Brief Papers

# The Connection Between Bayesian Estimation of a Gaussian Random Field and RKHS

Aleksandr Y. Aravkin, Bradley M. Bell, James V. Burke, and Gianluigi Pillonetto

***Abstract*— Reconstruction of a function from noisy data is key in machine learning and is often formulated as a regularized optimization problem over an infinite-dimensional reproducing kernel Hilbert space (RKHS). The solution suitably balances adherence to the observed data and the corresponding RKHS norm. When the data fit is measured using a quadratic loss, this estimator has a known statistical interpretation. Given the noisy measurements, the RKHS estimate represents the posterior mean (minimum variance estimate) of a Gaussian random field with covariance proportional to the kernel associated with the RKHS. In this brief, we provide a statistical interpretation when more general losses are used, such as absolute value, Vapnik or Huber. Specifically, for any finite set of sampling locations (that includes where the data were collected), the maximum *a posteriori* estimate for the signal samples is given by the RKHS estimate evaluated at the sampling locations. This connection establishes a firm statistical foundation for several stochastic approaches used to estimate unknown regularization parameters. To illustrate this, we develop a numerical scheme that implements a Bayesian estimator with an absolute value loss. This estimator is used to learn a function from measurements contaminated by outliers.**

***Index Terms*— Gaussian processes, kernel-based regularization, Markov chain Monte Carlo (MCMC), regularization networks, representer theorem, reproducing kernel Hilbert spaces (RKHSs), support vector regression.**

## I. Introduction

Reconstruction of a function $F : \mathscr{X} \to \mathbf{R}$ from noisy data is key in machine learning [1], [2]. A popular approach to this problem is minimizing a regularized functional with respect to a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ [3]–[6]. To be specific, regularization in $\mathscr{H}$ estimates $F$ using $\hat{F}$ defined by

$$\hat{F} = \arg\min_{F \in \mathscr{H}} \left( \sum_{i=1}^{N} V_i[y_i - F(x_i)] + \gamma \|F\|^2_{\mathscr{H}} \right) \tag{1}$$

where $\gamma \in \mathbf{R}^+$ is the regularization parameter, $\mathscr{X}$ is a set (finite or infinite), $x_i \in \mathscr{X}$ is the location where $y_i \in \mathbf{R}$ is measured, $V_i : \mathbf{R} \to \mathbf{R}^+$ is the loss function for $y_i$, and $\|\cdot\|_{\mathscr{H}}$ is the RKHS norm induced by the positive definite reproducing kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbf{R}$ [3]. Here, $y_i$ is the $i$th element of a vector $y$, which is the default meaning for subscripts. Note that $x_i$ is the $i$th measurement location (not the $i$th element of a vector $x$), and $V_i$ is the loss function corresponding to the $i$th residual.

One of the important features of the above approach is that, even if the dimension of $\mathscr{H}$ is infinite, the solution belongs to a finite-dimensional subspace. In fact, under mild assumptions on the loss, according to the representer theorem [7], [8], $\hat{F}$ in (1) is the sum of kernel sections $K_i : \mathscr{X} \to \mathbf{R}$ defined by $K_i(x) = K(x_i, x)$. To be specific

$$\hat{F}(\cdot) = \sum_{i=1}^{N} \hat{c}_i K_i(\cdot) \tag{2}$$

where $\hat{c}$ is defined by

$$\hat{c} = \arg\min_{c \in \mathbf{R}^N} \left( \sum_{i=1}^{N} V_i \left[ y_i - \sum_{j=1}^{N} K(x_i, x_j) c_j \right] + \gamma\, c^{\mathrm{T}} \overline{K} c \right). \tag{3}$$

Here and below, $\overline{K} \in \mathbf{R}^{N \times N}$ denotes the kernel matrix, or Gram matrix, defined by $\overline{K}_{ij} = K(x_i, x_j)$. When the component loss functions $V_i$ are quadratic, the problem in (1) admits the structure of a regularization network [1] and also has a statistical interpretation. Specifically, suppose that $F$ is a zero-mean Gaussian random field with a prior covariance proportional to $K$, and that $F$ is independent of white Gaussian measurement noise. Then, given the measurements, for every $x$, the value $\hat{F}(x)$ is the posterior mean, and hence the minimum variance estimate of $F(x)$ [9, Sec. 2.3]. This connection, briefly reviewed in Section III, is well known in the literature and was initially studied in [10] in the context of spline regression [5], [11], [12]. It can be proved using the representer theorem. In the case of quadratic loss functions, it also yields the following closed form expression for the coefficients $\hat{c}_i$ in (2):

$$\hat{c} = (\overline{K} + \gamma\, \mathbf{I}_N)^{-1} y \tag{4}$$

where $y \in \mathbf{R}^N$ is the vector of measurements $y_i$ and $\mathbf{I}_N$ is the $N \times N$ identity matrix. Typically, a closed from expression of this type is not available when the component loss functions $V_i$ are not quadratic.

A formal statistical model for more general loss functions (e.g., the Vapnik $\varepsilon$-insensitive loss used in support vector regression [13]–[15]) is missing from the literature. After interpreting the $V_i$ as alternative statistical models for the observation noise, many papers argue that $\hat{F}$ in (1) can be viewed as a maximum *a posteriori* (MAP) estimator assuming the *a priori* probability density of $F$ is proportional to $\exp(-\|F\|^2_{\mathscr{H}})$ [14, Sec. 7]. These kinds of statements are informal, since in an infinite-dimensional function space, the concept of probability density is not well defined, see [16] for a thorough treatment of Gaussian measures.

The main contribution of this note is to provide a rigorous statistical model that justifies $\hat{F}$ as an estimate of a Gaussian random field. This connection provides a firm statistical foundation for several stochastic approaches for estimating unknown regularization parameters. Examples of such parameters include $\gamma$ in (1) and possibly other parameters used to specify $K$. As an illustration of such approaches, we make

Manuscript received February 26, 2013; revised April 30, 2014; accepted July 3, 2014. Date of publication August 5, 2014; date of current version June 16, 2015. This work was supported in part by the European Community's Seventh Framework Programme [FP7/2007-2013] under Grant FP7-ICT-223866-FeedNetBack, in part the HYCON2 Network of Excellence under Grant 257462, and in part by the Basic Research Investment Fund Project entitled Learning meets time.

A. Y. Aravkin is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: saravkin@us.ibm.com).

B. M. Bell is with the Applied Physics Laboratory, Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA 98105 USA (e-mail: bradbell@uw.edu).

J. V. Burke is with the Department of Mathematics, University of Washington, Seattle, WA 98105 USA (e-mail: burke@math.washington.edu).

