Nonparametric curve fitting with roughness penalties

Zhang Jin

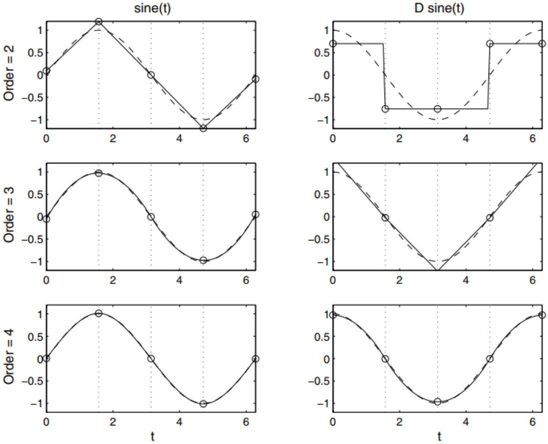
University of Science and Technology of China

September, 2022

Table of Contents

Spline

- A spline is a special function defined piecewise by polynomials.
  - Over each interval, a spline is a polynomial of specified order m
  - The function values and derivatives up to order $m-2$ must match up at the junctions.



B-Spline

To actually construct splines, we specify a system of basis functions $\phi_k(t)$, and these will have the following essential properties:

- Each basis function $\phi_k(t)$ is itself a spline function as defined by an order m and a knot sequence τ.
- Any spline function defined by m and τ can be expressed as a linear combination of these basis functions.

Although there are many ways that such systems can be constructed, the **B-spline basis system** is the most popular one.

B-Spline

Denote by $B_{i,m}(x)$ the i th B-spline basis function of order m for the knot-sequence $\tau, m \leq M$. They are defined recursively in terms of divided differences as follows:

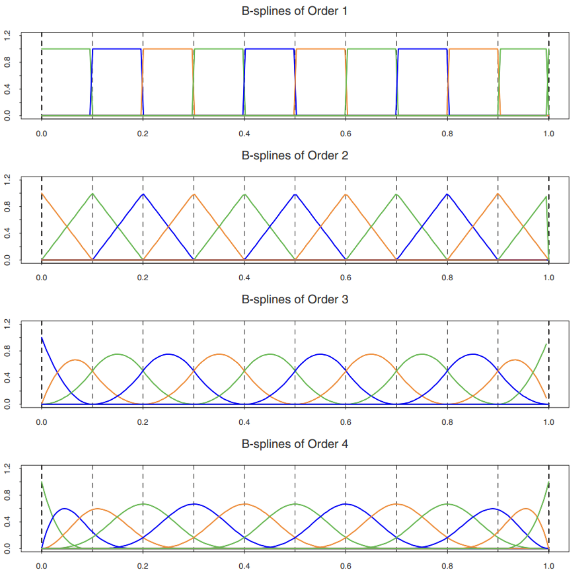
$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, K + 2M - 1$.

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x)$$

for $i = 1, \ldots, K + 2M - m$

B-Spline


B-splines of Order 1
B-splines of Order 2
B-splines of Order 3
B-splines of Order 4
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0

The aims of curve fitting

- What is it that make the functions shown in these two curves unsatisfactory as explanations of the given data?

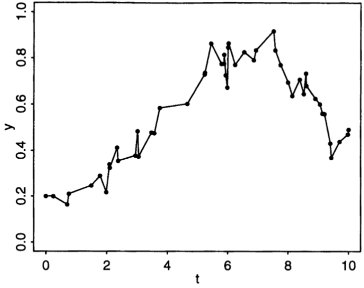


Figure 1.1. *Synthetic data joined by straight lines.*

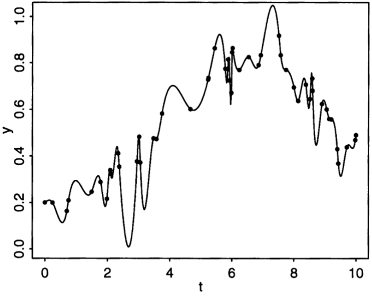


Figure 1.2. *Synthetic data interpolated by a curve with continuous second derivative.*

Roughness Penalties

- No restriction on the curve g leads to 'unnatural' estimation.
- There are many different ways of measuring how 'rough' or 'wiggly' the curve g is. Given the interval $[a, b]$, an intuitive measure is $\int_a^b \{g''(t)\}^2 \, dt$.
- Particularly in the context of regression, it is **natural** for any measure of roughness not to be affected by the addition of a constant or linear function, so that if two functions differ only by a constant or a linear function then their roughness should be identical.
- The roughness penalty approach to curve estimation is now easily stated. Given any twice-differentiable function g defined on $[a, b]$, and a smoothing parameter $\alpha > 0$, define the penalized sum of squares

$$S(g) = \sum_{i=1}^{n} \{Y_i - g(t_i)\}^2 + \alpha \int_a^b \{g''(x)\}^2 \, dx$$

- Moderate α, large α and small α

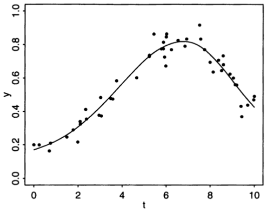


Figure 1.3. *Synthetic data with the curve that minimizes S(g) with* $\alpha = 1$.

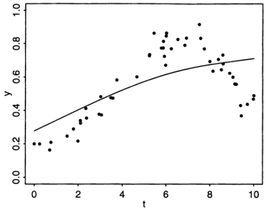


Figure 1.4. *Synthetic data with the curve that minimizes S(g) for a large value of* α.

