Bayesian Regression Using an Approximated Solution to the Penalized Least Squares Minimization Problem

Eduardo Antonio Trujillo-Rivera*

Abstract

In the context of semi-parametric regression with multiple covariates, it is known that the solution to the penalized least squares minimization problem can be interpreted as the mean of a Gaussian process arising from the posterior distribution of an empirical Bayesian approach. Using the Representer Theorem, we propose a Bayesian regression model with normal distributed errors at the response level and prove that conditionally to the variance, it defines the same Gaussian Process. A Gaussian process which approximates the solution to the penalized least squares minimization problem using its mean function is described. We study using simulations, the performance of the means of the posterior predictive as point estimates for the regression function and the empirical coverage of the pointwise credible intervals from the the posterior predictive distributions of the approximated Gaussian Process.

Key Words: Bayesian prediction, regularization, regression, Gaussian process, credible intervals.

1. Introduction

Let $\mathbb{X}$ be a non-empty set and $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{X} \times \mathbb{R}$ denote a sample of regressors $\mathbf{x}_i$ and response variables y_i. Let $\mathcal{H} \subset \{\eta : \mathbb{X} \to \mathbb{R}\}$ be a class of response curves for describing the response mean as a function of the regressors; in particular $\mathcal{H}$ will be a Reproducing Kernel Hilbert Space (*RKHS*) over $\mathbb{R}$. Let be J the square of the norm in $\mathcal{H}$. The penalized least squares minimization problem is formulated as finding the function $\eta \in \mathcal{H}$ that solves

$$\arg\min_{\eta \in \mathcal{H}} \sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i))^2 + n\lambda J(\eta) \tag{1}$$

with provided penalty parameter $\lambda > 0$. The solution to this problem has been widely studied by different authors using theory of linear operators in Hilbert spaces; see [Weidmann, 1980], [Akhiezer and Glazman, 1981a], [Akhiezer and Glazman, 1981b] from the perspective of *RKHS's* , and [Aronszajn, 1950], [Kimeldorf and Wahba, 1971], [Duchon, 1977], [Wahba and Craven, 1978], [Meinguet, 1979], [Wahba and Wendelberger, 1980], [Wahba, 1985], [Wahba, 1987], [Chen, 1993], [Wood, 2003] and [Gu, 2013].

Under some regularity conditions on J, a set of theorems jointly called the *Representer Theorems* [Schölkopf et al., 2001] states that a unique solution to (1) exists in a finite dimension subspace of $\mathcal{H}$ furthermore, a basis of functions in $\mathcal{H}$ is completely described by the representer Theorem. Hence, for given smoothing parameters, the general approach to solve (1) is to write the solution to (1) as a finite linear combination of the known basis set of functions that depend entirely on the sampling points $\{\mathbf{x}_i\}_{i=1}^n$ and on the associated reproducing kernel of $\mathcal{H}$. Using this finite representation, one can transform the functional minimization problem (1) into a simpler one: to solve a real linear finite system of equations. This system may still be very large and computationally intensive, but at least might be approximately solved and the rate of approximation can be studied.

The penalty term J reflects, up to the value of $n\lambda$, the smoothness desired for the solution to (1). In practice we must choose J in a case by case basis. In general,

*Biomedical Informatics, School of Medicine and Health Sciences, The George Washington University. 800 22nd St NW, Suite 8390. Washington, DC 20052

for a smooth solution in the sense of derivatives we can choose J to be in the family of the thin plate splines [Duchon, 1977], [Wahba and Wendelberger, 1980], [Wood, 2003], [Ruppert et al., 2003], [Gu, 2013] or we can construct new penalty terms using a tensor product of Hilbert spaces which leads to solutions to (1) as tensor product smoothing splines [Barry et al., 1986],[Wahba, 1987], [Gu and Wahba, 1993a], [Gu and Wahba, 1993b], [Chen, 1993] , [Barry et al., 1986], [Gu, 2013].

The choice of an appropriate smoothing parameters λ and other hidden smoothing terms inside the penalty term J is challenging and has been approached by defining score functions that need to be minimized. Examples of these approached are the unbiased parameter estimation that minimizes some loss function [Mallows, 1973] (UERL), cross validation and generalized cross validation [Wahba and Craven, 1978] (GCV), and scores obtained in the context of maximum likelihood estimation in certain models [Wecker and Ansley, 1983], [Wahba, 1985], [Li, 1986] (RML). When there are no smoothing parameters in J, λ may be interpreted as the ratio of two variance components in a linear mixed effects model [Ruppert et al., 2003] equivalent to (1); such linear mixed model is obtained from a perspective of the best linear unbiased prediction (BLUP) [Henderson, 1973], [Robinson, 1991].

For computational reasons, when ill-posed problems arise while inverting large matrices and expensive required computations, approximate solution to (1) has been proposed [Gu and Kim, 2002], [Wood, 2003] by assuming the solution is in an even smaller subspace contained in the subspace stated by the Representer Theorems. This solution has the same asymptotic convergence rate as the exact solution [Kim and Gu, 2004]. Some loss in the accuracy of point estimates of the function η is the price we pay for a fast algorithm that can be used in practice.

We review the needed theory to obtain the exact and the approximate solution to the penalized least square minimization problem (1) and describe the known interpretation of this solution as the mean of the posterior distribution arising in the context of an empirical Bayesian approach [Kim and Gu, 2004], [Gu, 2013] . The probability model in this approach has a Gaussian process as prior on the target regression function with covariance structure depending on the reproducing kernel of an associated reproducing kernel Hilbert space.

For our main Theoretical result, we propose a Bayesian linear regression model with multivariate normal priors on the coefficients and covariance structure depending on the reproducing kernel. We prove that with this model, the full conditional posterior estimators induce the same Gaussian process as in the previous formulation with the Gaussian process prior. In particular, the function defined by the mean the Gaussian process from the full conditional posterior distribution solves the minimization problem (1). Our approach has an advantage over its predecessor; in order to predict the value of the target function on any domain and to produce credible intervals for the predictions, we only need to evaluate the known basis functions using estimators of the coefficients. In contrast, when using the first Bayes formulation with Gaussian process prior, we first need to fix the points where the Gaussian process is to be estimated but subsequent evaluations of the process is done externally, ex., with interpolations.

We evaluate the performance of our method using simulation. We use our Bayesian model applied to existing methods proposed for the estimation in non-parametric regression in the frequentist setting. We include thin plate splines, a linear mixed model interpretation of thin plate splines, and tensor product splines with marginal thin plate splines. In all cases, we use the approximate solution to the optimization least squares problem.

We compare the various approaches of regression estimators by focusing on the frequentist properties of the mean of the process defined with posterior predictive distributions. In addition to point estimates, we study the empirical coverage of the pointwise credible

intervals for the regression function. In particular, we compute the average coverage rates of the credible intervals for all methods, where the average is taken over the prediction points.

Finally, we apply the proposed model to a real data set from a longitudinal study in African-American women of bone health and biomarkers. All the theoretical auxiliary results are given in the Appendix.

2. Background

Let $\mathbb{X}$, $\mathcal{H} \subset \{\eta : \mathbb{X} \to \mathbb{R}\}$ spaces, as before, and a given training set $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{X}\times\mathbb{R}$. The *RKHS* $\mathcal{H}$ can be written as tensor sum decomposition $\mathcal{H} = \bigoplus_{i=0}^p \mathcal{H}_i$ with $\mathcal{H}_i \subset \mathcal{H}$ closed sub-spaces which may be independent and have inner products $(\phi_i, \psi_i)_i$ with respective reproducing kernels R_i, where $\phi_i, \psi \in \mathcal{H}_i$ are the unique projections of $\phi, \psi \in \mathcal{H}$ onto $\mathcal{H}_i$ [Weidmann, 1980, Theorem 3.2]. For convenience, we write $(\phi_i, \psi_i)_i = (\phi, \psi)_i$. Lets write the square norm J for any $\eta \in \mathcal{H}$ as

$$J(\eta) = J(\eta, \eta) = \sum_{i=1}^{p} \theta_i^{-1} (\eta, \eta)_i , \tag{2}$$

The parameter λ controls the trade-off between smoothness as measured by J, and the discrepancy between the fitted function and the training set as measured by the quadratic loss function $\sum_{i=1}^n (y_i - \eta(\mathbf{x}_i))^2$. The positive tunning parameters $\{\theta_i\}_{i=1}^p$ allow for re-scaling of the metrics $(\cdot,\cdot)_i$. The smoothing parameters λ and $\{\theta_i\}_{i=1}^p$ need to be selected.

The bi-linear form $J(f,g) = \sum_{i=1}^p \theta_i^{-1} (f,g)_i$ is assumed to be an inner product in $\bigoplus_{i=1}^p \mathcal{H}_i = \mathcal{H} \ominus \mathcal{H}_0$ which has a reproducing kernel

$$R_J = \sum_{i=1}^{p} \theta_i R_i.$$

R_J is the reproducing kernel because it has the reproducing property, for $\mathbf{x} \in \mathbb{X}$:

$$\begin{aligned} J(R_J(\mathbf{x},\cdot), f) &= \sum_{i=1}^{p} \theta^{-1} \left(\left[\sum_{j=1}^{p} \theta_j R_j(\mathbf{x},\cdot) \right]_i , f_i \right)_i = \sum_{i=1}^{p} \theta^{-1} (\theta_i R_i(\mathbf{x},\cdot), f_i)_i \\ &= \sum_{i=1}^{p} (R_i(\mathbf{x},\cdot), f_i)_i = \sum_{i=1}^{p} f_i(\mathbf{x}) = f(\mathbf{x}); \end{aligned}$$

the reproducing kernel is unique[Aronszajn, 1950] , [Gu, 2013]. For given λ (and θ_i's) if the null space $\mathcal{N}_J = \mathcal{H}_0$ of J have finite dimension l then a unique solution to (1) exists [Gu, 2013, Theorem 2.5] and the solution has the form [Kimeldorf and Wahba, 1971], [Wahba and Wendelberger, 1980], [Schölkopf et al., 2001] and [Gu, 2013]:

$$\eta(\mathbf{x}) = \sum_{i=1}^{l} d_i \psi_i(\mathbf{x}) + \sum_{i=1}^{n} c_i R_J(\mathbf{x}_i, \mathbf{x}) , \tag{3}$$

where $\{\psi_\nu\}_{\nu=1}^l$ is a basis of the space $\mathcal{N}_J = \mathcal{H}_0$. An usual interpretation [Gu, 2013, Theorem 2.12] to the solution of (1) is that for a given smoothing parameter $\lambda > 0$, minimizing (1) in $\mathcal{H}$ is equivalent to finding the function $\hat{\eta} \in H$ that best fits the training data in the sense of minimizing the quadratic loss function $\mathsf{L}(\eta) := n^{-1} \sum_{i=1}^n (y_i - \eta(\mathbf{x}_i))^2$, where $\hat{\eta}$ is subject to the constraint that $\hat{\eta}$ is in the ball of radius ρ ($J(\hat{\eta}) \leq \rho^2$); ρ depending on both λ and the Gâteaux derivative of L.

Nevertheless, instead of using the solution (3), we use an approximate solution [Gu and Kim, 2002] by solving (1) in the space

$$\mathcal{H}^\star = \mathcal{N}_J \bigoplus span\{R_J(\mathbf{z}_i, \cdot), i = 1, ..., k\}. \tag{4}$$

Analogously to (3), any functions $\eta \in \mathcal{H}^\star$ can be written as

$$\eta(\mathbf{x}) = \sum_{i=1}^{l} d_i \psi_i(\mathbf{x}) + \sum_{i=1}^{k} c_i R_J(\mathbf{z}_i, \mathbf{x}) \tag{5}$$

with $\{\psi_\nu\}_{\nu=1}^{l}$ being a basis of $\mathcal{N}_J$ and

$$\mathbf{d} = (d_1 \cdots d_l)^\intercal, d_i \in \mathbb{R}$$
$$\mathbf{c} = (c_1 \cdots c_k)^\intercal, c_i \in \mathbb{R}$$

Plugging in (5) to functional inside (1) and using the reproducing kernel property, the next expression is obtained

$$(\mathbf{y} - S\mathbf{d} - R\mathbf{c})^\intercal(\mathbf{y} - S\mathbf{d} - R\mathbf{c}) + n\lambda\mathbf{c}^\intercal Q\mathbf{c}, \tag{6}$$

where $R \in \mathcal{M}_{n\times k}(\mathbb{R})$ with $(i,j)th$ entry $R_J(\mathbf{x}_i, \mathbf{z}_j)$, $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ with $(i,j)th$ entry $R_J(\mathbf{z}_i, \mathbf{z}_j)$ and $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ with $(i,j)th$ entry $\psi_j(x_i)$. Finally, taking first derivative with respect to $\mathbf{d}, \mathbf{c}$ the solutions to next system of linear equations provide candidates for the solution to (1)

$$\begin{pmatrix} S^\intercal S & S^\intercal R \\ R^\intercal S & R^\intercal R + n\lambda Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S^\intercal \mathbf{y} \\ R^\intercal \mathbf{y} \end{pmatrix}. \tag{7}$$

Observe that $span\{R_J(\mathbf{z}_i)\}_{i=1}^{k}$ is a closed subspace of $\mathcal{N}_J \ominus \mathcal{H}$ and furthermore it is a Hilbert space with the same inner product J and same reproducing kernel R_1 as $\mathcal{N}_J \ominus \mathcal{H}$. The functional $\sum_{i=1}^{n}(y_i - \eta(\mathbf{x}_i))^2$ is continuous and convex in $\mathcal{H}^\star$ and when S is of full column rank, the convexity is strict in $\mathcal{N}_J$ and the functional then has a minimizer in this space. A solution to (1) exists in $\mathcal{H}^\star$ as long as $\sum_{i=1}^{n}(y_i - \eta(\mathbf{x}_i))^2$ has a minimizer in $\mathcal{N}_J$ [Gu and Qiu, 1993]. The functional inside (1) is strictly convex in $\mathcal{H}^\star$ when S is of full column rank, and by Proposition 7, (1) has a unique solution in $\mathcal{H}^\star$. Then if S is full column rank, a solution to (7) will lead to the unique solution to (1) in $\mathcal{H}^\star$ through (5). Even when there were multiple solutions to (7), they yield the same $\eta \in \mathcal{H}^\star$. In practice, a solution to (7) is chosen as described in our main theoretical results stated in the next section.

3. Main Theoretical Results

We are ready to state the first of our main theoretical results in the next proposition.

Proposition 1 (A First Bayesian Model)
In previous context, let $\{\psi_i\}_{i=1}^{l}$ *be a basis of* $\mathcal{N}_J$*, the null space of the norm of* $\mathcal{H}$*. Consider* $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \subset \mathbb{X} \times \mathbb{R}$ *a training set, let* $\mathbf{Z} := \{\mathbf{z}_i\}_{i=1}^{k} \subset \{\mathbf{x}_i\}_{i=1}^{n} =: \mathbf{X}$, $\mathbf{d} := (d_1\, d_2 \cdots d_l)^\intercal$, $\mathbf{c} := (c_1\, c_2 \cdots c_k)^\intercal$. *Consider the model*

$$y_i = \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i) + \epsilon_i,$$
$$\eta_{\binom{\mathbf{d}}{\mathbf{c}}} = \sum_{i=1}^{l} d_i \psi_i + \sum_{i=1}^{k} c_i R_J(\mathbf{z}_i, \cdot)$$

$$\epsilon_i \overset{iid}{\sim} N_1(0, \sigma^2).$$

Let $\lambda > 0$, *and* $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ *with entries* $Q_{i,j} = R_J(\mathbf{z}_i, \mathbf{z}_j)$, $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ *with* $S_{i,j} = \psi_j(\mathbf{x}_i)$ *full column rank,* $R \in \mathcal{M}_{n\times k}(\mathbb{R})$, $R_{i,j} = R_J(\mathbf{x}_i, \mathbf{z}_j)$, $M = RQ^{+}R^{\intercal} + n\lambda I_n$ *and define* $b = \frac{\sigma^2}{n\lambda}$. *Consider the priors*

$$\begin{aligned} d_i &\overset{iid}{\sim} 1, \\ \mathbf{c}|\sigma^2 &\sim N_l\left(\mathbf{0}, b\,Q^{+}\right), \\ \sigma^2 &\sim Inv - Gamma(A_\epsilon, B_\epsilon), \\ \mathbf{d} &\perp \mathbf{c}, \\ \mathbf{d} &\perp \sigma^2 \\ \binom{\mathbf{d}}{\mathbf{c}} &\perp (\epsilon_1 \cdots \epsilon_n)^{\intercal}, \\ \sigma^2 &\perp \epsilon_i,\ i \in \{1, \cdots, n\}. \end{aligned}$$

Then the posterior of the parameters exists and the full conditional posteriors are

- $\binom{\mathbf{d}}{\mathbf{c}}|\mathbf{y},\ \sigma^2,\ b,\ \mathbf{X} \sim N_{l+k}\left(\mu_{\mathbf{dc}}, b\mathbf{\Sigma}_{\mathbf{dc}}\right)$, *where*

$$\mu_{\mathbf{dc}} = \begin{pmatrix} (S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1} \\ Q^{+}R^{\intercal}M^{-1}(I - S(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}) \end{pmatrix}\mathbf{y} \tag{8}$$

$$\mathbf{\Sigma}_{\mathbf{dc}} = \begin{pmatrix} (S^{\intercal}M^{-1}S)^{-1} & -(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}RQ^{+} \\ -Q^{+}R^{\intercal}M^{-1}S(S^{\intercal}M^{-1}S)^{-1} & Q^{+} - Q^{+}R^{\intercal}\{M^{-1} - M^{-1}S(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}\}RQ^{+} \end{pmatrix} \tag{9}$$

- $\sigma^2|\mathbf{y},\ \binom{\mathbf{d}}{\mathbf{c}},\ \mathbf{X} \sim Inv - Gamma\left(A_\epsilon + \frac{1}{2}n,\ \left[B_\epsilon^{-1} + \frac{1}{2}\sum_{i=1}^{n}(y_i - \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i))^2\right]^{-1}\right)$.

Proof.
Denote $\mathbf{\Theta} = \left(\mathbf{d}^{\intercal}\ \mathbf{c}^{\intercal}\ \sigma^2\right)^{\intercal}$. The posterior density of $\mathbf{\Theta}$ can be expressed as

$$[\mathbf{\Theta}|\mathbf{y},\ \mathbf{X}] \propto \left[\mathbf{y}|\binom{\mathbf{d}}{\mathbf{c}},\ \sigma^2,\ \mathbf{X}\right] \times \left[\binom{\mathbf{d}}{\mathbf{c}}\middle|\,\sigma^2,\ \mathbf{X}\right] \times [\sigma^2],$$

thus the posterior is proper if the right hand side of the previous expression is proper. The distribution $\left[\binom{\mathbf{d}}{\mathbf{c}}\middle|\,\mathbf{y}, \sigma^2,\ \mathbf{X}\right] \propto \left[\mathbf{y}|\binom{\mathbf{d}}{\mathbf{c}},\ \sigma^2,\ \mathbf{X}\right] \times \left[\binom{\mathbf{d}}{\mathbf{c}}\middle|\,\sigma^2,\ \mathbf{X}\right]$ may not integrate to 1, but the rest of the distributions are proper. This distribution can be shown to be proper by considering first the model with proper prior $d_i \overset{iid}{\sim} N(0, \tau^2)$ with the rest of the priors kept the same. For $\tau^2 \to \infty$ we can then prove that $\left[\binom{\mathbf{d}_\tau}{\mathbf{c}}\middle|\,\mathbf{y}, \sigma^2\,\mathbf{X}\right]$ converges in distribution. This is shown in Proposition 12.

