Neural Computing Research Group
Dept of Computer Science & Applied Mathematics
Aston University
Birmingham B4 7ET
United Kingdom
Tel: +44 (0)121 333 4631
Fax: +44 (0)121 333 4586
http://www.ncrg.aston.ac.uk/

Gaussian Regression and Optimal Finite Dimensional Linear Models

Huaiyu Zhu
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
zhuh@santafe.edu

Christopher K. I. Williams
c.k.i.williams@aston.ac.uk

Richard Rohwer
Prediction Company, 236 Montezuma Ave, Santa Fe, NM 87501, USA
rr@predict.com

Michal Morciniec
Hewlett-Packard Labs, Filton Road, Stoke Gifford, Bristol BS12 6QZ, UK
mmm@hplb.hpl.hp.com

Technical Report NCRG/97/011 July 3, 1997

Abstract

The problem of regression under Gaussian assumptions is treated generally. The relationship between Bayesian prediction, regularization and smoothing is elucidated. The ideal regression is the posterior mean and its computation scales as $O(n^3)$, where n is the sample size. We show that the optimal m-dimensional linear model under a given prior is spanned by the first m eigenfunctions of a covariance operator, which is a trace-class operator. This is an infinite dimensional analogue of principal component analysis. The importance of Hilbert space methods to practical statistics is also discussed.

Keywords: regression, Gaussian measures, linear model, principal component, spline, regularization, eigenfunctions.

1 Introduction

Many problems in computation and statistics can be generally described as fitting a "curve" from a discrete set of data. Here we allow a liberal interpretation of curve which could be any mapping from a finite dimensional space to a finite dimensional space. Such problems are usually studied under the name "regression" in statistics or "approximation" in numerical analysis.

A general treatment of such problems is this: Suppose the input and output spaces are X and Y. We assume there is a prior $P(f)$ which is a Gaussian measure of functions f in a function space $\mathcal{H}$ from X to Y. Denote the data as $z \in Z^n$ where $Z = X \times Y$. Then using a Gaussian noise model the posterior $P(f|z)$ is also a Gaussian measure on $\mathcal{H}$, and the posterior mean $\widehat{f}$ is the "ideal estimate".

However, $\widehat{f}$ usually lies in an infinite dimensional space and its computation involves inverting an n-dimensional matrix, where n is the sample size. Practical computation is usually performed within a finite dimensional model $M \subseteq \mathcal{H}$ of dimension m. Here we only consider linear models. Suppose we use $\|f - g\|$ to measure the discrepancy between the true function f and estimated function g, where $\|\cdot\|$ is a particular norm defined by an inner product. The optimal estimate $\widehat{g} \in M$ is the projection of $\widehat{f}$ onto M under this norm.

Two questions arise naturally from the fact that the actual computation is performed within M. The first question concerns how to compute $\widehat{g}$ (approximately) without computing $\widehat{f}$. Because $\widehat{f}$ is a sufficient statistic while $\widehat{g}$ is not, computation involving only $\widehat{g}$ inevitably loses information. The regression filter developed in [Zhu and Rohwer 1996] usually provides reasonable solutions with little computational cost for practical problems. Work of a similar nature is reported in [Hastie 1996], where a computationally-efficient finite-dimensional approximation to spline smoothing is developed (spline smoothing derives from a particular choice of Gaussian measure over functions). The second question relates to choosing a model which loses the least information. There are several different versions of this problem; here we concentrate on the choice of a fixed model which is optimal under a given prior. Basically we will show that if the approximation error is measured relative to a weight function $p(x)$ and $\int p(x)dx < \infty$, then the optimal m-dimensional linear model is spanned by the first m eigenfunctions of a particular covariance operator.

Intuitively the problem may be viewed like this: The distribution of the ideal estimate on the function space looks like an ellipsoid under the inner product used for approximation. The optimal m-dimensional linear space should then be spanned by the longest m axes of the ellipsoid. This is an infinite dimensional principal component analysis, and the solution is obtained by eigendecomposition of a trace-class operator. Although the solution is straightforward once the problem is properly formulated, its specific form illuminates interesting relations between estimation, approximation and smoothing.

We have adopted a complete function-space approach which enables us to summarize much of existing and new results in a notation almost as simple as that of finite dimensional algebra. Results presented here are not necessarily new, but they are given a more unified and concise exposition thanks to the new notation. For motivation and intuitive explanation of Hilbert space concepts see the appendix.

Our main contribution is not in generality or new technical results, but in the general framework and notation which allows us to collect together many results scattered across many fields, some quite abstract, with minimum complication of notations and in a form directly applicable in data analysis. It is not possible to cite all relevant previous works, but this work should make it easier to recognize related works which would otherwise be regarded as unrelated to each other.

2 Regression, Approximation and Smoothing

Let X be a d-dimensional manifold, and Y be an Euclidean space. Often X is simply an Euclidean space. Let $\mathcal{H}_0$ be a Hilbert space of functions from X to Y, with inner product P. We shall consider functions as infinite dimensional vectors, and linear operators as infinite dimensional matrices, in the sense of Schwartz's theory of distributions [Schwartz 1966], also called generalized functions [Gel'fand and Shilov 1964; Gel'fand and Vilenkin 1964]. See [Zemanian 1965] for an introduction. See the appendix for a discussion of the intuition behind considerations of infinite dimensional objects.

The framework adopted here is essentially the same but slightly more general than the reproducing kernel Hilbert space (rkhs) approach [Parzen 1961; Kailath 1971; Wahba 1990]. It has the advantage that every generalized function can be differentiated infinitely many times, and that there is a one-one correspondence between kernels and operators, according to Schwartz's kernel theorem [Gel'fand and Vilenkin 1964, p.18]. On the other hand, the rkhs are spaces of functions for which pointwise values are defined, according to the Aronsajn-Bergman theorem [Yosida 1965, p. 96]. The more general framework avoids asking existence questions and enables us to use notations of ordinary finite dimensional linear algebra for infinite dimensional objects. Certain variables are denoted in bold face to emphasize finite-dimensional-only features. We shall always use $L_2(X,Y)$ as the “pivotal space” for the definition of transpose and the correspondence between kernels and operators, that is,

$$f^T g := \int d\xi\, f(\xi)^T g(\xi), \qquad Af(\xi) := \int d\xi'\, A(\xi,\xi') f(\xi'), \tag{1}$$

$$A^T(\xi,\xi') = A(\xi',\xi)^T, \qquad f^T A g = \int d\xi\, d\xi'\, f(\xi)^T A(\xi,\xi') g(\xi'). \tag{2}$$

If Y is complex then ^{T} should be replaced by ^{H}, the conjugate transpose. The appearance of transpose in $f(\xi)^T$ is due to the fact that Y is itself an Euclidean space. To avoid complications in the notation we shall assume $Y = \mathbb{R}$ in the sequel, but all the results are applicable generally. If the kernel A is symmetric (Hermitian) and positive definite it defines an inner product. Note that not all operators can be represented by a proper kernel.