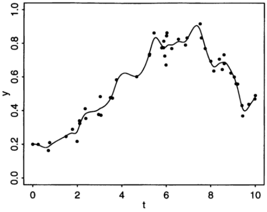
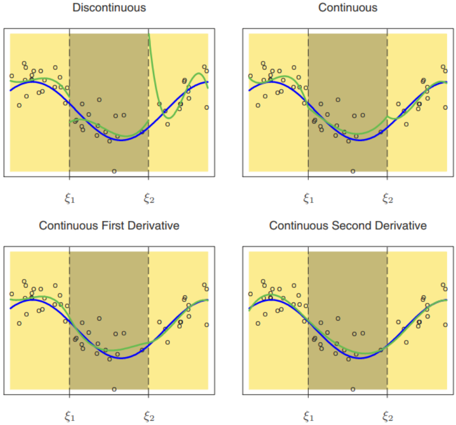


Figure 1.5. *Synthetic data with the curve that minimizes S(g) for a small value of* α.

Cubic spline

- Suppose we are given real numbers $t_1, \ldots, t_n$ on some interval $[a, b]$, satisfying $a < t_1 < t_2 < \ldots < t_n < b$. A function g defined on $[a, b]$ is a cubic spline if two conditions are satisfied
  - cubic polynomial on each of the intervals
  - continuous (itself, first and second derivatives) at each knots t_i.



Natural cubic spline

- Natural cubic spline (NCS): a cubic spline on $[a, b]$ with its second and third derivatives are 0 at a and b.
- Value-second derivative representation: Suppose that g is a NCS with knots $t_1 < \ldots < t_n$. Define

$$g_i = g\left(t_i\right) \text{ and } \gamma_i = g''\left(t_i\right) \text{ for } i = 1, \ldots, n$$

The vectors $\boldsymbol{g}$ and γ specify the curve g completely.

Natural cubic spline: definition of matrices Q,R,K

- Let $h_i = t_{i+1} - t_i$ for $i = 1, \ldots, n-1$. Let Q be the $n \times (n-2)$ matrix with entries q_{ij}, for $i = 1, \ldots, n$ and $j = 2, \ldots, n-1$, given by

$$q_{j-1,j} = h_{j-1}^{-1}, q_{jj} = -h_{j-1}^{-1} - h_j^{-1}, \text{ and } q_{j+1,j} = h_j^{-1}$$

  for $j = 2, \ldots, n-1$, and $q_{ij} = 0$ for $|i-j| \geq 2$. (The top left element of Q is q_{12})
- The symmetric matrix R is $(n-2) \times (n-2)$ with elements r_{ij}, for i and j running from 2 to $(n-1)$, given by

$$r_{ii} = \frac{1}{3}(h_{i-1} + h_i) \text{ for } i = 2, \ldots, n-1,$$
$$r_{i,i+1} = r_{i+1,i} = \frac{1}{6} h_i \text{ for } i = 2, \ldots, n-2,$$

  and $r_{ij} = 0$ for $|i-j| \geq 2$
- Define the matrix K by $K = QR^{-1}Q^T$

Natural cubic spline

Theorem 1

The vectors g *and* γ *specify a natural cubic spline* g *if and only if the condition*

$$Q^T\mathbf{g} = R\boldsymbol{\gamma}$$

is satisfied. If (2.4) is satisfied then the roughness penalty will satisfy

$$\int_a^b g''(t)^2 dt = \gamma^T R\gamma = \mathbf{g}^T K\mathbf{g}$$

Interpolating NCS

Theorem 2

Suppose $n \geq 2$ and that $t_1 < \ldots < t_n$. Given any values $z_1, \ldots, z_n$, there is a **unique** *natural cubic spline g with knots at the points t_i satisfying*

$$g\left(t_i\right) = z_i \text{ for } i = 1, \ldots, n$$

Proof: Since R is strictly positive-definite, there will be a unique γ, given by $\gamma = R^{-1}Q^T\mathbf{g}$, satisfying the required condition.

Optimality properties of the NCS interpolant

Theorem 3

Suppose $n \geq 2$, and that g is the natural cubic spline interpolant to the values $z_1, \ldots, z_n$ at points $t_1, \ldots, t_n$ satisfying $a < t_1 < \ldots < t_n < b$. Let $\tilde{g}$ be any function in $\mathcal{S}_2[a, b]$ for which $\tilde{g}(t_i) = z_i$ for $i = 1, \ldots, n$. Then

$$\int \tilde{g}''^2 \geq \int g''^2$$

, with equality only if $\tilde{g}$ and g are identical.

Smoothing spline

- Given any function g in $S_2[a, b]$, let $S(g)$ be the penalized sum of squares

$$\sum_{i=1}^{n}\{Y_i - g(t_i)\}^2 + \alpha \int_a^b \{g''(x)\}^2 dx$$

- To minimize $S(g)$, g must be a natural cubic spline, as a corollary of Theorem 3.
- Knowing that g is an NCS, we can rewrite $S(g)$ as

$$\begin{aligned} S(g) &= (\mathbf{Y} - \mathbf{g})^T(\mathbf{Y} - \mathbf{g}) + \alpha \mathbf{g}^T K \mathbf{g} \\ &= \mathbf{g}^T(I + \alpha K)\mathbf{g} - 2\mathbf{Y}^T\mathbf{g} + \mathbf{Y}^T\mathbf{Y}' \end{aligned}$$

which has a unique minimum, obtained by setting $\mathbf{g} = (I + \alpha K)^{-1}\mathbf{Y}$

Smoothing spline

Theorem 4

Suppose $n \geq 3$ and that $t_1, \ldots, t_n$ are points satisfying $a < t_1 < \ldots < t_n < b$. Given data points $Y_1, \ldots, Y_n$, and a strictly positive smoothing parameter α, let $\hat{g}$ be the natural cubic spline with knots at the points $t_1, \ldots, t_n$ for which $\mathrm{g} = (I + \alpha K)^{-1}\mathbf{Y}$. Then, for any g in $\mathcal{S}_2[a, b]$,

$$S(\hat{g}) \leq S(g)$$

with equality only if g and $\hat{g}$ are identical.

