Chapter 6 Relationships between GPs and Other Models, Part I

Randy

8/25/2021

1. 6.1 Reproducing Kernel Hilbert Spaces
2. 6.2 Regularization
3. 6.3 Spline Models

Section 1

6.1 Reproducing Kernel Hilbert Spaces

RKHS

Definition 6.1 (Reproducing kernel Hilbert space)

Let $\mathcal{H}$ be a Hilbert space of real functions f defined on an index set $\mathcal{X}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with an inner product $\langle ., \ . \rangle_{\mathcal{H}}$ (and norm $\|f\|_{\mathcal{H}} = \sqrt{\langle f, \ f \rangle_{\mathcal{H}}}$);
If there exists a function $k : \mathcal{X} \times X \to \mathfrak{R}$ with the following properties:
- $\forall \boldsymbol{x}, \ k(\boldsymbol{x}, \ \boldsymbol{x}')$ as a function of $\boldsymbol{x}'$ belongs to $\mathcal{H}$
- k has the reproducing property $\langle f(.), \ k(., x) \rangle_{\mathcal{H}} = f(\boldsymbol{x})$

Note also that as $k(\boldsymbol{x}, .)$ and $k(\boldsymbol{x}', .)$ are in $\mathcal{H}$ we have that $\langle k(\boldsymbol{x}, \ .), \ k(\boldsymbol{x}', \ .) \rangle_{\mathcal{H}} = k(\boldsymbol{x}, \ \boldsymbol{x}')$. The RKHS uniquely determines k, and vice versa.

Moore-Aronszajn Theorem

Theorem 6.1 (Moore-Aronszajn theorem, Aronszajn [1950])

Let $\mathcal{X}$ be an index set. Then for every positive definite function k($\langle ., \ . \rangle$) on $\mathcal{X} \times \mathcal{X}$ there exists a unique RKHS, and vice versa.

The Hilbert space L_2 (which has the dot product $\langle f, \ g \rangle_{L_2} = \int f(\boldsymbol{x}) g(\boldsymbol{x}) d\boldsymbol{x}$) contains many non-smooth functions.

In L_2 (which is not a RKHS) the delta function is the representer of evaluation, i.e. $f(\boldsymbol{x}) = \int f(\boldsymbol{x}') \delta(x - \boldsymbol{x}') d\boldsymbol{x}'$. Kernels are the analogues of delta functions within the smoother RKHS.

Note that the delta function is not itself in L_2; in contrast for a RKHS the kernel k is the representer of evaluation and is itself in the RKHS.

The key intuition behind the RKHS formalism is that the squared norm $\|f\|_{\mathcal{H}}^2$ can be thought of as a generalization to functions of the n-dimensional quadratic form $\boldsymbol{f}^\top K^{-1}\boldsymbol{f}$ we have seen in earlier chapters.

Consider a real positive semidefinite kernel $k(\boldsymbol{x},\ \boldsymbol{x}')$ with an eigenfunction expansion $k(\boldsymbol{x},\ \boldsymbol{x}') = \sum_{i=1}^{N} \lambda_i \phi_i(\boldsymbol{x})\phi_i(\boldsymbol{x}')$ relative to a measure μ.

Recall from Mercer's theorem that the eigenfunctions are orthonormal w.r.t. μ, i.e. we have $\int \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})d\mu(\boldsymbol{x}) = \delta_{ij}$.

We now consider a Hilbert space comprised of linear combinations of the eigenfunctions, i.e. $f(\boldsymbol{x}) = \sum_{i=1}^{N} f_i \phi_i(\boldsymbol{x})$ with $\sum_{i=1}^{N} f_i^2/\lambda_i < \infty$.

We assert that the inner product $\langle f,\ g \rangle_{\mathcal{H}}$ in the Hilbert space between functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x}) = \sum_{i=1}^{N} g_i \phi_i(\boldsymbol{x})$ is defined as

$$\langle f,\ g \rangle_{\mathcal{H}} = \sum_{i=1}^{N} \frac{f_i g_i}{\lambda_i} \qquad (6.1)$$

Thus this Hilbert space is equipped with a norm $\|f\|_{\mathcal{H}}$ where $\|f\|_{\mathcal{H}}^2 = \langle f,\ f \rangle_{\mathcal{H}} = \sum_{i=1}^{N} f_i^2/\lambda_i$.