Our statistical model is

$$y = f(x) + \eta(x), \tag{3}$$

where $f \in \mathcal{H}_0$ is the true function, $x \in X$ is the input point, $y \in Y$ is the output point, and η is the noise. We assume that

- f and η are random fields, and x is a random process [Gel'fand and Vilenkin 1964].
- f, x and η are independent of each other, $P(f, x, \eta) = P(f)P(x)P(\eta)$.
- f and η are Gaussian.

In this paper every random variable is a function of f, x, η. The notation $\langle a \rangle$ denotes the mean of a (averaged over $P(f, x, \eta)$), while $\langle a, b \rangle := \langle (a - \langle a \rangle)(b - \langle b \rangle) \rangle$ denotes the covariance of a and b. The notations $\langle a \rangle_c$ and $\langle a, b \rangle_c$ denote corresponding conditional mean and covariance (averaged over $P(f, x, \eta | c)$). The distributions $P(f)$ and $P(\eta)$ are uniquely specified by their mean functions and covariance kernels

$$\langle f \rangle = \widehat{f}_0, \qquad \left\langle f, f^T \right\rangle = V_0, \qquad \langle \eta \rangle = \widehat{\eta}_0, \qquad \left\langle \eta, \eta^T \right\rangle = R. \tag{4}$$

For simplicity we assume that $\widehat{f}_0 = \widehat{\eta}_0 = 0$ and that the covariances V_0 and R are finite. For uncorrelated noise R is diagonal.

Given a data set $z = [z_1, \ldots, z_n] \in (X \times Y)^n$ of size n, where $z_i = [x_i, y_i]$, the posterior $P(f|z)$ is a Gaussian with well-known mean and covariance (See, for example, [Lindley and Smith 1972])

$$\widehat{f}_n := \langle f \rangle_z = V_0 \boldsymbol{X}^T (\boldsymbol{V}_0 + \boldsymbol{R})^{-1} \boldsymbol{y}, \tag{5}$$

$$V_n := \left\langle f, f^T \right\rangle_z = V_0 - V_0 \boldsymbol{X}^T (\boldsymbol{V}_0 + \boldsymbol{R})^{-1} \boldsymbol{X} V_0, \tag{6}$$

where

$$\boldsymbol{X}(i, \xi) := \delta(\xi - x_i), \qquad V_{ij} := V_0(x_i, x_j), \qquad R_{ij} := R(x_i, x_j), \tag{7}$$

$$\boldsymbol{V}_0 = \boldsymbol{X} V_0 \boldsymbol{X}^T = \begin{bmatrix} V_{11} & \ldots & V_{1n} \\ \vdots & & \vdots \\ V_{n1} & \ldots & V_{nn} \end{bmatrix}, \tag{8}$$

$$\boldsymbol{R} = \boldsymbol{X} R \boldsymbol{X}^T = \begin{bmatrix} R_{11} & \ldots & R_{1n} \\ \vdots & & \vdots \\ R_{n1} & \ldots & R_{nn} \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{9}$$

The posterior mean is also the ideal Bayesian estimate under the norm $\|g\|_P^2 := g^T P g$,

$$\widehat{f}_n = \operatorname*{argmin}_{g \in \mathcal{H}_0} \left\langle \|f - g\|_P^2 \right\rangle_z , \tag{10}$$

but it is obviously independent of P. Furthermore, if $\boldsymbol{R} = 0$, then $\widehat{f}_n(\boldsymbol{x}) = \boldsymbol{X}\widehat{f}_n = \boldsymbol{y}$. In other words, the regression passes through the data points if we believe the data are noise-free.

For any positive operator V_0 there is an operator H as its inverse, $V_0 H = I$. If V_0 is integral operator (i.e. a proper kernel), then H is a (pseudo)differential operator and corresponds to the inner product of the reproducing kernel Hilbert space $\mathcal{H}$ with reproducing kernel V [Yosida 1965]. A pseudodifferential operator differs from a proper differential operator in that its Fourier transform is not necessarily a polynomial [Hörmander 1983]. Conversely, given differential operator H, the kernel V is called its Green's function. If H is certain forms of pseudodifferential operators the space $\mathcal{H}$ is usually called a (generalized) Sobolev space [Adams 1975; Triebel 1978].

A different but entirely equivalent point of view from Bayesian estimation is regularized approximation. Let D and $H = D^T D$ be pseudodifferential operators without null space, and V_0 be its Green's function. The objective function

$$\begin{aligned} J(g) &:= \sum_{ij} (y_i - g(x_i))(\boldsymbol{R}^{-1})_{ij}(y_j - g(x_j)) + \int_\xi (Dg(\xi))^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}g)^T \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}g) + g^T H g, \end{aligned} \tag{11}$$

attains its minimum at [Poggio and Girosi 1990]

$$\widehat{f}_n = V_0 \boldsymbol{X}^T (\boldsymbol{V}_0 + \boldsymbol{R})^{-1} \boldsymbol{y}, \tag{12}$$

which is the solution of the Euler equation of (11)

$$\boldsymbol{X}^T \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{f}_n) = H \widehat{f}_n. \tag{13}$$

This solution is identical to the Bayes posterior mean, given the assumed relation between the regularization operator H and the covariance kernel V_0.

Yet another equivalent point of view is that of smoothing with smoothing operator $\boldsymbol{K} = V_0 \boldsymbol{X}^T (\boldsymbol{V}_0 + \boldsymbol{R})^{-1}$. Obviously the solution is in the space $V_0 \boldsymbol{X}^T \mathbb{R}^n$ spanned by the basis $V_0 \boldsymbol{X}^T$. If H is an iterated Laplacian operator, then V_0 is given explicitly in [Gel'fand and Shilov 1964, p. 202] and $V_0 \boldsymbol{X}^T$ is known as the basis of "thin-plate splines" with $\boldsymbol{x}$ as nodes [Meinquet 1979; Wahba 1990]. If H is a proper differential operator, then its Fourier transform is a polynomial and $V\boldsymbol{X}^T$ is the basis of generalized splines (g-splines) [Ahlberg, Nilson, and Walsh 1967, chap. 6]. If H is a pseudo-differential operator, then V is a quite arbitrary kernel which can be regard as the furthest generalization of splines in this direction. If $H = (-\nabla^2)^s$ then $\mathcal{H} = H_0^s$, and if $H = (I - \nabla^2)^s$ then $\mathcal{H} = H^s$, both known as Sobolev spaces. It is rkhs (and V_0 is a proper kernel) if and only if $s > d/2$, by Sobolev embedding theorem. Another direction to generalize splines is to replace $\boldsymbol{X}$ by a general linear operator L and the solution $V_0 L^T$ is usually called L-splines [Wahba 1990].

The relation between the inversion of a differential operator, minimization in a Hilbert space, and smoothing with an integral operator is well known but much of the results are scattered in the literature. See [Donsker and Lions 1962] for the same relation in a different context. One advantage of Bayesian approach is that it naturally extends to the case that the random function f is also a process in time. The solution is given by a Kalman filter [Kalman 1960]. We can also relax the assumption that V_0 is finite, or equivalently that H does not have a null space, but the exposition becomes complicated [Wahba 1990, p.11–12]. See also [Kimeldorf and Wahba 1970; Thomas-Agnan 1991; Meinquet 1979; Kent and Mardia 1994].