However, it's in practice inefficient to use $\mathbf{g} = (I + \alpha K)^{-1}\mathbf{Y}$ to find $\mathbf{g}$ and hence g.

The Reinsch algorithm

- Rearrange the items, we can get $\mathbf{g} = \mathbf{Y} - \alpha Q\boldsymbol{\gamma}$, with $\left(R + \alpha Q^T Q\right)\gamma = Q^T\mathbf{Y}$
- The matrix $\left(R + \alpha Q^T Q\right)$ is a band matrix with bandwidth 5, symmetric and strictly PD. A Cholesky decomposition shows

$$R + \alpha Q^T Q = LDL^T$$

- The Reinsch algorithm can be set out, which can be solved in $O(n)$ operations:
  1. Evaluate the vector $Q^T\mathbf{Y}$
  2. Find the non-zero diagonals of $R + \alpha Q^T Q$, and hence the Cholesky decomposition factors L and D.
  3. Solve the equation for γ from $LDL^T\gamma = Q^T\mathbf{Y}$
  4. Use $\mathbf{g} = \mathbf{Y} - \alpha Q\boldsymbol{\gamma}$ to find $\mathbf{g}$.

Choosing the smoothing parameter α

- From the most well known cross-validation method, the overall efficacy of the procedure with the smoothing parameter α can be quantified by the cross-validation score function

$$CV(\alpha) = n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \hat{g}^{(-i)} (t_i; \alpha) \right\}^2$$

- From Theorem 4, the values of the smoothing spline $\hat{g}$ depend linearly on the data Y_i through the equation

$$\mathbf{g} = A(\alpha)\mathbf{Y}$$

where the matrix $A(\alpha)$ is defined by

$$A(\alpha) = \left(I + \alpha Q R^{-1} Q^T \right)^{-1}$$

Calculating the CV score

Theorem 5

The cross-validation score satisfies

$$CV(\alpha) = n^{-1} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{g}(t_i)}{1 - A_{ii}(\alpha)} \right)^2,$$

where $\hat{g}$ is the spline smoother calculated from the full data set $\{(t_i, Y_i)\}$ with smoothing parameter α.

Provided the diagonal entries $A_{ii}(\alpha)$ are known, the cross-validation score can be calculated from the residuals $Y_i - \hat{g}(t_i)$ about the spline smoother calculated from the full data set.

Generalized cross-validation

- To replace the factors in $CV(\alpha)$ by their average value, we obtain

$$GCV(\alpha) = n^{-1} \frac{\sum_{i=1}^{n} \{Y_i - \hat{g}(t_i)\}^2}{\{1 - n^{-1} \operatorname{tr} A(\alpha)\}^2}$$

- Degrees of freedom: $DF_\alpha = \operatorname{tr} A(\alpha)$
  - As $\alpha \to 0, \mathrm{DF}_\alpha \to N,$ and $A(\alpha) \to \mathbf{I}$
  - As $\alpha \to \infty, \mathrm{DF}_\alpha \to 2,$ and $A(\alpha) \to \mathbf{H}$, the hat matrix for linear regression.

Table of Contents

Motivations of the L-Spline criterion

The roughness penalty $PEN_2(x) = \left\|D^2x\right\|^2$ can be extended with a more general linear differential operator. Motivations includes:

- We may wish the class of functions that have zero roughness to be wider than, or otherwise different from, those that are of the form $a + bt$. For example, if we desire a smooth estimate of acceleration D^2x, we may well want to penalize the size of D^4x.
- We may have in mind that, locally at least, curves x should ideally satisfy a particular DE, and we may wish to penalize departure from this. For instance, if we were observing periodic data on an interval $[0, T]$, we know that $\omega^2x + D^2x = 0$ is the linear differential equation satisfied by this type of variation.

Example of the L: Sweden GDP

The long-range trend in GDP tends to be roughly exponential. This suggests the use of the order 4 composite operator

$$L = \left(-\gamma D + D^2\right)\left(\omega^2 I + D^2\right)$$

to annihilate $\mathbf{u}(t) = (1, \exp \gamma t, \sin \omega t, \cos \omega t)'$.

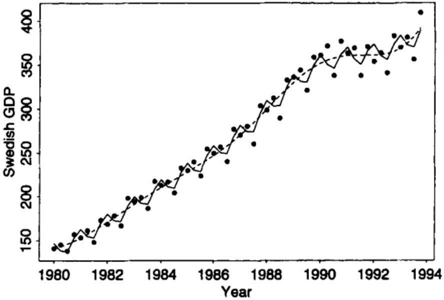


Figure 1. The gross domestic product for Sweden with seasonal variation. The solid line is the smooth using operator $L = (-\gamma D + D^2)(\omega^2 I + D^2)$, and the dashed line is the smooth for $L = D^4$, the smoothing parameter being determined by minimizing the *GCV* criterion in both cases.

Linear differential operator L

We will assume that the linear differential operator is in the form

$$Lx = \sum_{j=0}^{m-1} \beta_j D^j x + D^m x$$

Linear differential operators L of degree m have m linearly independent solutions ξ_j of the homogeneous equation $L\xi_j = 0$.

Consider both directions of $L\xi_j = 0$

- Finding L that annihilates known functions $\boldsymbol{\xi}$:
  The order m Wronskian matrix

  $$\mathbf{W}(t) = \left[\begin{array}{cccc} \boldsymbol{\xi}(t) & D\boldsymbol{\xi}(t) & \ldots & D^{m-1}\boldsymbol{\xi}(t) \end{array}\right]$$

  must be invertible. Then, the vector of weight functions $\boldsymbol{\beta} = (\beta_0(t), \ldots, \beta_{m-1}(t))'$ satisfy the system of m linear equations

  $$\mathbf{W}(t)\boldsymbol{\beta}(t) = -D^m\boldsymbol{\xi}(t)$$

- Finding the functions ξ_j satisfying $L\xi_j = 0$:
  A common procedure is to use a numerical differential equation solving algorithm, such as one of the Runge-Kutta methods, to solve the equation for initial value constraints.

