A BRIEF TREATISE ON BAYESIAN INVERSE REGRESSION

SUPPLEMENTARY MATERIAL

DEBASHIS CHATTERJEE

Indian Statistical Institute,
203, B. T. Road, Kolkata, India.

Contents

1

Supplement to: A Statistical Perspective on Inverse and Inverse Regression Problems

1.1 Traditional inverse problem

1.1.1 Examples of inverse problems

Vertical seismic profiling

In this scientific field, one wishes to learn about the vertical seismic velocity of the material surrounding a borehole. A source generates downward-propagating seismic wavefront at the surface, and in the borehole, a string of seismometers sense these seismic waves. The arrival times of the seismic wavefront at each instrument are measured from the recorded seismograms. These times provide information on the seismic velocity

for vertically traveling waves as a function of depth. The problem is nonlinear if it is expressed in terms of seismic velocities. However, we can linearize this problem via a simple change of variables, as follows. Letting z denote the depth, it is possible to parameterize the seismic structure in terms of slowness, $s(z)$, which is the reciprocal of the velocity $v(z)$. The observed travel time at depth z can then be expressed as:

$$t(z) = \int_0^z s(u)du = \int_0^\infty s(u)H(z-u)du, \tag{1.1.1}$$

where H is the Heaviside step function. The interest is to learn about $s(z)$ given observed $t(z)$. Theoretically, $s(z) = \frac{dt(z)}{dz}$, but in practice, simply differentiating the observations need not lead to useful solutions because noise is generally present in the observed times $t(z)$, and naive differentiation may lead to unrealistic features of the solution.

Estimation of buried line mass density from vertical gravity anomaly

Here the problem is to estimate an unknown buried line mass density $m(x)$ from data on vertical gravity anomaly, $d(x)$, observed at some height, h. The mathematical relationship between $d(x)$ and $m(x)$ is given by

$$d(x) = \int_{-\infty}^{\infty} \frac{h}{[(u-x)^2 + h^2]^{\frac{3}{2}}} m(u)du.$$

As before, noise in the data renders the above linear inverse problem difficult. Variations of the above example has been considered in Aster *et al.* (2013).

Estimation of incident light intensity from diffracted light intensity

Consider an experiment in which an angular distribution of illumination passes through a thin slit and produces a diffraction pattern, for which the intensity is observed. The data, $d(s)$, are measurements of diffracted light intensity as a function of the outgoing angle $-\pi/2 \leq s \leq \pi/2$. The goal here is to obtain the intensity of incident light on the

slit, $m(\theta)$, as a function of the incoming angle $-\pi/2 \leq \theta \leq \pi/2$, using the following mathematical relationship:

$$d(s) = \int_{-\pi/2}^{\pi/2} (\cos(s) + \cos(\theta))^2 \left(\frac{\sin(\pi\,(\sin(s) + \sin(\theta)))}{\pi\,(\sin(s) + \sin(\theta))} \right)^2 m(\theta) d\theta.$$

Groundwater pollution source history reconstruction problem

Consider the problem of recovering the history of groundwater pollution at a source site from later measurements of the contamination at downstream wells to which the contaminant plume has been transported by advection and diffusion. The mathematical model for contamination transport is given by the following advection-diffusion equation with respect to t and transported site x:

$$\begin{aligned} \frac{\partial C}{\partial t} &= D\frac{\partial^2 C}{\partial x^2} - \nu\frac{\partial C}{\partial x} \\ C(0,t) &= C_{in}(t) \\ C(x,t) &\to 0 \text{ as } x \to \infty. \end{aligned}$$

In the above, D is the diffusion coefficient, ν is the velocity of the groundwater flow, and $C_{in}(t)$ is the time history of contaminant injection at $x = 0$. The solution to the above advection-diffusion equation is given by

$$C(x,T) = \int_0^T C_{in}(t) f(x, T-t) dt,$$

where

$$f(x, T-t) = \frac{x}{2\sqrt{\pi D (T-t)^3}} \exp\left[\frac{(x - \nu(T-t))^2}{4D(T-t)} \right].$$

It is of interest to learn about $C_{in}(t)$ from data observed on $C(x,T)$.

Transmission tomography

The most basic physical model for tomography assumes that wave energy traveling between a source and receiver can be considered to be propagating along infinitesimally narrow ray paths. In seismic tomography, if the slowness at a point x is $s(x)$, and the ray path is known, then the travel time for seismic energy transiting along that ray path is given by the line integral along ℓ:

$$t = \int_{\ell} s(x(l))dl. \tag{1.1.2}$$

Learning of $s(x)$ from t is required. Note that (1.1.2) is a high-dimensional generalization of (1.1.1). In reality, seismic ray paths will be bent due to refraction and/or reflection, resulting in nonlinear inverse problem.

The above examples demonstrate the ubiquity of linear inverse problems. As a result, in the next section we take up the case of linear inverse problems and illustrate the Bayesian approach in details, also investigating connections with the deterministic approach employed by the general scientific community.

1.2 Linear inverse problem

The motivating examples and discussions in this section are based on Bui-Thanh (2012).

Let us consider the following one-dimensional integral equation on a finite interval as in equation (1.2.1):

$$G(x, \theta) = \int K(x, t)\ \theta(t)\ dt, \tag{1.2.1}$$

where $K(x, \cdot)$ is some appropriate, known, real-valued function given x Now, let the dataset be $\boldsymbol{y}_n = (y_1, y_2, \ldots, y_n)^T$. Then for a known system response $K(x_i, t)$ for the

dataset, the equation can be written as follows:

$$y_i = \int G(x_i, \theta) + \epsilon_i \ ; \quad i \in \{1, 2, \ldots, n\} \tag{1.2.2}$$

As a particular example, let $G(x, \theta) = \int_0^1 K(x, t)\ \theta(t)\ dt$, where $K(x, t) = \frac{1}{\sqrt{2\pi\psi^2}} \exp\left\{-(x-t)^2/2\psi^2\right\}$ is the Gaussian kernel and $\theta : [0, 1] \mapsto \mathbb{R}$ is to be learned given the data $\boldsymbol{y}_n$ and $\boldsymbol{x}_n = (x_1, \ldots, x_n)^T$. We first illustrate the Bayesian approach and draw connections with the traditional approach of Tikhonov's regularization when the integral in G is discretized. In this regard, let $x_i = (i-1)/n$, for $i = 1, \ldots, n$. Letting $\boldsymbol{\theta} = (\theta(x_1), \ldots, \theta(x_n))^T$ and $\boldsymbol{K}$ be the $n \times n$ matrix with the (i, j)-th element $K(x_i, x_j)/n$, and $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)^T$, the discretized version of (1.2.2) can be represented as

$$\boldsymbol{y}_n = \boldsymbol{K}\boldsymbol{\theta} + \boldsymbol{\epsilon}_n. \tag{1.2.3}$$

We assume that $\boldsymbol{\epsilon}_n \sim N_n\left(\boldsymbol{0}_n, \sigma^2 \boldsymbol{I}_n\right)$, that is, an n-variate normal with mean $\boldsymbol{0}_n$, an n-dimensional vector with all components zero, and covariance $\sigma^2 \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the n-th order identity matrix.