To compute the full conditional posterior $\left[\sigma^2\middle|\,\binom{\mathbf{d}}{\mathbf{c}}, \mathbf{X}, \mathbf{y}\right]$ we recall that

$$[\mathbf{\Theta}|\mathbf{y},\ \mathbf{X}] \propto \left[\,\mathbf{y}\,\middle|\binom{\mathbf{d}}{\mathbf{c}}, \sigma^2,\ \mathbf{X}\right]\left[\binom{\mathbf{d}}{\mathbf{c}}\middle|\,\sigma^2,\ b,\ \mathbf{X}\right]\left[\sigma^2|\mathbf{X}\right],$$

and obtain

$$\left[\sigma^2\middle|\,\binom{\mathbf{d}}{\mathbf{c}}, \mathbf{X}, \mathbf{y}\right] \propto \left[\,\mathbf{y}\,\middle|\binom{\mathbf{d}}{\mathbf{c}}, \sigma^2,\ \mathbf{X}\right]\left[\binom{\mathbf{d}}{\mathbf{c}}\middle|\,\sigma^2,\ b,\ \mathbf{X}\right]\left[\sigma^2|\mathbf{X}\right]. \tag{10}$$

But $\left[\binom{\mathbf{d}}{\mathbf{c}}\middle|\,\sigma^2,\ b,\ \mathbf{X}\right]$ depends on σ^2 only through the expression σ^2/b and

$$\frac{\sigma^2}{b} = \frac{\sigma^2}{\sigma^2/n\lambda} = n\lambda.$$

Therefore, we have $\left[\binom{\mathbf{d}}{\mathbf{c}}\middle|\,\sigma^2,\ b,\ \mathbf{X}\right] = \left[\binom{\mathbf{d}}{\mathbf{c}}\middle|\,\lambda,\ \mathbf{X}\right]$ which indicates that the conditional distribution is independent of σ^2. By (10), we have

$$\left[\sigma^2\middle|\,\binom{\mathbf{d}}{\mathbf{c}}, \mathbf{X}, \mathbf{y}\right] \propto \left[\,\mathbf{y}\,\middle|\binom{\mathbf{d}}{\mathbf{c}}, \sigma^2,\ \mathbf{X}\right]\left[\sigma^2|\mathbf{X}\right]$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}(\mathbf{x}_i)\right)^2-\frac{1}{2B_\epsilon\sigma^2}\right)\times\sigma^{-2(n/2+A_\epsilon+1)}$$

$$\propto Inv-Gamma\left(A_\epsilon+\frac{1}{2}n,\left[B_\epsilon^{-1}+\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}(\mathbf{x}_i)\right)^2\right]^{-1}\right).$$

■

Kim's *et.al.* results [Kim and Gu, 2004], is related to our own via the point estimates of η in any $\chi\in\mathbb{R}^d$. We claim that the full conditional posterior distribution of η in Proposition 1, as a process, is the same as the one proposed by *Kim, et.al.* We state and prove this result in Proposition 2.

Proposition 2 (Equivalence Bayesian Models)
In the context of Proposition 1, $\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}$ *is a Gaussian process with mean and covariance functions*

$$\begin{aligned}
\mathbb{E}\left[\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x})\right] =&\boldsymbol{\Psi}^\intercal(\mathbf{x})\hat{\mathbf{d}}+\boldsymbol{\Xi}^\intercal(\mathbf{x})\hat{\mathbf{c}},\\
b^{-1}Cov\left[\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x}),\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{y})\right] =&\boldsymbol{\Xi}(\mathbf{x})^\intercal Q^+\boldsymbol{\Xi}(\mathbf{y})+\boldsymbol{\Psi}(\mathbf{x})^\intercal\left(S^\intercal M^{-1}S^\intercal\right)^{-1}\boldsymbol{\Psi}(\mathbf{x})\\
&-\left[\boldsymbol{\Psi}(\mathbf{x})^\intercal\tilde{\mathbf{d}}(\mathbf{y})+\boldsymbol{\Psi}(\mathbf{y})^\intercal\tilde{\mathbf{d}}(\mathbf{x})\right]-\boldsymbol{\Xi}(\mathbf{x})^\intercal\tilde{\mathbf{c}}(\mathbf{y})\\
b^{-1}Var\left[\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x})\right] =&\boldsymbol{\Xi}(\mathbf{x})^\intercal Q^+\boldsymbol{\Xi}(\mathbf{x})+\boldsymbol{\Psi}^\intercal(\mathbf{x})\left(S^\intercal M^{-1}S\right)^{-1}\boldsymbol{\Psi}(\mathbf{x})\\
&-2\boldsymbol{\Psi}(\mathbf{x})^\intercal\tilde{\mathbf{d}}(\mathbf{x})-\boldsymbol{\Xi}(\mathbf{x})^\intercal\tilde{\mathbf{c}}(\mathbf{x}),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\Psi}(\mathbf{x}) &= (\psi_1(\mathbf{x})\cdots\psi_l(\mathbf{x}))^\intercal\\
\boldsymbol{\Xi}(\mathbf{x}) &= (R_J(\mathbf{z}_1,\mathbf{x})\cdots R_J(\mathbf{z}_k,\mathbf{x}))^\intercal\\
\hat{\mathbf{c}} &= Q^+R^\intercal\left(M^{-1}-M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\mathbf{y}, \qquad (11)\\
\hat{\mathbf{d}} &= \left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y}, \qquad (12)\\
\tilde{\mathbf{c}}(\mathbf{x}) &= Q^+R^\intercal\left(M^{-1}-M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)RQ^+\boldsymbol{\Xi}(\mathbf{x}),\\
\tilde{\mathbf{d}}(\mathbf{x}) &= \left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}RQ^+\boldsymbol{\Xi}(\mathbf{x}).
\end{aligned}$$

Proof. The proof is direct using Proposition 13, a result with a slightly different notation. ■

We have shown that the full conditional posterior distribution of the regression function η evaluated at any $\chi\in\mathbb{R}^d$, conditional on σ and the smoothing parameters λ, $\{\theta_i\}_{i=1}^p$, induce the same Gaussian process as [Kim and Gu, 2004] evaluated at any sampling point $\mathbf{x}_i$. It follows that for any $\chi\in\mathbb{R}^d$, the full conditional posterior predictive distribution generate the same processes. The advantage of our results lays in terms of storage, prediction, and that this model can be directly extended to a measurement error regression problem and to predict directly using the posterior predictive distribution in any $\chi\in\mathbb{R}^d$. In order to predict using the model from Proposition 1 we only need to save the values $\hat{\mathbf{d}}$ and $\hat{\mathbf{c}}$ from an MCMC sampler and then use them to evaluate the basis of functions $\{\psi_i\}_{i=1}^l$ and $\{R_J(\mathbf{z}_i,\cdot)\}_{i=1}^k$ to simulate from the marginal posterior predictive distribution $[\eta(\chi)|\mathbf{y},\mathbf{X},\lambda]$ for any new χ.

For the sake of completing the arguments we now show that $\mu_{\mathbf{dc}}$ in (8) satisfy equations (7). Therefore the mean of the Gaussian process $\eta(\chi)$ described by Proposition 1 and Proposition (2) as functions of $\mathbf{x} \in \mathbb{R}^d$ solve the regularized minimization problem (1) in the space $\mathcal{H}^\star$.

Proposition 3 (Full Conditional Posterior Mean as Smoothing Splines) *In the setting of Proposition 1, let S be full column rank $n \times l$ matrix. The mean of the full conditional posterior of $\begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix}$ described by (8), and the vector $\begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \end{pmatrix}$ from (11) – (12) satisfy equations (7) and thus the mean of the Gaussian process $\eta_{\begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}}$ in Proposition 1 are the unique solution to (1) in $\mathcal{H}^\star$.*

Proof.
The first expression of (8) satisfies equations (7) because:

$$\begin{aligned}
S^\intercal S\hat{\mathbf{d}} + S^\intercal R\hat{\mathbf{c}} - S^\intercal\mathbf{y} &= S^\intercal S\left[\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y}\right] \\
&\quad + S^\intercal R\left[Q^+R^\intercal\left(M^{-1} + M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\mathbf{y}\right] - S^\intercal\mathbf{y} \\
&= S^\intercal\left(RQ^+R^\intercal\right)M^{-1}\mathbf{y} + S^\intercal\left(I_n - RQ^+R^\intercal M^{-1}\right)S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y} - S^\intercal\mathbf{y} \\
&= S^\intercal\left(RQ^+R^\intercal M^{-1} - I_n\right)\mathbf{y} - S^\intercal\left(RQ^+R^\intercal M^{-1} - I_n\right)S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y} \\
&= S^\intercal\left(RQ^+R^\intercal M^{-1} - I_n\right)\left(I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\mathbf{y} \\
&= S^\intercal\left(-n\lambda M^{-1}\right)\left(I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\mathbf{y} \\
&= -\ n\lambda\left[S^\intercal M^{-1} - \left(S^\intercal M^{-1}S\right)\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= -\,n\lambda\left[S^\intercal M^{-1} - S^\intercal M^{-1}\right]\mathbf{y} \\
&= \mathbf{0}.
\end{aligned}$$

That the second expression of (8) satisfies equations (7) follows as:

$$\begin{aligned}
R^\intercal S\hat{\mathbf{d}} + \left(R^\intercal R + n\lambda Q\right)\hat{\mathbf{c}} - R^\intercal\mathbf{y} &= R^\intercal S\left[\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y}\right] - R^\intercal\mathbf{y} \\
&\quad + \left(R^\intercal R + n\lambda Q\right)\left[Q^+R^\intercal M^{-1}(I - S(S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1})\right] \\
&= R^\intercal S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y} - R^\intercal\mathbf{y} \\
&\quad + R^\intercal RQ^+R^\intercal M^{-1}\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&\quad + n\lambda QQ^+R^\intercal M^{-1}\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= R^\intercal S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y} - R^\intercal\mathbf{y} \\
&\quad + R^\intercal RQ^+R^\intercal M^{-1}\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&\quad + n\lambda R^\intercal M^{-1}\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \qquad (13) \\
&= \left[R^\intercal\left(RQ^+R^\intercal M^{-1}\right) + n\lambda R^\intercal M^{-1} - R^\intercal\right]\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= \left[R^\intercal\left(I_n - n\lambda M^{-1}\right) + n\lambda R^\intercal M^{-1} - R^\intercal\right]\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= \left[R^\intercal - n\lambda R^\intercal M^{-1} + n\lambda R^\intercal M^{-1} - R^\intercal\right]\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= \mathbf{0}.
\end{aligned}$$

Equality (13) is obtained using $QQ^+R^\intercal = R^\intercal$ which can be proven as follow. Let

$$\xi(\mathbf{x}) = \left(R_J(\mathbf{z}_1, \mathbf{x}) \cdots R_J(\mathbf{z}_k, \mathbf{x})\right)^\intercal.$$

Notice that by definition of the generalized inverse Q^+, we have $QQ^+Q = Q$, therefore QQ^+ is the projection matrix on the column space of Q; if we prove that $\xi(\mathbf{x})$ is in the column space of Q we would have proven $QQ^+\xi(\mathbf{x}) = \xi(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^l$, in particular we would have proven that $QQ^+R^\intercal = R^\intercal$. $\xi(\mathbf{x})$ is in the column space of Q by Proposition 6, and thus we have (13).

We have shown that $\hat{\mathbf{d}}$ and $\hat{\mathbf{c}}$ satisfy equation (7), but this by itself only shows that they are critical points of (6). There may be multiple critical points to (6) but all the critical points would produce a solution to (1) through expressions (5), along with the expected value in of $\mathbb{E}(\eta(\mathbf{x}))$, or $\mathbb{E}\left[\eta\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x})\right]$ from Proposition 1. Such solution would be in the space of functions $\{\eta = \sum_{i=1}^{l} d_i\psi_i + \sum_{i=1}^{k} R_J(\mathbf{z}_i,\cdot)\}$. That the solution is unique in $\mathcal{H}^\star$ is concluded by the representer Theorems which requires that S is full column rank. ■

Another interpretation of the deterministic function defined by the mean of the Gaussian process from the posterior predictive is that the minimizer function, in this case $\mathbb{E}\left[\eta\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x})\right]$ as function of $\mathbf{x} \in \mathbb{R}^d$, is the best interpolation as measured by the quadratic loss function $\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2$ subject to the constrain $J(f) \leq \rho(\lambda)$ for some non-decreasing function $\rho(\lambda) > 0$.

3.1 Models

In this section we describe the complete Bayes models. In every case we assume that we have a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ is available, and a set of knots $\{(\mathbf{z}_i)\}_{i=1}^k$, $k << n$ with similar distribution as $\{(\mathbf{x}_i)\}_{i=1}^n$. In the decomposition of the space $\mathcal{H}^\star$ (4), the Hilbert subspace $\mathcal{N}_J$ is of finite dimension with basis $\{\psi_i\}_{i=1}^l$ and orthonormal basis $\{\phi_i\}_{i=1}^l$. We have available a (semi) kernel $R_J : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ in the space $\mathcal{H}^*$ which we throughly describe in case by case.

3.1.1 Bayesian regression model using thin plate splines

The thin plate splines are a generalization of cubic splines to any dimensions and can be used to obtain a smooth estimate of a surface by data interpolation and smoothing. Define a semi-inner product [Wahba and Wendelberger, 1980] J in the space of functions $\mathbb{R}^d \to \mathbb{R}$ where it can be finitely computed as:

$$J(\eta,\zeta) := \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\cdots\alpha_d!} \int_{\mathbb{R}^d} \left(\frac{\partial^m \eta}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}\right)\left(\frac{\partial^m \zeta}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}\right) dx_1\cdots dx_d. \quad (14)$$

The objective is to estimate η by solving (1) for $\mathbb{X} = (-\infty,\infty)^d$ subject to the penalty $J(\eta) = J(\eta,\eta) = J(\eta)$ which can be written as

$$J(\eta) := \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\cdots\alpha_d!} \int_{\mathbb{R}^d} \left(\frac{\partial^m \eta}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}\right)^2 dx_1\cdots dx_d, \quad (15)$$

and $\mathcal{H} = \{\eta : \mathbb{X} \to \mathbb{R} \mid J(\eta) < \infty\}$. For η to be in $\mathcal{H}$ we have to be able to compute $J(\eta)$. The null space $N_J := \{\eta \in H : J(\eta) = 0\}$ of J consists of polynomials in d variables of order up to $m-1$. This space is of finite dimension $l = \binom{d+m-1}{d}$ [Rivera, 2017, Lemma 3]. In order for $[x] : \mathcal{H} \to \mathbb{R}$ and $x \in \mathbb{X}$ defined as $[x]\eta := \eta(x)$ to be continuous, we need that $2m > d$ in $\mathcal{H}$ [Duchon, 1977], [Meinguet, 1979], and thus $\mathcal{H}$ is a *RKHS* . In this context, J is a square semi norm [Rivera, 2017, Lemma 68] and hence $\mathcal{H} \ominus N_J$ is a *RKHS* .

In order to describe the reproducing kernel R_J needed in (3), (5) and in Proposition 1, first we present the radial basis functions introduced by [Wahba and Wendelberger, 1980].

Denote the euclidean norm as $\|\cdot\|$ and

$$E_{m,d}(r) = \begin{cases} \theta_{m,d} r^{2m-d} \log(r), & \text{d even, for } \theta_{m,d} = \frac{(-1)^{d/2+m+1}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!}, \\ \theta_{m,d} r^{2m-d}, & \text{d odd, for } \theta_{m,d} = \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} \end{cases} \tag{16}$$

so that for a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, a radial basis is $\{E_{m,d}\| \cdot -\mathbf{x}_i\|\}_{i=1}^n$.

Second, we need a reproducing kernel in the null space of J, $\mathcal{N}_J := \{\eta \in \mathcal{H} | J(\eta) = 0\}$ which is a complete linear space [Rivera, 2017, Theorem 58]. In order to describe the structure of *RKHS* for $\mathcal{N}_J$, we need an inner product: for some $N \in \mathbb{N}$, $\{u_i\}_{i=1}^N \subset \mathbb{R}^d$, $\{p_i\}_{i=1}^N \subset \mathbb{R}$ with $p_i > 0$ and $\sum_{i=1}^N p_i = 1$ define

$$(\eta, \zeta)_0 = \sum_{i=1}^N p_i \eta(u_i).\zeta(u_i). \tag{17}$$

Let $\{u_i\}_{i=1}^N$ and $\{p_i\}_{i=1}^N$ such that the matrix with $(i,j)th$ entry $(\psi_i, \psi_j)_0$ is non-singular, where $\{\psi_i\}_{i=1}^j$ a fixed basis of $\mathcal{N}_{J_m^d}$. These assumptions are sufficient for $\mathcal{N}_{J_m^d}$ to be a *RKHS* with (17) the inner product in $\mathcal{N}_{J_m^d}$ (Proposition 59 and Lemma 67 in [Rivera, 2017]).

In order to find and compute the reproducing kernel R_0 of $\mathcal{N}_{J_m^d}$ with inner product (17), we need an orthonormal basis $\{\phi_i\}_{i=1}^l \subset \mathcal{N}_{J_m^d}$ with $\phi_1(\mathbf{x}) = 1$. By the Gram-Schimdt normalization [Hoffman and Kunze, 1990, Golub and Van Loan, 2012], given $\{\psi_i\}_{i=1}^l$ a set of polynomials that span $\mathcal{N}_J$, we can transform them and find such orthonormal basis. Explicit expressions for and an example of orthonormal basis $\{\phi_i\}_{i=1}^l$ can be found in Proposition 62 [Rivera, 2017]. The reproducing kernel in $\mathcal{N}_J$ by Proposition 59 [Rivera, 2017] is then

$$R_0(x, y) = \sum_{i=1}^l \phi_i(x)\phi_i(y). \tag{18}$$

Given the inner product (17) (or the choices of $\{u_i\}_{i=1}^N$ and $\{p_i\}_{i=1}^N$), any orthonormal basis will lead to the same R_0 [Aronszajn, 1950], [Gu, 2013, Theorem 2.5], where it is stated that the reproducing kernel is unique provided it exists.