Constraint operator B

We require to identify a specific function x as the unique solution to $Lx = 0$. This operator B simply evaluates x or its derivatives in m different ways. Common examples such as

- Initial operator:

$$B_0 x = \begin{bmatrix} x(0) \\ Dx(0) \\ \vdots \\ D^{m-1}x(0) \end{bmatrix}$$

- Boundary operator:

$$B_B x = \begin{bmatrix} x(0) \\ x(T) \\ \vdots \\ D^{(m-2)/2}x(0) \\ D^{(m-2)/2}x(T) \end{bmatrix}$$

L and B can partition functions

- (Partition principle): Any function x having m derivatives can be expressed uniquely as

$$x = \xi + e \text{ where } L\xi = 0 \text{ and } Be = 0$$

- This happens if and only if $x = 0$ is the only function satisfying both $Bx = 0$ and $Lx = 0$. Or, in algebraic notation,

$$\ker B \cap \ker L = 0$$

- We can define a large family of inner products as follows:

$$\langle x, y \rangle_{B,L} = (Bx)'(By) + \int (Lx)(t)(Ly)(t)dt$$

with the corresponding norm

$$\|x\|_{B,L}^2 = (Bx)'(Bx) + \int (Lx)^2(t)dt$$

Nonhomogeneous equation

Consider the nonhomogeneous equation

$$Lx = u$$

for known L but arbitrary u. In effect, we want to reverse the effect of applying operator L because we have a **forcing function** u and we want to find x. In addtion, to promise the solution is unique, we add the constraints in the form

$$Bx = \mathbf{b}$$

for some known fixed m-vector $\mathbf{b}$.

Nonhomogeneous equation

Define the matrix $\mathbf{A}$ as the result of applying constraint operator B to each of the ξ_j 's in turn:

$$\mathbf{A} = B\boldsymbol{\xi}'$$

so that the element in row i and column j of $\mathbf{A}$ is the i th element of vector $B\xi_j$. Since every ξ in $\ker L$ can be written as

$$\xi(t) = \sum_j c_j \xi_j(t) = \boldsymbol{\xi}'\mathbf{c}$$

for an m-vector of coefficients $\mathbf{c}$, then by the definition of $\mathbf{A}$ we have that

$$B\xi = \mathbf{b} = \mathbf{A}\mathbf{c}.$$

The conditions we have specified ensure that $\mathbf{A}$ is invertible, and consequently we have that

$$\mathbf{c} = \mathbf{A}^{-1}\mathbf{b}$$

Nonhomogeneous equation

Now suppose that ν satisfies $L\nu = u$ and also $B\nu = 0$. That is, $\nu \in \ker B$, and in this sense is the complement of $\xi \in \ker L$. Then

$$x(t) = \xi(t) + \nu(t)$$

satisfies

$$Lx = u \text{ subject to } Bx = \mathbf{b}.$$

Consequently, if we can solve the problem

$$L\nu = u \text{ subject to } \nu \in \ker B,$$

we can find a solution subject to the more general constraint $Bx = \mathbf{b}$.

Green's function

It can be shown that there exists a bivariate function $G(t;s)$ called the Green's function, associated with the pair of operators (B, L) that satisfies

$$\nu(t) = \int G(t;s)L\nu(s)ds \text{ for } \nu \in \ker B.$$

Thus, for $L\nu = u$, the Green's function defines an integral transform

$$\mathcal{G}u = \int G(t;s)u(s)ds$$

that inverts the linear differential operator L. That is, $\mathcal{G}L\nu = \nu$, given that $B\nu = 0$

Green's function

A Green's function, $G(t, s)$, of a linear differential operator L acting on distributions over a subset of the Euclidean space $\mathbb{R}^n$, at a point s, is any solution of

$$LG(t, s) = \delta(s - t)$$

where δ is the Dirac delta function.

- Let's look at a few specific examples. The first is nearly trivial: If our interval is $[0, T]$ and our constraint operator is the initial value constraint $B_0 x = x(0)$, then for $L = D$,