1.2.1 Smooth prior on θ

To reflect the belief that the function θ is smooth, one may presume that

$$\theta(x_i) = \frac{\theta(x_{i-1}) + \theta(x_{i+1})}{2} + \tilde{\epsilon}_i, \tag{1.2.4}$$

where, for $i = 1, \ldots, n$, $\tilde{\epsilon}_i \overset{iid}{\sim} N\left(0, \tilde{\sigma}^2\right)$. Thus, *a priori*, $\theta(x_i)$ is assumed to be an average of its nearest neighbors to quantify smoothness, with an additive random perturbation

term. Letting

$$\boldsymbol{L} = \frac{1}{2}\begin{pmatrix} -1 & 2 & -1 & 0 & \cdots & \cdots \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \end{pmatrix}, \tag{1.2.5}$$

and $\tilde{\boldsymbol{\epsilon}} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n)^T$, it follows from (1.2.4) that

$$\boldsymbol{L\theta} = \tilde{\boldsymbol{\epsilon}}, \tag{1.2.6}$$

Now, noting that the Laplacian of a twice-differentiable real-valued function f with independent arguments $z_1, \ldots, z_k$ is given by $\Delta f = \sum_{i=1}^{k} \frac{\partial^2 f}{\partial z_i^2}$, we have

$$\Delta\theta(x_j) \approx n^2(\boldsymbol{L\theta})_j, \tag{1.2.7}$$

where $(\boldsymbol{L\theta})_j$ is the j-th element of $\boldsymbol{L\theta}$.

However, the rank of $\boldsymbol{L}$ is $n-1$, and boundary conditions on the Laplacian operator is necessary to ensure positive definiteness of the operator. In our case, we assume that $\theta \equiv 0$ outside $[0,1]$, so that we now assume $\theta(0) = \frac{\theta(x_1)}{2} + \tilde{\epsilon}_0$ and $\theta(x_n) = \frac{\theta(x_{n-1})}{2} + \tilde{\epsilon}_n$, where $\tilde{\epsilon}_0$ and $\tilde{\epsilon}_n$ are iid $N\left(0, \tilde{\sigma}^2\right)$. With this modification, the prior on $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\tilde{\sigma}^2}\|\tilde{\boldsymbol{L}}\boldsymbol{\theta}\|^2\right), \tag{1.2.8}$$

where $\|\cdot\|$ is the Euclidean norm and

$$\tilde{\boldsymbol{L}} = \frac{1}{2}\begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & \cdots \\ -1 & 2 & -1 & 0 & \cdots & \cdots \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}. \tag{1.2.9}$$

Rather than assuming zero boundary conditions, more generally one may assume that $\theta(0)$ and $\theta(x_n)$ are distributed as $N\left(0, \frac{\tilde{\sigma}^2}{\delta_0^2}\right)$ and $N\left(0, \frac{\tilde{\sigma}^2}{\delta_n^2}\right)$, respectively. The resulting modified matrix is then given by

$$\hat{\boldsymbol{L}} = \frac{1}{2}\begin{pmatrix} 2\delta_0 & 0 & 0 & 0 & \cdots & \cdots \\ -1 & 2 & -1 & 0 & \cdots & \cdots \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & \cdots & 0 & 0 & 2\delta_n \end{pmatrix}. \tag{1.2.10}$$

To choose δ_0 and δ_n, one may assume that

$$Var\left[\theta(0)\right] = \frac{\tilde{\sigma}^2}{\delta_0^2} = Var\left[\theta(x_n)\right] = \frac{\tilde{\sigma}^2}{\delta_n^2} = Var\left[\theta(x_{[n/2]})\right] = \tilde{\sigma}^2 \boldsymbol{\epsilon}_{[n/2]}^T \left(\hat{\boldsymbol{L}}^T \hat{\boldsymbol{L}}\right)^{-1} \boldsymbol{\epsilon}_{[n/2]},$$

where $[n/2]$ is the largest integer not exceeding $n/2$, and $\boldsymbol{\epsilon}_{[n/2]}$ is the $[n/2]$-th canonical

basis vector in $\mathbb{R}^{n+1}$. It follows that

$$\delta_0^2 = \delta_n^2 = \frac{1}{\boldsymbol{\epsilon}_{[n/2]}^T \left(\hat{\boldsymbol{L}}^T \hat{\boldsymbol{L}}\right)^{-1} \boldsymbol{\epsilon}_{[n/2]}}.$$

Since this requires solving a non-linear equation (since $\hat{\boldsymbol{L}}$ contains δ_0 and δ_n), for avoiding computational complexity one may simply employ the approximation

$$\delta_0^2 = \delta_n^2 = \frac{1}{\boldsymbol{\epsilon}_{[n/2]}^T \left(\tilde{\boldsymbol{L}}^T \tilde{\boldsymbol{L}}\right)^{-1} \boldsymbol{\epsilon}_{[n/2]}},$$

where $\tilde{\boldsymbol{L}}$ is given by (1.2.9).

1.2.2 Non-smooth prior on θ

To begin with, let us assume that θ has several points of discontinuities on the grid of points $\{x_0, \ldots, x_n\}$. To reflect this information in the prior, one may assume that $\theta(0) = 0$ and for $i = 1, \ldots, n$, $\theta(x_i) = \theta(x_{i-1}) + \tilde{\epsilon}_i$, where, as before, $\tilde{\epsilon}_i$ are $iid\ N\left(0, \tilde{\sigma}^2\right)$. Then, with

$$\boldsymbol{L}^* = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & \cdots \\ -1 & 1 & 0 & 0 & \cdots & \cdots \\ 0 & -1 & 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 & 0 \\ 0 & 0 & \cdots & 0 & - & 1 \end{pmatrix}, \tag{1.2.11}$$

the prior is given by

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\tilde{\sigma}^2} \|\boldsymbol{L}^* \boldsymbol{\theta}\|^2\right). \tag{1.2.12}$$