We now present the projection of $f \in \mathcal{H}$ onto $\mathcal{N}_J$ since we need it for the reproducing kernel of $\mathcal{H} \ominus \mathcal{N}_J$. Let $\eta \in \mathcal{H} = \mathcal{N}_{J_m^d} \bigoplus \left(\mathcal{H} \ominus \mathcal{N}_{J_m^d}\right)$ or equivalently $\eta = \eta_0 + \eta_1$ with $\eta_0 \in \mathcal{N}_J$ and $\eta_1 \in \mathcal{H} \ominus \mathcal{N}_{J_m^d}$ for unique η_0 and η_1 (existence and uniqueness is ensured by Theorem 3.2 [Weidmann, 1980]). The projection of η onto $\mathcal{N}_J$ is by definition $P\eta = \eta_0$. By Proposition 60 [Rivera, 2017], the projection of $\eta \in \mathcal{H}$ onto is $\mathcal{N}_J$ is

$$(Pf)(\mathbf{x}) = \sum_{\nu=1}^l (f, \phi_\nu)_0 \phi_\nu(\mathbf{x}). \tag{19}$$

Define now the bi-linear form R_1 as

$$R_1(\mathbf{x}, \mathbf{y}) = \left(I - P_{(\mathbf{x})}\right)\left(I - P_{(\mathbf{y})}\right) E\left(\|\mathbf{x} - \mathbf{y}\|\right), \tag{20}$$

where I is the identity operator and $P_{(x)}$ and $P_{(y)}$ are the projection operators defined by applying (19) to the arguments x and y, while E is given by (16). By Proposition 61 [Rivera, 2017], R_1 is the reproducing kernel of $\mathcal{H} \ominus \mathcal{N}_J$ with inner product J. R_1 is symmetric by the properties of $\|\cdot\|$ and the projections. To show that R_1 is non-negative definite we need to show that $J\left(R_1(x,\cdot), R_1(y,\cdot)\right) = R_1(x,y)$. The reproducing property $J\left((I-P)f, R_1(x,\cdot)\right) = (I-P)f(x)$ for $f \in \mathcal{H}$ is more challenging to demonstrate [Wahba and Wendelberger, 1980].

With the reproducing kernels (19) and (20) we can provide an specific Bayesian regression model. Proposition 1 describes the most troublesome part of the model in this

section, namely the priors and conditional posterior of $\mathbf{c}$ and $\mathbf{d}$ given the variance at the observation level σ^2 and the bandwidth parameter $\lambda > 0$ are required to solve (1).

For $\eta \in \mathcal{H}^*$ consider the model

$$y_i = \eta(\mathbf{x}_i) + \epsilon_i$$
$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Here $\eta = \sum_{j=1}^{l} d_j \phi_j + \sum_{j=1}^{q} c_j R_0(\mathbf{z}_j, \cdot)$ by hypothesis. Using the same notation as in Proposition 1, consider the priors on the parameters

$$\begin{aligned}
\mathbf{d} &\sim 1, \\
\mathbf{c}|\sigma^2, \lambda &\sim N_k\left(\mathbf{0}, \frac{\sigma^2}{n\lambda} Q^+\right), \\
\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X}, \sigma^2) &= \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{U}(x|\mathbf{y}, \mathbf{X}, \sigma^2)\right\} dF_{\mathbf{y}|\mathbf{X}}(\mathbf{y}), \qquad (21) \\
\sigma^2 &\sim Inv - Gamma(A_\epsilon, B_\epsilon), \\
\mathbf{d} \perp \mathbf{c}, \quad \mathbf{d} \perp \sigma^2, \quad \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} &\perp (\epsilon_1 \cdots \epsilon_n)^\intercal, \quad \sigma^2 \perp \epsilon_i,\ i \in \{1, \cdots, n\},
\end{aligned}$$

where every distribution is conditional on the observed $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, and A_ϵ and B_ϵ are hyperprior parameters. The prior on λ is conditional on σ^2, but two similar models with λ independent of σ^2 can be proposed. Here, in two different models, the priors on λ would only depend on $\{\mathbf{x}_i\}_{i=1}^n$ and can be chosen as

$$\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X}) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{V}(x|\mathbf{y}, \mathbf{X}, \alpha)\right\} dF_{\mathbf{y}|\mathbf{X}}(\mathbf{y}) \text{ or,} \qquad (22)$$

$$\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X}) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{M}(x|\mathbf{y}, \mathbf{X})\right\} dF_{\mathbf{y}|\mathbf{X}}(\mathbf{y}); \qquad (23)$$

where

$$\mathcal{U}(\lambda|\mathbf{y}, \mathbf{X}, \sigma^2) = \frac{1}{n}\mathbf{y}^\intercal (I - A(\lambda))^2 \mathbf{y} + 2\frac{\sigma^2}{n} tr A(\lambda), \qquad (24)$$

$$\mathcal{V}(\lambda|\mathbf{y}, \mathbf{X}, \alpha) = \frac{n^{-1}\mathbf{y}^\intercal (I - A(\lambda))^2 \mathbf{y}}{\{n^{-1} tr(I - \alpha A(\lambda))\}^2}, \qquad (25)$$

$$\mathcal{M}(\lambda|\mathbf{y}, \mathbf{X}) = \frac{\mathbf{y}^\intercal (I - A(\lambda))\mathbf{y}}{|I - A(\lambda)|_+^{1/(n-l)}}. \qquad (26)$$

$\mathcal{U}$ [Mallows, 1973] is the *Unbiased Estimate of Relative Loss* (UERL). $\mathcal{V}$ [Kim and Gu, 2004] is the *Generalized Cross Validation* (GCV) score to choose smoothing parameters. α is a ridge parameter empirically known that $\alpha = 1.4$ is adequate in normal cases. $\mathcal{M}$ [Wahba, 1985] is the *Restricted Maximum Likelihood*, (RML) estimate of λ. In all cases the matrix $A(\lambda) \in \mathcal{M}_{n\times n}^+(\mathbb{R})$

$$A(\lambda) = I - n\lambda M^{-1}\left(I - S\left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1}\right), \qquad (27)$$

where $M = RQ^+R^\intercal + n\lambda I_n$.

Let $\Theta = \left(\mathbf{d}^\intercal, \mathbf{c}^\intercal, \sigma^2_\epsilon, \lambda\right)$. In principle, we would need to show that the posterior $[\Theta|\mathbf{y}]$ exists because an improper prior was given to $\mathbf{d}$. The formal way to show the existence of the posterior is to propose the proper prior $\mathbf{d} \sim N_l(\mathbf{0}, \tau^2 I_l)$, and follow the proof of Proposition 10 by taking the limit $\tau \to \infty$. The details of the proof are exactly the same as in Proposition 10 with the addition of multiplicative terms independent of τ, where the multiplicative terms correspond to the joint prior distribution of σ^2 and λ using (21), (22) or (23). In this light, the joint posterior distribution of Θ with improper prior on $\mathbf{d}$ exists and is proportional to

$$\begin{aligned}
[\Theta|\mathbf{y}] &\propto \left[\mathbf{y}|\mathbf{c}, \mathbf{d}, \sigma^2, \lambda\right] \times \left[\mathbf{d}, \mathbf{c}|\lambda, \sigma^2\right] \times \left[\lambda|\sigma^2\right] \times \left[\sigma^2\right] \\
&= \left[\mathbf{y}|\mathbf{c}, \mathbf{d}, \sigma^2\right] \times \left[\mathbf{d}, \mathbf{c}|\lambda, \sigma^2\right] \times \left[\lambda|\sigma^2\right] \times \left[\sigma^2\right] \\
&= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i; \mathbf{d}, \mathbf{c}))^2) - \frac{n\lambda}{\sigma^2} \mathbf{c}^\intercal Q \mathbf{c} - \frac{1}{B_\epsilon \sigma^2} \right\} \times \sigma^{-2(n/2 + A_\epsilon + 1)} \times \left[\lambda|\sigma^2\right].
\end{aligned}$$

Remark 4
In the expressions above we are using the probability density functions of the respective conditional distributions, or in measure theory terminology, the Radon-Nikodyn derivatives [Athreya and Lahiri, 2006] of the respective probability functions with respect to Lebesgue measures. In this context, $\left[\lambda|\sigma^2\right]$ *would be the Radon-Nikodyn derivative of the measure defined by* (21) *with respect to Lebesgue measure. Formally, we would need to prove the existence of such derivative, using for example, the Radon-Nikodym Theorem [Athreya and Lahiri, 2006]. If such derivative does not exist then* $[\Theta|\mathbf{y}]$ *can not be analytically expressed as the product of densities, as we did above; instead, the use of the cumulative distributions would be needed. For simplicity in the notation, we keep using the probability density distributions. Furthermore, we do not use the existence of the density* $\left[\lambda|\sigma^2\right]$ *but we will only use that* $\left[\lambda|\sigma^2, \mathbf{y}\right] = \arg\min_{x>0} \mathcal{U}(x|\sigma^2, \mathbf{y})$ *almost surely, by construction of* (21)*. Similarly if we use the priors* (22) *or* (23)*.*

It is now desired to simulate from $[\Theta|\mathbf{y}]$ which can be accomplished using the Gibbs sampler algorithm [Gelman et al., 2014, p. 276 - 278] simulating sequentially from the following full conditional distributions.

$$\begin{aligned}
\left[(\mathbf{d}^\intercal \ \mathbf{c}^\intercal)^\intercal | \lambda, \sigma^2, \mathbf{y}\right] &\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i; \mathbf{d}, \mathbf{c}))^2) - \frac{n\lambda}{\sigma^2} \mathbf{c}^\intercal Q \mathbf{c} \right\} \\
&\sim N_{l+k}\left(\mu_{\mathbf{dc}}, \mathbf{\Sigma}_{\mathbf{dc}}\right) \text{ (following proof Proposition 1), and ,} \\
\left[\lambda, \sigma^2 | \mathbf{d}, \mathbf{c}, \mathbf{y}\right] &\propto \left[\lambda|\sigma^2, \mathbf{d}, \mathbf{c}, \mathbf{y}\right] \times \left[\sigma^2|\mathbf{d}, \mathbf{c}, \mathbf{y}\right],
\end{aligned}$$

for $\mu_{\mathbf{dc}}$ and $\mathbf{\Sigma}_{\mathbf{dc}}$ as in (8) and (9). It is straightforward to simulate from $\left[(\mathbf{d}^\intercal \ \mathbf{c}^\intercal)^\intercal | \lambda, \sigma^2, \mathbf{y}\right]$. In order to simulate from $\left[\lambda, \sigma^2|\mathbf{d}, \mathbf{c}, \mathbf{y}\right]$, first it is needed to simulate from $\left[\sigma^2|\mathbf{d}, \mathbf{c}, \mathbf{y}\right]$ and using the simulated value σ^2, one can simulate from $\left[\lambda|\sigma^2, \mathbf{d}, \mathbf{c}, \mathbf{y}\right]$. Observe that

$$\begin{aligned}
\left[\sigma^2|\mathbf{d}, \mathbf{c}, \mathbf{y}\right] &\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i; \mathbf{d}, \mathbf{c}))^2 - \frac{1}{B_\epsilon \sigma^2} \right\} \times \sigma^{-2(n/2 + A_\epsilon + 1)} \\
&\sim Inv - Gamma\left(A_\epsilon + \frac{1}{2}n, \left[B_\epsilon^{-1} + \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \eta(\mathbf{x}_i)\right)^2 \right]^{-1} \right)
\end{aligned}$$

For the distribution $[\lambda|\sigma^2, \mathbf{d}, \mathbf{c}, \mathbf{y}]$ observe that by the law of Total Probability we have

$$\mathbf{P}(\lambda \geq \lambda_0 | \sigma^2, \mathbf{y}) = \int \mathbf{P}(\lambda \geq \lambda_0 | \sigma^2, \mathbf{d}, \mathbf{c}, \mathbf{y}) dF_{\mathbf{d},\mathbf{c}}(\mathbf{d}, \mathbf{c}), \text{ then}$$

$$\mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{U}(x|\mathbf{y},\mathbf{X},\sigma^2)\right\} = \int \mathbf{P}(\lambda \geq \lambda_0|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y}) dF_{\mathbf{d},\mathbf{c}}(\mathbf{d},\mathbf{c}), \text{ then it must be that}$$
$$\mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{U}(x|\mathbf{y},\mathbf{X},\sigma^2)\right\} = \mathbf{P}(\lambda \geq \lambda_0|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y}),$$

therefore we can conclude that

$$(\lambda|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y}) = \mathcal{U}(x|\mathbf{y},\mathbf{X},\sigma^2) \text{ almost surely.}$$

If instead of using the prior (21), we decide to use (22) or (23), the full conditional posteriors of the parameters would be:

$$\left[(\mathbf{d}^\intercal\ \mathbf{c}^\intercal)^\intercal|\lambda,\sigma^2\right] \sim N_{l+k}\left(\mu_{\mathbf{dc}},\mathbf{\Sigma}_{\mathbf{dc}}\right),$$
$$\left[\sigma^2|\lambda,\mathbf{d},\mathbf{c}\right] \sim Inv-Gamma\left(A_\epsilon + \frac{1}{2}n, \left[B_\epsilon^{-1} + \frac{1}{2}\sum_{i=1}^{n}(y_i - \eta(\mathbf{x}_i))^2\right]^{-1}\right),$$

and λ (and θ_i's) fixed as

$$(\lambda|\mathbf{y},\mathbf{X}) = \arg\min_{x>0} \mathcal{V}(x|\mathbf{y},\mathbf{X},\alpha), \text{ if prior (22) was used, or}$$
$$(\lambda|\mathbf{y},\mathbf{X}) = \arg\min_{x>0} \mathcal{M}(x|\mathbf{y},\mathbf{X}) \text{ if prior (23) was used.}$$

3.1.2 Bayesian regression model using tensor thin plate splines

The model in this section has the same form as for the thin plate splines (Section 3.1.1) with the only differences observed on the penalty term (2) and the corresponding reproducing kernel R. Therefore, the Bayesian model is the same but the basis functions changes. We now describe the changes on the basis functions.

Lets consider the reproducing kernel R_1 of the thin plate spline minimization problem. The following expressions are reproducing kernels for the tensor thin plate spline setting [Rivera, 2017, Section 2.1.2]. In the case $\mathbb{X} = \mathbb{R}^2$, with $\mathbf{x} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)}) \in \mathbb{R}^2$, and using the notation $R_{J_m^d}$ to denote the dependency of (15),

$$R_{K_1}(\mathbf{x},\mathbf{y}) = \theta_1 R_{J_m^1}\left(\mathbf{x}_{(1)},\mathbf{y}_{(1)}\right) + \theta_2 R_{J_m^1}\left(\mathbf{x}_{(2)},\mathbf{y}_{(2)}\right), \tag{28}$$
$$\begin{aligned} R_{K_2}(\mathbf{x},\mathbf{y}) =& \theta_1 R_{J_m^1}\left(\mathbf{x}_{(1)},\mathbf{y}_{(1)}\right) + \theta_2 R_{J_m^1}\left(\mathbf{x}_{(2)},\mathbf{y}_{(2)}\right) \\ &+ \theta_3 R_{J_m^1}\left(\mathbf{x}_{(1)},\mathbf{y}_{(1)}\right) R_{0,m}\left(\mathbf{x}_{(1)},\mathbf{y}_{(1)}\right) + \theta_4 R_{J_m^1}\left(\mathbf{x}_{(2)},\mathbf{y}_{(2)}\right) R_{0,m}\left(\mathbf{x}_{(2)},\mathbf{y}_{(2)}\right) \\ &+ \theta_5 R_{J_m^1}\left(\mathbf{x}_{(1)},\mathbf{y}_{(1)}\right) R_{J_m^1}\left(\mathbf{x}_{(2)},\mathbf{y}_{(2)}\right), \end{aligned} \tag{29}$$

where $R_{J_m^1}$ is the reproducing kernel for the thin plate splines in the domain $\mathbb{X} = \mathbb{R}$ described by (20), and $R_{0,m}$ is the reproducing kernel of the space of polynomials in $\mathbb{R}$ with degree smaller than $m+1$; note $R_{0,m}$ is fully described by (18). The reproducing kernels (28) and (29) are the respective kernels for the tensor thin plate spline without interaction in the Anova decomposition of the associated Hilbert space, [Aronszajn, 1950, Akhiezer and Glazman, 1981a, Akhiezer and Glazman, 1981b, Gu, 2013], while the setting with interaction terms has reproducing kernel (29).

For the simulation study in Section 4 we use the tensor thin plate spline with interaction, hence the reproducing kernel (29). In this case we need to choose the smoothing parameters $\{\theta_i\}_{i=1}^p$ and set $\lambda = 1$ for identifiability reasons. The Bayesian regression model interpretation in this setting has the same form as in the thin plate spline, but now the matrices Q, R and the projection matrix $A(\lambda)$ in (27) depend on $\{\theta_i\}_{i=1}^p$ with $\lambda = 1$; the functions $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{M}$ depend as well on $\{\theta_i\}_{i=1}^p$ and the priors (21), (22), (23) are specified minimizing over the positive quadrant of $\mathbb{R}^p$ ($p = 5$ here).

3.1.3 Full Bayes linear mixed effects model

The full Bayes model follows directly from a linear mixed model. Consider the linear mixed model

$$\begin{aligned} \mathbf{y}|\mathbf{d},\, \mathbf{c},\, \mathbf{e} &= S\mathbf{d} + U\mathbf{c} + \mathbf{e}, \\ \begin{pmatrix} \mathbf{c} \\ \mathbf{e} \end{pmatrix} &\sim \mathbf{N}_{n+k}\left(\mathbf{0}, \begin{pmatrix} \sigma_c^2 I_k & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 I_n \end{pmatrix}\right), \end{aligned} \tag{30}$$

with $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ as before and $U = FV^{-1/2}$, $F_{i,j} = E_{m,d}\|\mathbf{x}_i - \mathbf{z}_j\|$, $V_{i,j} = E_{m,d}\|\mathbf{z}_i - \mathbf{z}_j\|$. Consider the priors

$$\begin{aligned} \mathbf{d} &\sim 1 \\ \sigma_c^2 &\sim Inv - Gamma\,(A_c, B_c) \\ \sigma_e^2 &\sim Inv - Gamma\,(A_e, B_e)\,. \end{aligned}$$

That the posterior distribution of the parameters is proper follows by similar arguments as in Proposition 12. The full conditional posterior distributions are

$$\begin{aligned} \left[(\mathbf{d}^\intercal\, \mathbf{c}^\intercal)^\intercal | \sigma_c^2, \sigma_e^2\right] &\sim N_{k+l}\left(\left([V\,U]^\intercal\,[V\,U] + \frac{\sigma_e^2}{\sigma_c^2}\mathbf{D}\right)^{-1}[V\,U]\,\mathbf{y}, \left([V\,U]^\intercal\,[V\,U] + \frac{\sigma_e^2}{\sigma_c^2}\mathbf{D}\right)^{-1}\right) \\ \left[\sigma_c^2 | \sigma_e^2, \mathbf{d}, \mathbf{c}\right] &\sim Inv - Gamma\left(A_c + k/2, \left(B_c + \frac{1}{2}\|\mathbf{c}\|^2\right)^{-1}\right) \\ \left[\sigma_e^2 | \sigma_c^2, \mathbf{d}, \mathbf{c}\right] &\sim Inv - Gamma\left(A_e + n/2, \left(B_e + \frac{1}{2}\|\mathbf{y} - S\mathbf{d} - U\mathbf{c}\|^2\right)^{-1}\right), \end{aligned}$$

where $\mathbf{D} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_l \end{pmatrix}$. Observe that the ratio σ_e^2/σ_c^2 plays the role of $n\lambda$, the smoothing parameter. It could be interpreted as assigning a prior to λ and observing the corresponding distribution on $\sigma_c^2 = \frac{\sigma_e^2}{n\lambda}$ which in this case is an inverse gamma. The next models use this approach, a prior is assigned to λ and σ_e^2, and σ_c^2 follows the corresponding induced prior.