G. Pillonetto is with the Department of Information Engineering, University of Padova, Padova 35131, Italy (e-mail: giapi@dei.unipd.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2337939

another contribution by developing a new Bayesian estimator. The estimator uses the absolute value ($\ell_1$) loss and the Markov chain Monte Carlo (MCMC) framework [17] to recover a function from measurements contaminated by outliers. It compares favorably with tuning approaches that rely on cross validation, and with recently proposed techniques in [18], where $\gamma$ is determined by combining Mallows $C_p$ statistic with the concept of equivalent degrees of freedom (EDF).

The structure of this brief is as follows. In Section II, we formulate the statistical model. In Section III, we review the connection between regularized estimation in RKHS and estimation in the quadratic case, and then extend this connection to more general losses. Section IV uses this connection to describe Bayesian approaches that estimate regularization parameters, in addition to the unknown function. A numerical experiment is then reported in Section V to illustrate the theoretical results. Section VI contains a summary and conclusion. The proofs are presented in Section VI.

## II. Statistical Model

Here and below, $\mathbf{E}[\cdot]$ indicates the expectation operator, and given (column) random vectors $u$ and $v$, we define

$$\mathrm{cov}[u, v] = \mathbf{E}[(u - \mathbf{E}[u])(v - \mathbf{E}[v])^{\mathrm{T}}].$$

We assume that the measurements $y_i$ are obtained by measuring the function $F$ at sampled points $x_i$ in the presence of additive noise

$$y_i = F(x_i) + e_i, \quad i = 1, \ldots, n \tag{5}$$

where each $x_i$ is a known sampling location. We make the following assumptions.

*Assumption 1:* We are given a known strictly positive definite[1] autocovariance function $K$ on $\mathscr{X} \times \mathscr{X}$ and a scalar $\lambda > 0$ such that for any sequence of points $\{x_j : j = 1, \ldots, J\}$, the vector $f = [F(x_1), \ldots, F(x_J)]$ is a Gaussian random variable with mean zero and covariance given by

$$\mathrm{cov}(f_j, f_k) = \lambda K(x_j, x_k). \qquad \blacksquare$$

A random function $F$ that satisfies Assumption 1 is often referred to as a zero-mean Gaussian random field on $\mathscr{X}$.

*Assumption 2:* We are given a sequence of measurement pairs $(x_i, y_i) \in \mathscr{X} \times \mathbf{R}$ and corresponding loss functions $V_i$ for $i = 1, \ldots, N$. In addition, we are given a scalar $\sigma > 0$ such that

$$\mathbf{p}(y|F) \propto \prod_{i=1}^{N} \exp\left(-\frac{V_i[y_i - F(x_i)]}{2\sigma^2}\right).$$

Furthermore, the measurement noise random variables $e_i = y_i - F(x_i)$ are independent of the the random function $F$. $\blacksquare$

For example, $V_i(r) = r^2$ corresponds to Gaussian noise, while using $V_i(r) = |r|$ corresponds to Laplacian noise. These loss functions (and corresponding standardized densities) are pictured in Fig. 1. The statistical interpretation of an $\epsilon$-insensitive $V_i$ in terms of Gaussians with mean and variance described by suitable random variables can be found in [19].

## III. Estimation in RKHSs

### A. Gaussian Measurement Noise

We first consider the case of Gaussian measurement noise, i.e., $V_i(r) = r^2$. This corresponds to modeling the $\{e_i\}$ as independent identically distributed. Gaussian random variables with variance $\sigma^2$. In view of the independence of $F$ and $e$, it turns out that $F(x)$ and $y$ are jointly Gaussian for any $x \in \mathscr{X}$. Hence, the posterior $\mathbf{p}[F(x)|y]$ is also Gaussian. The mean and variance for this posterior can be calculated using the following proposition [9, Example 3.6].

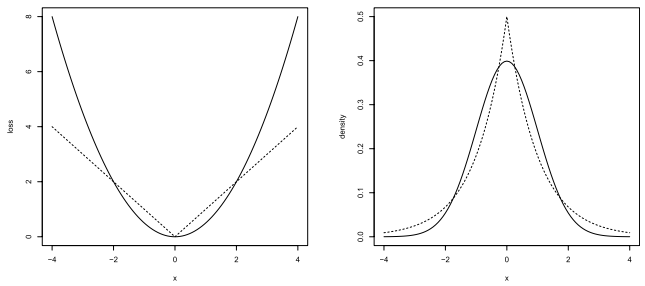


Fig. 1. Left: quadratic and absolute losses are solid and dashed lines. Right: mean zero variance one Gaussian (solid) and Laplace (dashed) densities. Note that Laplace has heavier tails than the Gaussian, which explains its robustness properties.

*Proposition 3:* Suppose $u$ and $v$ are jointly Gaussian random vectors. Then, $\mathbf{p}(u|v)$ is also Gaussian with mean and autocovariance given by

$$\begin{aligned}\mathbf{E}(u|v) &= \mathbf{E}(u) + \mathrm{cov}(u, v)\mathrm{cov}(v, v)^{-1}[v - \mathbf{E}(v)] \\ \mathrm{cov}(u, u|v) &= \mathrm{cov}(u, u) - \mathrm{cov}(u, v)\mathrm{cov}(v, v)^{-1}\mathrm{cov}(v, u).\end{aligned} \qquad \blacksquare$$

Suppose Assumptions 1 and 2 hold with $V_i(r) = r^2$ and $K_i$ as given in (2) for $i = 1, \ldots, N$. It follows that $y$ is Gaussian. Applying Proposition 3 with $u = F(x)$ and $v = y$, we obtain $\mathbf{E}(u) = 0$, $\mathbf{E}(v) = 0$, and

$$\mathbf{E}[F(x)|y] = \lambda[K_1(x) \ldots K_N(x)](\lambda\overline{K} + \sigma^2\mathbf{I}_N)^{-1}y.$$

Using the notation $\gamma = \sigma^2/\lambda$, one obtains

$$\begin{aligned}\mathbf{E}[F(x)|y] &= [K_1(x) \ldots K_N(x)](\overline{K} + \gamma\,\mathbf{I}_N)^{-1}y \\ &= \sum_{i=1}^{N} \hat{c}_i K_i(x)\end{aligned}$$

where $\hat{c}$ is computed using (4). This shows that in the Gaussian case the minimum variance estimate coincides with $\hat{F}$ defined by (1). We formalize this result in the following proposition.