One may also flexibly account for any particular big jump. For instance, if for some $\ell \in \{0, \ldots, n\}$, the jump $\theta(x_\ell) - \theta(x_{\ell-1})$ is particularly large compared to the other jumps, then it can be assumed that $\theta(x_\ell) = \theta(x_{\ell-1}) + \epsilon_\ell^*$, with $\epsilon_\ell^* \sim N\left(0, \frac{\tilde{\sigma}^2}{\xi^2}\right)$, where $\xi < 1$. Letting $\boldsymbol{D}_\ell$ be the diagonal matrix with ξ^2 being the ℓ-th diagonal element and 1 being the other diagonal elements, the prior is then given by

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\tilde{\sigma}^2}\|\boldsymbol{D}_\ell \boldsymbol{L}^* \boldsymbol{\theta}\|^2\right). \tag{1.2.13}$$

A more general prior can be envisaged where the number and location of the jump discontinuities are unknown. Then we may consider a diagonal matrix $\boldsymbol{D} = diag\{\xi_1, \ldots, \xi_n\}$, so that conditionally on the hyperparameters $\xi_1, \ldots, \xi_n$, the prior on $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}|\xi_1, \ldots, \xi_n) \propto \exp\left(-\frac{1}{2\tilde{\sigma}^2}\|\boldsymbol{D}\boldsymbol{L}^* \boldsymbol{\theta}\|^2\right). \tag{1.2.14}$$

Prior on $\xi_1, \ldots, \xi_n$ may be considered to complete the specification. These may also be estimated by maximizing the marginal likelihood obtained by integrating out $\boldsymbol{\theta}$, which is known as the ML-II method; see Berger (1985). Calvetti and Somersalo (2007) also advocate likelihood based methods.

1.2.3 Posterior distribution

For convenience, let us generically denote the matrices $\boldsymbol{L}$, $\tilde{\boldsymbol{L}}$, $\hat{\boldsymbol{L}}$, $\boldsymbol{L}^*$, $\boldsymbol{D}_\ell \boldsymbol{L}^*$, $\boldsymbol{D}\boldsymbol{L}^*$, by $\boldsymbol{\Gamma}^{-\frac{1}{2}}$. Then it can be easily verified that the posterior of θ admits the following generic form:

$$\pi\left(\boldsymbol{\theta}|\boldsymbol{y}_n, \boldsymbol{x}_n\right) \propto \exp\left\{-\left[\frac{1}{2\sigma^2}\|\boldsymbol{y}_n - \boldsymbol{K}\boldsymbol{\theta}\|^2 + \frac{1}{2\tilde{\sigma}^2}\|\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{\theta}\|^2\right]\right\}. \tag{1.2.15}$$

Note that the exponent of the posterior is of the form of the Tikhonov functional, which we denote by $T(\boldsymbol{\theta})$. The maximizer of the posterior, commonly known as the *maximum*

a posteriori (MAP) estimator, is given by

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \pi\left(\boldsymbol{\theta}|\boldsymbol{y}_n, \boldsymbol{x}_n\right) = \arg\min_{\boldsymbol{\theta}} T(\boldsymbol{\theta}). \tag{1.2.16}$$

In other words, the deterministic solution to the inverse problem obtained by Tikhonov's regularization is nothing but the Bayesian MAP estimator in our context.

Writing $\boldsymbol{H} = \frac{1}{\sigma^2}\boldsymbol{K}^T\boldsymbol{K} + \frac{1}{\tilde{\sigma}^2}\boldsymbol{\Gamma}^{-1}$, which is the Hessian of the Tikhonov functional (regularized misfit), and writing $\|\cdot\|_{\boldsymbol{H}} = \|\boldsymbol{H}^{\frac{1}{2}}\cdot\|$, it is clear that (1.2.15) can be simplified to the Gaussian form, given by

$$\pi\left(\boldsymbol{\theta}|\boldsymbol{y}_n, \boldsymbol{x}_n\right) \propto \exp\left\{-\left\|\boldsymbol{\theta} - \frac{1}{\sigma^2}\boldsymbol{H}^{-1}\boldsymbol{K}^{-1}\boldsymbol{y}_n\right\|_{\boldsymbol{H}}^2\right\}. \tag{1.2.17}$$

It follows from (1.2.17) that the inverse of the Hessian of the regularized misfit is the posterior covariance itself. From the above posterior it also trivially follows that

$$\hat{\boldsymbol{\theta}}_{MAP} = \frac{1}{\sigma^2}\boldsymbol{H}^{-1}\boldsymbol{K}^{-1}\boldsymbol{y}_n = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\boldsymbol{K}^T\boldsymbol{K} + \frac{1}{\tilde{\sigma}^2}\boldsymbol{\Gamma}^1\right)^{-1}\boldsymbol{K}^T\boldsymbol{Y}_n, \tag{1.2.18}$$

which coincides with the Tikhonov solution for linear inverse problems. The connection between the traditional deterministic Tikhonov regularization approach with Bayesian analysis continues to hold even if the likelihood is non-Gaussian.

1.2.4 Exploration of the smoothness conditions

For deeper investigation of the smoothness conditions, let us write

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\min_{\boldsymbol{\theta}} T(\boldsymbol{\theta}) = \sigma^2\left(\frac{1}{2}\|\boldsymbol{y}_n - \tilde{\boldsymbol{y}}_n\|^2 + \frac{1}{2}\varrho\|\tilde{\boldsymbol{\Gamma}}^{\frac{1}{2}}\boldsymbol{\theta}\|^2\right), \tag{1.2.19}$$

where $\tilde{\boldsymbol{y}}_n = \boldsymbol{K}\boldsymbol{\theta}$, $\varrho = \sigma^2/\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\Gamma}}^{\frac{1}{2}} = \boldsymbol{\Gamma}^{-\frac{1}{2}}$. Now, from (1.2.7) it follows that for the smooth priors with the zero boundary conditions, our Tikhonov functional discretizes

$$T_\infty(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y}_n - \tilde{\boldsymbol{y}}_n\|^2 + \frac{1}{2}\varrho\|\Delta\theta\|^2_{L^2(0,1)}, \tag{1.2.20}$$

where $\|\cdot\|^2_{L^2(0,1)} = \int_0^1 (\cdot)^2 dt$.

On the other hand, for the non-smooth prior (1.2.12), rather than discretizing $\Delta\theta$, $\nabla\theta$, that is, the gradient of θ, is discretized. In other words, for non-smooth priors, our Tikhonov functional discretizes

$$T_\infty(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y}_n - \tilde{\boldsymbol{y}}_n\|^2 + \frac{1}{2}\varrho\|\nabla\theta\|^2_{L^2(0,1)}. \tag{1.2.21}$$

Hence, realizations of prior (1.2.12) is less smooth compared to those of our smooth priors. However, the realizations (1.2.12) must be continuous. The priors given by (1.2.13) and (1.2.14) also support continuous functions as long as the hyperparameters are bounded away from zero. These facts, although clear, can be rigorously justified by functional analysis arguments, in particular, using the Sobolev imbedding theorem (see, for example, Arbogast and Bona (2008)).