The Bayesian model from this section was inspired by the full Bayes model to estimate function in the presence of errors in covariance by [Berry et al., 2002].

3.1.4 Bayesian linear mixed model interpretation and empirical bandwidth parameters

Consider the linear mixed model (30) and define $\lambda = \frac{\sigma_e^2}{n\sigma_c^2}$. Consider the priors

$$\begin{aligned} \mathbf{d} &\sim 1, \\ \lambda|\sigma_e^2 &= \arg\min_{x>0}\left\{\mathcal{U}(x|\sigma_e^2)\right\} \text{ almost surely}, \\ \sigma_e^2 &\sim Inv - Gamma\,(A_e, B_e)\,. \end{aligned}$$

The full conditional posterior of the parameters are

$$\begin{aligned} \left[(\mathbf{d}^\intercal\, \mathbf{c}^\intercal)^\intercal | \lambda, \sigma_e^2\right] &\sim N_{k+l}\left(\left([V\,U]^\intercal\,[V\,U] + \frac{\sigma_e^2}{\sigma_c^2}\mathbf{D}\right)^{-1}[V\,U]\,\mathbf{y}, \left([V\,U]^\intercal\,[V\,U] + \frac{\sigma_e^2}{\sigma_c^2}\mathbf{D}\right)^{-1}\right) \\ \lambda|\sigma_e^2 &= \arg\min_{x>0}\left\{\mathcal{U}(x|\sigma_e^2)\right\} \text{ almost surely}, \\ \left[\sigma_e^2 | \lambda, \mathbf{d}, \mathbf{c}\right] &\sim Inv - Gamma\left(A_e + n/2, \left(B_e + \frac{1}{2}\|\mathbf{y} - S\mathbf{d} - U\mathbf{c}\|^2\right)^{-1}\right), \end{aligned}$$

Alternatively, in two different models, λ can be chosen as well using (22) or (23) and the full conditional posteriors of the parameters would be similar to above but λ as the *argmin* of the respective function.

4. Simulation Study: Bayesian Models using Thin Plate Splines, Tensor Thin Plate Splines and Linear Mixed Model Interpretation

We perform a simulation study to compare the performance of the models in Section 3.1 in terms of point estimates and coverage of credible intervals for functions $\eta : \mathbb{R}^2 \to \mathbb{R}$. The algorithm and our methods work in theory for any number of covariates but we provide a robust simulation study for the case of two covariates. We tested our methods with some examples to estimate functions with domain in $\mathbb{R}^3$ and $\mathbb{R}^4$ obtaining good resulting estimation. The purpose of the simulation study is to observe the performance on the estimation provided by all 10 models measured in terms of the Bayes estimates of the function, Bayes prediction, and empirical coverage of the credible intervals for predictions of η.

There has been previous work on simulation studies for nonparametric regression models with Bayes interpretation, such in [Wahba, 1983, Nychka, 1988, Kim and Gu, 2004], but most of the studies are for estimation of univariate functions or examples of estimation for bivariate functions are simply shown. [Wahba, 1983] and [Nychka, 1988] report an average coverage probability across the region of estimation for the credible intervals that is similar to the level of the credibility, when using the GCV method to choose the bandwidth parameters. [Nychka, 1988] mentions that the main disadvantage of the approach is that the "confidence intervals" are only valid in an average sense over the region of estimation, and may not be reliable if evaluated for pointwise estimation or only evaluated at peaks or troughs in the estimate; the pointwise coverage of the credible intervals depend on the unknown function. With the simulation study we have designed, we compare four methods to choose the smoothing parameters. We can observe the impact of having more than one smoothing parameters in the model (tensor thin plate splines) or choosing the smoothing parameter by assigning a prior. We are able to observe the behavior of the point estimates and the empirical coverage of credible intervals in different regions of estimation varying in size and the concentration of observed data for bivariate function.

Table 1: Summary form of the models and smoothing methods to choose/estimate the smoothing parameters in the simulation study. LMM - linear mixed model interpretation, TPS - thin plate splines, TTPS - tensor thin plate splines with anova interaction. UERL - unbiased estimate relative loss, GCV - generalized cross validation, REML - restricted maximum likelihood under the Bayes model, Bayes - inverse gamma prior on both variances σ_e^2 and σ_c^2.

Smoothing Model	Bandwidth Method
LMM	UERL
LMM	GCV
LMM	RML
LMM	Bayes
TPS	UERL
TPS	GCV
TPS	RML
TTPS	UERL
TTPS	GCV
TTPS	RML

The form of the competing models and the methods to select the bandwidth parameters are summarized in Table 4. There, we use the abbreviations LMM for the linear mixed models described in Sections (3.1.3) and (3.1.4), TPS for the Bayesian models using thin

plate splines (Section (3.1.1)), and TTPS for the Bayesian models using tensor thin plate splines (Section (3.1.2)). With regard to the selection of the smoothing parameters, the notation UERL indicates that λ (and θ_i's) were assigned with the prior (21), GCV indicates that the prior (22) was used, and a Bayesian model with (23) as prior is denoted as RML. In the same table, the combination LMM and *Bayes* bandwidth method denotes the model described in Section (3.1.3), and rest of the LMM models were described in Section (3.1.4). The former model will denoted as *Full Bayes* model, and the later models will be jointly denoted as *Bayes Empirical* models.

For each combination of the values of the parameters, 200 simulated data sets were generated, with $\{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$, $n = 50,\ 100,\ 400,$ and 800; the responses $\{y_i\}_{i=1}^n$ were simulated with the form $y_i = \eta(\mathbf{x}_i) + \epsilon_i$, while $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ using $\sigma = 0.01,\ 0.1,\ 0.25,\ 0.5,$ and 1. The deterministic function to be estimated is

$$\eta((x_{(1)}, x_{(2)})) = \left[4 + \frac{\sin(\frac{1}{2}\pi x_{(1)})}{1 + 4x_{(1)}^2 \mathbf{1}(x_{(1)} > 0)}\right] \times \left[\sin(x_{(2)}) + \cos(x_{(2)}) + x_{(2)}\right]. \tag{31}$$

For each of the simulations, the priors on the variances were chosen to be $\sigma_\epsilon^2 \sim Inv - Gamma(1,1)$, $\sigma_c^2 \sim Inv - Gamma(1,1)$, and $\sigma_e^2 \sim Inv - Gamma(1,1)$ for the respective models. The values of the hyper-parameters for the priors were chosen in this way because the prior information they provide is not strong. It was found that estimation is insensitive to moderate modifications to these values.

Estimates for $\eta(\chi)$ for any $\chi \in \mathbb{R}^2$ are found as the mean of the posterior predictive distribution, [Gelman et al., 2014, pag 7]:

$$\Pi(\eta(\chi_i)|\mathbf{y}) = \int \Pi\left(\eta(\chi)|\theta\right) \Pi\left(\theta|\mathbf{y}\right) d\theta.$$

Observe that we are abusing of the notation, η is a deterministic functions while the notation $\Pi(\eta|\mathbf{Y})$ suggest that η is a random process with variance function, in principle, non zero. We use this notation keeping in mind that we are interested in the posterior mean of $\Pi(\eta(\chi_i)|\mathbf{Y}) = [\eta(\chi_i)|\mathbf{Y}]$ as point estimate of the deterministic $\eta(\chi_i)$.

The posterior distribution of the parameters in the models and the posterior predictive distribution $\Pi(\eta(\chi)|\mathbf{Y})$ do not have an analytically form. Samples were drawn using MCMC methods. Two independent chains with different initial overdispersed values for each parameter were drawn using Gibbs sampler, [Gelman et al., 2014, pag 276 - 278]. Each of the chains were run for $10,000$ iterations discarding the first $7,000$ realizations as burn in and thinning the rest of the sequences by keeping every 3 draws. In the simulation study is not possible to assess convergence of the MCMC chains for all simulated parameters and all data sets at the same time. Instead, convergence tests such as Geweke test, [Geweke et al., 1991] and Gelman test [Gelman et al., 2014, pag. 285] were used to separately test the convergence for the parameters.

Realizations of the posterior predictive distribution are achieved using the samples from the posteriors. Let $\hat{\eta}(\chi) := \mathbb{E}\left[\Pi(\eta(\chi)|\mathbf{Y})\right]$ denote the point estimator of $\eta(\chi)$ and $\tilde{\eta}(\chi) := \mathrm{sd}\left[\Pi(\eta(\chi)|\mathbf{Y})\right]$ denote the standard deviation of the posterior distribution. Then $\hat{\eta}(\chi)$ and $\tilde{\eta}(\chi)$ are approximated with the sample mean and the unbiased sample standard deviation from the realizations of the posterior predictive distributions.

Figure 1 shows an example of estimation using the model obtained from the Bayes interpretation of the thin plate splines in a grid of resolution 0.05×0.05 inside the square $[-2.25, 2.25]^2$. We denote this grid as $\{\chi_i\}_{i=1}^N \subset \mathbb{R}^2$. The square $[-2.25, 2.25]^2$ was chosen as the region of estimation for all the simulated data sets of the study because it has the property that it would contain about 95.17% of all points simulated from $N_2(\mathbf{0}, I_2)$; the grid was not chosen to be finer because of storage availability. The Bayes point estimator $\hat{\eta}$

is expected to behave similarly to frequentist estimation of the non-parametric regression obtained by (1) because of the interpretation of the mean of the full conditional posterior distribution as a solution. Observe that the pointwise standard deviation in the estimation $\tilde{\eta}$ is larger on the boundaries of the region of estimation because have less information about the regression function in that area, thought the standard deviation is smaller in the center of the region as expected.

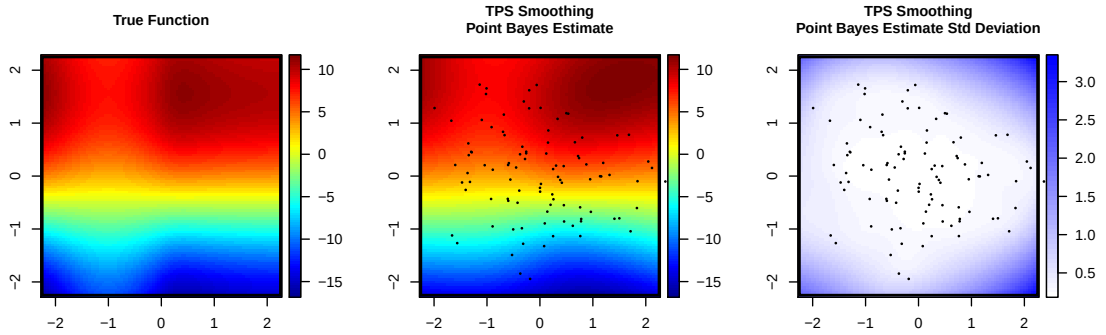

Figure 1: Level curves for the true function η equation (31) (left), point Bayes estimate $\hat{\eta}(\chi)$ (center), and pointwise standard deviation, $\tilde{\eta}(\chi)$ (right), for estimation using Bayes model interpretation of the thin plate splines with $m = 3$ and smoothing parameter chosen using the restricted maximum likelihood method, with $n = 100$ and $\sigma^2 = 0.5$. The dots in the plots are the observed values of the covariates $\{\mathbf{x}_i\}_{i=1}^{100}$.

Each model in Table 4 does not assume a parametric form for η, instead it is assumed that η is in the space $\mathcal{H}^\star$ (4); in this way we approximate the solution to (1) in $\mathcal{H} \supseteq \mathcal{H}^\star$, the *RKHS* generated by the reproducing kernel (20). Such a *RKHS* , for the estimation in Figure 1, does not contain functions that are not at least three ($m = 3$) times partially differentiable where the squares of the partial derivatives of degree $m = 3$ are integrable. Therefore the strongest assumptions made on η is that all of its third partial derivatives exist for every point in $\mathbb{R}^2$ and the integral over $\mathbb{R}^2$ of the square of each of the derivatives is finite. It is possible to impose a weaker assumption on η and the proposed methodology would still be theoretically justified and interpretations would be the same; it would be enough to have that all the partial derivatives of degree $m = 3$ are integrable in the square $[-2.25, 2.25]^2$, but we did not pursuit to prove this statement. The required properties of partial differentiabilty, integrability of the square of the partial derivatives, and that the mean of the full conditional distribution of η is the function that best interpolates the data as measured by the squared loss function $\frac{1}{n}\sum_{i=1}^{n}(y_i - \eta(\mathbf{x}_i))^2$ subject to the constrain that $J_3^2(\eta)$ is small, are the assumptions that lead us to compute $\hat{\eta}$ as observed in Figure 1.

4.1 Prediction and variability of prediction for the target regression function and discussion

We propose a summary to evaluate the performance of the estimated functions $\hat{\eta}$ as follow. By the interpretation of the mean of the full conditional posterior of the models as thin plate splines or tensor thin plate splines, and by the way the smoothing parameters were chosen using the Unbiased Estimate Relative Loss method (24) and the generalized cross validation method (25), $\hat{\eta}$ approximately minimizes the loss function $\frac{1}{n}\sum_{i=1}^{n}(\eta_\lambda(\mathbf{x}_i) - \eta(\mathbf{x}_i))^2$, up to the constrain [Gu, 2013, Theorem 2.12]. Instead of evaluating the performance of the estimation $\hat{\eta}$ using the loss function $N^{-1}\sum_{i=1}^{N}(y_i - \eta(\mathbf{x}_i))^2$ (as we know η_λ already minimizes a form like this), we propose to use a summary that is function of absolute differences: $|\eta(\mathbf{x}) - \hat{\eta}(\mathbf{x})|$. For consistency we use this summary for all models. The mean

absolute error $N^{-1}\sum_{i=1}^{N}|y_i - \eta(\mathbf{x}_i)|$ is a robust measure of predictive accuracy, it tends to prefer predictions procedures that on average are reasonably good and is less sensitive to large deviations than the square loss function.We propose the use of

$$\frac{1}{N \times \sigma}\sum_{i=1}^{N}|\eta(\chi_i) - [\eta(\chi_i)|\,\mathbf{Y}]| \tag{32}$$

to compute and compare deviations from the real function and to measure variability of the predictions around η for all models in Table 4. As the previous expression is a random variable, we approximate the mean

$$MAPE := \frac{1}{N \times \sigma}\mathbb{E}\left[\sum_{i=1}^{N}|\eta(\chi_i) - [\eta(\chi_i)|\,\mathbf{Y}]|\right] \tag{33}$$

and its respective variability

$$SDMAPE := \frac{1}{N \times \sigma}\sqrt{\mathbf{Var}\left[\sum_{i=1}^{N}|\eta(\chi_i) - [\eta(\chi_i)|\,\mathbf{Y}]|\right]}. \tag{34}$$

for each combination of parameters of the simulated data and each of the 200 repetitions. Above, MAPE stands for *Mean Absolute Prediction Error*. Estimation of (33) and (34) is achieved using the realizations of $[\eta(\chi_i)|\,\mathbf{Y}]$, computing (32) and finally obtaining the sample mean and sample standard deviation.

Table 4.1 summarizes a part of the results for the computations of $MAPE$ and $SDMAPE$. This table summarizes the case for $n = 100$ and $\sigma^2 = 0.5$. Column **Avg M** is the average across 200 computed $MAPE$ for each model, **M 25** and **M 75** are the 25% and 75% empirical percentiles of these 200 computed $MAPE$. **Avg SDM** is the average of the 200 computed $SDMAPE$ with their respective 25% and 75% percentiles. Bold number for column **M 25** indicates that the 25% percentile computed is larger than at least one of the 75% percentile computed for another model; the graphical interpretation of boxplots in Figure 2 is that boxes of these cases do not intersect. Bold numbers for column **M 75** indicate that there is at least one 25% percentile from another model that is larger than this 75% percentile for this specific combination of parameters $n = 100$ and $\sigma^2 = 0.5$. Similar graphical representations for the last three columns of Table 4.1 are shown in Figures 3.

The striking features of this table and its graphical displays (Figures 2 for the $MAPE$ columns is that the frequentist properties for the estimates of η as measured by (33) seems to be similar within TTPS, TPS, and *Empirical Bayes* models regardless of the score minimization criteria used to choose the smoothing parameters, with the exception of the *Bayes* bandwidth method (Table 4). The TPS models have as good estimates as the TTPS models when the variance σ^2 is not small ($\sigma^2 \neq 0.1^2$) and when comparing across different values of m. This suggests at least for this example, that one smoothing parameter λ is enough to provide similar estimates of η as when using five smoothing parameters $\{\theta_i\}_{i=1}^{5}$ in the TTPS models. Complete boxplots for the simulation are available in Figure C.1 from [Rivera, 2017].

The *Full Bayes* model performs in a similar way to the *Bayes Empirical* models in the sense that there is not practical difference in the median of the MAPE for the 200 repetitions when comparing across values of the parameter m for the same σ^2. For the *Full Bayes* model and the *Bayes Empirical*, we did not find practical differences for predicting in the square $[-2.25, 2.25]^2$. Similar conclusions hold comparing the predictions of the *Bayes Empirical* and TTPS or with the TPS models. For now, we leave the comparison

Table 2: Part summary simulation results for $MAPE$ and $SDMAPE$. Simulated data with link function (31), $n = 100$ and $\sigma^2 = 0.5$. Using 200 repetitions, **Avg M** is the average of the computed $MAPE$, **M 25** and **M 75** are the 25% and 75% empirical percentile of the 200 computed $MAPE$. **Avg SDM** is the average of the 200 computed $SDMAPE$ with their respective 25% and 75% percentiles. Black number for column **M 25** indicate that the 25% percentiles computed with the respective model is larger than at least one of the 75% $MAPE$ percentile computed for another model. Black numbers for column **M 75** indicates that there is at least one 25% percentile $MAPE$ from another model that is larger than this 75% percentile. Similarly for columns **SDM 25** and **SDM 75**.