Note that for $\|f\|_{\mathcal{H}}$ to be finite the sequence of coefficients $\{f_i\}$ must decay quickly; effectively this imposes a smoothness condition on the space.

Reproducing property

$$\langle f(.),\ k(.,\ \boldsymbol{x})\rangle_{\mathcal{H}} = \sum_{i=1}^{N} \frac{f_i \lambda_i \phi_i(\boldsymbol{x})}{\lambda_i} = f(\boldsymbol{x}) \qquad (6.2)$$

$$\langle k(\boldsymbol{x},\ .),\ k(\boldsymbol{x}',\ .)\rangle_{\mathcal{H}} = \sum_{i=1}^{N} \frac{\lambda_i \phi_i(\boldsymbol{x}) \lambda_i \phi_i(\boldsymbol{x}')}{\lambda_i} = k(\boldsymbol{x},\ \boldsymbol{x}') \qquad (6.3)$$

Notice also that $k(\boldsymbol{x},\ .)$ is in the RKHS as it has norm $\sum_{i=1}^{N} (\lambda_i \phi_i(\boldsymbol{x}))^2 / \lambda_i = k(\boldsymbol{x},\ \boldsymbol{x}) < \infty$.

The Hilbert space comprised of linear combinations of the eigenfunctions with the restriction $\sum_{i=1}^{N} f_i^2 / \lambda_i < \infty$ fulfills the two conditions given in Definition 6.1. As there is a unique RKHS associated with $k(.,\ .)$, this Hilbert space must be that RKHS.

The advantage of the abstract formulation of the RKHS

The eigenbasis will change with different measures μ in Mercer's theorem. However, the RKHS norm is in fact solely a property of the kernel and is invariant under this change of measure.

Notice the analogy between the RKHS norm $\|f\|_{\mathcal{H}}^2 = \langle f,\ f\rangle_{\mathcal{H}} = \sum_{i=1}^{N} f_i^2/\lambda_i$ and the quadratic form $\boldsymbol{f}^\top K^{-1}\boldsymbol{f}$ if we express K and f in terms of the eigenvectors of K we obtain exactly the same form (but the sum has only n terms if f has length n).

If we sample the coefficients f_i in the eigenexpansion $f(\boldsymbol{x}) = \sum_{i=1}^{N} f_i \phi_i(\boldsymbol{x})$ from $\mathcal{N}(0,\ \lambda_i)$ then

$$E[\|f\|_{\mathcal{H}}^2] = \sum_{i=1}^{N} E[f_i^2]\lambda_i = \sum_{i=1}^{N} 1 \quad (6.4)$$

Thus if N is infinite the sample functions are not in $\mathcal{H}$ (with probability 1) as the expected value of the RKHS norm is infinite

However, note that although sample functions of this Gaussian process are not in $\mathcal{H}$, the posterior mean after observing some data will lie in the RKHS, due to the smoothing properties of averaging.

Another view of the RKHS can be obtained from the reproducing kernel map construction. We consider the space of functions f defined as

$$\{f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x},\ \boldsymbol{x}_i) : n \in N,\ \boldsymbol{x}_i \in \mathcal{X},\ \alpha_i \in \mathfrak{R}\} \quad (6.5)$$

Now let $g(\boldsymbol{x}) = \sum_{j=1}^{n'} \alpha'_j k(\boldsymbol{x}, \boldsymbol{x}'_j)$. Then we define the inner product

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \alpha'_j k(\boldsymbol{x}_i,\ \boldsymbol{x}'_j) \quad (6.6)$$

Clearly condition 1 of Definition 6.1 is fulfilled under the reproducing kernel map construction. We can also demonstrate the reproducing property, as $\langle k(.,\ x),\ f(.) \rangle_{\mathcal{H}} = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x},\ \boldsymbol{x}_i) = f(\boldsymbol{x}) \quad (6.7)$

Section 2

6.2 Regularization

Inferring from a finite dataset without any assumption is clearly "ill posed". For example, in the noise-free case, any function that passes through the given data points is acceptable.

