http://dx.doi.org/10.1214/11-AOS920. 220

Knapik, B. T., Szabó, B., van der Vaart, A., and van Zanten, J. (2013). "Bayes procedures for adaptive inference in inverse problems for the white noise model." *Probability Theory and Related Fields*, 1–43. 220

Kohn, R. and Ansley, C. F. (1987). "A new algorithm for spline smoothing based on smoothing a stochastic process." *SIAM Journal on Scientific and Statistical Computing*, 8: 33–48. MR0873922. doi: http://dx.doi.org/10.1137/0908004. 220

Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press. MR2038227. doi: http://dx.doi.org/10.1017/CBO9780511550683. 222

Krivobokova, T. (2013). "Smoothing parameter selection in two frameworks for penalized splines." *Journal of the Royal Statistical Society. Series B*, 75: 725–741. MR3091656. doi: http://dx.doi.org/10.1111/rssb.12010. 220, 225, 226, 230

Low, M. (1997). "On nonparametric confidence intervals." *Annals of Statistics*, 25(6): 2547–2554. MR1604412. doi: http://dx.doi.org/10.1214/aos/1030741084. 226, 229

Lukas, M. (1998). "Assessing regularised solutions." *Journal of the Australian Mathematical Society, Series B*, 30: 24–42. MR0943480. doi: http://dx.doi.org/10.1017/S0334270000006019. 222

McAuliffe, J., Blei, D., and Jordan, M. (2006). "Nonparametric empirical Bayes for the Dirichlet process mixture model." *Statistics and Computing*, 5–14. MR2224185. doi: http://dx.doi.org/10.1007/s11222-006-5196-2. 220

Picard, D. and Tribouley, K. (2000). "Adaptive confidence interval for pointwise curve estimation." *Annals of Statistics*, 28(1): 298–335. MR1762913. doi: http://dx.doi.org/10.1214/aos/1016120374. 226

Robbins, H. (1955). "An empirical Bayes approach to Statistics." In *Proc. 3rd Berkeley Symp. on Math. Statist. and Prob.*, volume 1, 157–164. Berkeley: Univ. of California Press. MR0084919. 223

Shen, W. and Ghosal, S. (2015). "Adaptive Bayesian procedures using random series priors." *Scandinavian Journal of Statistics*, 43: 1194–1213. doi: http://dx.doi.org/10.1111/sjos.12159. 220

Shen, X. and Wasserman, L. (2001). "Rates of convergence of posterior distributions." *Annals of Statistics*, 687–714. MR1865337. doi: http://dx.doi.org/10.1214/aos/1009210686. 220

Serra, P. and Krivobokova, T. (2016). "Supplementary materials: adaptive empirical Bayesian smoothing splines"." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA997SUPP. 221

Speckman, P. (1985). "Spline smoothing and optimal rates of convergence in nonparametric regression models." *Annals of Statistics*, 13: 970–983. MR0803752. doi: http://dx.doi.org/10.1214/aos/1176349650. 222

Speckman, P. and Sun, D. (2003). "Fully Bayesian spline smoothing and intrinsic autoregressive priors." *Biometrika*, 90(2): 289–302. MR1986647. doi: http://dx.doi.org/10.1093/biomet/90.2.289. 220, 222

Szabó, B., van der Vaart, A., and van Zanten, H. (2014). "Frequentist coverage of adaptive nonparametric Bayesian credible sets." *Annals of Statistics*, 43(4): 1391–1428. MR3357861. doi: http://dx.doi.org/10.1214/14-AOS1270. 220, 227, 228, 230

Unser, M. and Blu, T. (2000). "Fractional splines and wavelets." *SIAM Review*, 42(1): 43–67. MR1738098. doi: http://dx.doi.org/10.1137/S0036144598349435. 224

van der Vaart, A. and van Zanten, J. (2008). "Rates of contraction of posterior distributions based on Gaussian process priors." *Annals of Statistics*, 36(3): 1435–1463. MR2418663. doi: http://dx.doi.org/10.1214/009053607000000613. 220

van der Vaart, A. W. and van Zanten, J. H. (2009). "Adaptive Bayesian Estimation Using A Gaussian Random Field With Inverse Gamma Bandwidth." *Annals of Statistics*, 37(5B): 2655–2675. MR2541442. doi: http://dx.doi.org/10.1214/08-AOS678. 220

Wahba, G. (1985). "A comparison of GCV and GML for choosing the smoothing parameter in the generalised spline smoothing problem." *The Annals of Statistics*, 1378–1402. MR0811498. doi: http://dx.doi.org/10.1214/aos/1176349743. 224

Wahba, G. (1990). *Spline models for observational data*, volume 59. SIAM. MR1045442. doi: http://dx.doi.org/10.1137/1.9781611970128. 219, 221, 228

Yue, Y. R., Simpson, D., Lindgren, F., and Rue, H. (2014). "Bayesian Adaptive Smoothing Splines Using Stochastic Differential Equations." *Bayesian Analysis*, 397–424. MR3217001. doi: http://dx.doi.org/10.1214/13-BA866. 220

Zhang, C.-H. (2005). "General empirical Bayes wavelet methods and exactly adaptive minimax estimation." *Annals of Statistics*, 54–100. MR2157796. doi: http://dx.doi.org/10.1214/009053604000000995. 220

Acknowledgments

We would like to thank the editor, the associate editor and both referees for the extremely useful remarks, that helped to improve the paper substantially. We express also our gratitude to Francisco Rosales for his implementation of the Demmler–Reinsch basis and helpful discussions.