For later purposes we also need to study $P(\widehat{f}_n)$, the prior distribution of the posterior mean $\widehat{f}_n$ over all samples z. This should not be confused with the posterior $P(f|z)$ itself, which is the distribution of the true function conditional on a given sample. Since $\widehat{f}_n$ depends on z, its distribution depends on $P(z) = P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$. We have

$$\left\langle \widehat{f}_n \right\rangle_{\boldsymbol{x}} = \boldsymbol{K} \left\langle \boldsymbol{y} \right\rangle_{\boldsymbol{x}} = 0, \tag{14}$$

$$\left\langle \widehat{f}_n, \widehat{f}_n^T \right\rangle_{\boldsymbol{x}} = \boldsymbol{K} \left\langle \boldsymbol{y}, \boldsymbol{y}^T \right\rangle_{\boldsymbol{x}} \boldsymbol{K}^T = V_0 \boldsymbol{X}^T (\boldsymbol{V}_0 + \boldsymbol{R})^{-1} \boldsymbol{X} V_0 = V_0 - V_n, \tag{15}$$

$$V := \left\langle \widehat{f}_n, \widehat{f}_n^T \right\rangle = V_0 \left\langle \boldsymbol{X}^T (\boldsymbol{V}_0 + \boldsymbol{R})^{-1} \boldsymbol{X} \right\rangle V_0 = V_0 - \langle V_n \rangle \leq V_0, \tag{16}$$

where we have used

$$\left\langle \boldsymbol{y}, \boldsymbol{y}^T \right\rangle_{\boldsymbol{x}} = \left\langle \boldsymbol{X} f + \boldsymbol{X} \eta, (\boldsymbol{X} f + \boldsymbol{X} \eta)^T \right\rangle_{\boldsymbol{x}} = \boldsymbol{V}_0 + \boldsymbol{R}. \tag{17}$$

It is interesting to note that the shrinkage from V_0 to V, instead of from V_0 to V_n as often thought, was in fact the origin of the term "regression" introduced by Galton more than a century ago [Stigler 1986]. In general, if we have any way to divide a Gaussian into subgroups, then the mean of the subgroups cannot vary as much as individuals. In fact, the variance of the whole population is the sum of the variance of the means of the subgroups and the average variance within each subgroup.

For infinitely large n, generally we have $\widehat{f}_n \approx f$ so that $P(\widehat{f}_n) \approx P(f)$ and $V \approx V_0$. If the sample input $\boldsymbol{x}$ can be approximated by a continuous distribution $p_0(x)$, then $\boldsymbol{X}^T$ can be regarded as an invertible operator so that

$$V \approx V_0 (V_0 + R)^{-1} V_0 = V_0, \tag{18}$$

because R is a finite kernel which is zero except on the diagonal (which has zero measure according to $p_0(x)$).

In reality, the sample input only defines a discrete distribution so R cannot be ignored this way. It is difficult to obtain an explicit expression for V for small n, as it will necessarily depend on the actual distribution of $\boldsymbol{x}$. However, for large n we can obtain an asymptotic expression more accurate than the above, assuming that for large n the distribution of $\boldsymbol{x}$ approaches $p_0(x)$, that the variability of $\boldsymbol{x}$ has little effect on the estimate of covariances, and that R is diagonal. We have

$$\langle V_n \rangle \approx \left\langle V_n^{-1} \right\rangle^{-1} = \left\langle V_0^{-1} + \boldsymbol{X}^T \boldsymbol{R}^{-1} \boldsymbol{X} \right\rangle^{-1} \approx (V_0^{-1} + n\mathcal{R}^{-1})^{-1}, \tag{19}$$

$$V = V_0 - \langle V_n \rangle \approx V_0 - (V_0^{-1} + n\mathcal{R}^{-1})^{-1} = V_0(V_0 + \mathcal{R}/n)^{-1} V_0, \tag{20}$$

where

$$\mathcal{R}(\xi, \xi') := \int_\tau R(\tau, \tau) p_0(\tau)^{-1} \delta(\xi - \tau)\delta(\xi' - \tau), \tag{21}$$

$$\mathcal{R}^{-1}(\xi, \xi') = \int_\tau R(\tau, \tau)^{-1} p_0(\tau) \delta(\xi - \tau)\delta(\xi' - \tau), \tag{22}$$

$$\begin{aligned} \left\langle \boldsymbol{X}^T \boldsymbol{R}^{-1} \boldsymbol{X} \right\rangle (\xi, \xi') &= \left\langle \sum_i \delta(\xi - x_i) R(x_i, x_i)^{-1} \delta(\xi' - x_i) \right\rangle \\ &\approx n \int_\tau p_0(\tau) R(\tau, \tau)^{-1} \delta(\xi - \tau)\delta(\xi' - \tau) = n\mathcal{R}^{-1}(\xi, \xi'). \end{aligned} \tag{23}$$

3 Finite Dimensional Models

We see that once the input points $\boldsymbol{x}$ are fixed, the solution is confined to a finite dimensional linear space $V_0 \boldsymbol{X}^T \mathbb{R}^n$, which is the optimal n-dimensional model. Our main interest in this paper is to find a fixed optimal m-dimensional linear model independent of the input data $\boldsymbol{x}$, where the dimension m is also independent of n.

An m-dimensional linear model $M \subseteq \mathcal{H}_0$ may be represented as $M = \Phi \mathbb{R}^m$ where $\Phi := [\phi_1, \ldots, \phi_m]$ is a basis of M. It is well-known that under the inner product P, the dual basis Ψ, the projection operator P_M and the remainder operator R_M (the orthogonal projections to M and $M^\perp$, respectively), are given by

$$\Psi := \Phi(\Phi^T P \Phi)^{-1}, \qquad \mathsf{P}_M := \Phi \Psi^T P, \qquad \mathsf{R}_M := I - \mathsf{P}_M, \tag{24}$$

with the following properties,

$$\mathsf{R}_M^T P \mathsf{P}_M = 0, \qquad \mathsf{P}_M^T P \mathsf{P}_M = P\mathsf{P}_M = \mathsf{P}_M^T P, \qquad \mathsf{R}_M^T P \mathsf{R}_M = P\mathsf{R}_M = \mathsf{R}_M^T P, \tag{25}$$

$$\mathsf{P}_M^2 = \mathsf{P}_M, \qquad \mathsf{R}_M^2 = \mathsf{R}_M, \qquad \|\mathsf{P}_M f\|_P^2 + \|\mathsf{R}_M f\|_P^2 = \|f\|_P^2. \tag{26}$$