Under a Bayesian approach our assumptions are characterized by a prior over functions, and given some data, we obtain a posterior over functions.

The problem of bringing prior assumptions to bear has also been addressed under the regularization viewpoint, where these assumptions are encoded in terms of the smoothness of f: $J[f] = \frac{\lambda}{2}\|f\|^2_{\mathcal{H}} + Q(\boldsymbol{y},\ \boldsymbol{f})$ (6.8)

- The first term is regularizer and represents smoothness assumptions on f as encoded by a suitable RKHS
- The second term is a data-fit term assessing the quality of the prediction $f(x_i)$ for the observed datum y_i, e.g. the negative log likelihood.

Ridge Regression

Ridge regression can be seen as a particular case of regularization: $\|f\|^2_{\mathcal{H}} = \sum_{i=1}^{N} f_i^2/\lambda_i$ where f_i is the coefficient of eigenfunction $\phi_i(\boldsymbol{x})$, we see that we are penalizing the weighted squared coefficients.

This is taking place in feature space, rather than simply in input space, as per the standard formulation of ridge regression, so it corresponds to **kernel ridge regression**.

Hastie and Tibshirani eq(3.41) Ridge regression

The ridge coefficients minimized the penalized residual sum of squares:

$$\hat{\beta}^{ridge} = argmin_{\beta}\left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum \beta_j^2 \right\}$$

The representer theorem shows that each minimizer $f \in \mathcal{H}$ of $J[f]$ has the form $f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i)$. The representer theorem was first stated by **Kimeldorf and Wahba [1971]** for the case of squared error. **O'Sullivan et al. [1986]** showed that the representer theorem could be extended to likelihood functions arising from generalized linear models.

If the data-fit term is convex (see section A.9) then there will be a unique minimizer $\widehat{f}$ of $J[f]$. (we can regard the $J[f]$ as the penalized likelihood, in ESL [Hastie and Tibshirani] the $J[f]$ is the regularization term)

For Gaussian process prediction with likelihoods that involve the values of f at the n training points only (so that $Q(y, f)$ is the negtive log likelihood up to some terms not involving f), the analogue of the representer theorem is obvious.

This is because the predictive distribution of $f(x_*) \stackrel{\triangle}{=} f_*$ at test point x_* given the data $\boldsymbol{y}$ is $p(f_*|\boldsymbol{y}) = \int p(f_*|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{y})d\boldsymbol{f}$. As derived in eq. (3.22) we have $E[f_*|\boldsymbol{y}] = \boldsymbol{k}(\boldsymbol{x}_*)^\top K^{-1} E[\boldsymbol{f}|\boldsymbol{y}] \quad (6.9)$ due to the formulae for the conditional distribution of a multivariate Gaussian.

Thus $E[f_*|\boldsymbol{y}] = \sum_{i=1}^N \alpha_i k(\boldsymbol{x}_*, \boldsymbol{x}_i)$, where $\boldsymbol{\alpha} = K^{-1} E[\boldsymbol{f}|\boldsymbol{y}]$.

6.2.1 Regularization Defined by Differential Operators

For $\boldsymbol{x} \in \Re^D$ define $\|O^m f\|^2 = \int \sum_{j_1+\ldots+j_D=m} \left(\frac{\partial^m f(\boldsymbol{x})}{\partial x_1^{j_1} \ldots x_j^{j_D}} \right)^2 dx \quad (6.10)$

For example for $m = 2$ and $D = 2$:

$$\|O^2 f\|^2 = \int \left[\left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 \right] dx_1 dx_2 \quad (6.11)$$

Now set $\|Pf\|^2 = \sum_{m=0}^{M} a_m \|O^m f\|^2$ with non-negative coefficients a_m.
Notice that $\|Pf\|^2$ is translation and rotation invariant.

$\|O^m f\|$ is the penalty term; $\|Pf\|$ is the a projection in null space

In this section we assume that $a_0 > 0$; if this is not the case and a_k is the first non-zero coefficient, then there is a null space of functions that are unpenalized. For example if $k = 2$ then constant and linear functions are in the null space. This case is dealt with in section 6.3.