$$G(t; s) = 1, s \leq t, \text{ and 0 otherwise.}$$

  That is, for ν such that $\nu(0) = 0$,

$$\nu(t) = \int_0^t D\nu(s)ds = \int_0^t u(s)ds$$

Green's function

- Now consider the first order constant coefficient equation

$$Dx(t) = -\beta x(t) + u(t)$$

In this problem, $L = D + \beta$. The well-known solution is

$$x(t) = Ce^{-\beta t} + \alpha \int_0^t e^{-\beta(t-s)} u(s) ds$$

We see by inspection that

$$G(t; s) = e^{-\beta(t-s)}, s \leq t, \text{ and 0 otherwise.}$$

Green's function

Differential operator L	Green's function G	Example of application
∂_t^{n+1}	$\frac{t^n}{n!}\Theta(t)$	
$\partial_t+\gamma$	$\Theta(t)e^{-\gamma t}$	
$(\partial_t+\gamma)^2$	$\Theta(t)te^{-\gamma t}$	
$\partial_t^2+2\gamma\partial_t+\omega_0^2$ where $\gamma<\omega_0$	$\Theta(t)e^{-\gamma t}\,\frac{\sin(\omega t)}{\omega}$ with $\omega=\sqrt{\omega_0^2-\gamma^2}$	1D underdamped harmonic oscillator
$\partial_t^2+2\gamma\partial_t+\omega_0^2$ where $\gamma>\omega_0$	$\Theta(t)e^{-\gamma t}\,\frac{\sinh(\omega t)}{\omega}$ with $\omega=\sqrt{\gamma^2-\omega_0^2}$	1D overdamped harmonic oscillator
$\partial_t^2+2\gamma\partial_t+\omega_0^2$ where $\gamma=\omega_0$	$\Theta(t)e^{-\gamma t}t$	1D critically damped harmonic oscillator
2D Laplace operator $\nabla^2_{\text{2D}}=\partial_x^2+\partial_y^2$	$\frac{1}{2\pi}\ln\rho$ with $\rho=\sqrt{x^2+y^2}$	2D Poisson equation
3D Laplace operator $\nabla^2_{\text{3D}}=\partial_x^2+\partial_y^2+\partial_z^2$	$\frac{-1}{4\pi r}$ with $r=\sqrt{x^2+y^2+z^2}$	Poisson equation
Helmholtz operator $\nabla^2_{\text{3D}}+k^2$	$\frac{-e^{-ikr}}{4\pi r}=i\sqrt{\frac{k}{32\pi r}}H^{(2)}_{1/2}(kr)=i\frac{k}{4\pi}h^{(2)}_0(kr)$	stationary 3D Schrödinger equation for free particle
∇^2-k^2 in n dimensions	$-(2\pi)^{-n/2}\left(\frac{k}{r}\right)^{n/2-1}K_{n/2-1}(kr)$	Yukawa potential, Feynman propagator
$\partial_t^2-c^2\partial_x^2$	$\frac{1}{2c}\Theta(t-\lvert x/c\rvert)$	1D wave equation
$\partial_t^2-c^2\,\nabla^2_{\text{2D}}$	$\frac{1}{2\pi c\sqrt{c^2t^2-\rho^2}}\Theta(t-\rho/c)$	2D wave equation
D'Alembert operator $\square=\frac{1}{c^2}\partial_t^2-\nabla^2_{\text{3D}}$	$\frac{\delta(t-\frac{r}{c})}{4\pi r}$	3D wave equation

A matrix analogue of the Green's function

- Suppose that we have, for $n > m$ an $n - m$ by n matrix $\mathbf{L}$ of rank $n - m$. Then there exists a subspace of n-vectors $\boldsymbol{\xi} \in \ker \mathbf{L}$ such that

$$\mathbf{L}\xi = 0$$

  and that space is of dimension m. We can construct a n by m matrix $\mathbf{Z}$ whose columns span this subspace such that $\mathbf{LZ} = 0$.
- Also, we can always find an m by n matrix $\mathbf{B}$ of rank m such that there exists a space of dimension m of n vectors $\boldsymbol{\nu}$ such that

$$\mathbf{B}\nu = 0$$

  We can find an n by $n - m$ matrix $\mathbf{N}$ such that $\mathbf{BN} = 0$

A matrix analogue of the Green's function

Now suppose that we have an arbitrary n-vector $\mathbf{u}$. Then it follows that

$$\boldsymbol{\nu} = \mathbf{N}(\mathbf{LN})^{-1}\mathbf{u}$$

solves the equation

$$\mathbf{L}\boldsymbol{\nu} = \mathbf{u}$$

and, moreover, $\boldsymbol{\nu} \in \ker \mathbf{B}$ since $\mathbf{BN} = 0$. Matrix

$$\mathbf{G} = \mathbf{N}(\mathbf{LN})^{-1}$$

is the analogue of the Green's function $G(s;t)$.

A recipe for the Green's function

- We can now offer a recipe for constructing the Green's function for any linear differential operator L and the initial value constraint B_I of the corresponding order
  1. First, compute the Wronskian matrix $\mathbf{W}(t)$.
  2. Secondly, define the functions
  $$\mathbf{v}(t) = (v_1(t), \ldots, v_m(t))'$$
  to be the vector containing the elements of the last row of $\mathbf{W}^{-1}$.
  3. The initial value constraint Green's function $G_0(t; s)$ is
  $$G_0(t; s) = \sum_{j=1}^{m} \xi_j(t) v_j(s) = \boldsymbol{\xi}(t)' \mathbf{v}(s), s \leq t, \text{ and 0 otherwise.}$$

A recipe for the Green's function

Let's see how this works for

$$L = \beta D + D^2.$$

The space $\ker L$ is spanned by the two functions $\xi_1(t) = 1$ and $\xi_2(t) = \exp(-\beta t)$. The Wronskian matrix is

$$\mathbf{W}(t) = \left[\begin{array}{cc} \xi_1(t) & D\xi_1(t) \\ \xi_2(t) & D\xi_2(t) \end{array} \right] = \left[\begin{array}{ll} 1 & 0 \\ \exp(-\beta t) & -\beta \exp(-\beta t) \end{array} \right]$$

and consequently

$$\mathbf{W}^{-1}(t) = \left[\begin{array}{ll} 1 & 0 \\ \beta^{-1} & -\beta^{-1} \exp(\beta t) \end{array} \right]$$

from which we have

$$\mathbf{v}(s) = -\beta^{-1}[-1, \exp(\beta s)]'$$

and finally

$$G_0(t; s) = -\beta^{-1} \left[e^{-\beta(t-s)} - 1 \right], s \leq t, \text{ and } 0 \text{ otherwise.}$$