The optimal solution $\widehat{g}_n \in M$ which minimizes $J(g) = \langle \|y' - g\|_P^2 \rangle_z$, where $y' = f + \eta$, is $\widehat{g}_n = \mathsf{P}_M \widehat{f}_n$. It depends essentially on the norm $\|\cdot\|_P$, because in general the ideal estimate $\widehat{f}_n \notin M$. The mean squared test error of an arbitrary estimate $g \in M$ is decomposed as

$$\begin{aligned} J(g) &= \left\langle \|\eta\|_P^2 \right\rangle_z + \left\langle \|f - \widehat{f}_n\|_P^2 \right\rangle_z + \left\langle \|\widehat{f}_n - \widehat{g}_n\|_P^2 \right\rangle_z + \left\langle \|\widehat{g}_n - g\|_P^2 \right\rangle_z \\ &= \mathrm{tr}(PR) + \mathrm{tr}(PV_n) + \|\widehat{f}_n - \widehat{g}_n\|_P^2 + \|\widehat{g}_n - g\|_P^2. \end{aligned} \tag{27}$$

The four terms on the right hand side are, in the given order: the intrinsic noise of test data, the uncertainty of ideal regression, the approximation error due to model deficiency, and the computational error. It is not possible to do anything about the first two, and by choosing $g = \widehat{g}_n$ the the last term can be made to vanish.

Our goal in this paper is to find an optimal model M independent of the training data z. This is achieved by minimizing the mean of the third term above, $\left\langle \|\mathsf{R}_M \widehat{f}_n\|_P^2 \right\rangle$, or equivalently, to maximize $\left\langle \|\mathsf{P}_M \widehat{f}_n\|_P^2 \right\rangle$. This depends on V which in turn depends on $\boldsymbol{x}$. Since we want to obtain the optimal model independently of $\boldsymbol{x}$, it seems reasonable to make the conservative assumption $V = V_0$; as we have shown in §2, this is approximately true for large n. This has the advantage that the optimal model will also be optimal for the prior. The disadvantage is that it does not utilize the fact that V is shrunk from V_0. In other words, the model wastes some capability to represent the variability of f which could not be detected by the regression $\widehat{f}_n$ for small n.

For any given V, we have

$$\begin{aligned} &\left\langle \|\mathsf{P}_M \widehat{f}_n\|_P^2 \right\rangle = \left\langle \widehat{f}_n^T \mathsf{P}_M^T P \mathsf{P}_M \widehat{f}_n \right\rangle = \operatorname{tr}\left(\left\langle \widehat{f}_n \widehat{f}_n^T \right\rangle \mathsf{P}_M^T P \mathsf{P}_M \right) \\ &= \operatorname{tr}(V \mathsf{P}_M^T P) = \operatorname{tr}((\Phi^T P \Phi)^{-1} \Phi^T P V P \Phi), \end{aligned} \tag{28}$$

which further reduces to $\operatorname{tr}(\Phi^T P V P \Phi)$, if we set $\Phi^T P \Phi = I$ using the Gram-Schmidt procedure. Since $P \in SPD$, the set of symmetric positive definite operators, it has a uniquely defined square root $P^{1/2}$ [Riesz and Nagy 1955]. Define

$$V_P := P^{1/2} V P^{1/2}, \qquad \Phi_P := P^{1/2} \Phi. \tag{29}$$

Then our problem finally reduces to

$$\operatorname*{Max}_{\Phi_P^T \Phi_P = I} \operatorname{tr}(\Phi_P^T V_P \Phi_P) \le \operatorname{tr} V_P. \tag{30}$$

Since the space $\mathcal{H}_0$ is infinite dimensional we need an extra assumption to guarantee that this trace above is finite. In other words, we want $V_P \in L_1(\mathcal{H}_0)$, the trace-class [Kuo 1975]. A trace-class operator (also called a nuclear operator) has a spectrum composed entirely of a countable number of eigenvalues with a finite sum [Gel'fand and Vilenkin 1964]. In practice, the norm $\|\cdot\|_P$ is usually defined through test data $z' = [x', y']$ by

$$\|g\|_P^2 := \int dx'\, p(x') g(x')^2, \tag{31}$$

with the interpretation that $p(x') = p_1(x') p_2(x')$, where p_1 is the test data input distribution and p_2 is a weighting function. Because

$$\begin{aligned} \operatorname{tr} V_P &= \operatorname{tr}(VP) = \operatorname{tr}(\left\langle \widehat{f}_n \widehat{f}_n^T \right\rangle P) = \left\langle \widehat{f}_n^T P \widehat{f}_n \right\rangle \\ &= \int dx\, p(x) \left\langle \widehat{f}_n(x)^2 \right\rangle \le \int dx\, p(x) \max_x V(x, x), \end{aligned} \tag{32}$$

we see that V_P is a trace-class operator if $\int dx\, p(x) < \infty$ and $\max_x V(x,x) < \infty$. In the rest of the paper we shall assume that this is so, which does not impose much restriction in practice.

According to Rayleigh's principle [Nef 1967], which underlies principal component analysis, the optimal model $\Phi_P \mathbb{R}^m$ for (30) is spanned by the eigenfunctions $\Phi_P = [u_i : i = 1 \ldots m]$ corresponding to the m largest eigenvalues of V_P. As $V_P \in L_1 \cap SPD$, its spectrum consists of a countable set of positive eigenvalues λ_k with $\sum_k \lambda_k < \infty$. They can be ordered, taking into account multiplicity, as $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, with corresponding orthogonal eigenfunctions $u_1, u_2, \ldots$. The optimal model $\Phi\mathbb{R}^m$ for the original problem is therefore spanned by the orthonormal (with respect to P) basis $\Phi = P^{-1/2}\Phi_P$, with

$$\left\langle \|\widehat{f}_n\|_P^2 \right\rangle = \sum_{i=1}^{\infty} \lambda_i, \qquad \left\langle \|\mathsf{P}_M \widehat{f}_n\|_P^2 \right\rangle = \sum_{i=1}^{m} \lambda_i, \qquad \left\langle \|\mathsf{R}_M \widehat{f}_n\|_P^2 \right\rangle = \sum_{i=m+1}^{\infty} \lambda_i. \tag{33}$$

After completing this work we became aware of the paper [Castro, Lawton, and Sylvestre 1986], which treats the problem of optimal finite linear models in much the same way, although the problem is restricted to the third example towards the end of the next section. Most of its results can be easily generalized to the situation as treated here. The advantage of our more general framework is that the measure p, corresponding to the weight function w in their paper, is explicitly specified so that it will be invariant to the training data positions. This explains why the distinction between principal component analysis and eigenexpansion as emphasized there disappears here.