$\|Pf\|^2$ penalizes f in terms of the variability of its function values and derivatives up to order M. The key is to recognize that the complex exponentials $\exp(2\pi i \boldsymbol{s} \cdot \boldsymbol{x})$ are eigenfunctions of the differential operator if $\mathcal{X} = \mathfrak{R}^D$.

$$\|Pf\|^2 = \int \sum_{m=0}^{M} a_m (4\pi^2 \boldsymbol{s} \cdot \boldsymbol{s})^m |\tilde{f}(\boldsymbol{s})|^2 ds \quad (6.12)$$

where $\tilde{f}(\boldsymbol{s})$ is the Fourier transform of $f(\boldsymbol{x})$. Comparing eq. (6.12) with eq. (6.1) we see that the kernel has the power spectrum $S(s) = \frac{1}{\sum_{m=0}^{M} a_m (4\pi^2 \boldsymbol{s} \cdot \boldsymbol{s})^m} \quad (6.13)$ and thus by Fourier inversion we obtain the stationary kernel $k(\boldsymbol{x}) = \int \frac{e^{2\pi \boldsymbol{s} \cdot \boldsymbol{x}}}{\sum_{m=0}^{M} a_m (4\pi^2 \boldsymbol{s} \cdot \boldsymbol{s})^m} ds \quad (6.14)$, by Bochner Theorem and Wiener-Khintchin Theorem (explained why penalize higher frequnecy functions)

A slightly different approach to obtaining the kernel is to use calculus of variations to minimize $J[f]$ with respect to f. The Euler-Lagrange equation leads to: $f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i G(\boldsymbol{x} - \boldsymbol{x}_i) \quad (6.15)$ and $\sum_{m=0}^{M} (-1)^m a_m \nabla^{2m} G = \delta(\boldsymbol{x} - \boldsymbol{x}') \quad (6.16)$, where $G(\boldsymbol{x}, \boldsymbol{x}')$ is known as a Green's function. Notice that the Green's function also depends on the boundary conditions.

For the case of $\mathcal{X} = \Re^D$ by Fourier transforming eq. (6.16) we recognize that G is in fact the kernel k.

The differential operator $\sum_{m=0}^{M} (-1)^m a_m \nabla^{2m}$ and the integral operator $k(\langle ., . \rangle)$ are in fact inverses, as shown by eq. (6.16). Arfken [1985] provides an introduction to calculus of variations and Green's functions. RKHSs for regularizers defined by differential operators are **Sobolev spaces**.

Sidenotes for Green's function

If L is the linear differential operator, then:

- the Green's function G is the solution of the equation $LG = \delta$, where δ is Dirac's delta function;
- the solution of the initial-value problem $Ly = f$ is the convolution $(G * f)$, where G is the Green's function.

Green's function

$G(x,\ s)$ of a linear differential operator $L = L(x)$ acting on distributions over a subset of the Euclidean space $\mathfrak{R}^n$, at a point s, is any solution of $L\ G(x,\ s) = \delta(s - x)$
where δ is the Dirac delta function.
This property of a Green's function can be exploited to solve differential equations of the form $L\ u(x) = f(x)$

If the operator is translation invariant, that is, when L has constant coefficients with respect to x, then the Green's function can be taken to be a convolution kernel, $G(x,\ s) = G(x - s)$.

Motivation

A function G can be found for the operator L, then multiply the Green's function by $f(s)$, and then integrate with respect to s, we obtain $\int L\ G(x,\ s)\ f(s)\ ds = \int \delta(x - s)\ f(s)\ ds = f(x)$.
Because the operator $L = L(x)$ is linear and acts only on the variable x (and not on the variable of integration s), one may take the operator L outside of the integration, yielding $L\ \int G(x,\ s)\ f(s)\ ds = f(x)$
This means that $u(x) = \int G(x,\ s)\ f(s)\ ds$ is a solution to the equation $L\ u(x) = f(x)$

Example 1

Set $a_0 = \alpha^2$, $a_1 = 1$ and $a_m = 0$ for $m \geq 2$ in $D = 1$. Using the Fourier pair $e^{-\alpha|x|} \Leftrightarrow 2\alpha/(\alpha^2 + 4\pi^2 s^2)$ we obtain $k(x - x') = \frac{1}{2\alpha} e^{-\alpha|x-x'|}$. Note that this is the covariance function of the Ornstein-Uhlenbeck process, see section 4.2.1.