1.3 Links between Bayesian inverse problems based on Gaussian process prior and deterministic regularizations

In this section, based on Rasmussen and Williams (2006), we illustrate the connections between deterministic regularizations such as those obtained from differential operators as above, and Bayesian inverse problems based on the very popular Gaussian process prior on the unknown function. A key tool for investigating such relationship is the reproducing kernel Hilbert space (RKHS).

1.3.1 RKHS

We adopt the following definition of RKHS provided in Rasmussen and Williams (2006):

Definition 1 (RKHS) *Let $\mathcal{H}$ be a Hilbert space of real functions θ defined on an index set $\mathcal{X}$. Then $\mathcal{H}$ is called an RKHS endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ (and norm $\|\theta\|_{\mathcal{H}} = \langle \theta, \theta \rangle_{\mathcal{H}}$) if there exists a function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ with the following properties:*

(a) for every x, $\mathcal{K}(\cdot, x) \in \mathcal{H}$, and

(b) $\mathcal{K}$ has the reproducing property $\langle \theta(\cdot), \mathcal{K}(\cdot, x) \rangle_{\mathcal{H}} = \theta(x)$.

Observe that since $\mathcal{K}(\cdot, x), \mathcal{K}(\cdot, x') \in \mathcal{H}$, it follows that $\langle \mathcal{K}(\cdot, x), \mathcal{K}(\cdot, x') \rangle_{\mathcal{H}} = \mathcal{K}(x, x')$. The Moore-Aronszajn theorem asserts that the RKHS uniquely determines $\mathcal{K}$, and vice versa. Formally,

Theorem 2 (Aronszajn (1950)) *. Let $\mathcal{X}$ be an index set. Then for every positive definite function $\mathcal{K}(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$ there exists a unique RKHS, and vice versa.*

Here, by positive definite function $\mathcal{K}(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$, we mean $\int \mathcal{K}(x, x') g(x) g(x') d\nu(x) d\nu(x') > 0$ for all non-zero functions $g \in L_2(\mathcal{X}, \nu)$, where $L_2(\mathcal{X}, \nu)$ denotes the space of functions square-integrable on $\mathcal{X}$ with respect to the measure ν.

Indeed, the subspace $\mathcal{H}_0$ of $\mathcal{H}$ spanned by the functions $\{\mathcal{K}(\cdot, \boldsymbol{x}_i);\ i = 1, 2, \ldots\}$ is dense in $\mathcal{H}$ in the sense that every function in $\mathcal{H}$ is a pointwise limit of a Cauchy sequence from $\mathcal{H}_0$.

To proceed, we require the concepts of eigenvalues and eigenfunctions associated with kernels. In the following section we provide a briefing on these.

1.3.2 Eigenvalues and eigenfunctions of kernels

We borrow the statements of the following definition of eigenvalue and eigenfunction, and the subsequent statement of Mercer's theorem from Rasmussen and Williams (2006).

Definition 3 *A function $\psi(\cdot)$ that obeys the integral equation*

$$\int_{\mathcal{X}} \mathcal{C}(x, x')\psi(x)d\nu(x) = \lambda\psi(x'), \tag{1.3.1}$$

is called an eigenfunction of the kernel $\mathcal{C}$ with eigenvalue λ with respect to the measure ν.

We assume that the ordering is chosen such that $\lambda_1 \geq \lambda_2 \geq \cdots$. The eigenfunctions are orthogonal with respect to ν and can be chosen to be normalized so that $\int_{\mathcal{X}} \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x})d\nu(x) = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

The following well-known theorem (see, for example, König (1986)) expresses the positive definite kernel $\mathcal{C}$ in terms of its eigenvalues and eigenfunctions.

Theorem 4 (Mercer's theorem) *Let $(\mathcal{X}, \nu)$ be a finite measure space and $\mathcal{C} \in L_\infty\left(\mathcal{X}^2, \nu^2\right)$ be a positive definite kernel. By $L_\infty\left(\mathcal{X}^2, \nu^2\right)$ we mean the set of all measurable functions $\mathcal{C} : \mathcal{X}^2 \mapsto \mathbb{R}$ which are essentially bounded, that is, bounded up to a set of ν^2-measure zero. For any function $\mathcal{C}$ in this set, its essential supremum, given by $\inf\{C \geq 0 : |\mathcal{C}(x_1, x_2)| < C,$ for almost all $(x_1, x_2) \in \mathcal{X} \times \mathcal{X}\}$ serves as the norm $\|\mathcal{C}\|$.*

Let $\psi_j \in L_2(\mathcal{X}, \nu)$ be the normalized eigenfunctions of $\mathcal{C}$ associated with the eigenvalues $\lambda_j(\mathcal{C}) > 0$. Then

(a) the eigenvalues $\{\lambda_j(\mathcal{C})\}_{j=1}^{\infty}$ are absolutely summable.

(b) $\mathcal{C}(x, x') = \sum_{j=1}^{\infty} \lambda_j(\mathcal{C})\psi_j(\boldsymbol{x})\bar{\psi}_j(x')$ holds ν^2-almost everywhere, where the series converges absolutely and uniformly ν^2-almost everywhere. In the above, $\bar{\psi}_j$ denotes the complex conjugate of ψ_j.

It is important to note the difference between the eigenvalue $\lambda_j(\mathcal{C})$ associated with the kernel $\mathcal{C}$ and $\lambda_j(\boldsymbol{\Sigma}_n)$ where $\boldsymbol{\Sigma}_n$ denotes the $n \times n$ Gram matrix with (i, j)-th element $\mathcal{C}(x_i, x_j)$. Observe that (see Rasmussen and Williams (2006)):

$$\lambda_j(\mathcal{C})\psi_j(x') = \int_{\mathcal{X}} \mathcal{C}(x, x')\psi_j(x)d\nu(x) \approx \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(x_i, x')\psi_j(x_i, x'), \tag{1.3.2}$$

where, for $i = 1, \ldots, n$, $\boldsymbol{x}_i \sim \nu$, assuming that ν is a probability measure. Now substituting $x' = x_i$; $i = 1, \ldots, n$ in (1.3.2) yields the following approximate eigen system for the matrix $\boldsymbol{\Sigma}_n$:

$$\boldsymbol{\Sigma}_n \boldsymbol{u}_j \approx n\lambda_j(\mathcal{C})\boldsymbol{u}_j, \tag{1.3.3}$$

where the i-th component of $\boldsymbol{u}_j$ is given by

$$u_{ij} = \frac{\psi_j(x_i)}{\sqrt{n}}. \tag{1.3.4}$$

Since ψ_j are normalized to have unit norm, it holds that

$$\boldsymbol{u}_j^T \boldsymbol{u}_j = \frac{1}{n}\sum_{i=1}^{n} \psi_j^2(x_i) \approx \int_{\mathcal{X}} \psi^2(x) d\nu(x) = 1. \tag{1.3.5}$$