4 Examples

Although the optimal model depends on V which in turn depends on n, for large enough n we can assume $V \approx V_0$. All the examples given here rely on this assumption. As our first example consider

$$p(\xi) = \exp(-2a\xi^2), \qquad V(\xi_1, \xi_2) = V(\xi_1 - \xi_2), \qquad V(\xi) = \exp(-b\xi^2), \tag{34}$$

$$V_P(\xi_1, \xi_2) = \exp\left(-a\xi_1^2 - b(\xi_1 - \xi_2)^2 - a\xi_2^2\right). \tag{35}$$

The eigenvalues λ_k, eigenfunctions u_k and basis ϕ_k (for convenience let $k = 0, 1, \ldots$) are given by

$$\lambda_k = \sqrt{\frac{\pi}{A}} B^k, \qquad u_k(x) = \exp(-cx^2) H_k(\sqrt{2c}x), \tag{36}$$

$$\phi_k(x) = \exp(-(c-a)x^2) H_k(\sqrt{2c}x), \tag{37}$$

where H_k is the kth order Hermite polynomial, and

$$c = \sqrt{a^2 + 2ab}, \qquad A = a + b + c, \qquad B = b/A. \tag{38}$$

This can be proved by using equation 7.374.8 in [Gradshteyn and Ryzhik 1980],

$$\int_{-\infty}^{\infty} e^{-(x-y)^2} H_n(\alpha x)\, dx = \sqrt{\pi}(1-\alpha^2)^{n/2} H_n\left(\frac{\alpha y}{(1-\alpha^2)^{1/2}}\right). \tag{39}$$

The total prior variance and the residual ratio are

$$\left\langle \|\widehat{f}_n\|_P^2 \right\rangle = \operatorname{tr} V_P = \sum_{k=0}^{\infty} \lambda_k = \sqrt{\frac{\pi}{A}} \frac{1}{1-B}, \tag{40}$$

$$\frac{\left\langle \|\mathsf{R}_M \widehat{f}_n\|_P^2 \right\rangle}{\left\langle \|\widehat{f}_n\|_P^2 \right\rangle} = \frac{\sum_{k=m}^{\infty} \lambda_k}{\sum_k \lambda_k} = B^m. \tag{41}$$

That is, the optimal m-dimensional model catches a portion $1 - B^m$ of the variance of $\widehat{f}_n$, while leaving portion B^m as the residual approximation error. For $a = 1, b = 3$, we have $B = 0.4514$. The first 15 eigenvalues $\{\lambda_k\}$ are shown in Figure 1(a). The first 6 eigenfunctions $\{u_k\}$ are shown in Figure 1(b). The 15th basis function is shown in Figure 1(c). The first 6 basis functions $\phi_k = P^{-1/2} u_k$ are shown in Figure 1(d). Note that the high-eigenvalue components correspond to features typically of more interest, namely the low frequency features toward the center of $p(x)$. The infinite dimensional principal component analysis is possible because there is an effective ordering of the "components" determined by the forms of p and V.

In fact, for this particular example, the first 6 dimensions catch 99.15% of variance and leave out only 0.85%, while the 12 dimensional model would catch 99.993% variance, leaving out only 0.007%. This is despite the fact that the distribution $P(\widehat{f}_n)$ has power at all frequencies (the power spectrum of f is the Fourier transform of V which extends to all frequencies). and that the weighting is non-zero over an infinite interval. This is clearly advantageous compared with the assumption of uniform prior on a finite dimensional model over a finite interval. To illustrate the effectiveness of the optimal m-dimensional model even with a small m, we pick a typical true function f from the prior, and generate a typical sample z. In Figure 2 we plot the effect of sample size n and model size m on the ideal regression $\widehat{f}_n$ and the optimal regression $\widehat{g}_n$ of the model.

It is interesting to observe how the eigenvalues and eigenfunctions change with the parameters a and b. As b/a increases toward infinity, B approaches unity. The eigenvalues become more tightly clustered. To get a reasonable approximation a larger m is needed, with many basis functions looking essentially like sines and cosines within the range where p is not negligible. As B approaches unity the spectrum essentially becomes continuous and the principal component analysis becomes less useful, similarly to the transition from Fourier series to Fourier transform. In the limit all the "components" are equally important, and there are infinitely many of them, so that the PCA completely breaks down. This also reveals the important role played by the inner product even for finite dimensional PCA.

Two drawbacks of the optimal linear model are that the basis functions are non-local, and each new dimension introduces an entirely different shape of eigenfunction. It is possible to overcome these two shortcomings by trading-off some optimality of approximation with ease of computation. The classical method of obtaining local basis functions is the finite element method [Ciarlet 1978]. A currently more fashionable method is wavelet analysis which provides (almost) local basis functions spanning almost the optimal eigen-subspace, yet all the basis functions can be obtained from one "mother wave-shape". There is a large and rapidly expanding literature on wavelet analysis [Benassi and Jaffard 1994]. A more recent reference is [Wang 1996].

The next example, which is perhaps more familiar, describes a prior more welcoming to non-smoothness. See [Zhu and Rohwer 1996] for more examples of other priors. Suppose p is the "gate function"

$$p(x) = \begin{cases} 1/2a, & |x| < a, \\ 0, & |x| \geq a. \end{cases} \tag{42}$$

Let V be the Green's function corresponding to the differential operator $H: f \to -f''$ with homogeneous Dirichlet boundary condition on $[-a, a]$ [Yosida 1960]

$$V(\xi_1, \xi_2) = a^2 - \xi_1\xi_2 - \frac{1}{2}|\xi_1 - \xi_2|. \tag{43}$$

This is the covariance kernel of the "Brownian bridge" [Grimmett and Stirzaker 1992]. Note that for our analysis the covariance kernel need not be in a translation invariant form $V(\xi_1, \xi_2) = V(\xi_1 - \xi_2)$. The reproducing kernel Hilbert space with reproducing kernel V is the Sobolev space

$$H_0^1 := \left\{ f : \int_{-a}^{a} dx\, f'(x)^2 < \infty, f(a) = f(-a) = 0 \right\}. \tag{44}$$

The ideal estimate $\widehat{f}$ for any given data set z is a piecewise linear function. If we assume the data to be error-free, $R = 0$, then $\widehat{f}$ is simply the Lagrange interpolation with piecewise linear functions passing all the data points. Intuitively, given any two points on a sample path, the best bet for any point in between is on the straight line connecting these two points.

The eigenfunctions and the eigenvalues are

$$u_k(\xi) = \sin\left(\frac{k\pi}{2a}(\xi + a)\right), \qquad \lambda_k = \frac{k^2\pi^2}{4a^2}. \tag{45}$$

The optimal m-dimensional model is simply the truncated Fourier series. In this case, since p is uniform over the fixed finite range $[-a, a]$, further eigenfunctions are only concerned with highly frequency features.