Model	**m**	**Avg M**	**M 25**	**M 75**	**Avg SDM**	**SDM 25**	**SDM 75**
Bayes Empcal GCV	2	0.96	**0.86**	**1.04**	0.10	0.09	**0.1**
Bayes Empcal GCV	3	0.96	**0.83**	**1.09**	0.14	**0.11**	**0.16**
Bayes Empcal GCV	4	1.06	**0.9**	**1.14**	0.19	**0.15**	**0.22**
Bayes Empcal RML	2	0.94	**0.84**	**1.03**	0.10	0.09	**0.11**
Bayes Empcal RML	3	0.96	**0.82**	**1.08**	0.15	**0.12**	**0.17**
Bayes Empcal RML	4	1.07	**0.92**	**1.14**	0.20	**0.16**	**0.23**
Bayes Empcal UERL	2	0.94	**0.85**	**1.02**	0.10	0.09	**0.11**
Bayes Empcal UERL	3	0.96	**0.82**	**1.08**	0.15	**0.12**	**0.17**
Bayes Empcal UERL	4	1.11	**0.95**	**1.2**	0.22	**0.16**	**0.24**
Full Bayes	2	0.94	**0.84**	**1.03**	0.10	0.09	**0.11**
Full Bayes	3	1.12	**0.91**	**1.28**	0.17	**0.11**	**0.17**
Full Bayes	4	1.51	**1.34**	1.59	0.15	**0.11**	**0.17**
Tensor TPS GCV	2	0.65	0.55	**0.74**	0.60	**0.55**	**0.65**
Tensor TPS GCV	3	0.87	0.67	**0.98**	0.84	**0.7**	**0.94**
Tensor TPS GCV	4	1.37	**0.92**	1.69	1.30	**0.99**	1.48
Tensor TPS RML	2	0.65	0.53	**0.72**	0.66	**0.6**	**0.71**
Tensor TPS RML	3	0.84	0.62	**0.96**	0.90	**0.74**	**1**
Tensor TPS RML	4	1.39	**0.91**	1.76	1.41	**1.08**	1.59
Tensor TPS UERL	2	0.66	0.56	**0.76**	0.62	**0.57**	**0.66**
Tensor TPS UERL	3	0.88	0.68	**0.98**	0.87	**0.74**	**0.95**
Tensor TPS UERL	4	1.43	**0.91**	1.77	1.34	**1.02**	1.51
TPS GCV	2	0.81	0.72	**0.93**	0.54	**0.51**	**0.56**
TPS GCV	3	0.78	0.64	**0.9**	0.66	**0.61**	**0.71**
TPS GCV	4	0.81	0.66	**0.95**	0.81	**0.7**	**0.88**
TPS RML	2	0.77	0.67	**0.86**	0.58	**0.55**	**0.61**
TPS RML	3	0.75	0.61	**0.87**	0.72	**0.66**	**0.77**
TPS RML	4	0.80	0.65	**0.94**	0.89	**0.77**	**0.95**
TPS UERL	2	0.78	0.69	**0.88**	0.56	**0.54**	**0.58**
TPS UERL	3	0.74	0.61	**0.86**	0.73	**0.66**	**0.77**
TPS UERL	4	0.81	0.65	**0.94**	0.96	**0.82**	**1.04**

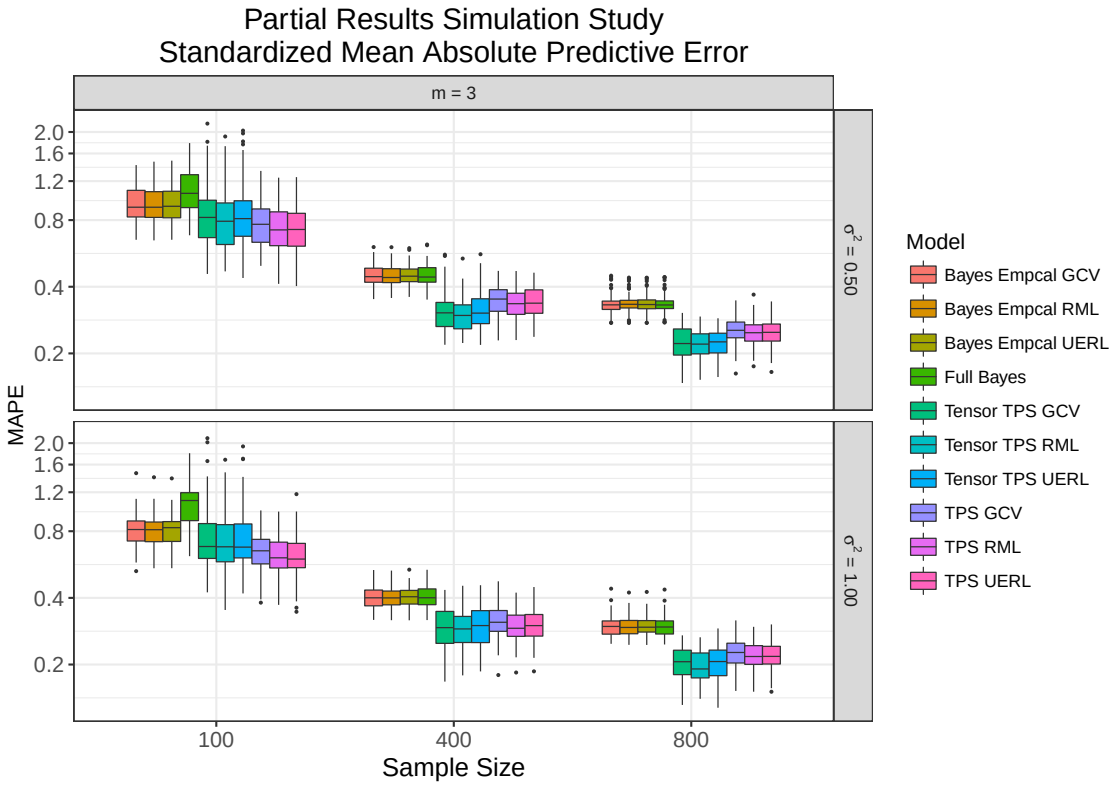

Figure 2: Boxplots part of the simulation results for the standardized mean absolute predictive error (33) (MAPE) for the multivariate regression problem predicting over the grid of resolution 0.05×0.05 inside the square $[-2.25, 2.25]^2$. Observe that the y-axis is in the log_{10} scale. The rows indicates the true observation-error variance σ^2. The columns indicate the degree of derivative m for the penalty in (14). The models are described in Table 4. Complete simulation results appear in Figure C.1 [Rivera, 2017].

of the MAPE between the *Full Bayes* model and the TTPS and TPS models until after we discuss the SDMAPE for the *Bayes Empirical* models, TTPS and TPS.

The variability of the marginal posterior process $[\eta|\mathbf{y}]$ around η is positively related with the $SDMAPE$ and discussed now. The most striking feature in the last three columns of Table 4.1 and represented in Figure 3 and Figure C.2 [Rivera, 2017], is that the average variability of the predictions of η over the grid in the square $[-2.25, 2.25]^2$ is not statistically significant or different between methods to choose the smoothing parameters and within the models TTPS, TPS or Bayes Empirical. For the cases shown in Table $C.2$ [Rivera, 2017] and Figure 3, there is no statistically significant difference in the average variability around η of the predictions in the grid for the TTPS and TPS models (median of SDMAPE for 200 repetitions). For the few cases when there is a statistical difference in the values of SDMAPE as seen in the appendix Figure $C.2$ in [Rivera, 2017], that is to say, for the cases when the predictions from the TPS have smaller variability than those from the TTPS model as measured by the SDMAPE, the $MAPE$ is statistically smaller for TTPS than for TPS. Based on the results of the point predictions for η and their variability around the true link function η, we have shown evidence at least for this simulated setting that the TTPS and TPS models are equally competitive, at least in a practical sense of predicting and extrapolating in the square $[-2.25, 2.25]^2$. For now, the advantage of TPS models over TTPS is that less computational effort is required to estimate the single bandwidth parameter λ for the TPS model than the five bandwidth parameters $\{\theta_i\}_{i=1}^5$ for the TTPS model.

We admit that a function η can always be constructed such that a TTPS model with five

smoothing parameters is required over the TPS model but, in similar way, another artificial function ψ can be constructed such that the TTPS over fits the data. The construction would follow from the criticism of [Barry et al., 1986]. For the example function η (31) that was chosen without being influenced a priori by the form of the models and that we use in this simulation study, it seems that both TPS and TTPS behave practically equal when predicting.

Now that we have discussed the SDMAPE, we come back to the MAPE summary for the *Full Bayes* model. The median of the MAPE over the 200 repetitions of the *Full Bayes* model in comparison with the medians of the TTPS model or with the TPS model is statistically larger. However, the difference is not clearly practical different as we argue now. While the median value of the MAPE seems to be larger than with the TTPS or TPS models, the median of the SDMAPE for these same cases of the *Full Bayes* model are large indicating that the variability of the point predictions for the $FullBayes$ model for each fitting, is large around the true function. This large variability is observed because the *Full Bayes* model produce larger pointwise credible intervals for $[\eta(\chi_i)|\mathbf{y}]$ than any other model. In order to evaluate if the credible intervals are too large we evaluate the empirical coverage of the credible intervals produced by each model in Section 4.2.

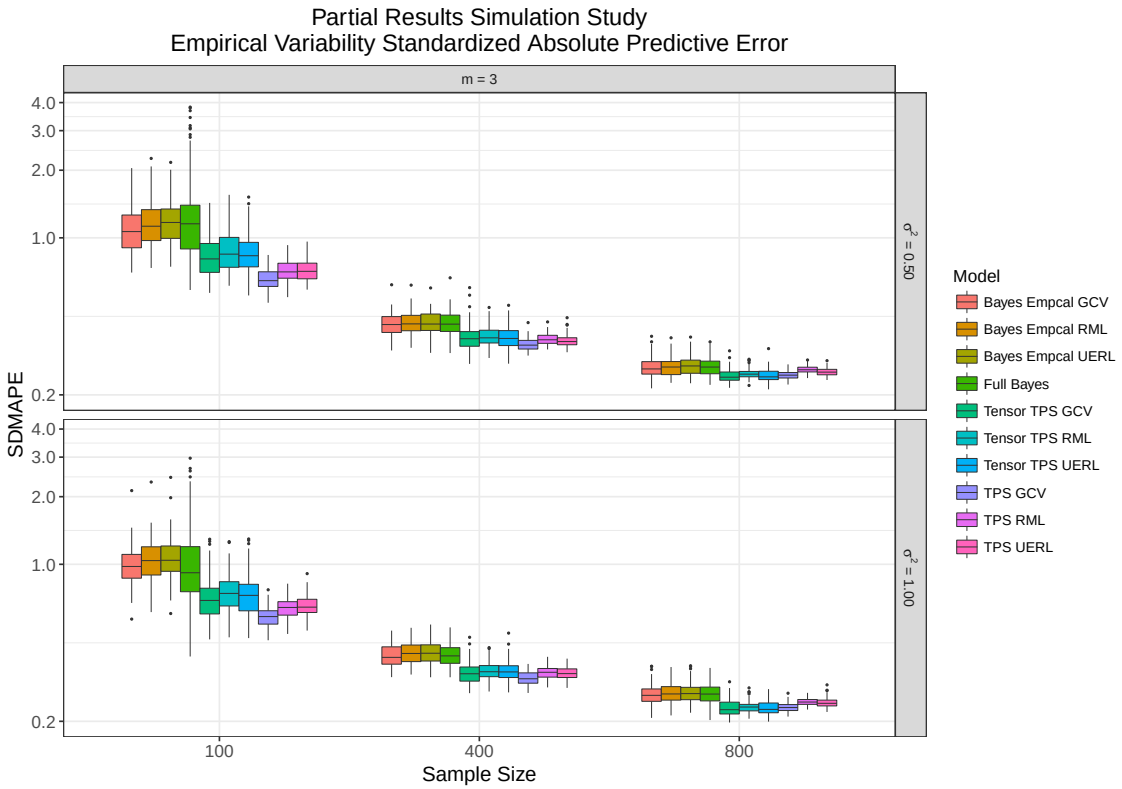

Figure 3: Boxplots part of the simulation results for the standardized standard deviation of the absolute predictive error (SDMAPE) for the multivariate regression problem predicting over the grid of resolution 0.05×0.05 inside the square $[-2.25, 2.25]^2$. Observe that the y-axis is in the log_{10} scale. The rows indicates the true observation-error variance σ^2. The columns indicate the degree of derivative m for the penalty in (14). The models are described in Table 4. Complete simulation results appear in Figure $C.2$ [Rivera, 2017].

Recall the interpretation of λ in the *Full Bayes* model, from Section 3.1.3, as $n\lambda = \frac{\sigma_e^2}{\sigma_c^2}$. It is possible that by providing a more informative prior over σ_c^2 or σ_e^2 or even changing the family of the distribution of the priors for these variance parameters, the bias and the variability of the predictions for η decrease. We did not pursue this objective.

4.2 Empirical coverage of credible intervals from predictive posterior distribution for the target regression function and discussion

We evaluate now the empirical pointwise coverage of the credible intervals for $\eta(\chi_i)$ for all χ_i in the grid and C-level pointwise credible intervals for each $\eta(\chi_i)$. For each independent simulation $j = 1, \cdots, 200$ define the variable

$$\xi_i^{(j)} = \begin{cases} 1 & \text{if } \eta(\chi_i) \text{ is contained in the } C\% \text{ centered credible intval of } \Pi(\eta(\chi_i)|\mathbf{Y}) \text{ from simulation } j, \\ 0 & \text{if } \eta(\chi_i) \text{ is not contained in the } C\% \text{ centered credible intval of } \Pi(\eta(\chi_i)|\mathbf{Y}) \text{ from simulation } j. \end{cases}$$

The random variables $\xi_i^{(j)} \sim Bernoulli(\rho_i)$ are Bernoulli distributed, and the random variables ρ_i have the same expected value $\zeta \in [0, 1]$. If the points $\{\chi_i\}_{i=1}^N$ where we try to predict η are the observation points $\{\mathbf{x}_i\}_{i=1}^N$ then the defined ζ is known as the *average coverage probability* (ACP) [Wahba, 1983]. Wahba estimates ACP using a single data set ($j = 1$) for the $C\%$ level using the expression

$$\hat{\zeta} = ACP(C) := \frac{1}{n} \# \left\{ i : |\eta(\mathbf{x}_i) - \hat{\eta}_{\lambda_v}(\mathbf{x}_i)| < z_{\alpha/2} \sigma_v^2 \sqrt{[A(\lambda_v)]_{ii}} \right\} = \frac{1}{n} \sum_{i=1}^{n} \xi_i^{(1)}, \tag{35}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quartile of the standard normal distribution, $\alpha = 1 - C/100$. The ACP(C) (35) is obtained from a centered credible interval from Wahba's Bayesian model where the prior on σ^2 is degenerately σ_v^2 or equivalently $\sigma^2 = \sigma_v^2$ is assumed and $\lambda = \lambda_v$. We use the same definition of ACP for any grid $\{\chi_i\}_{i=1}^N$ and we estimate it using the credible intervals from the marginal posterior distributions $\{[\eta(\chi_i)|\mathbf{y}]\}_{i=1}^N$ or equivalently it can be estimated by obtaining the sample mean of $\{\rho_i\}_{i=1}^N$:

$$\begin{aligned} \hat{\zeta} = ACP(C) &= \frac{1}{N \times 200} \sum_{i=1}^{N} \sum_{j=1}^{200} \xi_i^{(j)} \\ &= \frac{1}{N} \sum_{i=1}^{N} \rho_i^{(j)}. \end{aligned} \tag{36}$$

It was investigated by [Nychka, 1988], in the setting of [Wahba, 1983] using large-sample approximation theory and simulation, that (35) is close to the nominal. [Nychka, 1988] reports *"From a frequency point of view, this agreement occurs because the average posterior variance for the spline is similar to a consistent estimate of the average squared error and because the average squared bias is a modest fraction of the total average squared error. These properties are independent of the Bayesian assumptions used to derive this confidence procedure"*. In our setting, we can only use the same arguments for the full conditional posterior distribution $[\eta|\sigma^2, \mathbf{y}, \lambda]$, as these arguments do not apply to the marginal posterior distribution and also not to the *Full Bayes* and *Bayes Empirical* models. However, part of the explanation about we obtaining similar results regarding to (36) being close to the nominal value for some models, as we will describe, must follow Nychka's arguments closely. Our purpose now is to evaluate the nominal level achieved by each of the ρ_i for each combination of parameters and models, and to test the nominal level achieved by ζ or ACP in our setting of bivariate regression with different algorithms to choose the bandwidth parameters.

A graphical display obtained from the computation of the ρ_i over the grid in the region $[-2.25, 2.25]$ using the 200 simulated data sets is shown in Figure 4 for the credibility levels 95%, 65% and 35%. A more comprehensive example using TPS model and RML for the bandwidth parameter is presented in Figure $C.4$ [Rivera, 2017]. We would like that each ρ_i is close to the respective nominal level.

Observe in Figure 4 and in Figure $C.4$ [Rivera, 2017], that the nominal level for the ρ_i's seems to be approximately achieved in the center of the region of estimation while the empirical coverage decreases at the boundary of the regions of estimation. This effect is because we are extrapolating in extreme parts of the regions of estimation. We may evaluate the change in the empirical coverage distribution of the ρ_i's as we reduce the area of prediction of the function to concentric smaller circles instead of on the square region $[-2.25, 2.25]$. The true function and an example of observed values $\{\mathbf{x}_i\}$ for Figure 4 are presented in Figure 1.

For each of the models in Table 4 and all combinations of the parameters $m \in \{2, 3, 4\}$, $\sigma^2 \in \{0.1^2, 0.1, 0.5^5, 0.5, 1\}$, we computed the empirical coverages ρ_i of 95% credible intervals from the respective marginal posteriors (left plot Figure 4). We summarize each plot as in Figure 4 or $C.4$ [Rivera, 2017] using boxplots of the ρ_i's. Each box in Figure 5 was computed from the respective ρ_i. A more comprehensive summary is shown in the appendix Figure $C.5$ [Rivera, 2017].

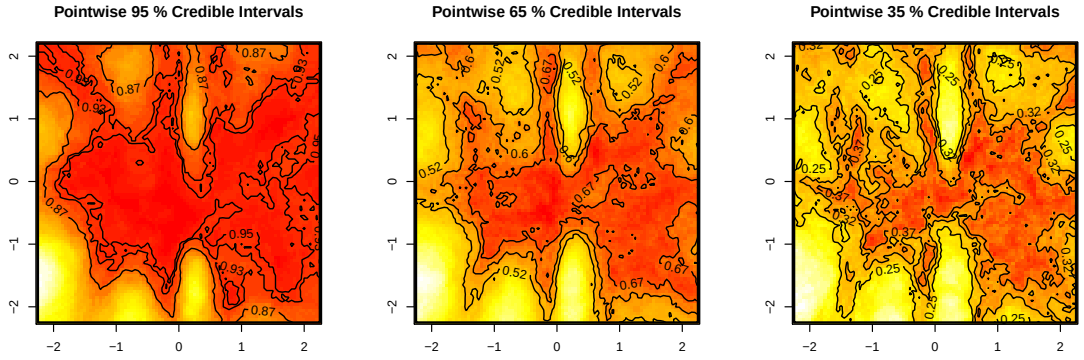

Figure 4: Level curves of the empirical coverage for the pointwise 95%, 65% and 35% credible intervals using the Bayesian model with thin plate splines, $m = 3$, and smoothing parameter λ chosen with the restricted maximum likelihood method. Each value in the level plot is an estimate of ρ_i for the mean coverage of $\eta(\chi_i)$ in the grid $\{\chi_i\}_{i=1}^{N}$ using 200 different simulated data sets and computing the respective credible intervals. Each data set was simulated with $n = 100$, $\sigma^2 = 0.5$ and $\mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$. The target function is described by (31) and plotted in Figure 1. Figure $C.4$ [Rivera, 2017] shows a more complete simulation for the credibility levels $C = 95\%$.