From (1.3.5) it follows that

$$\lambda_j(\boldsymbol{\Sigma}_n) \approx n\lambda_j(\mathcal{C}). \tag{1.3.6}$$

Indeed, Theorem 3.4 of Baker (1977) shows that $n^{-1}\lambda_j(\boldsymbol{\Sigma}_n) \to \lambda_j(\mathcal{C})$, as $n \to \infty$.

For our purposes the main usefulness of the RKHS framework is that $\|\theta\|_{\mathcal{H}}^2$ can be perceived as a generalization of $\boldsymbol{\theta}^T \boldsymbol{K}^{-1}\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta(x_1), \ldots, \theta(x_n))^T$ and $\boldsymbol{K} = (\mathcal{K}(x_i, x_j))_{i,j=1,\ldots,n}$, is the $n \times n$ matrix with (i, j)-th element $\mathcal{K}(x_i, x_j)$.

1.3.3 Inner product

Consider a real positive semidefinite kernel $\mathcal{K}(x, x')$ with an eigenfunction expansion $\mathcal{K}(x, x') = \sum_{i=1}^{N} \lambda_i \phi_i(x)\phi_i(x')$ relative to a measure μ. Mercer's theorem ensures that the eigenfunctions are orthonormal with respect to μ, that is, we have $\int \phi_i(x)\phi_j(x)d\mu(x) = \delta_{ij}$. Consider a Hilbert space of linear combinations of the eigenfunctions, that is, $\theta(x) = \sum_{i=1}^{N} \theta_i \phi_i(x)$ with $\sum_{i=1}^{N} \frac{\theta_i^2}{\lambda_i} < \infty$. Then the inner product $\langle \theta_1, \theta_2 \rangle_{\mathcal{H}}$ between

$\theta_1 = \sum_{i=1}^N \theta_{1i}\phi_i(x)$, and $\theta_2 = \sum_{i=1}^N \theta_{2i}\phi_i(x)$ is of the form

$$\langle \theta_1, \theta_2 \rangle_{\mathcal{H}} = \sum_{i=1}^{N} \frac{\theta_{1i}\theta_{2i}}{\lambda_i}. \tag{1.3.7}$$

This induces the norm $\|\cdot\|_{\mathcal{H}}$, where $\|\theta\|_{\mathcal{H}}^2 = \sum_{i=1}^N \frac{\theta_i^2}{\lambda_i}$. A smoothness condition on the space is immediately imposed by requiring the norm to be finite – the eigenvalues must decay sufficiently fast.

The Hilbert space defined above is a unique RKHS with respect to $\mathcal{K}$, in that it satisfies the following reproducing property:

$$\langle \theta, \mathcal{K}(\cdot, x) \rangle = \sum_{i=1}^{N} \frac{\theta_i \lambda_i \phi_i(x)}{\lambda_i} = \theta(x). \tag{1.3.8}$$

Further, the kernel satisfies the following:

$$\langle \mathcal{K}(x, \cdot), \mathcal{K}(x', \cdot) \rangle = \sum_{i=1}^{N} \frac{\lambda_i^2 \phi_i(x)}{\lambda_i} = \mathcal{K}(x, x'). \tag{1.3.9}$$

Now, with reference to (1.3.6), observe that the square norm $\|\theta\|_{\mathcal{H}}^2 = \sum_{i=1}^N \theta_i^2/\lambda_i$ and the quadratic form $\boldsymbol{\theta}^T \boldsymbol{K} \boldsymbol{\theta}$ have the same form if the latter is expressed in terms of the eigenvectors of $\boldsymbol{K}$, albeit the latter has n terms, while the square norm has N terms.

1.3.4 Regularization

The ill-posed-ness of inverse problems can be understood from the fact that for any given data set $\boldsymbol{y}_n$, all functions that pass through the data set minimize any given measure of discrepancy $\mathbb{D}(\boldsymbol{y}_n, \boldsymbol{\theta})$ between the data $\boldsymbol{y}_n$ and $\boldsymbol{\theta}$. To combat this, one considers minimization of the following regularized functional:

$$R(\theta) = \mathbb{D}(\boldsymbol{y}_n, \boldsymbol{\theta}) + \frac{\tau}{2}\|\theta\|_{\mathcal{H}}^2, \tag{1.3.10}$$

where the second term, which is the regularizer, controls smoothness of the function and τ is the appropriate Lagrange multiplier.

The well-known representer theorem (see, for example, Kimeldorf and Wahba (1971), O'Sullivan *et al.* (1986), Wahba (1990), Schölkopf and Smola (2002)) guarantees that each minimizer $\theta \in \mathcal{H}$ can be represented as $\theta(x) = \sum_{i=1}^{n} c_i \mathcal{K}(x, x_i)$, where $\mathcal{K}$ is the corresponding reproducing kernel. If $\mathbb{D}(\boldsymbol{y}_n, \boldsymbol{\theta})$ is convex, then there is a unique minimizer $\hat{\theta}$.

1.3.5 Gaussian process modeling of the unknown function θ

For simplicity, let us consider the model

$$y_i = \theta(x_i) + \epsilon_i, \tag{1.3.11}$$

for $i = 1, \ldots, n$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, where we assume σ to be known for simplicity of illustration. Let $\theta(x)$ be modeled by a Gaussian process with mean function $\mu(x)$ and covariance kernel $\mathcal{K}$ associated with the RKHS. In other words, for any $x \in \mathcal{X}$, $E[\theta(x)] = \mu(x)$ and for any $x_1, x_2 \in \mathcal{X}$, $Cov(\theta(x_1), \theta(x_2)) = \mathcal{K}(x_1, x_2)$.