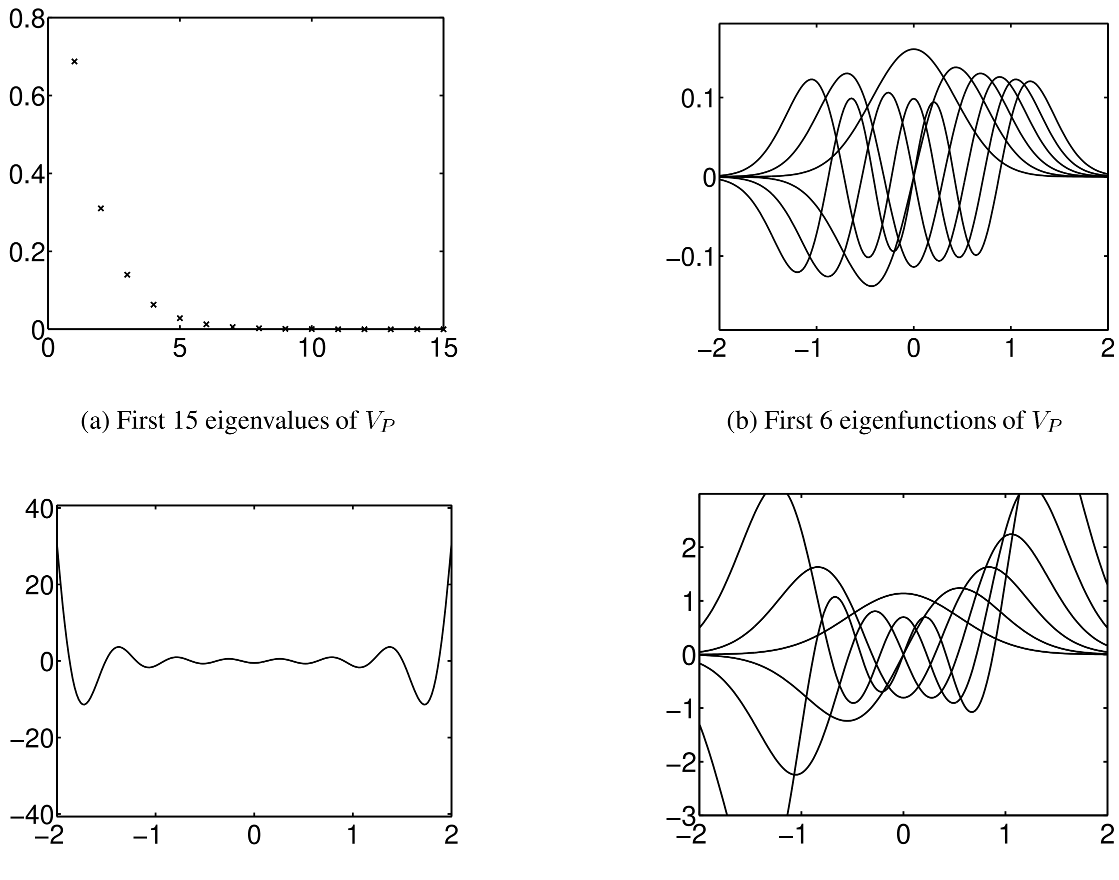
This can be regarded as a rigorous expression of the idea that Fourier series give "best" finite dimensional representation of an "arbitrary" function.

Let $g(x)$ be the standard Brownian motion (Wiener process). Then

$$f(x) = g(x) - g(a) + \frac{x - a}{2a}(g(-a) - g(a)) \tag{46}$$

is a typical sample from the prior. It is well known that the samples are, with probability one, continuous but nowhere differentiable. In fact, H_0^1 is very much like a space of $1/2$-order differentiable functions [Adams 1975]. Several samples are given in Figure 3. It is important to note that $f \notin H_0^1$ because $\|f\|_1 = \infty$. Given any orthonormal basis in H_0^1, the projection of f on each component is distributed as a standard Gaussian so that the total norm is infinite [Kuo 1975]. This is further explained in the appendix.

As a third example, suppose $p(x) = \sum_k \delta(x - x_k)$. That is, assume the test data may only come from a finite number of specified points. Then the optimal model becomes $V\boldsymbol{X}^T\mathbb{R}^n$. Therefore splines with preselected nodes are optimal under the assumption that the test data only come from these points. This example was studied in detail in [Castro, Lawton, and Sylvestre 1986].



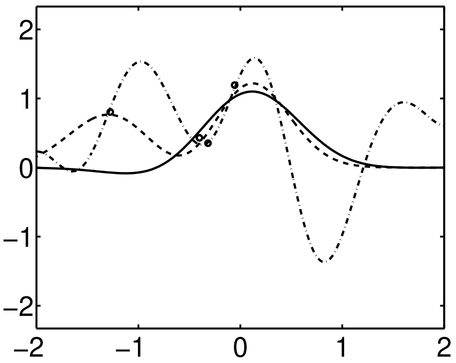
(a) First 15 eigenvalues of V_P

(b) First 6 eigenfunctions of V_P

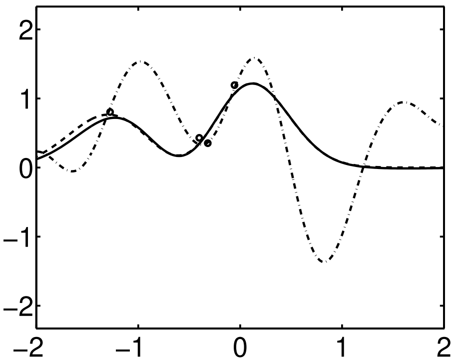
(c) The 15th basis function

(d) First 6 basis functions

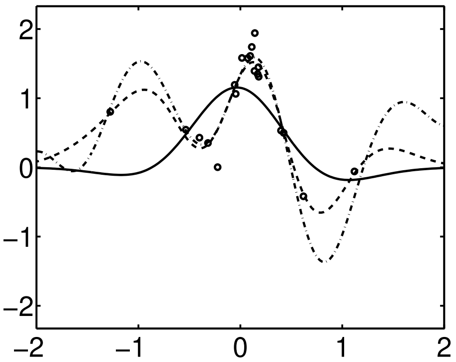
Figure 1: Eigen-decomposition ($a = 1$, $b = 2$).



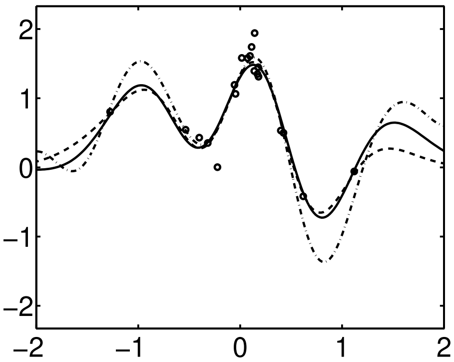
(a) $m = 3, n = 4$



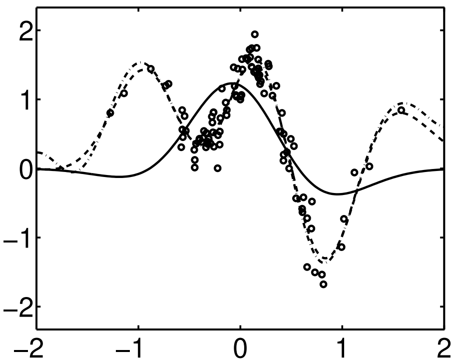
(b) $m = 6, n = 4$



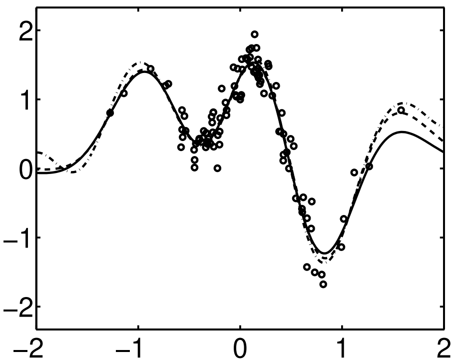
(c) $m = 3, n = 20$



(d) $m = 6, n = 20$



(e) $m = 3, n = 100$



(f) $m = 6, n = 100$

Figure 2: The effect of data size n and model dimension m. Legend: f—dash-dot; $\widehat{f}$—dashed; $\widehat{g}$—solid; z—circles.