*Proposition 4:* Suppose that $F$ satisfies Assumption 1 and $\mathbf{p}(y|F)$ satisfies Assumption 2 with $V_i(r) = r^2$. Then, the minimum variance estimate of $F(x)$ given $y$ is $\hat{F}(x)$ defined by (1), with $\gamma = \sigma^2/\lambda$ and $\mathscr{H}$, the RKHS induced by $K$. $\blacksquare$

### B. Non-Gaussian Measurements: MAP Estimate

We now consider what happens when the Gaussian assumptions on $e_i$ are removed. If the probability density function for $F$ was well defined and given by

$$\mathbf{p}(F) \propto \exp\left(-\frac{\|F\|^2_{\mathscr{H}}}{2\lambda}\right).$$

Then, the posterior density conditional on the data would be

$$\mathbf{p}(F|y) \propto \exp\left(-\sum_{i=1}^{N} \frac{V_i[y_i - F(x_i)]}{2\sigma^2} - \frac{\|F\|^2_{\mathscr{H}}}{2\lambda}\right).$$

In this case, the negative log of $\mathbf{p}(F|y)$ would be proportional to the objective in (1). Hence, one could immediately conclude that $\hat{F}$ is the MAP estimator. Unfortunately, the posterior density of $F$ on a function space is not well defined. However, one can consider the MAP estimates corresponding to any finite sample of $F$ that includes

[1]This means that the kernel matrix is invertible given any set of distinct input locations. This assumption is made to simplify the exposition. With minor modifications, all the results reported in the sequel hold also assuming that the covariance is just positive definite.

the observations $y_i$ (since these are finite-dimensional estimation problems). The following proposition shows that $\hat{F}$ solves all such problems.

*Proposition 5:* Suppose that $F$ satisfies Assumption 1 and $\mathbf{p}(y|F)$ satisfies Assumption 2. Let $\{x_i : i = N+1, \ldots, N+M\}$ be an arbitrary set of points in $\mathscr{X}$ where $M$ is a given nonnegative integer, and define

$$f = [F(x_1), \ldots, F(x_{N+M})]^{\mathrm{T}}.$$

Then, the MAP estimate for $f$ given $y$ is

$$\arg\max_f \mathbf{p}(y|f)\mathbf{p}(f) = [\hat{F}(x_1), \ldots, \hat{F}(x_{N+M})]^{\mathrm{T}}$$

where $\hat{F}$ is defined by (1), with $\gamma = \sigma^2/\lambda$, and $\mathscr{H}$ is the RKHS induced by $K$. ■

### C. Non-Gaussian Measurements: Minimum Variance Estimate

When considering non-Gaussian measurement loss functions, the minimum variance estimate $\mathbf{E}[F(\cdot)|y]$ and the MAP estimate $\hat{F}(\cdot)$ are different.

*Example 6:* Consider the case where $N = 1$, $M = 0$, $V_1(r) = |r|$, $y = 1$, and $\lambda = 1$, $\sigma = 1$, $K(x_1, x_1) = 1$. For this case, $f = F(x_1)$, and the MAP estimate for $f$ given $y$ is

$$\hat{f} = \arg\min_f (f^2 + |1 - f|) = 1/2.$$

Define $A > 0$ by

$$A = \int_{-\infty}^{+\infty} \exp(-f^2 - |1 - f|)\mathbf{d}f.$$

The difference between the minimum variance estimate and the MAP estimate is (see Appendix D for details)

$$\mathbf{E}(f|y) - \hat{f} = \frac{\exp(-3/4)}{A} \int_{1/2}^{+\infty} s\, \frac{\exp(1-2s) - 1}{\exp(s^2)} \mathbf{d}s. \tag{6}$$

For $s > 1/2$, the integrand in (6) is negative, so the right-hand side is negative, and $\mathbf{E}(f|y) < \hat{f}$.

The linear span of the kernel sections $K_i$ contains every possible MAP estimate of $F$; see Proposition 5 and (2). The following proposition shows that the minimum variance estimate $\mathbf{E}[F(\cdot)|y]$ also belongs to this subspace of $\mathscr{H}$.

*Proposition 7:* Suppose that $F$ satisfies Assumption 1 and $\mathbf{p}(y|F)$ satisfies Assumption 2. Define

$$g = [F(x_1), \ldots, F(x_N)]^{\mathrm{T}}$$
$$\hat{d} = \overline{K}^{-1}\mathbf{E}(g|y).$$

For each $x \in \mathscr{X}$, the minimum variance estimate of $F(x)$ is

$$\mathbf{E}[F(x)|y] = \sum_{i=1}^{N} \hat{d}_i K_i(x). \tag{7}$$

■

Note that, given $\sigma$ and $\lambda$, the vector $\mathbf{E}(g|y)$ can be approximated using the relation

$$\mathbf{p}(g|y) \propto \exp\left(-\sum_{i=1}^{N} \frac{V_i[y_i - g_i]}{2\sigma^2} - \frac{g^T \overline{K}^{-1} g}{2\lambda}\right)$$

together with random sampling technique, such as MCMC.

## IV. Function and Regularization Parameter Estimation

In real applications, the regularization parameter $\gamma = \sigma^2/\lambda$ is typically unknown and needs to be inferred from data. In the case of Gaussian measurement noise, this problem is often solved by exploiting the stochastic interpretation given by Proposition 4. For example, following an empirical Bayes approach, the marginal likelihood can be computed analytically and the unknown parameters (often called hyperparameters) can be estimated by optimizing this likelihood [12, Sec. 5.4.1], [20]. The regularization parameter $\gamma$ is then set to its estimated value, and $\hat{F}$ in (1) is obtained using (4) and (2). Propositions 5 and 7 provide the statistical foundations that extend this technique to non-Gaussian measurement noise.

In the more general case of Assumption 2 (non-Gaussian measurement noise), the marginal likelihood cannot be computed analytically. Let $\eta$ denote the vector of unknown hyperparameters ($\sigma$ and/or $\lambda$) and recall the notation $g = [F(x_1), \ldots, F(x_N)]^{\mathrm{T}}$. Following a Bayesian approach, we model $\eta$ as a random vector with prior probability density $\mathbf{p}(\eta)$. The conditional density for the data $y$ and the unknown function samples $g$, given the hyperparameters $\eta$ is

$$\mathbf{p}(y, g|\eta) \propto \prod_{i=1}^{N} \exp\left(-\frac{V_i(y_i - g_i)}{2\sigma^2} - \frac{g^T \overline{K}^{-1} g}{2\lambda}\right)$$

where the proportionality factor may depend on $\eta$. The difficulty underlying the estimation of $\eta$ is that $\mathbf{p}(\eta|y)$ is not, in general, available in closed form. One possibility is to use stochastic simulation techniques, e.g., MCMC [17] or particle filters [21], which can sample from $\mathbf{p}(\eta, g|y)$ provided that a suitable proposal density for $\eta$ and $g$ can be designed. An MCMC scheme for sampling from the posterior for $g$ and $\eta$ (corresponding to the $\ell_1$ measurement model) is described in Appendix E and applied in Section V. Proposition 7 is especially important because it shows how to compute $\mathbf{E}[F(x)|y]$ for any $x$ from the minimum variance estimate for $g$. Similarly, given an estimate of $\eta$, we can use Proposition 5 to compute the corresponding $\hat{F}(x)$ for any $x$.