Assuming for convenience that $\mu(x) = 0$ for all $x \in \mathcal{X}$, it follows that the posterior distribution of $\theta(x^*)$ for any $x^* \in \mathcal{X}$ is given by

$$\pi(\theta(x^*)|\boldsymbol{y}_n, \boldsymbol{x}_n) \equiv N\left(\hat{\mu}(x^*), \hat{\sigma}^2(x^*)\right), \tag{1.3.12}$$

where, for any $x^* \in \mathcal{X}$,

$$\hat{\mu}(x^*) = \boldsymbol{s}^T(x^*)\left(\boldsymbol{K} + \sigma^2 \mathbb{I}_n\right)^{-1} \boldsymbol{y}_n; \tag{1.3.13}$$

$$\hat{\sigma}^2(x^*) = \mathcal{K}(x^*, x^*) - \boldsymbol{s}^T(x^*)\left(\lfloor K + \sigma^2 \mathbb{I}_n\right)^{-1} \boldsymbol{s}(x^*), \tag{1.3.14}$$

with $\boldsymbol{s}(x^*) = (\mathcal{K}(x^*, x_1), \ldots, \mathcal{K}(x^*, x_n))^T$.

Observe that the posterior mean admits the following representation:

$$\hat{\mu}(x^*) = \sum_{i=1}^{n} \tilde{c}_i \mathcal{K}(x^*, x_i), \tag{1.3.15}$$

where $\tilde{c}_i$ is the i-th element of $\left(\boldsymbol{K} + \sigma^2 \mathbb{I}_n\right)^{-1} \boldsymbol{y}_n$.

In other words, the posterior mean of the Gaussian process based model is consistent with the representer theorem.

1.4 Regularization using differential operators and connection with Gaussian process

For $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, let

$$\|\mathcal{L}^m \theta\|^2 = \int \sum_{j_1 + \cdots + j_d = m} \left(\frac{\partial^m \theta(x)}{\partial x_1^{j_1} \cdots \partial x_d^{j_d}} \right)^2, \tag{1.4.1}$$

and

$$\|\mathcal{P}\theta\|^2 = \sum_{m=0}^{M} b_m \|\mathcal{L}^m \theta\|^2, \tag{1.4.2}$$

for some $M > 0$, where the co-efficients $b_m \geq 0$. In particular, we assume for our purpose that $b_0 > 0$. It is clear that $\|\mathcal{P}\theta\|^2$ is translation and rotation invariant. This norm penalizes θ in terms of its derivatives up to order M.

1.4.1 Relation to RKHS

It can be shown, using the fact that the complex exponentials $\exp(2\pi i s^T x)$ are eigen functions of the differential operator, that

$$\|\mathcal{P}\theta\|^2 = \int \sum_{m=0}^{M} b_m \left(4\pi^2 s^T s\right)^m \left|\tilde{\theta}(s)\right|^2 ds, \tag{1.4.3}$$

where $\tilde{\theta}(s)$ is the Fourier transform of $\theta(s)$. Comparison of (1.4.3) with (1.3.7) yields the power spectrum of the form $\left[\sum_{m=0}^{M} b_m \left(4\pi^2 s^T s\right)^m\right]^{-1}$ which yields the following kernel by Fourier inversion:

$$\mathcal{K}(x, x') = \mathcal{K}(x - x') = \int \frac{\exp(2\pi i s^T (x - x'))}{\sum_{m=0}^{M} b_m \left(4\pi^2 s^T s\right)^m} ds. \tag{1.4.4}$$

Calculus of variations can also be used to minimize $R(\theta)$ with respect to θ, which yields (using the Euler-Lagrange equation)

$$\theta(x) = \sum_{i=1}^{n} b_i \mathcal{G}(x - x_i), \tag{1.4.5}$$

with

$$\sum_{i=1}^{m} (-1)^m b_m \nabla^m \mathcal{G} = \delta_{x-x'}, \tag{1.4.6}$$

where $\mathcal{G}$ is known as the Green's function. Using Fourier transform on (1.4.6) it can be shown that the Green's function is nothing but the kernel $\mathcal{K}$ given by (1.4.4). Moreover, it follows from (1.4.6) that $\sum_{i=1}^{m} (-1)^m b_m \nabla^m$ and $\mathcal{K}$ are inverses of each other.

Examples of kernels derived from differential operators are as follows. For $d = 1$, setting $b_0 = b^2$, $b_1 = 1$ and $b_m = 0$ for $m \geq 2$, one obtains $\mathcal{K}(x, x') = \mathcal{K}(x - x') = \frac{1}{2b} \exp\left(-b|x - x'|\right)$, which is the covariance of the Ornstein-Uhlenbeck process. For general d dimension, setting $b_m = b^{2m}/(m! 2^m)$, yields $\mathcal{K}(x, x') = \mathcal{K}(x - x') = \frac{1}{(2\pi b^2)^{d/2}} \exp\left[-\frac{1}{2b^2}(x - x')^T (x - x')\right]$.

Considering a grid $\boldsymbol{x}_n$, note that

$$\|\mathcal{P}\theta\|^2 \approx \sum_{m=0}^{M} b_m \left(D_m \boldsymbol{\theta}\right)^T \left(D_m \boldsymbol{\theta}\right) = \boldsymbol{\theta}^T \left(\sum_{m=0}^{M} D_m^T D_m\right) \boldsymbol{\theta}, \tag{1.4.7}$$

where D_m is a suitable finite-difference approximation of the differential operator. Note that such finite-difference approximation has been explored in Section 1.2, which we now

investigate in a rigorous setting. Also, since (1.4.7) is quadratic in $\boldsymbol{\theta}$, assuming a prior for $\boldsymbol{\theta}$, the logarithm of which has this form, and further assuming that $\log[\mathbb{D}(\boldsymbol{y}_n, \boldsymbol{\theta})]$ is a log-likelihood quadratic in $\boldsymbol{\theta}$, a Gaussian posterior results.

1.4.2 Spline models and connection with Gaussian process

Let us consider the penalty function to be $\|\mathcal{L}^m\theta\|^2$. Then polynomials up to degree $m-1$ are not penalized and so, are in the null space of the regularization operator. In this case, it can be shown that a minimizer of $R(\theta)$ is of the form

$$\theta(x) = \sum_{j=1}^{k} d_j \psi_j(x) + \sum_{i=1}^{n} c_i G(x, x_i), \tag{1.4.8}$$

where $\{\psi_1, \ldots, \psi_k\}$ are polynomials that span the null space and the Green's function G is given by (see Duchon (1977), Meinguet (1979))

$$G(x, x') = G(x - x') = \begin{cases} c_{m,d}|x - x'|^{2m-d} \log|x - x'| & \text{if } 2m > d \text{ and } d \text{ even} \\ c_{m,d}|x - x'|^{2m-d} & \text{otherwise.} \end{cases}, \tag{1.4.9}$$

where $c_{m,D}$ are constants (see Wahba (1990) for the explicit form).