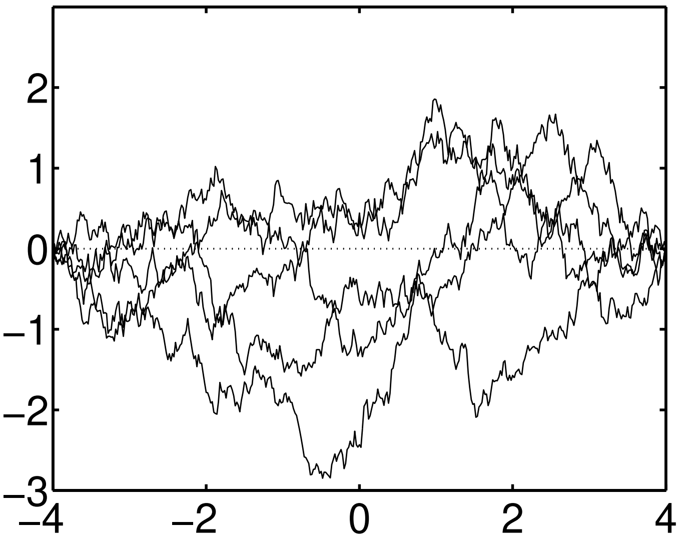


Figure 3: Three samples from the Brownian bridge

5 Summary

In most statistical problems the objects under consideration cannot be completely represented by a finite dimensional model, yet most of their properties of interest can be adequately approximated using a finite dimensional model. One such problem particularly important in practice is analyzed here, giving the conditions and optimality of finite dimensional approximation.

We have shown that the optimal finite dimensional model under a Gaussian prior may be obtained by an infinite dimensional principal component analysis. The condition for its applicability is given, which essentially requires the "length of axes" to sum to a finite number, despite the fact there are infinitely many of them. When this condition is not satisfied, the corresponding operator usually has a continuous spectrum so that no finite dimensional model could do a reasonable job. The relation between smoothing, approximation and estimation is also elucidated.

Essentially, to estimate an infinite dimensional object, such as a function, from a finite amount of data, we must assume an infinite amount of prior information, leaving out only a finite amount of uncertainty. Here the amount of data and information are measured in units of real numbers. The classical way to do this is to assume that we are absolutely certain in all but a finite number of directions; in other words, to assume a finite dimensional model. The alternative given here is more general. We assume that we may not be certain in any direction, but the uncertainty on these infinitely many directions as measured by variance sums to a finite number. This cannot be achieved if we assume the prior distribution is spherical relative to the inner product which measures the importance we assign to each direction.

Acknowledgment

The work of HZ and RR at Aston was supported by EPSRC grant GR/J17814. The work of CW was partially supported by EPSRC grant GR/J75425. The work of HZ at SFI was supported by TXN, Inc. The work of MM was supported by NCRG at Aston.

A Motivation for Hilbert space methods

The general treatment of Gaussian measures on function spaces has been known since the work of Wiener and Kolmogorov, which usually involves consideration of several norms or inner products. However, it appears not to have been intuitively introduced to practical statisticians, and therefore has been largely ignored by them. This is quite unfortunate since it is the direct cause of the well-known phenomenon of "over-fitting".

For motivation let us first consider Gaussian measures on a finite dimensional space (multivariate Gaussians). Let x be an m-dimensional Gaussian variable with zero mean and covariance matrix $V = \left\langle xx^T \right\rangle$. Its mean squared length is

$$\left\langle x^T x \right\rangle = \left\langle \operatorname{tr}(xx^T) \right\rangle = \operatorname{tr} V. \tag{47}$$

It is often convenient to linearly transform x to Lx such that $P(Lx)$ is of the standard form (spherical and of unit variance). The inner product on Lx induces an inner product on x, $H = V^{-1} = L^T L$, which is the Fisher information matrix for x. Under H the squared length of x has a χ^2 distribution of m degrees of freedom with mean

$$\left\langle x^T H x \right\rangle = \left\langle \operatorname{tr}(Hxx^T) \right\rangle = \operatorname{tr}(HV) = \operatorname{tr}(I_m) = m. \tag{48}$$

Now we see that this causes a problem for a Gaussian $P(f)$ on an infinite dimensional space: If we use an inner product H by which $P(f)$ looks spherical, the mean squared length of the random element f will be infinity, because it is the sum of variances in all the (infinitely many) orthogonal directions, all of them equal to one another. In other words, the collection of samples from a Gaussian with covariance kernel V does not form a Hilbert space with inner product H under which the covariance is spherical. Therefore H is unsuitable as an objective for approximation. For under it we start from infinite error and would retain infinite error given any finite amount of data.

Since it is desirable to be able both to represent Gaussians in a spherical form and to use a finite norm, Gaussian measures $P(f)$ in a Hilbert space $\mathcal{H}_0$ are generally defined by two inner products, necessarily not equivalent to each other [Kuo 1975]. One inner product P of Hilbert space $\mathcal{H}_0$ is used to measure distance by $f^T P f$. A stronger inner product H of Hilbert space $\mathcal{H}_1 := \left\{ f : f^T H f < \infty \right\} \subset \mathcal{H}_0$ makes $P(f)$ "appear" spherical.

Let us analyze the spectrum of V under both inner products. The eigenvalues of V under H are unity for all eigenfunctions; The distribution is spherical but the mean squared length is infinity. The eigenvalues of V under P form a decreasing sequence λ_k such that $\sum_k \lambda_k < \infty$; The mean squared length is finite but the distribution is not spherical.

Usually H is a pseudodifferential operator, of which the covariance kernel V is its Green's function. If V is proper kernel, i.e., if Dirac's measures belong to the dual space $\mathcal{H}_1'$ (i.e., they are bounded linear functionals), then the space $\mathcal{H}_1$ is called the reproducing kernel Hilbert space (rkhs) with reproducing kernel V [Aronszajn 1950; Parzen 1963; Kailath 1971; Yosida 1965]. For further references see [Wegman 1988]. The reason to single out rkhs from all function spaces is that they guarantee the regression based on a finite amount of data to be a proper function, as point values at finite many points form a multivariate Gaussian with finite covariance. The appearance of δ-measures in these considerations is due to $\boldsymbol{X}$ in our statistical model $\boldsymbol{y} = \boldsymbol{X} f + \boldsymbol{X} \eta$. If $\boldsymbol{X}$ is replaced by a linear operator L the solution will be L-splines not necessarily in a rkhs.