Example 2

By setting $a_m = \frac{\sigma^{2m}}{m! 2^m}$ and using the power series $e^y = \sum_{k=0}^{\infty} y^k/k!$ we obtain

$$k(\boldsymbol{x} - \boldsymbol{x}') = \int \exp\left(2\pi i \boldsymbol{s} \cdot (\boldsymbol{x} - \boldsymbol{x}')\right) \exp\left(- \frac{\sigma^2}{2}(4\pi^2 \boldsymbol{s} \cdot \boldsymbol{s})\right) d\boldsymbol{s} \quad (6.17)$$

$$k(\boldsymbol{x} - \boldsymbol{x}') = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(- \frac{1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{x}')^\top (\boldsymbol{x} - \boldsymbol{x}')\right) \quad (6.18)$$

as shown by Yuille and Grzywacz [1989]. This is the squared exponential covariance function that we have seen earlier

6.2.2 Obtaining the Regularized Solution

The representer theorem tells us the general form of the solution to eq. (6.8). We now consider a specific functional

$$J[f] = \frac{1}{2}\|f\|2_{\mathcal{H}} + \frac{1}{2\sigma_n^2}\sum_{i=1}^{n}\left(y_i - f(\boldsymbol{x}_i)\right)^2 \qquad (6.19)$$

which uses a squared error data-fit term (corresponding to the negative log likelihood of a Gaussian noise model with variance σ_n^2).

The solution $f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i)$ that minimizes eq. (6.19) was called a regularization network regularization network in Poggio and Girosi [1990].

Substituting $f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x},\ \boldsymbol{x}_i)$ and using $\langle k(.,\ x_i),\ k(.,\ x_j)\rangle_{\mathcal{H}} = k(x_i,\ x_j)$ we obtain

$$
\begin{aligned}
J[\boldsymbol{\alpha}] &= \frac{1}{2}\boldsymbol{\alpha}^\top K\boldsymbol{\alpha} + \frac{1}{2\sigma_n^2}|\boldsymbol{y} - K\boldsymbol{\alpha}|^2 \\
&= \frac{1}{2}\boldsymbol{\alpha}^\top(K + \frac{1}{\sigma_n^2}K^2)\boldsymbol{\alpha} - \frac{1}{\sigma_n}^2\boldsymbol{y}^\top K\boldsymbol{\alpha} + \frac{1}{2\sigma_n^2}\boldsymbol{y}^\top\boldsymbol{y} \qquad (6.20)
\end{aligned}
$$

Minimizing J by differentiating w.r.t. the vector of coefficients $\boldsymbol{\alpha}$ we obtain $\hat{\boldsymbol{\alpha}} = (K + \sigma_n^2 I)^{-1}\boldsymbol{y}$, so that the prediction for a test point x_* is $\widehat{f}(x_*) = k(x_*)^\top(K + \sigma_n^2 I)^{-1}\boldsymbol{y}$. This should look very familiar-it is exactly the form of the predictive mean obtained in eq. (2.23).

6.2.3 The Relationship of the Regularization View to Gaussian Process Prediction

The regularization method returns $\widehat{f} = argmin_f J[f]$. For a Gaussian process predictor we obtain a posterior distribution over functions.

In fact we shall see in this section that $\widehat{f}$ can be viewed as the maximum a posteriori (MAP) function under the posterior.

Following Szeliski [1987] and Poggio and Girosi [1990] we consider

$$\exp(-J[f]) = \exp\left(-\frac{\lambda}{2}\|Pf\|^2\right) \times \exp(-Q(\boldsymbol{y},\ \boldsymbol{f})) \quad (6.21)$$

- The first term on the RHS is a Gaussian process prior on f
- The second is proportional to the likelihood
- As $\widehat{f}$ is the minimizer of $J[f]$, it is the MAP function.