We now specialize the above arguments to the spline set-up. As before, let us consider the model $y_i = \theta(x_i) + \epsilon_i$, where, for $i = 1, \ldots, n$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. For simplicity, we consider the one-dimensional set-up, and consider the cubic spline smoothing problem that minimizes

$$R(\theta) = \sum_{i=1}^{n} (y_i - \theta(x_i))^2 + \tau \int_0^1 \left[\theta''(x)\right]^2 dx, \tag{1.4.10}$$

where $0 < x_1 < \cdots < x_n < 1$. The solution to this minimization problem is given by

$$\theta(x) = \sum_{j=0}^{1} d_j x^j + \sum_{i=1}^{n} c_i (x - x_i)_+^3, \tag{1.4.11}$$

where, for any x, $(x)_+ = x$ if $x > 0$ and zero otherwise.

Following Wahba (1978), let us consider

$$f(x) = \sum_{j=0}^{1} \beta_j x^j + \theta(x), \tag{1.4.12}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T \sim N\left(\mathbf{0}, \sigma_\beta^2 \mathbb{I}_2\right)$, and θ is a zero mean Gaussian process with covariance

$$\sigma_\theta^2 \mathcal{K}(x, x') = \int_0^1 (x-u)_+ (x'-u)_+ du = \sigma_\theta^2 \left(\frac{|x-x'|v^2}{2} + \frac{v^3}{3} \right), \tag{1.4.13}$$

where $v = \min\{x, x'\}$.

Taking $\sigma_\beta^2 \to \infty$ makes the prior of $\boldsymbol{\beta}$ vague, so that penalty on the polynomial terms in the null space is effectively washed out. It follows that

$$E\left[\theta(x^*) | \boldsymbol{y}_n, \boldsymbol{x}_n\right] = \boldsymbol{h}(x^*)^T \hat{\boldsymbol{\beta}} + \boldsymbol{s}(x^*)^T \hat{\boldsymbol{K}}^{-1} \left(\boldsymbol{y}_n - \boldsymbol{H}^T \hat{\boldsymbol{\beta}} \right), \tag{1.4.14}$$

where, for any x, $\boldsymbol{h}(x) = (1, x)^T$, $\boldsymbol{H} = (\boldsymbol{h}(x_1), \ldots, \boldsymbol{h}(x_n))$, $\hat{\boldsymbol{K}}$ is the covariance matrix corresponding to $\sigma_\theta^2 \mathcal{K}(x_i, x_j) + \sigma^2 \delta_{ij}$, and $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{H} \hat{\boldsymbol{K}}^{-1} \boldsymbol{H} \right)^{-1} \boldsymbol{H} \hat{\boldsymbol{K}}^{-1} \boldsymbol{y}_n$.

Since the elements of $\boldsymbol{s}(x^*)$ are piecewise cubic polynomials, it is easy to see that the posterior mean (1.4.14) is also a piecewise cubic polynomial. It is also clear that (1.4.14) is a first order polynomial on $[0, x_1]$ and $[x_n, 1]$.

Connection with the ℓ-fold integrated Wiener process

Shepp (1966) considered the ℓ-fold integrated Wiener process, for $\ell = 0, 1, 2 \ldots$, as follows:

$$W_\ell(x) = \int_0^1 \frac{(x-u)_+^\ell}{\ell!} Z(u) du, \tag{1.4.15}$$

where Z is a Gaussian white noise process with covariance $\delta(u - u')$. As a special case,

note that W_0 is the standard Wiener process. In our case, note that

$$\mathcal{K}(x, x') = Cov\left(W_1(x), W_1(x')\right). \tag{1.4.16}$$

The above ideas can be easily extended to the case of the regularizer $\int \left[f^{(m)}(x)\right]^2 dx$, for $m \geq 1$ by replacing $(x-u)_+$ with $(x-u)_+^{m-1}/(m-1)!$ and letting $\boldsymbol{h}(x) = \left(1, x, \ldots, x^{m-1}\right)^T$.

1.5 The Bayesian approach to inverse problems in Hilbert spaces

We assume the following model

$$y = G(\theta) + \epsilon, \tag{1.5.1}$$

where y, θ and ϵ are in Banach or Hilbert spaces.

1.5.1 Bayes theorem for general inverse problems

We will consider the model stated by equation (1.5.1). Let $\mathcal{Y}$ and Θ denote the sample spaces for y and θ, respectively. Let us first assume that both are separable Banach spaces. Assume μ_0 to be the prior measure for θ. Assuming well-defined joint distribution for (y, θ), let us denote the posterior of θ given y as μ_y. Let $\epsilon \sim Q_0$ where Q_0 such that ϵ and θ are independent. Let Q_0 be the distribution of ϵ. Let us denote the conditional distribution of y given θ by Q_θ, obtained from a translation of Q_0 by $G(\theta)$. Assume that $Q_\theta \ll Q_0$. Thus, for some potential $\Phi : \Theta \times \mathcal{Y} \mapsto \mathbb{R}$,

$$\frac{dQ_\theta}{dQ_0} = \exp\left(-\Phi(\theta, y)\right). \tag{1.5.2}$$

Thus, for fixed θ, $\Phi(\theta, \cdot) : \mathcal{Y} \mapsto \mathbb{R}$ is measurable and $E_{Q_0}\left[\exp\left(-\Phi(\theta, y)\right)\right] = 1$. Note that $-\Phi(\cdot, y)$ is nothing but the log-likelihood.

Let ν_0 denote the product measure

$$\nu_0(d\theta, dy) = \mu_0(d\theta)Q_0(dy), \tag{1.5.3}$$

and let us assume that Φ is ν_0-measurable. Then $(\theta, y) \in \Theta \times \mathcal{Y}$ is distributed according to the measure $\nu(d\theta, dy) = \mu_0(d\theta)Q_\theta(dy)$. It then also follows that $\nu \ll \nu_0$, with

$$\frac{d\nu_\theta}{d\nu_0}(\theta, y) = \exp\left(-\Phi(\theta, y)\right). \tag{1.5.4}$$

Then we have the following statement of Bayes' theorem for general inverse problems:

Theorem 5 (Bayes theorem for general inverse problems) *Assume that $\Phi : \Theta \times \mathcal{Y} \mapsto \mathbb{R}$ is ν_0-measurable and*

$$C = \int_\Theta \exp\left(-\Phi(\theta, y)\right) \mu_0(dy) > 0, \tag{1.5.5}$$

for Q_0-almost surely all y. Then the posterior of θ given y, which we denote by μ^y, exists under ν. Also, $\mu^y \ll \mu_0$ and for all y ν_0-almost surely,

$$\frac{d\mu_\theta^y}{d\mu_0}(\theta) = \frac{1}{C}\exp\left(-\Phi(\theta, y)\right). \tag{1.5.6}$$

Now assume that Θ and $\mathcal{Y}$ are Hilbert spaces. Suppose $\epsilon \sim \mathbf{N}(0, \Gamma)$. Then the following theorem holds:

Theorem 6 (Vollmer (2013))

$$\frac{d\mu^y}{d\mu_0} \propto \exp\left(-\frac{1}{2}\|G(\theta)\|_\Gamma^2 + \langle y, G(\theta)\rangle_\Gamma\right), \tag{1.5.7}$$

where $\langle \cdot, \cdot \rangle_\Gamma = \langle \Gamma^{-1} \cdot, \cdot \rangle$, and $\| \cdot \|_\Gamma$ is the norm induced by $\langle \cdot, \cdot \rangle_\Gamma$.