In fact, generalized functions can also be regarded as rkhs if δ_a is replaced by smoother test functions [Kailath 1971]. From this point of view, classical rkhs is such that point values of each member functions are well defined, while the general rkhs is such that locally smoothed values are well defined.

So why is this important to practical statisticians? In practice all the data are sampled at finite number of points, and it is well known that all the finite dimensional norms are equivalent. It may appear that the idealized infinite dimensional objects would have no practical consequence. However, it happens that although the norms are equivalent, the ratio between them generally depends on the discretization. It therefore may happen that as we increase the "precision" of data to get a better fit in one norm $\|\cdot\|_0$ it actually get worse in another norm $\|\cdot\|_1$.

As a concrete example, consider the Sobolev spaces

$$\mathcal{H}^0 := \left\{ f : \ \|f\|_0^2 := \int dx\, f(x)^2 < \infty \right\}, \tag{49}$$

$$\mathcal{H}^1 := \left\{ f : \ \|f\|_1^2 := \int dx\, (f(x)^2 + f'(x)^2) < \infty \right\}. \tag{50}$$

It is well-known in the function approximation literature that if the function f is discretized in the usual ways then $\|f\|_0 \le \|f\|_1 \le (C/h)\|f\|_0$, where h is the steplength of the discretization. So if we do not guarantee the increase of precision in $\|\cdot\|_0$ to be faster than the decrease in h, which is generally impossible anyway, there is no way to infer convergence in $\|\cdot\|_1$ from that of $\|\cdot\|_0$. In practice, $\|\cdot\|_0$ may be used to measure the approximation while $\|\cdot\|_1$ may come from the covariance of our prior. A good fit according to $\|\cdot\|_0$ may be so wiggly that we do not believe it to be true. In fact, a typical sample from a Gaussian which looks spherical under $\|\cdot\|_0$ is a sample from white noise! In this case we say "over-fitting" occurs. Of course, the term "over-fitting" is a misnomer: If our goal is to fit, how could we over-do it? This only happens because our implicit goal is different from the norm we tell the machine to minimize. The importance of norm was also emphasized in [Kailath 1971, §4]. The term corresponding to $\|\cdot\|_1$ is usually called a regularizer.

References

Adams, R. A. (1975). *Sobolev Spaces*. New York: Academic Press.

Ahlberg, J. H., E. N. Nilson, and J. L. Walsh (1967). *Theory of Splines and Their Applications*. New York: Academic Press.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**, 337–404.

Benassi, A. and S. Jaffard (1994). Wavelet decomposition of one and several dimensional Gaussian processes. In L. L. Schumaker and G. Webb (Eds.), *Recent Advances in Wavelet Analysis*, pp. 119–154. New York: Academic Press.

Castro, P. E., W. H. Lawton, and E. A. Sylvestre (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28**(4), 329–337.

Ciarlet, P. G. (1978). *The Finite Element Method for Elliptic Problems*. Amsterdam: North Holland.

Donsker, M. D. and J. L. Lions (1962). Fréchet-Volterra variational equations, boundary value problems, and function space integrals. *Acta Math.* **108**, 147–228.

Gel' fand, I. M. and G. E. Shilov (1964). *Generalized Functions*, Volume 1. Properties and operations. New York: Academic Press. Traslated by E. Saletan from Russian, Moscow, 1958.

Gel' fand, I. M. and N. Y. Vilenkin (1964). *Generalized Functions*, Volume 4. Applications of harmonic analysis. New York: Academic Press. Traslated by A. Feinstein from Russian, Moscow, 1961.

Gradshteyn, I. S. and I. M. Ryzhik (1980). *Tables of Integrals, Series and Products*. New York: Academic Press. Corrected and enlarged edition prepared by A. Jeffrey.

Grimmett, G. R. and D. R. Stirzaker (1992). *Probability and Random Processes* (2nd ed.). Oxford: Clarendon Press.

Hastie, T. (1996). Pseudosplines. *J. R. Statist. Soc., B* **58**(2), 379–396.

Hörmander, L. (1983). *Analysis of Linear Partial Differential Operators*. New York: Springer-Verlag.

Kailath, T. (1971). RKHS approach to detection and estimation problems—Part I: Deterministic signals in Gaussian noise. *IEEE Trans. Info. Th.* **17**(5), 530–549.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME, Series D, J. Basic Eng.* **82**, 35–45.

Kent, J. K. and K. V. Mardia (1994). The link between Kriging and thin-plate splines. In F. P. Kelly (Ed.), *Probability, Statistics and Optimization*. New York: J. Wiley.

Kimeldorf, G. S. and G. Wahba (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **42**(2), 495–502.

Kuo, H.-H. (1975). *Gaussian Measures in Banach Spaces*, Volume 463 of *Lect. Notes in Math.* New York: Springer-Verlag.

Lindley, D. V. and A. F. M. Smith (1972). Bayes estimation for the linear model. *J. R. Statist. Soc., B* **34**, 1–41.

Meinquet, J. (1979). Multivariate interpolation at arbitrary points made simple. *J. Appl. Math. Phys. (ZAMP)* **30**, 292–302.

Nef, W. (1967). *Linear Algebra*. New York: McGraw-Hill.

Parzen, E. (1961). An approach to time series analysis. *Ann. Math. Statist.* **32**, 951–989.

Parzen, E. (1963). Probability density functionals and reproducing kernel Hilbert spaces. In M. Rosenblatt (Ed.), *Proc. Symp. on TIme Series Analysis*, New York, pp. 155–169. J. Wiley.

Poggio, T. and F. Girosi (1990). Networks for approximation and learning. *Proc. IEEE* **78**(9), 1481–1497.

Riesz, F. and B. S. Nagy (1955). *Functional Analysis*. New York: Frederick Ungar.

Schwartz, L. (1966). *Theorie des Distributions*. Paris: Herman. (First ed. 1950, 1951).

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap.

Thomas-Agnan, C. (1991). Spline function and stochastic filtering. *Ann. Statist.* **19**(3), 1512–1527.

Triebel, H. (1978). *Interpolation Theory, Function Spaces, Differential Operators*. Amsterdam: North Holland.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, PA: SIAM.

Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Ann. Statist.* **24**(2), 466–484.

Wegman, E. J. (1988). Reproducing kernel Hilbert spaces. In S. Kotz and N. L. Johnson (Eds.), *Encyclopedia of Statistical Siences*, Volume 8, pp. 81–84. New York: J. Wiley.

Yosida, K. (1960). *Lectures on Differential and Integral Equations*. New York: Interscience.

Yosida, K. (1965). *Functional Analysis*. Grundlehren, Vol. 123. Berlin: Springer-Verlag.

Zemanian, A. H. (1965). *Distribution Theory and Transform Analysis: An Introduction to Generalized Functions*. New York: McGraw-Hill.

Zhu, H. and R. Rohwer (1996). Bayesian regression filters and the issue of priors. *Neural Comp. Appl.* **4**(3), 130–142.