To get some intuition for the Gaussian process prior, imagine $f(\boldsymbol{x})$ being represented on a grid in $\boldsymbol{x}$-space, so that f is now an (infinite dimensional) vector $\boldsymbol{f}$. Thus we obtain

$$\|Pf\|^2 \simeq \sum_{m=0}^{M} a_m (D_m \boldsymbol{f})^\top (D_m \boldsymbol{f}) = \boldsymbol{f}^\top (\sum_m a_m D_m^\top D_m) \boldsymbol{f}$$

where D_m is an appropriate finite-difference approximation of the differential operator O_m. Observe that this prior term is a quadratic form in $\boldsymbol{f}$.

The MAP relationship

1. when $Q(y, f)$ is quadratic (corresponding to a Gaussian likelihood);
2. when $Q(y, f)$ is not quadratic but convex
3. when $Q(y, f)$ is not convex.

The MAP relationship

1. In case 1 the posterior mean function can be obtained exactly, and the posterior is Gaussian. As the mean of a Gaussian is also its mode this is the MAP solution.
2. In case 2 we have seen in chapter 3 for classification problems using the logistic, probit or softmax response functions that $Q(y, f)$ is convex.
3. In case 3 there will be more than one local minimum of $J[f]$ under the regularization approach. One could check these minima to find the deepest one. However, in this case the argument for MAP is rather weak (especially if there are multiple optima of similar depth) and suggests the need for a fully Bayesian treatment.

While the regularization solution gives a part of the Gaussian process solution, there are the following limitations:

1. It does not characterize the uncertainty in the predictions, nor does it handle well multimodality in the posterior.
2. The analysis is focussed at approximating the first level of Bayesian inference, concerning predictions for f. It is not usually extended to the next level, e.g. to the computation of the marginal likelihood. The marginal ikelihood is very useful for setting any parameters of the covariance function, and for model comparison (see chapter 5).

In addition, we find the specification of smoothness via the penalties on derivatives to be not very intuitive. The regularization viewpoint can be thought of as directly specifying the inverse covariance rather than the covariance.

As marginalization is achieved for a Gaussian distribution directly from the covariance (and not the inverse covariance) it seems more natural to us to specify the covariance function. Also, while non-stationary covariance functions can be obtained from the regularization viewpoint, e.g. by replacing the Lebesgue measure in eq. (6.10) with a non-uniform measure $\mu(\boldsymbol{x})$, calculation of the corresponding covariance function can then be very difficult.

Section 3

6.3 Spline Models

6.3 Spline Models

In section 6.2 we discussed regularizers which had $a_0 > 0$ in eq. (6.12). We now consider the case when $a_0 = 0$; in particular we consider the regularizer to be of the form $\|O^m f\|^2$, as defined in eq. (6.10). In this case polynomials of degree up to $m - 1$ are in the null space of the regularization operator, in that they are not penalized at all.

In the case that $\mathcal{X} = \Re^D$ we can again use Fourier techniques to obtain the Green's function G corresponding to the Euler-Lagrange equation $(-1)^m \nabla^{2m} G(\boldsymbol{x}) = \delta(\boldsymbol{x})$. The result, as shown by Duchon [1977] and Meinguet [1979] is

$$G(\boldsymbol{x} - \boldsymbol{x}') = \begin{cases} c_{m,D} |\boldsymbol{X} - \boldsymbol{x}|^{2m-D} \log |\boldsymbol{x} - \boldsymbol{x}'| & if\ 2m > D\ and\ even \\ c_{m,D} |\boldsymbol{X} - \boldsymbol{x}|^{2m-D} & otherwise \end{cases} \quad (6.22)$$

where $c_{m,\ D}$ is a constant (Wahba [1990, p. 31] gives the explicit form). Note that the constraint $2m > D$ has to be imposed to avoid having a Green's function that is singular at the origin.

Explicit calculation of the Green's function for other domains $\mathcal{X}$ is sometimes possible.

Because of the null space, a minimizer of the regularization functional has the form

$$f(\boldsymbol{x}) = \sum n_{i=1} \alpha_i G(\boldsymbol{x},\ \boldsymbol{x}_i) + \sum_{j=1}^{k} \beta_j h_j(\boldsymbol{x}) \qquad (6.23)$$

where $h_1(\boldsymbol{x})$, ... , $h_k(\boldsymbol{x})$ are polynomials that span the null space.