RKHS

- If the evaluation map $\rho_t(x) = x(t)$ is continuous on a Hilbert space, then it is called reproducing kernel Hilbert space (RKHS).
- From Riesz representation theorem, since the evaluation map $\rho_t(x)$ is linear and continuous, there must exist a bivariate function $k(s,t)$ such that $k(\cdot,t)$ is in the space for any t, and that

$$\rho_t(x) = \langle x, k(\cdot,t)\rangle$$

- The term reproducing kernel comes from the consequence that

$$k(s,t) = \langle k(\cdot,s), k(\cdot,t)\rangle.$$

The reproducing kernel for ker B

- Given any two functions x and y in $\ker B$, let us define the L-inner product

$$\langle x, y\rangle_L = \langle Lx, Ly\rangle = \int Lx(s)Ly(s)ds$$

- Let G_I be the Green's function, and define a function $k_2(t,s)$ such that, for all t,

$$Lk_2(t,\cdot) = G_I(t;\cdot) \text{ and } Bk_2(t,\cdot) = 0$$

  By the defining properties of Green's functions, this means that

$$k_2(t,s) = \int G_I(s;w)G_I(t;w)dw$$

The reproducing kernel for ker B

The function k_2 has an interesting property. Suppose that ν is any function in $\ker B$, and consider the L-inner product of $k_2(t, \cdot)$ and ν. We have, for all t

$$\langle k_2(t,\cdot), \nu\rangle_L = \int Lk_2(t,s)L\nu(s)ds = \int G_I(t;s)L\nu(s)ds = \nu(t)$$

Thus, in the space $\ker B$ equipped with the L-inner product, taking the L-inner product of k_2 using its second argument with any function ν yields the value of ν at its first argument.

Overall, taking the inner product with k_2 **reproduces** the function ν, and k_2 is called the **reproducing kernel** for this function space and inner product.

$$\langle k_2(s,\cdot), k_2(t,\cdot)\rangle_L = k_2(s,t)$$

We can put the expression in a slightly more convenient form for the purpose of calculation. Recalling the definitions of the vector-valued functions $\boldsymbol{\xi}$ and $\mathbf{v}$, assuming that $s \leq t$, that

$$k_2(s,t) = \int_0^s \left[\boldsymbol{\xi}(s)'v(w)\right]\left[v(w)'\boldsymbol{\xi}(t)\right] dw = \boldsymbol{\xi}(s)'\mathbf{F}(s)\boldsymbol{\xi}(t),$$

where the order m symmetric matrix-valued function $\mathbf{F}(s)$ is

$$\mathbf{F}(s) = \int_0^s v(w)v(w)'dw.$$

To deal with the case $s > t$, we use the property that $k_2(s,t) = k_2(t,s)$.

The reproducing kernel for ker L

Suppose now that $f = \sum a_i \xi_i$ and $g = \sum b_i \xi_i$ are elements of ker L. We can consider the B-inner product on the finite-dimensional space ker L, defined by

$$\langle f, g \rangle_B = (Bf)' Bg = a' \mathbf{A}' \mathbf{A} b.$$

Define a function $k_1(t, s)$ by

$$k_1(t, s) = \boldsymbol{\xi}(t)' \left(\mathbf{A}' \mathbf{A} \right)^{-1} u(s).$$

It is now easy to verify that, for any $f = \sum_i a_i \xi_i$,

$$\langle k_1(t, \cdot), f \rangle_B = \boldsymbol{\xi}(t)' \left(\mathbf{A}' \mathbf{A} \right)^{-1} \mathbf{A}' \mathbf{A} a = \boldsymbol{\xi}(t)' a = a' \boldsymbol{\xi}(t) = f(t).$$

So k_1 is the reproducing kernel for the space ker L equipped with the B inner product.

The reproducing kernel for a larger space

Finally, we consider the space of more general functions x equipped with the inner product

$$\langle x, y \rangle_{B,L} = (Bx)'(By) + \int (Lx)(t)(Ly)(t)dt$$

It is easy to check from the properties we have set out that the reproducing kernel in this space is given by

$$k(s,t) = k_1(s,t) + k_2(s,t).$$

More general roughness penalties

We propose using the criterion

$$\text{PENSSE}(x) = \sum_{j}^{n} [y_j - x(t_j)]^2 + \lambda \times \text{PEN}_L(x)$$

where

$$\text{PEN}_L(x) = \int (Lx)^2(t)dt$$

Next, we give a theorem that states that the optimal basis for spline smoothing in the context of operators (B, L) is defined by the reproducing kernel k_2.

Optimal Basis Theorem

Theorem 6

For any $\lambda > 0$, the function x minimizing the spline smoothing criterion (21.2) defined by a linear differential operator L of order m has the expansion

$$x(t) = \sum_{j=1}^{m} d_j \xi_j(t) + \sum_{i=1}^{n} c_i k_2(t_i, t).$$

It can be put a bit more compactly. Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)'$; define another vector function

$$\tilde{k}(t) = \{k_2(t_1, t), k_2(t_2, t), \ldots, k_2(t_n, t)\}'.$$

Then the optimal basis theorem says that the function x has to be of the form $x = \mathbf{d}'\boldsymbol{\xi} + \mathbf{c}'\tilde{k}$, where $\mathbf{d}$ is a vector of m coefficients d_j and $\mathbf{c}$ is the corresponding vector of n coefficients c_i.

Proof of Theorem 6

Suppose x^* is any function having square-integrable derivatives up to order m. The strategy for the proof is to construct a function $\tilde{x}$ of the form in theorem 5 such that

$$\text{PENSSE}(\tilde{x}) \leq \text{PENSSE}\left(x^*\right)$$

with equality only if $\tilde{x} = x^*$.

- First of all, write $x^* = u^* + e^*$ where $u^* \in \ker L$ and $e^* \in \ker B$. Let $\mathcal{K}$ be the subspace of $\ker B$ spanned by the n functions $k_2\left(t_i, \cdot\right)$, and let $\tilde{e}$ be the projection of e^* onto $\mathcal{K}$ in the L-inner product. This means that $e^* = \tilde{e} + e^{\perp}$, where

$$\tilde{e} = c' \tilde{k}$$

  for some vector c, and the residual $e^{\perp}$ in $\ker B$ satisfies the orthogonality relation

$$\left\langle e, e^{\perp} \right\rangle_L = \int (Le)\left(Le^{\perp}\right) = 0 \text{ for all } e \text{ in } \mathcal{K}.$$

Proof of Theorem 6 (Cont'd)

We now define our function $\tilde{x} = u^* + \tilde{e}$, meaning that $\tilde{x}$ is necessarily of the required form (21.5), and $x^* - \tilde{x}$ is equal to the residual $e^\perp$.