The first and most evident feature about Figure 5 is that for most of the models and methods to choose the smoothing parameters, $\hat{\zeta}$ is close to the nominal value 95% given that the boxes contain the value 0.95. The center 50% of the $\hat{\rho}_i$'s, in most cases, are within 20% points of the nominal value 0.95; or that the center 50% of the $\hat{\rho}_i$'s are in the interval $(0.78, 0.98)$. The few cases shown in this Figure are not enough to explain the distribution of pointwise coverages, so we must observe the more comprehensive Figure $C.5$ [Rivera, 2017].

Let us first consider the cases when the simulated data was generated with $\sigma^2 > 0.1$. We can say that the ζ is close to the nominal value and the center 50% ρ_i's are fairly close the the nominal value as well, around 20% points. For the rest of the ρ_i's, the ones contained in the whiskers, we have a diverse range of cases, some of them are about 30% points from the nominal level while some others have a empirical coverage as low as 25%. The ρ_i's that are considered as outliers could have a value as low as 10% or 0%; of course these ρ_i's correspond to the regions of estimation where we are extrapolating the function far away from the observed data. It may be surprising that the TTPS models provide fairly good pointwise empirical coverage, observe that the whiskers of the boxes for these cases are rarely below 50% and in most of the cases they are above 60%.

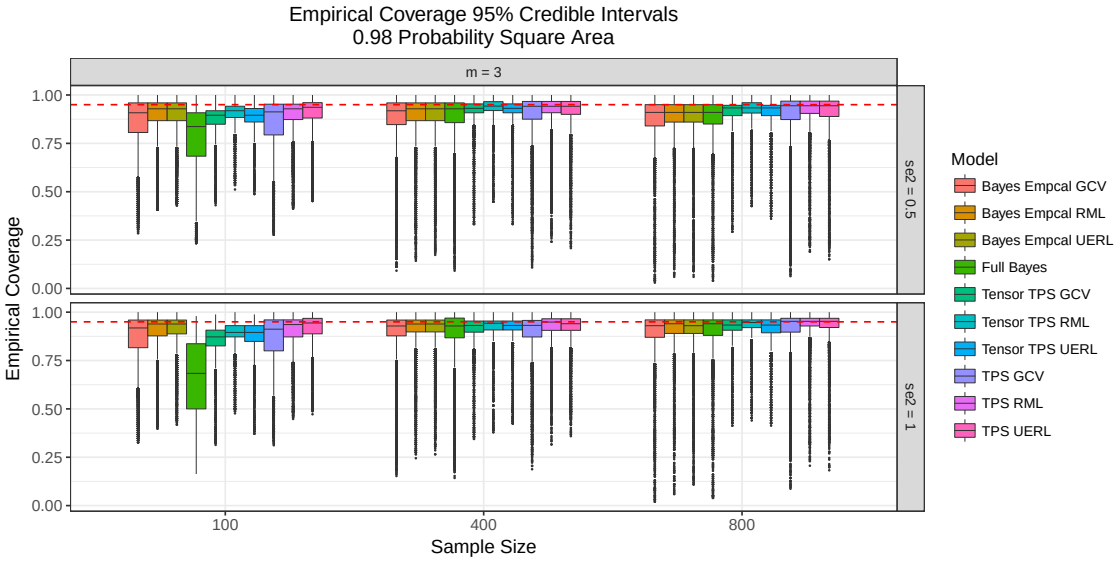

Figure 5: Boxplots partial simulation results, empirical coverage of pointwise 95% credible intervals for prediction of multivariate regression functions. Each box is the summary of the empirical coverages $\{\hat{\rho}_i\}_{i=1}^N$. $\hat{\rho}_i \in [0,1]$ is the empirical coverage of the 95% pointwise credible interval for the prediction of $\eta(\chi_i)$ computed after fitting the model to 200 different simulated data sets. The vectors $\{\chi_i\}_{i=1}^N \subset \mathbb{R}^2$ form a grid of resolution 0.05×0.05 in the square $[-2.5, 2.5]^2$. Complete Results in Figure $C.5$ [Rivera, 2017].

If $m < 4$ and the simulated data were generated with $\sigma^2 \leq 0.1$ there is statistical evidence that ζ is different (smaller) than the nominal value for most of the cases. In the cases in which we do not have evidence to reject $\zeta = 0.95$ against $\zeta < 0.95$, we have large variability in the pointwise empirical coverage. Again, the TTPS models provide less variability on the ρ_i's.

The cases $m = 4$ and $\sigma^2 \leq 0.1$ have their one story in terms of empirical coverage, but in practice we would probably not choose such smooth models with $m = 4$ to predict η because these produce the worst predictions and largest variability in the predictions as we discussed from Figures 2, 3, $C.1$ and $C.2$. Hence in a model selection procedure we would prefer models with $m < 4$.

In general we observe that the ACP, ζ, is close to the nominal value in this simulated example, but we cannot always trust the pointwise empirical coverages to have the nominal levels and, in any of the simulated cases, at most we can expect that half of the pointwise credible intervals have a coverage fairly close to but smaller than the nominal value. Similar results can be observed in plots $C.6$ and $C.7$ [Rivera, 2017] where pointwise 60% and 35% credible intervals for the predictions $\eta(\chi_i)$ were computed and the respective coverages were estimated.

We have described the empirical coverage of the pointwise credible intervals for η estimated over the square $[-2.25, 2.25]^2$. As was mentioned before, the square was chosen because it would contain about 98% of the covariates from the simulated data generated using $N_2(\mathbf{0}, I_2)$. Because of this large area of prediction, we have included in our discussions the credible intervals from extrapolation. Next, we reduce the area of estimation of η and compute the empirical coverages for predictions in centered circles around $\mathbf{0} \in \mathbb{R}^2$. We chose the circles of estimation with radius r in a way that they contain about α% of the covariates $\{\mathbf{x}_i\}_{i=1}^n$ generated with the bivariate standard normal distribution. The radius r and α have the relationship $P(z \leq r^2) = \alpha$ with $z \sim \chi^2(2)$ (chi square distribution).

We analyze the distribution of the empirical coverages $\hat{\rho}_i$'s of the credible intervals in these regions as α changes. Plots 6 and 7 show some of the estimation results. Both sets of

boxplots were computed using $m = 3$ in all models and the data generated have a variance error component of $\sigma^2 = 0.5$, 95% credible intervals for the first graphs and 60% credible intervals for the second graphs. More complete results are shown in the Figures $C.8$ and $C.9$ and $C.10$ [Rivera, 2017].

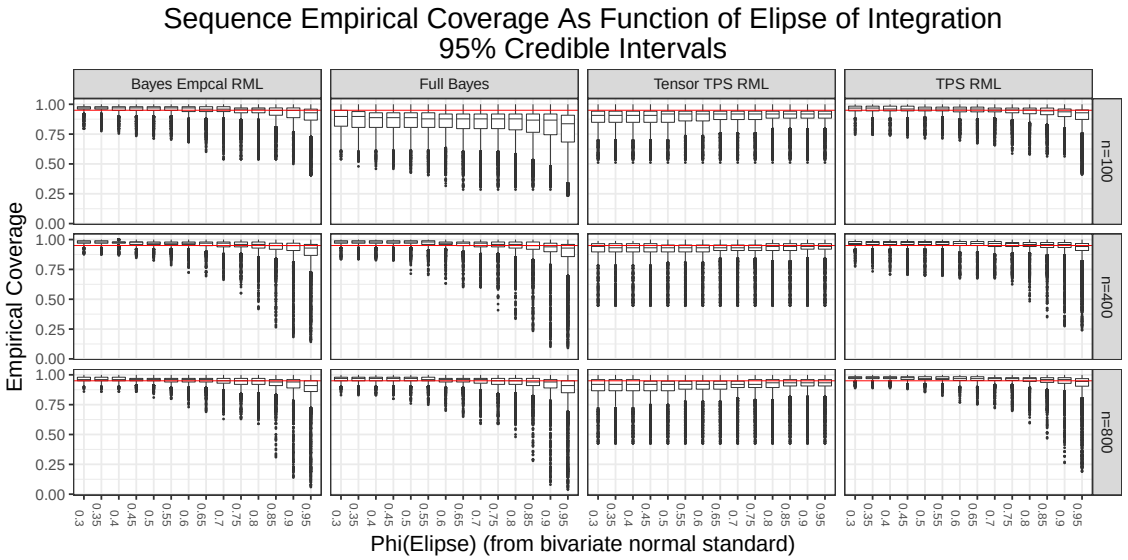

Figure 6: Sequential empirical coverage of pointwise 95% credible intervals. *Ellipse* is the ellipse region in $\mathbb{R}^2$ that would contain about $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution. $\alpha = Phi(Ellipse)$. Each box is the the summary of the empirical coverages $\hat{\rho}_i$'s for 95% pointwise credible intervals for the values of $\eta(\chi_i)$ and χ_i in the ellipse region. The sequence is in the sense of observing the distribution of the $\hat{\rho}_i$'s as α changes. The horizontal red line has a value in the vertical axis of 0.95. The simulated data were obtained using $\sigma^2 = 0.5$ and the models were fitted using $m = 3$.

As one can expect when the region of estimation is smaller and is less likely that we are extrapolating (as α decreases, $\alpha = Phi(Ellipse)$ in figures), the certainty of the prediction increases. This property is inherited from the mean of the full conditional posterior of the process η which was constructed with a non-parametric regression method. The method we are using takes advantage of the unbiased estimation of the thin plate splines and tensor thin plate splines which is reflected in the credible intervals being centered in the respective value $\eta(\chi_i)$. But that the empirical coverages $\hat{\rho}_i$'s of the intervals are closer to the nominal value as $Phi(Elipse)$ decreases is a property of the model as a whole. Observe that the extreme whiskers of the boxplots get closer to the nominal value as $Phi(Elipse)$ decreases, especially for the TTPS and TPS models, implying that the pointwise credible intervals for $\eta(\chi_i)$ have the approximate nominal coverage only for small-medium α. From the box plots, α should be smaller of about 0.5 but α should become smaller as n decreases in order to have a fairly close empirical coverage to the nominal value; it is expected that the non-parametric regression methods estimate incorrectly when n is small. The *Full Bayes* model produces specially under-covering credible intervals when $n = 50$.

We computed and analyzed the empirical coverage of credible intervals with different levels. We only show here and in the appendix the sequence coverages for the 95% and 60% credible intervals but we also computed the same summaries for a finer grid of values for the credibility level C. For the TTPS and TPS models and any method to choose the smoothing parameters we found similar behavior of the empirical coverages regardless of the levels of the intervals when the variance error σ^2 of the response y_i's is not too small ($> 0.1^2$). For $\sigma^2 = 0.1^2$ the intervals from TTPS and TPS models have undercoverage in such degree that not even ζ is close to the nominal value for any value of α, the undercoverage becomes

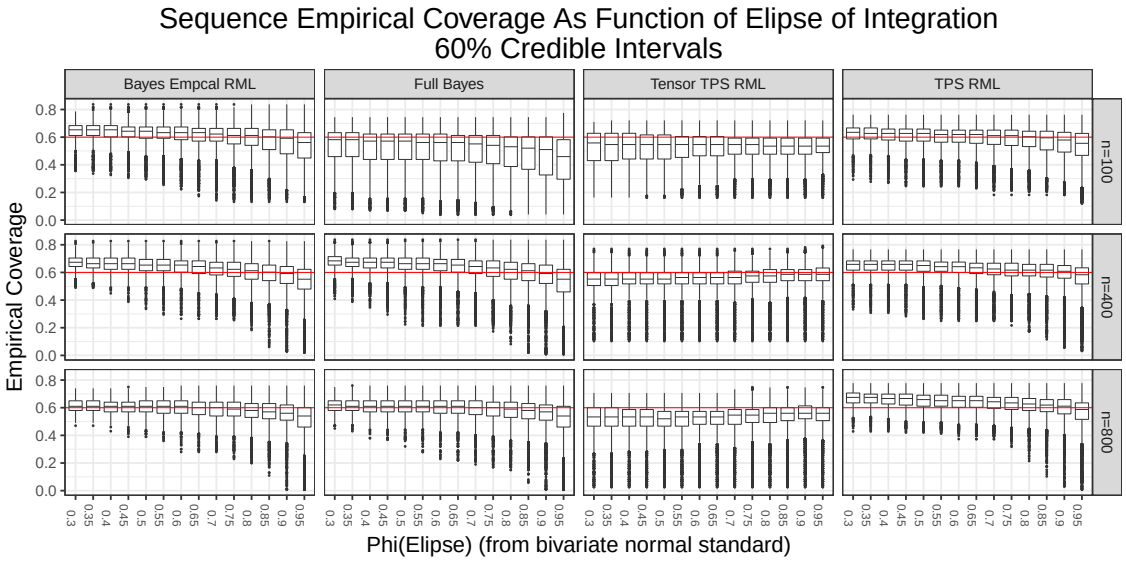

Figure 7: Sequential empirical coverage of pointwise 60% credible intervals. *Ellipse* is the ellipse region in $\mathbb{R}^2$ that would contain about $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution. $\alpha = Phi(Ellipse)$. Each box is the the summary of the empirical coverages $\hat{\rho}_i$'s for 95% pointwise credible intervals for the values of $\eta(\chi_i)$ and χ_i in the ellipse region. The sequence is in the sense of observing the distribution of the $\hat{\rho}_i$'s as α changes. The horizontal red line has a value in the vertical axis of 0.60 indicating the nominal coverage. The simulated data were obtained using $\sigma^2 = 0.5$ and the models were fitted using $m = 3$.

more extreme as σ^2 decreases. We observe a variety of more extreme undercoverage and overcoverage with the pointwise credible intervals and ζ when using the *Empirical Bayes* models and *Full Bayes* models as σ^2 decreases and for moderate or small sample sizes. The behavior of the credible intervals was hinted at but not evident in Figures 5, $C.5$, $C.6$ and $C.7$ [Rivera, 2017]. The apparent reason for the undercoverage of the credible intervals with TPS and TTPS regardless of the region of estimation (size of α) is that when σ^2 is small we have a fairly good estimation of the function η but the model is overconfident about the estimation and the posterior variance of every $\eta(\chi_i)$ is too small. For now, we did not find a way to modify the variability of the posterior predictive to fix the undercoverage of the credible intervals and preserve the interpretation of the full conditional pointwise estimation as a non parametric regression.

5. Conclusions

We reviewed thin plate splines, tensor thin plate splines and linear mixed models to obtain multivariate nonparametric regression methods with normal responses. We reviewed literature for approximation algorithms to faster compute the non parametric regression methods and we used these with four different techniques to choose the bandwidth parameters. We described the frequentist interpretation of all our Bayesian regression methods related to the minimization of a least squared penalized problem; a non-parametric regression method in the frequentist setting. We claim and proved that our estimators of the multivariate regression function, have the same theoretical properties to predict the link functions as the non-parametric regression method. The advantage of our proposed methods is that, besides providing good point estimators, we are able to produce credible intervals for predictions of η. Furthermore we show that our method has the advantage of requiring less storage of the realization of the posterior parameters needed for prediction

in relation to other similar Bayesian models.

We set a large simulation study to describe and compare the performance of all the Bayes models to predict the regression function η of two continuous covariates and to study the coverages of the point credible intervals for η evaluated in different regions of the domain. We varied sample sizes and variance errors in the response.

We found that the TTPS and TPS models produce better predictions for the link function using the mean of the posterior predictive distribution and estimate better the variance error component than the rest of the models. There is no evidence of differences in the performance of predictions within models when using different selections for the bandwidth parameter. We did not find significant difference in the average variability of the predictions around the true function using the variance of the posterior predictive distributions.

The TTPS and TPS models show unbiased estimation of the variance error but only when the sample size is not too small. It is not a surprise that the methods do not perform well for small sample sizes because we are using nonparametric regression methods for the mean of the posterior predictive distribution of η. Poor estimation of the link function induce bad estimation for the variance error parameter. We found as well that for σ^2 to be estimated using any model, it is required that the sample size n increases as σ^2 decreases. In the opposite cases when σ^2 is small and the n is not sufficiently large, the minimum of the realizations of the corresponding posterior distribution were always larger than σ^2. This is equivalent to say, at least from the empirical perspective, that 100% of the empirical credible intervals for σ^2 (not shown in this dissertation) never contained the real value of the variances. It is unfortunate to find this bias in the estimation of the variance produced by all models, but it is a property coming from the rate of covergence of the non parametric methods requiring large sample size.

In the study of the pointwise empirical coverage for the prediction of η we found as well that the average empirical probability ACP is not close to the nominal level when the variance of the errors is too small regardless of the area of prediction. But the TTPS and TPS models seems to produce ACP closer to the nominal value faster than the rest of the models as n increases.

Problem 5 *What is the rate at which the ACP is the nominal value as n increases? does it converge to the nominal value? what happens as* $\sigma^2 \to 0$ *and* $n \to \infty$.

An important feature was observed regarding the coverage of the credible intervals, as these can not be trusted to obtain the nominal level unless the area of prediction is really within the observed covariates (within a radius of 1.2 to the center of the observed covariates in our setting of simulated data, the area corresponds to a region covering about 50% of the data generated from a bivariate standard normal distribution), $n > 100$ and $\sigma^2 > 0.1$. Within this area, about $80\% - 90\%$ of the credible intervals have an empirical coverage close to the nominal value. One cannot expect the credible intervals in areas of extrapolation to have coverages close to the nominal value but it is necessary to predict within the region of observed data in order to have the desired coverage. Even while all models have these properties, the *Full Bayes* model has specially more deviation from the nominal level to the degree that not even the ACP is close to the nominal value. The *Bayes Empirical* seems to have less such deviations than the *Full Bayes* and is similar to the TTPS and TPS models.

The deviation from the nominal coverage of the credible intervals are in both directions with a tendency of the TTPS and TPS models to undercover. The other two models tend to produce large credible intervals such that they have over coverage in the case $n \leq 100$ and to produce under covering credible intervals when $n > 100$.

By the observed results of the simulations, we have provided numerical evidence, at least for the study setting, that the TTPS and TPS models produce better predictions with similar variability on the predictions. The credible intervals for η when using these models preserve the nominal average coverage probability, but the individual intervals do not have the nominal level unless the area of prediction is well within the observed data. We discussed that the TTPS produce better point predictions for η and both methods have similar variability on the predictions. Both models have similar coverages of the credible intervals with the TTPS having a slightly better statistical performance, but as we argued, the difference in the performance of the predictions is not a practical one. We argued that, even when there are statistically significant differences, such differences seem to be of no practical importance: both models produce predictions that detect the general form of the function and the small features of the target function at a similar degree.