## V. Simulation Example

We consider the simulated problem in [18, Sec. 5.1]. The unknown function to be estimated is

$$F_0(x) = \exp[\sin(8x)]\,, \quad 0 \le x \le 1$$

which is displayed as the thick line in the bottom two panels of Fig. 2. This function is reconstructed from the measurements

$$y_i = F_0(x_i) + e_i \quad \text{with} \quad x_i = (i-1)/63\,, \quad i = 1, \ldots, 64.$$

We include two Monte Carlo experiments, each consisting of 300 function reconstructions. In the first experiment, for each reconstruction, measurements $y_i$ are generated using $e_i \sim \mathbf{N}(0, 0.09)$. A typical data set is plotted as circles ∘ in the bottom left panel of Fig. 2. In the second experiment, we simulate measurement outliers by adding, with probability 0.1, a random offset equal to ±3 to each measurement generated in the first experiment. A typical data set is plotted as circles in the bottom right panel of Fig. 2.

Both experiments compare five different methods for modeling the measurement noise and estimating the kernel scale factor $\lambda$ (described below). All the methods model the function correlations using a cubic spline kernel [5, Ch. 1], shifted by 1 to deal with $f(0) \neq 0$. To be specific

$$K(x_i, x_j) = (x_i + 1)(x_j + 1)\min(x_i + 1, x_j + 1)/2 - \min(x_i + 1, x_j + 1)^3/6.$$

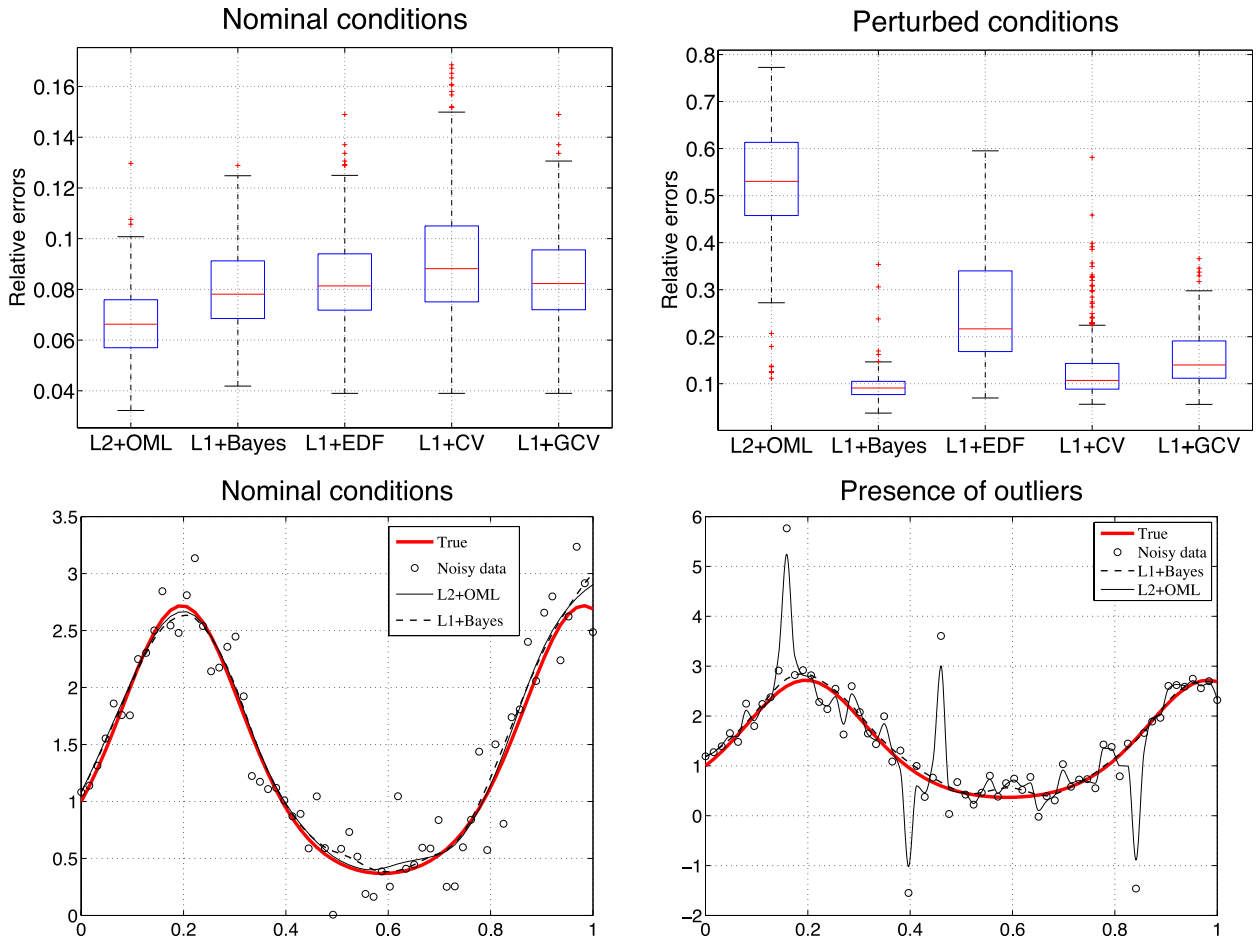


Fig. 2. Simulation. Top: boxplot of relative errors, nominal conditions (left), and in the presence of outliers (right). Bottom: true function (thick line), noisy output samples (∘), estimate using $L_2 + \text{OML}$ (solid line), estimate using $L_1 + \text{Bayes}$ (dotted line), nominal conditions (left), and in the presence of outliers (right).

In addition, once an estimate for $\lambda$ is determined, all methods use the MAP estimator (1) to reconstruct the function $F_0(x)$ by solving the problem in (3).