The exact values of the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for a specific problem can be obtained in an analogous manner to the derivation in section 6.2.2;

The solution is equivalent to that given in eq. (2.42). To gain some more insight into the form of the Green's function we consider the equation $(-1)^m \nabla^{2m} G(\boldsymbol{x}) = \delta(\boldsymbol{x})$ in Fourier space, leading to $\tilde{G}(\boldsymbol{s}) = (4\pi^2 \boldsymbol{s} \cdot \boldsymbol{s})^{-m}$. $\tilde{G}(\boldsymbol{s})$ plays a role like that of the power spectrum in eq. (6.13), but notice that $\int \tilde{G}(\boldsymbol{s}))ds$ is infinite, which would imply that the corresponding process has infinite variance.

The problem is of course that the null space is unpenalized; for example any arbitrary constant function can be added to f without changing the regularizer.

Because of the null space we have seen that one cannot obtain a simple connection between the spline solution and a corresponding Gaussian process problem.

However, by introducing the notion of an intrinsic random function (IRF) one can define a generalized covariance;

The basic idea is to consider linear combinations of $f(\boldsymbol{x})$ of the form $g(\boldsymbol{x}) = \sum_{i=1}^{k} a_i f(\boldsymbol{x} + \boldsymbol{\delta}_i)$ for which $g(\boldsymbol{x})$ is second-order stationary and where $(h_j(\boldsymbol{\delta}1), \ldots, h_j(\boldsymbol{\delta}_k))\boldsymbol{a} = 0$ for $j = 1, \ldots, k$. A careful description of the equivalence of spline and IRF prediction is given in Kent and Mardia [1994].

The power-law form of $\tilde{G}(\boldsymbol{s})) = (4\pi^2 \boldsymbol{s} \cdot \boldsymbol{s})^{-m}$ means that there is no characteristic length-scale for random functions drawn from this (improper) prior.

Some authors argue that the lack of a characteristic length-scale is appealing. This may sometimes be the case, but if we believe there is an appropriate length-scale (or set of length-scales) for a given problem but this is unknown in advance, we would argue that a hierarchical Bayesian formulation of the problem (as described in chapter 5) would be more appropriate.

Splines were originally introduced for one-dimensional interpolation and smoothing problems, and then generalized to the multivariate setting.

Schoenspline interpolation berg [1964] considered the problem of finding the function that minimizes $\int_a^b (f^{(m)}(\boldsymbol{x}))^2 dx \quad (6.24)$, where $f(m)$ denotes the mth derivative of f, subject to the interpolation constraints $f(x_i) = f_i$, $x_i \in (a,\ b)$ for $i = 1,\ \ldots\ ,\ n$ and for f in an appropriate natural polynomial Sobolev space. He showed that the solution is the natural polynomial spline, which is a piecewise polynomial of order $2m - 1$ in each interval $[x_i,\ x_{i+1}]$, $i = 1,\ \ldots,\ n - 1$, and of order $m - 1$ in the two outermost intervals.

The pieces are joined so that the solution has $2m-2$ continuous derivatives. Schoenberg also proved that the solution to the univariate smoothing problem (see eq. (6.19)) is a natural polynomial spline.

A common choice is $m=2$, leading to the cubic spline. One possible way of writing this solution is $f(\boldsymbol{x}) =$

$$\sum_{j=0}^{1} \beta_j x^j + \sum_{i=1}^{n} \alpha_i (x - x_i)_+^3, \; (\boldsymbol{x})_+ = \begin{cases} (x)_+ & if \; x > 0 \\ 0 & otherwise \end{cases} \qquad (6.25)$$

It turns out that the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be computed in time $\mathcal{O}(n)$ using an algorithm due to Reinsch.

Splines were first used in regression problems. However, by using generalized linear modelling **[McCullagh and Nelder, 1983]** they can be extended to classification problems and other non-Gaussian likelihoods, as we did for GP classification in section 3.3.

* 6.3.1 A 1-d Gaussian Process Spline Construction

In this section we will further clarify the relationship between splines and Gaussian processes by giving a GP construction for the solution of the univariate cubic spline smoothing problem whose cost functional is

$$\sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2 + \lambda \int_0^1 \left(f''(x)\right)^2 dx \quad (6.26)$$

where the observed data are $\{(x_i,\ y_i)|i = 1,\ ...,\ n,\ 0 < x_1 < ... < x_n < 1\}$ and λ is a smoothing parameter controlling the trade-off between the first term, the data-fit, and the second term, the regularizer, or complexity penalty.