For the model $y_i = \theta(x_i) + \epsilon_i$ for $i = 1, \ldots, n$, with $\epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$, the posterior is of

the form

$$\frac{d\mu^y}{d\mu_0} \propto \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \theta(x_i))^2}{2\sigma^2}\right). \tag{1.5.8}$$

1.5.2 Connection with regularization methods

It is not immediately clear if the Bayesian approach in the Hilbert space setting has connection with the deterministic regularization methods, but Vollmer (2013) prove consistency of the posterior assuming certain stability results which are used to prove convergence of regularization methods; see Engl *et al.* (1996).

1.6 Conclusion

In this chapter, we have clarified the similarities and dissimilarities between the traditional inverse problems and the inverse regression problems. In particular, we have argued that only the latter class of problems qualify as authentic inverse problems in they have significantly different goals compared to the corresponding forward problems. Moreover, they include the traditional inverse problems on learning unknown functions as a special case, as exemplified by our palaeoclimate and Milky Way examples. We advocate the Bayesian paradigm for both classes of problems, not only because of its inherent flexibility, coherency and posterior uncertainty quantification, but also because the prior acts as a natural penalty which is very important to regularize the so-called ill-posed inverse problems. The well-known Tikhonov regularizer is just a special case from this perspective.

References

Arbogast, T. and Bona, J. L. (2008). Methods of Applied Mathematics. University of Texas at Austin.

Aronszajn, N. (1950). Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, **68**, 337–404.

Aster, R. C., Borchers, B., and Thurber, C. H. (2013). *Parameter Estimation and Inverse Problems*. Academic Press, Oxford, UK.

Baker, C. T. H. (1977). *The Numerical Treatment of Integral Equations*. Clarendon Press, Oxford.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Bui-Thanh, T. (2012). A Gentle Tutorial on Statistical Inversion Using the Bayesian Paradigm. ICES Report 12-18. Available at http://users.ices.utexas.edu/ tanbui/PublishedPapers/BayesianTutorial.pdf.

Calvetti, D. and Somersalo, E. (2007). *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing*. Springer, New York.

Duchon, J. (1977). Splines Minimizing Rotation-Invariant Semi-norms in Sobolev Spaces. In W. Schempp and K. Zellner, editors, *Constructive Theory of Functions of Several Variables*, pages 85–100, New York. Springer-Verlag.

Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems*. Kluwer Academic Publishers Group, Dordrecht. Volume 375 of Mathematics and its Applications.

Kimeldorf, G. and Wahba, G. (1971). Some Results on Tchebycheffian Spline Functions. *Journal of Mathematical Analysis and Applications*, **33**, 82–95.

König, H. (1986). *Eigenvalue Distribution of Compact Operators*. Birkhäuser.

Meinguet, J. (1979). Multivariate Interpolation at Arbitrary Points Made Simple. *Journal of the Applied Mathematics and Physics*, **30**, 292–304.

O'Sullivan, F., Yandell, B. S., and Raynor, W. J. (1986). Automatic Smoothing of Regression Functions in Generalized Linear Models. *Journal of the American Statistical Association*, **81**, 96–103.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, Massachusetts.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, USA.

Shepp, L. A. (1966). Radon-Nikodym Derivatives of Gaussian Measures. *Annals of Mathematical Statistics*, **37**, 321–354.

Vollmer, S. (2013). Posterior Consistency for Bayesian Inverse Problems Through Stability and Regression Results. *Inverse Problems*, **29**. Article number 125011.

Wahba, G. (1978). Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression. *Journal of the Royal Statistical Society B*, **40**, 364–372.

Wahba, G. (1990). Spline Functions for Observational Data. CBMS-NSF Regional Conference series, SIAM. Philadelphia.

List of Publications from the Ph.D. Dissertation

Published Papers

- **Chatterjee, D. Maitra, T and Bhattacharya, S. (2020).** A Short Note on Almost Sure Convergence of Bayes Factors in the General Set-Up. *The American Statistician*, 74(1), 17–20.
  (DOI Link: https://doi.org/10.1080/00031305.2017.1397548)
  (Also available at https://arxiv.org/abs/1703.04956)

- **Chatterjee, D. and Bhattacharya, S. (2017).** A Statistical Perspective of Inverse and Inverse Regression Problems. *RASHI*, 2, 67–82
  Available at http://www.sasaa.org/complete_journal/vol2__9.pdf
  The latest version is available at https://arxiv.org/abs/1707.06852

ArXiv Preprints

- **Chatterjee, D. and Bhattacharya, S. (2020)**. Posterior Convergence of Gaussian and General Stochastic Process Regression Under Possible Misspecifications.
  Available at https://arxiv.org/abs/1810.10495

- **Chatterjee, D. and Bhattacharya, S. (2020)**. Posterior Convergence of Nonparametric Binary and Poisson Regression Under Possible Misspecifications.
  Available at https://arxiv.org/abs/2005.00234

- **Chatterjee, D. and Bhattacharya, S. (2020)**. Posterior Consistency of Bayesian Inverse Regression and Inverse Reference Distributions.
  Available at https://arxiv.org/abs/2005.00236

- **Chatterjee, D. and Bhattacharya, S. (2020)**. Convergence of Pseudo-Bayes Factors in Forward and Inverse Regression Problems.
  Available at https://arxiv.org/abs/2006.06020

- **Chatterjee, D. and Bhattacharya, S. (2020)**. A Bayesian Multiple Testing Paradigm for Model Selection in Inverse Regression Problems.
  Available at https://arxiv.org/abs/2007.07847

- **Chatterjee, D. and Bhattacharya, S. (2020)**. How Ominous is the Future Global Warming Premonition?
  Available at https://arxiv.org/abs/2008.11175