- To show that $\text{PENSSE}(\tilde{x}) \leq \text{PENSSE}\left(x^*\right)$, note first that, by the defining property of the reproducing kernel, for each i,

$$x^*\left(t_i\right) - \tilde{x}\left(t_i\right) = e^\perp\left(t_i\right) = \left\langle k_2\left(t_i, \cdot\right), e^\perp\right\rangle_L = 0$$

- Since $Lx^* = Le^*$ and $L\tilde{x} = L\tilde{e}$, we have

$$\begin{aligned} \text{PEN}_L\left(x^*\right) - \text{PEN}_L(\tilde{x}) &= \text{PEN}_L\left(e^*\right) - \text{PEN}_L(\tilde{e}) \\ &= \left\langle \tilde{e} + e^\perp, \tilde{e} + e^\perp \right\rangle_L - \langle \tilde{e}, \tilde{e} \rangle_L \\ &= \left\langle e^\perp, e^\perp \right\rangle_L + 2\left\langle \tilde{e}, e^\perp \right\rangle_L = \left\langle e^\perp, e^\perp \right\rangle_L \end{aligned}$$

- Consequently $\text{PENSSE}\left(x^*\right) \geq \text{PENSSE}(\tilde{x})$. Equality holds only if $e^\perp \in \ker L$; since we already know that $e^\perp \in \ker B$, this implies that $e^\perp = 0$ and that $x^* = \tilde{x}$. This completes the proof.

Solution of L-spline

Since we know that the required function is of the form $x = d'u + c'\tilde{k}$, we need only express PENSSE(x) in terms of c and d and minimize to find the best values of c and d.

Let $\mathbf{K}$ be the matrix with values $k_2(t_i, t_j)$. From equation (20.14) it follows that

$$\mathrm{PEN}_L(x) = \left\langle c'\tilde{k}, c'\tilde{k} \right\rangle_L = c'\mathbf{K}c.$$

The vector of values $x(t_i)$ is $\mathbf{U}d + \mathbf{K}c$, where $\mathbf{U}$ is the matrix with values $\xi_j(t_i)$. Hence, at least in principle, we can find x by minimizing the quadratic form

$$\mathrm{PENSSE}(x) = (y - \mathbf{U}d - \mathbf{K}c)'(y - \mathbf{U}d - \mathbf{K}c) + \lambda c'\mathbf{K}c$$

to find the vectors c and d.

Calculation of coefficients

Differentiating

$$\text{PENSSE}(x) = (y - \mathbf{U}d - \mathbf{K}c)'(y - \mathbf{U}d - \mathbf{K}c) + \lambda c'\mathbf{K}c$$

with respect to c and d and setting the derivatives to 0, one gets

$$\begin{aligned} \mathbf{K}\{(\mathbf{K} + \lambda I)c + \mathbf{U}d - y\} &= 0 \\ \mathbf{U}'\{\mathbf{K}c + \mathbf{U}d - y\} &= 0 \end{aligned}$$

Unfortunately the matrix **K** is in practice usually extremely badly conditioned (the ratio of its largest eigenvalue to its smallest explodes). The computations required to minimize the quadratic form are likely to be unstable or impossible.

Calculation of coefficients

From Theorem 2.9 in Chong Gu (2012), the minimizer x uniquely exists as long as $\mathbf{U}$ to be of full column rank. When $\mathbf{K}$ is singular, there may have multiple solutions for c and d, all that satisfy

$$\begin{aligned} \mathbf{K}\{(\mathbf{K}+\lambda I)c+\mathbf{U}d-y\} &= 0 \\ \mathbf{U}'\{\mathbf{K}c+\mathbf{U}d-y\} &= 0 \end{aligned}$$

However, all the solutions yield the same function estimate

$$x(t)=\sum_{i=1}^{m} d_j\xi_j(t)+\sum_{i=1}^{n} c_i k_2\left(t_i, t\right)$$

For definiteness in the numerical calculation, we shall compute a particular solution by solving the linear system

$$\begin{aligned} (\mathbf{K}+\lambda I)c+\mathbf{U}d &= y \\ \mathbf{U}'\mathbf{c} &= 0 \end{aligned}$$

Calculation of coefficients

Suppose **U** is of full column rank. Let

$$\mathbf{U} = FR^* = (F_1, F_2)\begin{pmatrix} \tilde{R} \\ O \end{pmatrix} = F_1\tilde{R}$$

be the QR-decomposition of **U** with F orthogonal and $\tilde{R}$ upper-triangular.

From $\mathbf{U}'c = 0$, one has $F_1^T c = 0$, so $c = F_2F_2'c$. Simple algebra leads to

$$\begin{aligned} c &= F_2\left(F_2'\mathbf{K}F_2 + \lambda I\right)^{-1}F_2'y \\ d &= \tilde{R}^{-1}\left(F_1'y - F_1'\mathbf{K}c\right) \end{aligned}$$

Calculation of coefficients

Some algebra yields

$$
\begin{aligned}
\hat{y} &= \mathbf{K}c + \mathbf{U}d \\
&= \left(F_1F_1' + F_2F_2'\mathbf{K}F_2\left(F_2'\mathbf{K}F_2 + \lambda I\right)^{-1}F_2'\right)y \\
&= \left(I - F_2\left(I - F_2'\mathbf{K}F_2\left(F_2'\mathbf{K}F_2 + \lambda I\right)^{-1}\right)F_2'\right)y \\
&= \left(I - \lambda F_2\left(F_2'\mathbf{K}F_2 + \lambda I\right)^{-1}F_2'\right)y.
\end{aligned}
$$