The original objective of this chapter was to set the foundations to estimate the link function in a regression problem with errors in the covariates. For the error in the covariates problem, we needed models that can be computed fast and produce good estimation. The proposed model, is ready to be extended solve the regression problem with error in the covariates setting. We found here that the best model to predict in our simulation setting was the TTPS, but this model has the computational disadvantage of requiring to estimate five smoothing parameters while the TPS model requires only one bandwidth parameter. Given our conclusions that there is no practical difference in the performance of the predictions between these models, we pick the TPS model to extend to the regression problem with measurement errors in the covariates.

Acknowledgement
This work was completed as part of Trujillo Rivera's doctoral dissertation at the Department of Statistics, Iowa State University. We thank Alicia Carriquiry, Daniel Norman and Kris De Brabanter for assistance with the development of the results.

A. Auxiliar Results and Proofs

Proposition 6
In the context of Proposition 3, $\xi(\mathbf{x}) \in Im(Q)$.

Proof.
Let $\mathbb{A} = \{c : c^\intercal Q c = 0\}$. Let $\mathbf{x} \in \mathbb{R}^l$ and $\mathbf{c} \in \mathbb{A}$.

Observe that $0 = \mathbf{c}^\intercal Q \mathbf{c} = J(\xi(\mathbf{x})^\intercal \mathbf{c})$ implies $\xi(\mathbf{x})^\intercal \mathbf{c} = 0$ because J is a norm in $span\left\{R_J(\mathbf{z}_i, \cdot)\right\}_{i=1}^k = \mathcal{H}^\star \ominus \mathcal{N}_J$.

Then $\mathbf{c} \perp \xi(\mathbf{x})\, \forall \mathbf{x} \in \mathbb{R}^l$ and $\forall \mathbf{c} \in \mathbb{A}$. Therefore $\xi(\mathbf{x}) \in \mathbb{A}^\perp$.

On another side, from definitions we observe that $\ker(Q) \subset \mathbb{A}$, then $Im(Q) = \ker(Q)^\perp \supseteq \mathbb{A}^\perp$; thus concluding that $\xi(\mathbf{x}) \in Im(Q)$ ■

Proposition 7 *If $\mathcal{A}$ is a strictly convex functional in a Hilbert space H with a local minimum, then $\mathcal{A}$ has a global minimum.*

Proof.
Let η be a minimum of $\mathcal{A}$ and pick $f \in \mathcal{A}$ where $\eta \neq f$.
By definition of local minimum there must be an open set $U \subset \mathcal{H}$ around η such that $\mathcal{A}(\eta) \leq \mathcal{A}(g), \forall g \in U$. We can take $g = \eta + t(f - \eta) = tf + (1-t)\eta$ since for small enough $t > 0$ we have $g \in U$ (we use the completeness of $\mathcal{H}$ as well). Then

$$\begin{aligned} \mathcal{A}(\eta) &\leq \mathcal{A}(g) \\ &= \mathcal{A}\left(tf + (1-t)\eta\right) \end{aligned}$$

$$< t\mathcal{A}(f) + (1-t)\mathcal{A}(\eta)$$

for small $t > 0$, where the last inequality is because $\mathcal{A}$ is strictly convex.
Then $t\mathcal{A}(\eta) < t\mathcal{A}(f)$ and $\mathcal{A}(\eta) < \mathcal{A}(f)$ follows.

Since $f \in \mathcal{H}$ was arbitrary we have shown that η is a global minimum of $\mathcal{A}$, which is unique. ■

Lemma 8 *Let $\{\phi\}_{i=1}^{M}$ be a collection of functions with domain $\mathbb{X} \neq \emptyset$ and range in $\mathbb{R}$. Define $\eta(\mathbf{x}) := \sum_{i=1}^{M} d_i\phi_i(\mathbf{x})$. If $\{d_i\}_{i=1}^{M} \overset{iid}{\sim} N(0,\tau^2)$ then η is a Gaussian Process with mean 0 and $\mathbb{E}(\eta(\mathbf{x})\eta(\mathbf{y})) = \tau^2 \sum_{i=1}^{M} \phi_i(\mathbf{x})\phi_i(\mathbf{y})$.*

Proof.
First we prove that η is indeed a Gaussian process by showing that $\sum_{i=1}^{n} a_i\eta(\mathbf{x}_i)$, $\{a_i\}_{i=1}^{n} \subset \mathbb{R}$ has normal distribution:

$$\begin{aligned}
\sum_{i=1}^{n} a_i\eta(\mathbf{x}_i) &= \sum_{i=1}^{n} a_i \sum_{j=1}^{M} d_j\phi_j(\mathbf{x}_i) \\
&= \sum_{j=1}^{M} d_j \left(\sum_{i=1}^{n} a_i\phi_j(\mathbf{x}_i) \right) \\
&\sim N\left(0, \tau^2 \sum_{j=1}^{M} \left(\sum_{i=1}^{n} a_i\phi_j(\mathbf{x}_i) \right)^2 \right).
\end{aligned}$$

Therefore η is a Gaussian process.

We have as well that $\mathbb{E}(\eta(\mathbf{x})) = \sum_{i=1}^{M} \mathbb{E}(d_i)\phi_i(\mathbf{x}) = 0$ and

$$\begin{aligned}
\mathbb{E}(\eta(\mathbf{x})\eta(\mathbf{y})) &= \mathbb{E}\left(\sum_{i=1}^{M} d_i^2\phi_i(\mathbf{x})\phi_i(\mathbf{y}) + \sum_{i\neq j}^{M} d_i d_j \phi_i(\mathbf{x})\phi_j(\mathbf{y}) \right) \\
&= \mathbb{E}\left(\sum_{i=1}^{M} d_i^2\phi_i(\mathbf{x})\phi_i(\mathbf{y}) \right) + \mathbb{E}\left(\sum_{i\neq j}^{M} d_i d_j \phi_i(\mathbf{x})\phi_j(\mathbf{y}) \right) \\
&= \sum_{i=1}^{M} \mathbb{E}\left(d_i^2\right) \phi_i(\mathbf{x})\phi_i(\mathbf{y}) \\
&= \sum_{i=1}^{M} \tau^2 \phi_i(\mathbf{x})\phi_i(\mathbf{y}).
\end{aligned}$$

■

Lemma 9 *Let $\{\psi\}_{i=1}^{n}$ be a collection of functions with domain $\mathbb{X} \neq \emptyset$ and range in $\mathbb{R}$. Let $\eta(\mathbf{x}) := \sum_{i=1}^{n} c_i\psi_i(\mathbf{x})$. If $(c_1 \dots c_n)^\intercal \sim N_n(\mathbf{0}, \mathbf{\Sigma})$ then η is a Gaussian Process with mean 0 and*

$$\mathbb{E}(\eta(\mathbf{x})\eta(\mathbf{y})) = (\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x}))\, \mathbf{\Sigma}\, (\psi_1(\mathbf{y}) \dots \psi_n(\mathbf{y}))^\intercal .$$

Proof.
Let $(\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x}))^\intercal = \mathbf{\Sigma}^{\frac{1}{2}} (\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x}))^\intercal$ where $\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Sigma}$. We have that

$$\begin{aligned}
\eta(\mathbf{x}) &= \sum_{i=1}^{n} c_i\psi_i(\mathbf{x}) \\
&= (c_1 \dots c_n)\, (\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x}))^\intercal
\end{aligned}$$

$$
\begin{aligned}
&= (c_1 \dots c_n)\, \boldsymbol{\Sigma}^{-\frac{1}{2}} (\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x}))^\intercal \\
&= \left(\boldsymbol{\Sigma}^{-\frac{1}{2}} (c_1 \dots c_n)^\intercal\right)^\intercal (\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x}))^\intercal \\
&= (d_1 \dots d_n)\, (\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x}))^\intercal \\
&= \sum_{i=1}^{n} d_i \phi_i(\mathbf{x}),
\end{aligned}
$$

where $(d_1 \dots d_n)^\intercal = \boldsymbol{\Sigma}^{-\frac{1}{2}} (c_1 \dots c_n)^\intercal$ and thus $(d_1 \dots d_n)^\intercal \sim N_n(\mathbf{0}, \mathbf{I}_n)$. Then by Lemma 8, η is a Gaussian process, with mean 0 and

$$
\begin{aligned}
\mathbb{E}(\eta(\mathbf{x})\eta(\mathbf{y})) &= \sum_{i=1}^{n} \phi_i(\mathbf{x})\phi_i(\mathbf{y}) \\
&= (\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x}))\, (\phi_1(\mathbf{y}) \dots \phi_n(\mathbf{y}))^\intercal \\
&= (\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x}))\, \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\Sigma}^{\frac{1}{2}} (\psi_1(\mathbf{y}) \dots \psi_n(\mathbf{y}))^\intercal \\
&= (\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x}))\, \boldsymbol{\Sigma}\, (\psi_1(\mathbf{y}) \dots \psi_n(\mathbf{y}))^\intercal .
\end{aligned}
$$

■

Proposition 10 *[Full Conditional Posterior of Coefficients $\binom{\mathbf{d}}{\mathbf{c}}$. Version 1.] For the observed pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ consider the model with $\sigma^2 > 0$ known given by*

$$
\begin{aligned}
y_i &= \sum_{j=1}^{k} d_j \phi_j(\mathbf{x}_i) + \sum_{j=1}^{l} d_j \psi_j(\mathbf{x}_i) + \epsilon_i, \\
\epsilon_i &\overset{iid}{\sim} N_1(0, \sigma^2),
\end{aligned}
$$

where $\{\phi\}_{i=1}^k$ and $\{\psi\}_{i=1}^l$ are known functions. For $b > 0, \tau > 0$ and Q a $l \times l$ positive definite matrix, consider the priors

$$
\begin{aligned}
d_i &\overset{iid}{\sim} N_1(0, \tau^2) \\
(c_1\, c_2\, \cdots\, c_l)^\intercal &\sim N_l(\mathbf{0}, b\, Q) \\
(d_1\, d_2\, \cdots\, d_k)^\intercal &\perp (c_1\, c_2\, \cdots\, c_l)^\intercal \\
(d_1\, \cdots\, d_k\, c_1\, \cdots\, c_l)^\intercal &\perp (\epsilon_1 \cdots \epsilon_n)^\intercal .
\end{aligned}
$$

The posterior of $\binom{\mathbf{d}}{\mathbf{c}}$ is $N_{k+l}\left(\mu_{\mathbf{dc}_\rho}, b\boldsymbol{\Sigma}_{\mathbf{dc}_\rho}\right)$ where

$$
\mu_{\mathbf{dc}_\rho} = \begin{pmatrix} \rho S^\intercal (\rho S S^\intercal + M)^{-1} \\ Q R^\intercal (\rho S S^\intercal + M)^{-1} \end{pmatrix} \mathbf{y} \tag{37}
$$

$$
\boldsymbol{\Sigma}_{\mathbf{dc}_\rho} = \begin{pmatrix} \left(\rho I_k - \rho S^\intercal (\rho S S^\intercal + M)^{-1} \rho S\right) & \left(-\rho S^\intercal (\rho S S^\intercal + M)^{-1} R Q\right) \\ \left(-Q R^\intercal (\rho S S^\intercal + M)^{-1} \rho S\right) & \left(Q - Q R^\intercal (\rho S S^\intercal + M)^{-1} R Q\right) \end{pmatrix}. \tag{38}
$$

and S is a $n \times l$ matrix with entry $S_{i,j} = \phi_j(\mathbf{x}_i)$, R is a $n \times l$ matrix with entries $R_{i,j} = \psi_j(\mathbf{x}_i)$, $M = RQR^\intercal + n\lambda I_n$, $n\lambda = \frac{\sigma^2}{b}$ and $\rho = \frac{\tau^2}{b}$.

Proof.
Observe that

$$
\mathbb{E}(\mathbf{d}) = \mathbf{0}_k,
$$

$$
\begin{aligned}
\mathbb{E}(\mathbf{c}) &= \mathbf{0}_l, \\
\mathbb{E}(\mathbf{y}) &= \mathbb{E}(S\mathbf{d} + R\mathbf{c} + \epsilon) = S\,\mathbb{E}(\mathbf{d}) + R\,\mathbb{E}(\mathbf{c}) + \mathbb{E}(\epsilon) = \mathbf{0}_n, \\
Var(\mathbf{y}) &= Var(S\mathbf{d} + R\mathbf{c} + \epsilon) = SVar(\mathbf{d})\,S^\intercal + RVar(\mathbf{c})\,R^\intercal + Var(\epsilon) \\
&= \tau^2 SS^\intercal + bRQR^\intercal + \sigma^2 I_n, \\
Var\left[\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}\right] &= \begin{pmatrix}\tau^2 I_k & \mathbf{0}_{k\times l}\\ \mathbf{0}_{l\times k} & bQ\end{pmatrix},
\end{aligned}
$$

$$
\begin{aligned}
Cov(y_i, d_j) &= Cov\left(\sum_{\nu=1}^{k} d_\nu \phi_\nu(\mathbf{x}_i) + \sum_{\nu=1}^{l} d_\nu \psi_\nu(\mathbf{x}_i) + \epsilon_i, d_j\right) \\
&= \phi_j(\mathbf{x}_i)\,Cov(d_j, d_j) = \tau^2 \phi_j(\mathbf{x}_i) = \tau^2 S_{i,j}, \\
Cov(y_i, c_j) &= Cov\left(\sum_{\nu=1}^{k} d_\nu \phi_\nu(\mathbf{x}_i) + \sum_{\nu=1}^{l} c_\nu \psi_\nu(\mathbf{x}_i) + \epsilon_i, c_j\right) \\
&= \sum_{\nu=1}^{l} \psi_\nu(\mathbf{x}_i)\,Cov(c_\nu, c_j) = b\sum_{\nu=1}^{l} R_{i,\nu} Q_{\nu,j} = b(RQ)_{i,j}, \\
\pi\left[\begin{pmatrix}\mathbf{y}\\ \mathbf{d}\\ \mathbf{c}\end{pmatrix}\right] &= \pi\left[\mathbf{y} \mid \begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}\right]\pi\left[\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}\right],
\end{aligned}
$$

then $\pi\left[\begin{pmatrix}\mathbf{y}\\ \mathbf{d}\\ \mathbf{c}\end{pmatrix}\right] = N_{n+l+k}(\mathbf{0}, \mathbf{\Sigma}_{ydc})$ where

$$
\mathbf{\Sigma}_{ydc} = \begin{pmatrix} (\tau^2 SS^\intercal + bRQR^\intercal + \sigma^2 I_n) & \begin{pmatrix}\tau^2 S & bRQ\end{pmatrix} \\ \begin{pmatrix}\tau^2 S^\intercal \\ bQR^\intercal\end{pmatrix} & \begin{pmatrix}\tau^2 I_k & \mathbf{0}\\ \mathbf{0} & bQ\end{pmatrix} \end{pmatrix}.
$$

By a known result for multivariate normal distributions ([Bilodeau and Brenner, 2008]), we obtain that the posterior distribution $\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right) \mid \mathbf{y}\right]$ is normal with mean and covariance described below:

$$
\begin{aligned}
\mathbb{E}\left[\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} \mid \mathbf{y}\right] &= \begin{pmatrix}\tau^2 S^\intercal \\ bQR^\intercal\end{pmatrix}(\tau^2 SS^\intercal + bRQR^\intercal + \sigma^2 I_n)^{-1}\mathbf{y} \\
&= \begin{pmatrix}\frac{\tau^2}{b} S^\intercal \\ QR^\intercal\end{pmatrix}\left(\frac{\tau^2}{b} SS^\intercal + RQR^\intercal + \frac{\sigma^2}{b} I_n\right)^{-1}\mathbf{y} \\
&= \begin{pmatrix}\rho S^\intercal \\ QR^\intercal\end{pmatrix}(\rho SS^\intercal + M)^{-1}\mathbf{y} \\
&= \begin{pmatrix}\rho S^\intercal (\rho SS^\intercal + M)^{-1} \\ QR^\intercal (\rho SS^\intercal + M)^{-1}\end{pmatrix}\mathbf{y},
\end{aligned}
$$

and

$$
\begin{aligned}
Var\left[\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} \mid \mathbf{y}\right] =& \begin{pmatrix}\tau^2 I_k & \mathbf{0}\\ \mathbf{0} & bQ\end{pmatrix} \\
& - \begin{pmatrix}\tau^2 S^\intercal \\ bQR^\intercal\end{pmatrix}(\tau^2 SS^\intercal + bRQR^\intercal + \sigma^2 I_n)^{-1}\begin{pmatrix}\tau^2 S & bRQ\end{pmatrix} \\
=& b\begin{pmatrix}\frac{\tau^2}{b} I_k & \mathbf{0}\\ \mathbf{0} & Q\end{pmatrix} - b\begin{pmatrix}\frac{\tau^2}{b} S^\intercal \\ QR^\intercal\end{pmatrix}\left(\frac{\tau^2}{b} SS^\intercal + M\right)^{-1}\begin{pmatrix}\frac{\tau^2}{b} S & RQ\end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
=& b\begin{pmatrix} \rho I_k & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix} - b\begin{pmatrix} \rho S^\intercal \\ QR^\intercal \end{pmatrix}(\rho SS^\intercal + M)^{-1}\begin{pmatrix} \rho S & RQ \end{pmatrix} \\
=& b\left[\begin{pmatrix} \rho I_k & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix} - \begin{pmatrix} \rho S^\intercal (\rho SS^\intercal + M)^{-1}\begin{pmatrix} \rho S & RQ \end{pmatrix} \\ QR^\intercal (\rho SS^\intercal + M)^{-1}\begin{pmatrix} \rho S & RQ \end{pmatrix} \end{pmatrix}\right] \\
=& b\left\{\begin{pmatrix} \rho I_k & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix} - \begin{pmatrix} \left[\rho S^\intercal (\rho SS^\intercal + M)^{-1}\rho S\right] & \left[\rho S^\intercal (\rho SS^\intercal + M)^{-1} RQ\right] \\ \left[QR^\intercal (\rho SS^\intercal + M)^{-1}\rho S\right] & \left[QR^\intercal (\rho SS^\intercal + M)^{-1} RQ\right] \end{pmatrix}\right\} \\
=& b\begin{pmatrix} \left[\rho I_k - \rho S^\intercal (\rho SS^\intercal + M)^{-1}\rho S\right] & \left[-\rho S^\intercal (\rho SS^\intercal + M)^{-1} RQ\right] \\ \left[-QR^\intercal (\rho SS^\intercal + M)^{-1}\rho S\right] & \left[Q - QR^\intercal (\rho SS^\intercal + M)^{-1} RQ\right] \end{pmatrix}.
\end{aligned}
$$