1) $L_2 + OML$: The measurement noise is modeled by a quadratic loss with $\sigma^2 = 0.09$. (During the second experiment, measurement outliers represent unexpected model noise.) For each reconstruction, the kernel scale factor $\lambda$ is estimated using marginal likelihood optimization [12, Sec. 5.4.1]. Once the estimate for $\lambda$ is determined, the solution of (3) is obtained using (4).
2) $L_1 + Bayes$: The measurement noise is modeled by the $\ell_1$ loss with $\sigma$ chosen so the variance of the corresponding Laplace distribution is 0.09. The kernel scale factor $\lambda$ is estimated by following the Bayesian approach discussed (for non-Gaussian noise) in Section IV. More details can be found in Appendix E. Once the estimate for $\lambda$ is determined, the problem in (3) is solved using the interior point (IP) method described in [22].
3) $L_1 + EDF$: The measurement noise is modeled by the $\ell_1$ loss with $\sigma$ chosen so the variance of the corresponding Laplace distribution is 0.09. The kernel scale factor $\lambda$ is estimated using the approach described in [18], i.e., relying on $C_p$-like statistic and the concept of EDF. The notation $C$ in [18, eq. (1)], corresponds to $\sigma^{-2}/2$ in this brief. The objective in [18, eq. (19)] is optimized on a grid containing 50 values of $\log_{10}(C)$ uniformly distributed on $[1, 6]$. The number of degrees of freedom entering [18, eq. (19)], as a function of $C$, is determined at every run as described in [18, Remark 1] (with $\epsilon = 0$). Once the estimate for $\lambda$ is determined, the problem in (3) is solved using the IP method [22].
4) $L_1 + CV$: The measurement noise is modeled by the $\ell_1$ loss and the regularization parameter $\gamma$ is estimated using cross validation. To be specific, given a value for $\gamma$, the data with odd indices $\{x_i, y_i\}_{i=1,3,\ldots}$ are used to define $\hat{F}_\gamma(x)$ by solving problem (3) using the IP method [22]. Given $\hat{F}_\gamma$, the data with even indices $\{x_i, y_i\}_{i=2,4,\ldots}$ define the corresponding prediction errors $y_i - \hat{F}_\gamma(x_i)$. The regularization parameter is chosen to minimize the sum of the squares of the prediction errors. Given this optimal $\gamma$, the final function estimate $\hat{F}(x)$ is obtained by solving problem (3) using all 64 measurements.
5) $L_1 + GCV$: The regularization parameter $\gamma$ is chosen to optimize the generalized cross validation score (defined, e.g., [23, p. 325, eq. (9.20)]) with the EDF computed as described in [18, Remark 1] (with $\epsilon = 0$). The parameter $\gamma$ is searched on the same grid as used for the $L_1 + \text{EDF}$ above. The final function estimate is then obtained by setting $\gamma$ to its optimal value and solving (3) using the IP method [22].

If these tests were a real data examples, we would do an out of sample prediction comparison. Since they are simulations, we know the true function $F_0$, and hence use it to measure the accuracy of the estimated function $\hat{F}$. The top panels of Fig. 2 are boxplots, for the

five methods, of the 300 relative errors defined by

$$\sqrt{\frac{\sum_{i=1}^{64}[F_0(x_i)-\hat{F}(x_i)]^2}{\sum_{i=1}^{64}F_0^2(x_i)}}.$$

Each boxplot has a line for the median, box surrounding the 25th and 75th percentiles, and whiskers that extend to the most extreme data. Relative errors, that the plotting system considers outliers, are plotted beyond the whiskers. In the absence of measurement outliers (top left panel), all the methods provide accurate function reconstructions, and the $L_2$+OML method performs best. The bottom left panel contains the results of a single reconstruction using the $L_2$ + OML and $L_1$ + Bayes methods.

The situation dramatically changes in the presence of measurement outliers (top right panel). As expected, the errors for the $L_2$ + OML method increase significantly. The estimate obtained by the $L_2$ + OML method for a single reconstruction is displayed in the bottom-right panel (solid line). It is apparent that the quadratic loss is very vulnerable to unexpected model deviations. On the other hand, the estimate obtained by $L_1$ + Bayes method is much closer to the truth. This remarkable performance is confirmed by the top-right panel. The errors corresponding to the $L_1$ + Bayes method with outliers is similar to the performance obtained in the absence of outliers. In addition, the $L_1$+Bayes method outperforms all the other estimators.

*Remark 8:* The MCMC scheme discussed in the last part of Appendix E was also used to compute the minimum variance estimate of $F$. The performance of this estimator is virtually identical to that of $L_1$ + Bayes. Once the MCMC samples are computed, there is very little extra computation required to obtain the minimum variance estimate of $F$. In addition, it does not require the somewhat complex optimization procedure described in [22].

*Remark 9:* We also considered a third experiment, where the true value of the noise variance, i.e., $\sigma^2 = 0.99$, is provided to $L_2$ + OML, $L_1$ + EDF and $L_1$ + Bayes methods. The average error of the $L_2$ + OML method decreases from 0.53 to 0.22, while that of the $L_1$ + EDF method decreases from 0.25 to 0.15. The average error of the $L_1$ + Bayes method does not change significantly, staying around 0.1 in both the second and third experiments. As a final note, both $L_1$ + CV and $L_1$ + GCV do not use (or require) the level of the noise variance, so their average errors (0.14 and 0.15, respectively) are not influenced.

## VI. Conclusion

When the RKHS induced by $K$ is infinite dimensional, the realizations of the Gaussian random field with autocovariance $K$ do not fall in $\mathscr{H}$ with probability one, see [24, eq. (34)] and also [25]–[27] for generalizations. A simple heuristic argument illustrating this fact can be also found in [5, Ch. 1]. The intuition here is that the realizations of $F$ are much less regular than functions in the RKHS whose kernel is equal to the autocovariance $K$. On the other hand, in the case of Gaussian measurement noise, $\hat{F}$ defined in (1) is the minimum variance estimate, see Proposition 4. In this note, we proved a formal connection between Bayesian estimation and the more general case prescribed by Assumption 2. Given the training set $\{(x_i, y_i)\}$, for any finite set of locations that include the training locations $\{x_i\}$, the MAP estimate of $F$ at the locations is the RKHS estimate evaluated at these locations. We have also shown that every possible MAP estimate of $F$, in the sense of Proposition 5, belongs to a finite-dimensional subspace of $\mathscr{H}$. In addition, the minimum variance estimate of $F$ is also in this subspace. These results can be extended to more general cases using more general versions of the representer theorem (2). This link between statistical estimation and RKHS regularization provides a foundation for the application of statistical approaches to joint estimation of the function and the regularization parameters. The simulation example in this brief illustrates the utility of this connection.