Recall that the solution is a piece-wise polynomial as in eq. (6.25). Following Wahba [1978], we consider the random function

$g(x) = \sum_{j=0}^{1} \beta_j x^j + f(x) (6.27)$

where $\beta \sim \mathcal{N}(0, \sigma^2 I)$ and $f(x)$ is a Gaussian process with covariance

To complete the analogue of the regularizer in eq. (6.26), we need to remove any penalty on polynomial terms in the null space by making the prior vague, i.e. by taking the limit $\sigma_\beta^2 \to \infty$. Notice that the covariance has the form of contributions from explicit basis functions, $\boldsymbol{h}(\boldsymbol{x}) = (1,\ x)^\top$ and a regular covariance function $k_{sp}(x,\ x')$, a problem which we have already studied in section 2.7.

Indeed we have computed the limit where the prior becomes vague $\sigma_\beta^2 \to \infty$, the result is given in eq. (2.42).

Plugging into the mean equation from eq. (2.42), we get the predictive mean $\bar{f}(x_*) = \boldsymbol{k}(x_*)^\top K_y^{-1}(\boldsymbol{y} - H^\top \bar{\beta}) + \boldsymbol{h}(x_*)^\top \bar{\beta}$ (6.29), where K_y is the covariance matrix corresponding to $\sigma_f^2 k_{sp}(x_i,\ x_j) + \sigma_n^2 \delta_{ij}$ evaluated at the training points, H is the matrix that collects the $\boldsymbol{h}(x_i)$ vectors at all training points, and $\bar{\beta} = (HK_y^{-1}H^\top)^{-1} HK_y^{-1}\boldsymbol{y}$ is given below eq. (2.42).

It is not difficult to show that this predictive mean function is a piecewise cubic polynomial, since the elements of $k(x_*)$ are piecewise cubic polynomials.

So far k_{sp} has been produced rather mysteriously "from the hat";

Shepp [1966] defined the l-fold integrated Wiener process as
$W_l(\boldsymbol{x}) = \int_0^1 \frac{(x-u)_+^l}{l!} Z(u)du,\ l = 0,\ 1,\ \ldots \quad (6.30)$

where $Z(u)$ denotes the Gaussian white noise process with covariance $\delta(u - u')$. Note that W_0 is the standard Wiener process.

It is easy to show that $k_{sp}(x,\ x')$ is the covariance of the once-integrated Wiener process by writing $W_1(x)$ and $W_1(x')$ using eq. (6.30) and taking the expectation using the covariance of the white noise process.

Note that W_l is the solution to the stochastic differential equation (SDE) $X^{(l+1)} = Z$; see Appendix B for further details on SDEs.

Thus for the cubic spline we set $l = 1$ to obtain the SDE $\boldsymbol{x}'' = Z$, corresponding to the regularizer $\int (f''(x))^2 dx$.

Consider the family of random functions given by

$$f_N(x) = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} \gamma_i (x - \tfrac{i}{N})_+ \qquad (6.31)$$

where $\boldsymbol{\gamma}$ is a vector of parameters with $\gamma \sim \mathcal{N}(\mathbf{0},\ I)$. Note that the sum has the form of evenly spaced "ramps" whose magnitudes are given by the entries in the $\boldsymbol{\gamma}$ vector.

$$E[f_N(x) f_N(x')] = \frac{1}{N} \sum_{i=0}^{N-1} (x - \frac{i}{N})_+ (x' - \frac{i}{N})_+ \qquad (6.32)$$

Taking the limit $N \to \infty$, we obtain eq. (6.28), a derivation which is also found in [Vapnik, 1998, sec. 11.6].

Notice that the covariance function k_{sp} given in eq. (6.28) corresponds to a Gaussian process which is MS continuous but only once MS differentiable.

Thus samples from the prior will be quite "rough", although (as noted in section 6.1) the posterior mean, eq. (6.25), is smoother.