■

Lemma 11
Suppose M is a symmetric and nonsingular matrix and S is a full column rank matrix, then

$$
\begin{aligned}
\lim_{\rho\to\infty} (\rho SS^\intercal + M)^{-1} &= M^{-1} - M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1} \\
\lim_{\rho\to\infty} \rho S^\intercal (\rho SS^\intercal + M)^{-1} &= \left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1} \\
\lim_{\rho\to\infty} \rho I - \rho^2 S^\intercal (\rho SS^\intercal + M)^{-1}S^\intercal &= \left(S^\intercal M^{-1}S\right)^{-1}.
\end{aligned}
$$

Proof.
We describe the idea of the proof; the details can be found in [Wahba, 1978, p. 367] and [Wahba, 1983, p. 137]. The proof is based on the identity

$$
(\rho SS^\intercal + M)^{-1} = M^{-1} - M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}\left(I + \rho^{-1}\left(S^\intercal M^{-1}S\right)^{-1}\right)^{-1}S^\intercal M^{-1} \tag{39}
$$

which can be proven directly. Using (39) with some algebra and taking the limit $\rho \to \infty$ the lemma follows easily. ■

Proposition 12 *[Full Conditional Posterior of Coefficients $\binom{\mathbf{d}}{\mathbf{c}}$. Version 2.]*
Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be observed, with $\sigma^2 > 0$ known, and set $\mathbf{d} := (d_1\, d_2\, \cdots\, d_l)^\intercal$, $\mathbf{c} := (c_1\, c_2\, \cdots\, c_k)^\intercal$. Consider the model

$$
\begin{aligned}
y_i &= \sum_{j=1}^{l} d_j\phi_j(\mathbf{x}_i) + \sum_{j=1}^{k} c_j\psi_j(\mathbf{x}_i) + \epsilon_i, \\
\epsilon_i &\overset{iid}{\sim} N_1(0, \sigma^2),
\end{aligned}
$$

where $\{\phi\}_{i=1}^l$ and $\{\psi\}_{i=1}^k$ are known functions. For $b > 0, \tau > 0$ and $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ positive definite matrix, consider the priors

$$
\begin{aligned}
d_i &\overset{iid}{\sim} 1 \\
\mathbf{c} &\sim N_l(\mathbf{0}, b\,Q) \\
\mathbf{d} &\perp \mathbf{c} \\
\tbinom{\mathbf{d}}{\mathbf{c}} &\perp (\epsilon_1 \cdots \epsilon_n)^\intercal.
\end{aligned}
$$

Then the posterior of $\binom{\mathbf{d}}{\mathbf{c}}$ is $N_{l+k}(\mu_{\mathbf{dc}}, b\boldsymbol{\Sigma}_{\mathbf{dc}})$ where

$$
\mu_{\mathbf{dc}} = \begin{pmatrix} (S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1} \\ QR^\intercal M^{-1}(I - S(S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1}) \end{pmatrix}\mathbf{y} \tag{40}
$$

$$\boldsymbol{\Sigma}_{\mathbf{dc}} = \begin{pmatrix} (S^\intercal M^{-1} S)^{-1} & -(S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1} RQ \\ -QR^\intercal M^{-1} S (S^\intercal M^{-1} S)^{-1} & Q - QR^\intercal \{M^{-1} - M^{-1} S (S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1}\} RQ \end{pmatrix} \tag{41}$$

and $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ *with entry* $S_{i,j} = \phi_j(\mathbf{x}_i)$, $R \in \mathcal{M}_{n\times k}(\mathbb{R})$ *with entries* $R_{i,j} = \psi_j(\mathbf{x}_i)$, $M = RQR^\intercal + n\lambda I_n$, $n\lambda = \frac{\sigma^2}{b}$.

Proof.
First we compute the limits as $\rho \to \infty$ of the mean and covariance matrix (37) and (38). Introducing the limits in the column vector and using Lemma 11 we have

$$\begin{aligned} \lim_{\rho\to\infty} \mu_{\mathbf{dc}_\rho} &= \lim_{\rho\to\infty} \begin{pmatrix} \rho S^\intercal \left(\rho SS^\intercal + M\right)^{-1} \\ QR^\intercal \left(\rho SS^\intercal + M\right)^{-1} \end{pmatrix} \mathbf{y} \\ &= \begin{pmatrix} \lim_{\rho\to\infty} \rho S^\intercal \left(\rho SS^\intercal + M\right)^{-1} \\ \lim_{\rho\to\infty} QR^\intercal \left(\rho SS^\intercal + M\right)^{-1} \end{pmatrix} \mathbf{y} \\ &= \begin{pmatrix} \left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1} \\ QR^\intercal M^{-1} \left(I_n - S\left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1}\right) \end{pmatrix} \mathbf{y} \\ &= \mu_{\mathbf{dc}}, \end{aligned} \tag{42}$$

while for the covariance, matrix we proceed in a similar way introducing the limits inside the matrix and using Lemma 11:

$$\begin{aligned} \lim_{\rho\to\infty} \boldsymbol{\Sigma}_{\mathbf{dc}_\rho} &= \lim_{\rho\to\infty} b \begin{pmatrix} \left(\rho I_k - \rho S^\intercal \left(\rho SS^\intercal + M\right)^{-1} \rho S\right) & \left(-\rho S^\intercal \left(\rho SS^\intercal + M\right)^{-1} RQ\right) \\ \left(-QR^\intercal \left(\rho SS^\intercal + M\right)^{-1} \rho S\right) & \left(Q - QR^\intercal \left(\rho SS^\intercal + M\right)^{-1} RQ\right) \end{pmatrix} \\ &= b \begin{pmatrix} (S^\intercal M^{-1} S)^{-1} & -(S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1} RQ \\ -QR^\intercal M^{-1} S (S^\intercal M^{-1} S)^{-1} & Q - QR^\intercal \{M^{-1} - M^{-1} S (S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1}\} RQ \end{pmatrix} \\ &= b\boldsymbol{\Sigma}_{\mathbf{dc}}. \end{aligned} \tag{43}$$

On the other side, the characteristic function of a multivariate normal distribution with mean $\mu_{\mathbf{dc}_\rho}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{dc}_\rho}$ is $\phi_\rho(\mathbf{t}) = \exp\left(i\mathbf{t}^\intercal \mu_{\mathbf{dc}_\rho} + \frac{1}{2} b\mathbf{t}^\intercal \boldsymbol{\Sigma}_{\mathbf{dc}_\rho} \mathbf{t}\right)$ and $\mathbf{t} \in \mathbb{R}^{k+l}$. For fixed $\mathbf{t}$, taking the limits and using (42) and (43) we have that

$$\begin{aligned} \lim_{\rho\to\infty} \phi_\rho(\mathbf{t}) &= \lim_{\rho\to\infty} \exp\left(i\mathbf{t}^\intercal \mu_{\mathbf{dc}_\rho} + \frac{1}{2} b\mathbf{t}^\intercal \boldsymbol{\Sigma}_{\mathbf{dc}_\rho} \mathbf{t}\right) \\ &= \exp\left(i\mathbf{t}^\intercal \left(\lim_{\rho\to\infty} \mu_{\mathbf{dc}_\rho}\right) + \frac{1}{2} b\mathbf{t}^\intercal \left(\lim_{\rho\to\infty} \boldsymbol{\Sigma}_{\mathbf{dc}_\rho}\right) \mathbf{t}\right) \\ &= \exp\left(i\mathbf{t}^\intercal \mu_{\mathbf{dc}} + \frac{1}{2} \mathbf{t}^\intercal \left(b\boldsymbol{\Sigma}_{\mathbf{dc}}\right) \mathbf{t}\right), \end{aligned} \tag{44}$$

where expression (44) is the characteristic function of a random vector with distribution $N_{k+l}\left(\mu_{\mathbf{dc}}, b\boldsymbol{\Sigma}_{\mathbf{dc}}\right)$. By the Lévy's continuity Theorem ([Athreya and Lahiri, 2006]), ([Fristedt and Gray, 2013]), a sequence of random vectors $\left\{\binom{\mathbf{d}}{\mathbf{c}}_\rho\right\}_{\rho=1}^{\infty}$ such $\binom{\mathbf{d}}{\mathbf{c}}_\rho \sim N_{k+l}\left(\mu_{\mathbf{dc}_\rho}, b\boldsymbol{\Sigma}_{\mathbf{dc}_\rho}\right)$ converge in distribution as $\rho \to \infty$ to $\binom{\mathbf{d}}{\mathbf{c}} \sim N_{k+l}\left(\mu_{\mathbf{dc}}, b\boldsymbol{\Sigma}_{\mathbf{dc}}\right)$. We have shown that if in the prior of $d_i \overset{iid}{\sim} N_1(0, \tau^2)$ in Lemma 10 we let $\frac{\tau^2}{b} = \rho \to \infty$, the posterior distribution (or the cumulative distribution function) of $\binom{\mathbf{d}}{\mathbf{c}}_\rho$ converges to $N_{k+l}\left(\mu_{\mathbf{dc}}, b\boldsymbol{\Sigma}_{\mathbf{dc}}\right)$. Therefore, if we take $d_i \overset{iid}{\sim} 1$ as prior for d_i the posterior $\left[\binom{\mathbf{d}}{\mathbf{c}} \middle| \mathbf{y}\right]$ is $N_{k+l}\left(\mu_{\mathbf{dc}}, b\boldsymbol{\Sigma}_{\mathbf{dc}}\right)$. ■

Proposition 13 (Equivalence Bayesian Models Auxiliary)
In the context of Lemma 12 we have $\binom{\mathbf{d}}{\mathbf{c}}|\mathbf{y} \sim N_{l+k}\left(\mu_{\mathbf{dc}}, \boldsymbol{\Sigma}_{\mathbf{dc}}\right)$, $\mu_{\mathbf{dc}}$ *and* $\boldsymbol{\Sigma}_{\mathbf{dc}}$ *described by* (40) *and* (41). *Define* $\eta : \mathbb{X} \to \mathbb{R}$ *as*

$$\eta(x) := \sum_{j=1}^{l} d_j \phi_j(x) + \sum_{j=1}^{k} c_j \psi_j(x)$$

then $\eta|\mathbf{y}$ *is a Gaussian process with mean and covariance*

$$\mathbb{E}\left[\eta(x)|\mathbf{y}\right] = \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}} \tag{45}$$

$$\begin{aligned} b^{-1}Cov\left(\eta(x),\eta(y)|\mathbf{y}\right) =& \Psi(x)^{\intercal}Q\Psi(y) + \Phi(x)^{\intercal}\left(S^{\intercal}M^{-1}S^{\intercal}\right)^{-1}\Phi(x) \\ & - \left[\Phi(x)^{\intercal}\tilde{\mathbf{d}}(y) + \Phi(y)^{\intercal}\tilde{\mathbf{d}}(x)\right] - \Psi(x)^{\intercal}\tilde{\mathbf{c}}(y) \end{aligned} \tag{46}$$

where

$$\begin{aligned} \Phi(x) &= (\phi_1(x)\cdots\phi_l(x))^{\intercal}, \\ \Psi(x) &= (\psi_1(x)\cdots\psi_k(x))^{\intercal}, \\ \tilde{\mathbf{d}}(x) &= \left(S^{\intercal}M^{-1}S\right)^{-1}S^{\intercal}M^{-1}RQ\Psi(x), \\ \tilde{\mathbf{c}}(x) &= QR^{\intercal}\left(M^{-1} - M^{-1}S\left(S^{\intercal}M^{-1}S\right)^{-1}S^{\intercal}M^{-1}\right)RQ\Psi(x). \end{aligned}$$

In particular

$$b^{-1}Var\left(\eta(x)|\mathbf{y}\right) = \Psi(x)^{\intercal}Q\Psi(x) + \Phi(x)^{\intercal}\left(S^{\intercal}M^{-1}S^{\intercal}\right)^{-1}\Phi(x) - 2\Phi(x)^{\intercal}\tilde{\mathbf{d}}(x) - \Psi(x)^{\intercal}\tilde{\mathbf{c}}(x). \tag{47}$$

Proof.
Consider

$$\begin{aligned} \begin{pmatrix}\mathbf{d}^{\star}\\ \mathbf{c}^{\star}\end{pmatrix} &= \begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} - \mu_{\mathbf{dc}} \\ \eta^{\star}(x) &= \sum_{j=1}^{l} d_j^{\star}\phi_j(x) + \sum_{j=1}^{k} c_j^{\star}\psi_j(x). \end{aligned}$$

By Lemma 9, $\eta^{\star}$ is a Gaussian process with mean 0 and covariance $Cov\left(\eta^{\star}(x),\eta^{\star}(y)\right) = \left(\begin{smallmatrix}\Phi(x)\\ \Psi(x)\end{smallmatrix}\right)^{\intercal}\mathbf{\Sigma_{dc}}\left(\begin{smallmatrix}\Phi(y)\\ \Psi(y)\end{smallmatrix}\right)$. Therefore, we have

$$\begin{aligned} b^{-1}Cov\left(\eta(x),\eta(y)\right) =& \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\begin{pmatrix}(S^{\intercal}M^{-1}S)^{-1} & -(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}RQ \\ -QR^{\intercal}M^{-1}S(S^{\intercal}M^{-1}S)^{-1} & Q - QR^{\intercal}\{M^{-1} - M^{-1}S(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}\}RQ\end{pmatrix}\begin{pmatrix}\Phi(y)\\ \Psi(y)\end{pmatrix} \\ =& \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}\Phi(y) - \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}RQ\Psi(y) \\ & - \Psi(x)^{\intercal}QR^{\intercal}M^{-1}S(S^{\intercal}M^{-1}S)^{-1}\Phi(y) \\ & + \Psi(x)^{\intercal}\left[Q - QR^{\intercal}\{M^{-1} - M^{-1}S(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}\}RQ\right]\Psi(y) \\ =& \Psi(x)^{\intercal}Q\Psi(y) + \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}\Phi(y) - \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}RQ\Psi(y) \\ & - \left[\Phi(y)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}RQ\Psi(x)\right]^{\intercal} \\ & - \Psi(x)QR^{\intercal}\left(M^{-1} - M^{-1}S\left(S^{\intercal}M^{-1}S\right)^{-1}S^{\intercal}M^{-1}\right)RQ\Psi(y) \\ =& \Psi(x)^{\intercal}Q\Psi(y) + \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}\Phi(y) - \phi(x)^{\intercal}\tilde{\mathbf{d}}(y) - \left[\phi(y)^{\intercal}\tilde{\mathbf{d}}(x)\right]^{\intercal} - \Psi(x)^{\intercal}\tilde{\mathbf{c}}(y) \\ =& \Psi(x)^{\intercal}Q\Psi(y) + \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}\Phi(y) - \left[\phi(x)^{\intercal}\tilde{\mathbf{d}}(y) + \phi(y)^{\intercal}\tilde{\mathbf{d}}(x)\right] - \Psi(x)^{\intercal}\tilde{\mathbf{c}}(y). \end{aligned}$$

Since $\eta(x) = \eta^{\star}(x) + \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}$ we have

$$\mathbb{E}\left[\eta(x)|\mathbf{y}\right] = \mathbb{E}\left[\eta^{\star}(x) + \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}\middle|\mathbf{y}\right] = \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}$$

and

$$Cov\left(\eta(x),\eta(y)|\,\mathbf{y}\right) = Cov\left(\eta^{\star}(x) + \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}, \eta^{\star}(y) + \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}\middle|\,\mathbf{y}\right) = Cov\left(\eta^{\star}(x),\eta^{\star}(y)|\mathbf{y}\right).$$

The expression for $Var(\eta|\,\mathbf{y})$ follows directly ■

References

[Akhiezer and Glazman, 1981a] Akhiezer, N. I. and Glazman, I. M. (1981a). *Theory of linear operators in Hilbert space Volume 1*. Boston: Pitman Pub.

[Akhiezer and Glazman, 1981b] Akhiezer, N. I. and Glazman, I. M. (1981b). *Theory of linear operators in Hilbert space Volume 2*. Boston: Pitman Pub.

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

[Athreya and Lahiri, 2006] Athreya, K. B. and Lahiri, S. N. (2006). *Measure theory and probability theory*. Springer Science & Business Media.

[Barry et al., 1986] Barry, D. et al. (1986). Nonparametric bayesian regression. *The Annals of Statistics*, 14(3):934–953.

[Berry et al., 2002] Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97(457):160–169.

[Bilodeau and Brenner, 2008] Bilodeau, M. and Brenner, D. (2008). *Theory of multivariate statistics*. Springer Science & Business Media.

[Chen, 1993] Chen, Z. (1993). Fitting multivariate regression functions by interaction spline models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 473–491.

[Duchon, 1977] Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer.

[Fristedt and Gray, 2013] Fristedt, B. E. and Gray, L. F. (2013). *A modern approach to probability theory*. Springer Science & Business Media.

[Gelman et al., 2014] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

[Geweke et al., 1991] Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

[Golub and Van Loan, 2012] Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

[Gu, 2013] Gu, C. (2013). *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media.

[Gu and Kim, 2002] Gu, C. and Kim, Y.-J. (2002). Penalized likelihood regression: general formulation and efficient approximation. *Canadian Journal of Statistics*, 30(4):619–628.

[Gu and Qiu, 1993] Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *The Annals of Statistics*, pages 217–234.

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

[Gu and Wahba, 1993b] Gu, C. and Wahba, G. (1993b). Smoothing spline anova with component-wise bayesian "confidence intervals". *Journal of Computational and Graphical Statistics*, 2(1):97–117.

[Henderson, 1973] Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, 1973(Symposium):10–41.

[Hoffman and Kunze, 1990] Hoffman, K. and Kunze, R. (1990). Linear algebra, 2nd.

[Kim and Gu, 2004] Kim, Y.-J. and Gu, C. (2004). Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):337–356.

[Kimeldorf and Wahba, 1971] Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

[Li, 1986] Li, K.-C. (1986). Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, pages 1101–1112.

[Mallows, 1973] Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.

[Meinguet, 1979] Meinguet, J. (1979). Multivariate interpolation at arbitrary points made simple. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 30(2):292–304.

[Nychka, 1988] Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, 83(404):1134–1143.

[Rivera, 2017] Rivera, E. A. T. (2017). *Nonparametric Regression Models With and Without Measurement Error in the Covariates for Univariate and Vector Responses: A Bayesian Approach*. PhD thesis, Iowa State University.

[Robinson, 1991] Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, pages 15–32.

[Ruppert et al., 2003] Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.

[Schölkopf et al., 2001] Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer.

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

[Wahba, 1983] Wahba, G. (1983). Bayesian" confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–150.

[Wahba, 1985] Wahba, G. (1985). A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, pages 1378–1402.

[Wahba, 1987] Wahba, G. (1987). Partial and interaction spline models for the semiparametric estimation of functions of several variables.

[Wahba and Craven, 1978] Wahba, G. and Craven, P. (1978). Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–404.

[Wahba and Wendelberger, 1980] Wahba, G. and Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly weather review*, 108(8):1122–1143.

[Wecker and Ansley, 1983] Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, 78(381):81–89.

[Weidmann, 1980] Weidmann, J. (1980). Linear operators in hilbert spaces.

[Wood, 2003] Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.