## Appendix

### A. Lemmas

We begin the appendix with two lemmas that are instrumental in proving Proposition 5.

*Lemma 10:* Suppose that $g$ and $h$ are jointly Gaussian random vectors. It follows that:

$$\max_h \log \mathbf{p}(h|g) = -\log\det\{2\pi[\mathrm{cov}(h,h)-\mathrm{cov}(h,g)\mathrm{cov}(g,g)^{-1}\mathrm{cov}(g,h)]\}/2$$

and this maximum does not depend on the value of $g$.

*Proof:* The proof comes from well-known properties of joint Gaussian vectors [9]. The conditional density $\mathbf{p}(h|g)$ is Gaussian and is given by

$$-2\log\mathbf{p}(h|g) = \log\det[2\pi\,\mathrm{cov}(h,h|g)] + [h-\mathbf{E}(h|g)]^{\mathrm{T}}\mathrm{cov}(h,h|g)^{-1}[h-\mathbf{E}(h|g)]$$

where recalling also Proposition 3

$$\mathrm{cov}(h,h|g) = \mathrm{cov}(h,h) - \mathrm{cov}(h,g)\mathrm{cov}(g,g)^{-1}\mathrm{cov}(g,h).$$

Thus, $\mathrm{cov}(h,h|g)$ does not depend on the value of $g$ (and it would not make sense for it to depend on the value of $h$). Hence, one has

$$\arg\max_h \mathbf{p}(h|g) = \mathbf{E}(h|g)$$
$$\max_h \log\mathbf{p}(h|g) = -\log\det[2\pi\,\mathrm{cov}(h,h|g)]/2.$$

This equation, and the representation for $\mathrm{cov}(h,h|g)$ above completes the proof of this lemma. ∎

*Lemma 11:* Assume that $g$ and $h$ are jointly Gaussian random vectors and that $y$ is a random vector such that $\mathbf{p}(y|g,h) = \mathbf{p}(y|g)$, and suppose we are given a value for $y$. Define the corresponding estimates for $g$ and $h$ by

$$(\hat{g},\hat{h}) = \arg\max_{g,h}\mathbf{p}(y,g,h)$$

and assume the above maximizers are unique. It follows that:

$$\hat{g} = \arg\max_g \mathbf{p}(y|g)\mathbf{p}(g) \tag{8}$$

$$\hat{h} = \arg\max_h \mathbf{p}(h|g=\hat{g}). \tag{9}$$

*Proof:* We have

$$\mathbf{p}(y,g,h) = \mathbf{p}(y|g,h)\,\mathbf{p}(h|g)\,\mathbf{p}(g) = \mathbf{p}(y|g)\,\mathbf{p}(g)\,\mathbf{p}(h|g)$$
$$\max_{g,h}\mathbf{p}(y,g,h) = \max_g\{[\mathbf{p}(y|g)\,\mathbf{p}(g)]\ \max_h\mathbf{p}(h|g)\}.$$

It follows from Lemma 10 that $\max_h \mathbf{p}(h|g)$ is constant with respect to $g$. Hence $\hat{g} = \arg\max_g[\mathbf{p}(y|g)\ \mathbf{p}(g)]$, which completes the proof of (8), and

$$\max_{g,h}\mathbf{p}(y,g,h) = \mathbf{p}(y|\hat{g})\,\mathbf{p}(\hat{g})\max_h\mathbf{p}(h|g=\hat{g})$$

which completes the proof of (9). ∎

### B. Proof of Proposition 5

The kernel matrix $\overline{K}$ is invertible (Assumption 1). Define the random vectors $g$ and $h$ by

$$g = [F(x_1), \ldots, F(x_N)]^{\mathrm{T}}$$
$$h = [F(x_{N+1}), \ldots, F(x_{N+M})]^{\mathrm{T}}.$$

It follows that $f$ in Proposition 5 is given by $f = (g^{\mathrm{T}}, h^{\mathrm{T}})^{\mathrm{T}}$. Notice that $\mathbf{p}(y|f) = \mathbf{p}(y|g)$ and that Lemma 11 can be applied. From (8) and the hypotheses above, we obtain

$$\begin{aligned}\hat{g} &= \arg\max_g \mathbf{p}(y|g)\mathbf{p}(g) \\ &= \arg\max_g \left( \frac{1}{2\sigma^2} \sum_{i=1}^{N} V_i[y_i - g_i] + \frac{g^{\mathrm{T}} \overline{K}^{-1} g}{2\lambda} \right).\end{aligned}$$

Using the representation $g = \overline{K}c$, we obtain

$$\hat{c} = \arg\max_c \left( \frac{1}{2\sigma^2} \sum_{i=1}^{N} V_i \left[ y_i - \sum_{j=1}^{N} K(x_i, x_j) c_j \right] + \frac{c^{\mathrm{T}} \overline{K} c}{2\lambda} \right).$$

This agrees with (3), because $\gamma = \sigma^2/\lambda$, and thereby shows

$$\hat{g} = [\hat{F}(x_1), \ldots, \hat{F}(x_N)]^{\mathrm{T}}.$$

Finally, by Proposition 3 and Lemma 10 in conjunction with (2) and (9), and the expression for $\hat{g}$ above, we obtain

$$\begin{aligned}\hat{h} &= \mathrm{cov}(h, g)\mathrm{cov}(g, g)^{-1}\hat{g} \\ &= \mathrm{cov}(h, g)(\lambda \overline{K})^{-1}(\overline{K}\hat{c}) \\ &= \begin{pmatrix} K_1(x_{N+1}) & \ldots & K_N(x_{N+1}) \\ \vdots & \ddots & \vdots \\ K_1(x_{N+M}) & \ldots & K_N(x_{N+M}) \end{pmatrix} \begin{pmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_N \end{pmatrix} \\ &= [\hat{F}(x_{N+1}), \ldots, \hat{F}(x_{N+M})]^{\mathrm{T}}.\end{aligned}$$

Combining this with the formula for $\hat{g}$ above, we conclude

$$[\hat{F}(x_1), \ldots, \hat{F}(x_{N+M})]^{\mathrm{T}} = \arg\max_f \mathbf{p}(y, f)$$

which completes the proof of Proposition 5.