The need for a good algorithm:

- In smoothing long sequences of observations, it is critical to devise a smoothing procedure that requires only $O(n)$ operations.
- The algorithm is a natural extension of Reinsch algorithm, which apply to the cubic polynomial smoothing case ($L = D^2$)

Requirements and phases of the algorithm

- The algorithm requires the computation of values of two types of function (user-supplied):
  - $\xi_j, j = 1, \ldots, m$: a set of m linearly independent functions satisfying $L\xi_j = 0$, that is, spanning ker L. As before, we refer to these collectively as the vector-valued function $\boldsymbol{\xi}$.
  - k_2 : the reproducing kernel function for the subspace of functions e satisfying $B_I e = 0$, where B_I is the initial value constraint operator.
- The algorithm splits into three phases:
  1. an initial setup phase that does not depend on the smoothing parameter
  2. a smoothing phase in which we smooth the data
  3. a summary phase in which we compute performance measures for the smooth

Initial setup phase

- In the initial phase, we define two symmetric $(n-m) \times (n-m)$ band-structured matrices $\mathbf{H}$ and $\mathbf{C}'\mathbf{C}$ where m is the order of L.
- For each $i = 1, \ldots, n-m$, define the $(m+1) \times m$ matrix $\mathbf{U}^{(i)}$ to have (l, j) element $\xi_j(t_{i+l})$, for $l = 0, \ldots, m$. Thus $\mathbf{U}^{(i)}$ is the submatrix of $\mathbf{U}$ consisting only of rows $i, i+1, \ldots, i+m$. Find the QR decomposition
$$\mathbf{U}^{(i)} = \mathbf{Q}^{(i)}\mathbf{R}^{(i)}$$
where the matrix $\mathbf{Q}^{(i)}$ is square, of order $m+1$, and orthonormal, and where the matrix $\mathbf{R}^{(i)}$ is $(m+1) \times m$ and upper triangular. Let the vector $c^{(i)}$ be the last column of $\mathbf{Q}^{(i)}$; this vector is orthogonal to all the columns of $\mathbf{U}^{(i)}$.

Initial setup phase

- Now define the $n \times (n-m)$ matrix $\mathbf{C}$ so that its i th column has the $m+1$ values $c^{(i)}$ starting in row i; elsewhere the matrix contains zeroes. The band structure of $\mathbf{C}$ immediately implies that $\mathbf{C}'\mathbf{C}$ has the required band structure, and can be found in $O(n)$ operations for fixed m.
- The other setup-phase matrix $\mathbf{H}$ is the $(n-m) \times (n-m)$ symmetric matrix

$$\mathbf{H} = \mathbf{C}'\mathbf{K}\mathbf{C}$$

where $\mathbf{K}$ is the matrix of values $k_2(t_i, t_j)$. It turns out that $\mathbf{H}$ is also band-structured with band width $2m-1$.

Smoothing phase

The actual smoothing consists of two steps:

1. Compute the vector z, of length $n-m$, that solves

$$\left(\mathbf{H}+\lambda\mathbf{C}'\mathbf{C}\right)z=\mathbf{C}'y,$$

   where the vector y contains the values to be smoothed.

2. Compute the vector of n values $\hat{y}_i = x\left(t_i\right)$ of the smoothing function x at the n argument values using

$$\hat{y}=y-\lambda\mathbf{C}z.$$

Because of the band structure of $\left(\mathbf{H}+\lambda\mathbf{C}'\mathbf{C}\right)$ and of $\mathbf{C}$, both of these steps can be computed in $O(n)$ operators.

Performance assessment phase

The vector of smoothed values $\hat{y}$ and the original values y that were smoothed are related as follows:

$$\hat{y} = y - \lambda \mathbf{C} \left(\mathbf{H} + \lambda \mathbf{C}'\mathbf{C} \right)^{-1} \mathbf{C}' y$$

The matrix $\mathbf{S}$ defined by

$$\mathbf{S} = \mathbf{I} - \lambda \mathbf{C} \left(\mathbf{H} + \lambda \mathbf{C}'\mathbf{C} \right)^{-1} \mathbf{C}'$$

is often called the hat matrix, and in effect defines a linear transformation that maps the unsmoothed data into its smooth image by

$$\hat{y} = \mathbf{S} y$$

Performance assessment phase

Various measures of performance depend on the diagonal values in $\mathbf{S}$. Of these, the most popular are currently

$$\mathrm{GCV} = \mathrm{SSE} / \left(1 - n^{-1} \operatorname{trace} \mathbf{S}\right)^2 ,$$

where

$$\mathrm{SSE} = \sum_{i=1}^{n} \left[y_i - x\left(t_i\right)\right]^2 = \|y - \hat{y}\|^2$$

and

$$\mathrm{CV} = \sum_{i=1}^{n} \left[\left\{y_i - x\left(t_i\right)\right\} / \left\{1 - s_{ii}\right\}\right]^2$$

where s_{ii} is the i th diagonal entry of $\mathbf{S}$. We can compute both measures GCV and CV in $O(n)$ operations given the band-structured nature of the matrices defining $\mathbf{S}$.

Table of Contents

References

- P.J.Green & B.W.Silverman (1995), **Nonparametric Regression and Generalized Linear Models**, Chapman & Hall (Sections 1-3)
- J.O.Ramsay & B.W.Silverman (2005), **Functional Data Analysis**, Springer (Sections 18-21)
- Chong Gu (2012), **Smoothing Spline ANOVA Models**, Springer (Sections 2-3)
- T.Hastie, R.Tibshirani & J.Friedman (2017), **The Elements of Statistical Learning**, Springer (Section 5)