### C. Proof of Proposition 7

To obtain the representation (7), we compute $\mathbf{E}[F(x)|y]$ by first projecting $F(x)$ onto $g$ and then onto $y$, i.e., using the equivalence

$$\mathbf{E}[F(x)|y] = \mathbf{E}(\, \mathbf{E}[F(x)|g] \,|y).$$

Exploiting Proposition 3, and recalling that $\mathrm{cov}(g, g) = \overline{K}$, the first projection is given by

$$\mathbf{E}[F(x)|g] = \mathrm{cov}[F(x), g]\mathrm{cov}(g, g)^{-1} g = a^{\mathrm{T}} \overline{K}^{-1} g$$

where $a \in \mathbf{R}^N$ and $a_i = \mathrm{cov}[F(x), g_i] = K_i(x)$. The second projection yields

$$\mathbf{E}(\, \mathbf{E}[F(x)|g] \,|y) = a^{\mathrm{T}} \overline{K}^{-1} \mathbf{E}(g|y) = \sum_{i=1}^{N} \hat{d}_i K_i(x)$$

where $\hat{d} = \overline{K}^{-1} \mathbf{E}(g|y)$, which completes the proof.

### D. Proof of (6)

It follows from $N = 1$, $\gamma = 1$, that $c$ is a scalar, $f = F(x_1) = c$, and using (3), we have:

$$\hat{f} = \hat{c} = \arg\min_c |1 - c| + c^2 = 1/2.$$

It also follows that:

$$\mathbf{p}(y|f)\mathbf{p}(f) \propto \exp(-f^2 - |1 - f|).$$

The minimum variance estimate $\mathbf{E}(f|y)$, and its difference from the map estimate $\hat{f}$, are given by

$$\begin{aligned}\mathbf{E}(f|y) &= \frac{1}{A} \int_{-\infty}^{+\infty} f \exp(-f^2 - |1 - f|) \mathbf{d}f \\ \mathbf{E}(f|y) - \hat{f} &= \frac{1}{A} \int_{-\infty}^{1} (f - 1/2) \exp(-f^2 - 1 + f) \mathbf{d}f \\ &\quad + \frac{1}{A} \int_{1}^{+\infty} (f - 1/2) \exp(-f^2 + 1 - f) \mathbf{d}f.\end{aligned}$$

Multiplying both sides of the equation by $A$ and using the change of variables $s = f - 1/2$, we obtain

$$\begin{aligned}&A(\mathbf{E}(f|y) - \hat{f}) \\ &\quad = \int_{-\infty}^{1/2} s e^{-(s+1/2)^2 + s - 1/2} \mathbf{d}s \\ &\qquad + \int_{1/2}^{+\infty} s e^{-(s+1/2)^2 - s + 1/2} \mathbf{d}s \\ &\quad = \int_{-\infty}^{1/2} s e^{-s^2 - 3/4} \mathbf{d}s + \int_{1/2}^{+\infty} s e^{-s^2 - 2s + 1/4} \mathbf{d}s \\ &\quad = \int_{-\infty}^{-1/2} s e^{-s^2 - 3/4} \mathbf{d}s + \int_{1/2}^{+\infty} s e^{-s^2 - 2s + 1/4} \mathbf{d}s \\ &\quad = \int_{1/2}^{+\infty} s e^{-s^2 - 3/4} [e^{1 - 2s} - 1] \mathbf{d}s.\end{aligned}$$

This completes the proof of (6).

### E. Details of the MCMC Scheme for $L_1 + Bayes$

If Assumptions 1 and 2 hold with $V_i(r) = 2(2)^{1/2}\sigma|r|$, the noise $e_i$ is Laplacian with variance $\sigma^2$. In this case, it can be difficult to build an efficient MCMC scheme to sample from the posterior of $\eta$ and $g$. This is because, *a posteriori*, the components of $g$ are generally strongly correlated. It is useful to use a scaled mixture of normal representation because for each normal, the posterior distribution can be represented in closed form. To be specific, each $\mathbf{p}(e_i)$ admits the representation [28]

$$\begin{aligned}\mathbf{p}(e_i) &= \frac{1}{\sqrt{2}\sigma} \exp(-\sqrt{2}\,|e_i|\,/\,\sigma) \\ &= \int_0^{+\infty} \frac{1}{\sqrt{2\pi\,\tau_i}} \exp\left(-\frac{e_i^2}{2\tau_i}\right) \frac{1}{\sigma^2} \exp\left(-\frac{\tau_i}{\sigma^2}\right) \mathbf{d}\tau_i.\end{aligned}$$

Hence, we can model Laplacian noise $e_i$ as a mixture of Gaussians with variances $\tau_i$ that are exponential random variables with probability density

$$\begin{aligned}\mathbf{p}(\tau_i) &= \begin{cases} \exp\left(-\tau_i\,/\,\sigma^2\right)\,/\sigma^2 & \text{if } \tau_i \geq 0 \\ 0 & \text{otherwise} \end{cases} \\ \mathbf{p}(\tau) &= \mathbf{p}(\tau_1) \cdots \mathbf{p}(\tau_N). \qquad (10)\end{aligned}$$

We restrict our attention to the case where $\eta = \lambda$, and use $\tau = (\tau_1, \ldots, \tau_N)^{\mathrm{T}}$ to denote the independent random variables (which are also independent of $\lambda$). We have

$$\mathbf{p}(\tau, \lambda|y) \propto \mathbf{p}(y|\tau, \lambda)\mathbf{p}(\tau)\mathbf{p}(\lambda).$$

Given $\tau$ and $\lambda$, we have the linear Gaussian model $y = g + \xi$, where $g$ and $\xi$ are independently distributed according to

$$g \sim \mathbf{N}(0, \lambda\overline{K}) \quad \text{and} \quad \xi \sim \mathbf{N}[0, \text{diag}(\tau)]$$

where $\text{diag}(\tau)$ is the diagonal matrix with $\tau$ along its diagonal. Note that $\mathbf{p}(y|\tau, \lambda)$ can be computed in closed from using the classical Gaussian marginal likelihood result [12, Sec. 5.4.1]. To be specific, using the notation $C(\tau, \lambda) = \lambda\overline{K} + \text{diag}(\tau)$

$$\mathbf{p}(y|\tau, \lambda) = \frac{1}{\sqrt{2\pi \det[C(\tau, \lambda)]}} \exp\left[-\frac{1}{2} y^{\mathrm{T}} C(\tau, \lambda)^{-1} y\right]. \quad (11)$$

Using an improper flat prior on $\lambda \geq 0$, we obtain

$$\mathbf{p}(\tau, \lambda|y) \propto \begin{cases} \mathbf{p}(y|\tau, \lambda)\mathbf{p}(\tau) & \text{if } \lambda \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

were $\mathbf{p}(y|\tau, \lambda)$ can be computed using (11) and $\mathbf{p}(\tau)$ can be computed using (10). We are now in a position to describe the MCMC scheme used for the $L_1$+Bayes method in Section V. The scale factor $\lambda$, and all the components of $\tau$ are simultaneously updated using a random walk Metropolis scheme [17]. The proposal density is independent normal increments with standard deviation 30 and $\sigma^2/30$ for $\lambda$ and $\tau_i$, respectively. This simple scheme has always led to an acceptance rate over 20%. We have assessed that this follows from the fact that the components of $\lambda$ and $\tau$ have low correlation *a posteriori*. For each function reconstruction, $L = 10^6$ MCMC realizations from $\mathbf{p}(\tau, \lambda|y)$ were obtained by the MCMC scheme (which we denote by $\{\tau^\ell, \lambda^\ell\}$ below). Using the convergence diagnostics described in [29], this allowed us to estimate the quantiles $q = 0.025, 0.25, 0.5, 0.75, 0.975$ of the marginal posterior of $\lambda$ with precision $r = 0.02, 0.05, 0.01, 0.05, 0.02$, respectively, with probability 0.95.

Now consider recovering the minimum variance estimate $\mathbf{E}[F(x)|y]$. We have seen from Proposition 7 that this reduces to computing $\mathbf{E}(g|y)$. Note that, given a value for $\lambda$ and $\tau$, $g$, and $y$ are jointly Gaussian. Applying Proposition 3

$$\begin{aligned} \mathbf{E}(g|y, \tau, \lambda) &= \text{cov}(g, y|\tau, \lambda)\text{cov}(y, y|\tau, \lambda)^{-1} y \\ &= \lambda\overline{K} C(\tau, \lambda)^{-1} y. \end{aligned}$$

Hence, it follows that $\mathbf{E}(g|y)$ can be approximated as:

$$\mathbf{E}(g|y) \approx \overline{K}\, \frac{1}{L} \sum_{\ell=1}^{L} \lambda^\ell C(\tau^\ell, \lambda^\ell)^{-1} y$$

where $\{\tau^\ell, \lambda^\ell\}_{\ell=1}^{L}$ are the realizations from $\mathbf{p}(\tau, \lambda|y)$ achieved by the MCMC scheme above described above.

*Remark 12:* In general, MCMC is more computationally expensive than optimization because MCMC needs to represent the entire support of a probability density to reconstruct it in sampled form. Another complexity of the $L_1$ + Bayes MCMC method is the need to evaluate the marginal likelihood in (11), every time the proposal density is evaluated (which requires $O(N^3)$ operations). Note that the same strategies used to reduce the computational load of optimization methods can be used with MCMC methods. In particular, one can replace the marginal likelihood in (11) with suitable approximations as described in [30] and [31].

## REFERENCES


[1] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proc. IEEE*, vol. 78, no. 9, pp. 1481–1497, Sep. 1990.

[2] H. Zhang, Y. Xu, and J. Zhang, "Reproducing kernel Banach spaces for machine learning," *J. Mach. Learn. Res.*, vol. 10, pp. 2741–2775, Dec. 2009.

[3] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, May 1950.

[4] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2001.

[5] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA, USA: SIAM, 1990.

[6] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architecture," *Neural Comput.*, vol. 7, no. 2, pp. 219–269, Mar. 1995.

[7] G. Wahba, "Support vector machines, reproducing kernel Hilbert spaces and randomized GACV," Dept. Statist., Univ. Wisconsin, Madison, WI, USA, Tech. Rep. 984, 1998.

[8] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," *Neural Netw. Comput. Learn. Theory*, vol. 81, pp. 416–426, Jul. 2001.

[9] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1979.

[10] G. Kimeldorf and G. Wahba, "A correspondence between Bayesan estimation of stochastic processes and smoothing by splines," *Ann. Math. Statist.*, vol. 41, no. 2, pp. 495–502, Apr. 1971.

[11] F. Girosi, M. Jones, and T. Poggio., "Regularization theory and neural networks architectures," *Neural Comput.*, vol. 7, no. 2, pp. 219–269, Mar. 1995.

[12] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[13] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[14] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–150, 2000.

[15] L. Gunter and J. Zhu, "Computing the solution path for the regularized support vector regression," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA, USA: MIT Press, 2006, pp. 483–490.

[16] V. Bogachev, *Gaussian Measures*. Providence, RI, USA: AMS, 1998.

[17] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. London, U.K.: Chapman & Hall, 1996.

[18] F. Dinuzzo, M. Neve, G. De Nicolao, and U. P. Gianazza, "On the representer theorem and equivalent degrees of freedom of SVR," *J. Mach. Learn. Res.*, vol. 8, no. 10, pp. 2467–2495, 2007.

[19] M. Pontil, S. Mukherjee, and F. Girosi, "On the noise model of support vector machine regression," in *Proc. 11th Int. Conf. Algorithmic Learn. Theory (ALT)*, Sydney, Australia, Dec. 2000.

[20] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, May 1992.

[21] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *J. Royal Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 72, no. 3, pp. 269–342, Jun. 2010.

[22] A. Aravkin, J. Burke, and G. Pillonetto, "Nonsmooth regression and state estimation using piecewise quadratic log-concave densities," in *Proc. 51st IEEE Conf. Decision Control (CDC)*, Dec. 2012, pp. 4101–4106.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer-Verlag, 2013, http://statweb.stanford.edu/~tibs/ElemStatLearn/.

[24] E. Parzen, "Probability density functionals and reproducing kernel Hilbert spaces," in *Proc. Symp. Time Ser. Anal.*, 1963, pp. 155–169.

[25] G. Kallianpur, "The role of reproducing kernel Hilbert spaces in the study of Gaussian processes," in *Advances in Probability and Related Topics*. New York, NY, USA: Marcel Dekker, 1970, pp. 49–83.

[26] M. F. Driscoll, "The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 26, no. 4, pp. 309–316, 1973.

[27] M. Lukic and J. Beder, "Stochastic processes with sample paths in reproducing kernel Hilbert spaces," *Trans. Amer. Math. Soc.*, vol. 353, no. 10, pp. 3945–3969, May 2001.

[28] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. Roy. Statist. Soc., Ser. B*, vol. 36, no. 1, pp. 99–102, 1974.

[29] A. Raftery and S. Lewis, "Implementing MCMC," in *Markov Chain Monte Carlo in Practice*. London, U.K.: Chapman & Hall, 1996, pp. 115–130.

[30] J. Quinonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1939–1959, Dec. 2005.

[31] M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *J. Mach. Learn. Res.*, vol. 11, pp. 1865–1881, Mar. 2010.