$p=1$ Smoothing Splines

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

Function optimization problem

A way of avoiding the direct selection of knots for a regression spline is to instead, for a smoothing parameter $\lambda > 0$, consider the problem of finding (for $a \leq \min\{x_i\}$ and $\max\{x_i\} \leq b$)

$$\hat{f}_\lambda = \underset{\text{functions } h \text{ with 2 derivatives}}{\arg\min} \left(\sum_{i=1}^{N} (y_i - h(x_i))^2 + \lambda \int_a^b (h''(x))^2 \, dx \right)$$

In a surprising piece of mathematics, it turns out that this seemingly abstract problem has a tractable solution.

Natural cubic spline solution

As it turns out, $\hat{f}_\lambda$ is a natural cubic spline with knots at the distinct values x_i in the training set. That is, for a set of (now data-dependent, as the knots come from the training data) basis functions for such splines

$$h_1, h_2, \ldots, h_N$$

(here we're tacitly assuming that the N values of the input variable in the training set are all different)

$$\hat{f}_\lambda(x) = \sum_{j=1}^{N} \widehat{\beta}_{\lambda j} h_j(x)$$

where the $\widehat{\beta}_{\lambda j}$ are yet to be identified.

Development of the coefficient vector

For

$$g(x) = \sum_{j=1}^{N} \theta_j h_j(x) \tag{1}$$

it is the case that

$$\left(g''(x)\right)^2 = \sum_{j=1}^{N} \sum_{l=1}^{N} \theta_j \theta_l h_j''(x) h_l''(x)$$

So for $\boldsymbol{\theta}' = (\theta_1, \theta_2, \ldots, \theta_N)$ and

$$\underset{N \times N}{\boldsymbol{\Omega}} = \left(\int_a^b h_j''(t) h_l''(t) \, dt \right)$$

we then have that

$$\int_a^b \left(g''(x)\right)^2 dx = \boldsymbol{\theta}' \boldsymbol{\Omega} \boldsymbol{\theta}$$

(Since for every $\boldsymbol{\theta}$ this is non-negative, $\boldsymbol{\Omega}$ is non-negative definite.)

Development of the coefficient vector cont.

Then with the notation

$$\underset{N \times N}{\mathbf{H}} = \left(h_j \left(x_i \right) \right)$$

(i indexing rows and j indexing columns) the criterion to be optimized to find $\hat{f}_\lambda$ is for functions of the form (1)

$$\left(\mathbf{Y} - \mathbf{H}\theta \right)' \left(\mathbf{Y} - \mathbf{H}\theta \right) + \lambda \theta' \mathbf{\Omega} \theta$$

and some vector calculus shows that the optimizing θ is

$$\widehat{\beta}_\lambda = \left(\mathbf{H}'\mathbf{H} + \lambda \mathbf{\Omega} \right)^{-1} \mathbf{H}'\mathbf{Y} \qquad (2)$$

a kind of "generalized ridge regression" coefficient vector.

Fitted values and smoother matrix

Corresponding to (2) is a vector of smoothed output values

$$\widehat{\mathbf{Y}}_\lambda = \mathbf{H}\left(\mathbf{H}'\mathbf{H} + \lambda\Omega\right)^{-1}\mathbf{H}'\mathbf{Y}$$

and the matrix

$$\mathbf{S}_\lambda \equiv \mathbf{H}\left(\mathbf{H}'\mathbf{H} + \lambda\Omega\right)^{-1}\mathbf{H}'$$

is called a **smoother matrix**. As it turns out, $\mathbf{S}_\lambda$ (is non-negative definite symmetric of rank N and) has the property that

$$\mathbf{S}_\lambda\mathbf{S}_\lambda \preceq \mathbf{S}_\lambda$$

meaning that $\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda$ is non-negative definite.

Contrast with OLS

Consider a case where some fairly small number, p, of *fixed* basis functions are employed in a regression context. That is, for basis functions $b_1, b_2, \ldots, b_p$ suppose

$$\underset{N \times p}{\mathbf{B}} = \left(b_j\left(x_i\right)\right)$$

OLS produces the vector of fitted values

$$\widehat{\mathbf{Y}} = \mathbf{B}\left(\mathbf{B}'\mathbf{B}\right)^{-1}\mathbf{B}'\mathbf{Y}$$

and the projection matrix onto the column space of $\mathbf{B}$, $C\left(\mathbf{B}\right)$, is $\mathbf{P_B} = \mathbf{B}\left(\mathbf{B}'\mathbf{B}\right)^{-1}\mathbf{B}'$. $\mathbf{P_B}$ is (non-negative definite symmetric of rank p) and has the property that

$$\mathbf{P_B}\mathbf{P_B} = \mathbf{P_B}$$

i.e. $\mathbf{P_B}$ is idempotent.

Effective df and the Reinsch form

In analogy to the ridge regression case, one might define effective degrees of freedom for $\mathbf{S}_\lambda$ by

$$\mathrm{df}(\lambda) \equiv \mathrm{tr}(\mathbf{S}_\lambda) \quad (3)$$

and we proceed to develop motivation and a formula for this quantity.

For

$$\mathbf{K} = (\mathbf{H}')^{-1} \Omega \mathbf{H}^{-1}$$

it is the case that

$$\begin{aligned} \mathbf{S}_\lambda &= \mathbf{H}(\mathbf{H}'\mathbf{H} + \lambda\Omega)^{-1}\mathbf{H}' \\ &= \mathbf{H}\left(\mathbf{H}'\left(\mathbf{I} + \lambda\mathbf{H}'^{-1}\Omega\mathbf{H}\right)\mathbf{H}\right)^{-1}\mathbf{H}' \\ &= \mathbf{H}\mathbf{H}^{-1}\left(\mathbf{I} + \lambda\mathbf{H}'^{-1}\Omega\mathbf{H}\right)^{-1}\mathbf{H}'^{-1}\mathbf{H}' \\ &= (\mathbf{I} + \lambda\mathbf{K})^{-1} \end{aligned}$$

This is the so-called **Reinsch form** for $\mathbf{S}_\lambda$, from whence $\mathbf{S}_\lambda^{-1} = \mathbf{I} + \lambda\mathbf{K}$.

Penalty interpretation of the K matrix

Some vector calculus shows that $\widehat{\mathbf{Y}}_\lambda = \mathbf{S}_\lambda \mathbf{Y}$ is a solution to the minimization problem

$$\underset{\mathbf{v}\in\Re^N}{\text{minimize}} \left((\mathbf{Y}-\mathbf{v})' (\mathbf{Y}-\mathbf{v}) + \lambda \mathbf{v}'\mathbf{K}\mathbf{v} \right) \tag{4}$$

so that the matrix $\mathbf{K}$ $\left(= \left(\mathbf{H}'\right)^{-1} \Omega \mathbf{H}^{-1}\right)$ can be thought of as defining a "penalty" in fitting a smoothed version of $\mathbf{Y}$. (There is more on this to come.)

Eigen decomposition of the smoother matrix

Then, since $\mathbf{S}_\lambda$ is symmetric non-negative definite, it has an eigen decomposition as

$$\mathbf{S}_\lambda = \mathbf{U}\mathbf{D}\mathbf{U}' = \sum_{j=1}^{N} d_j \mathbf{u}_j \mathbf{u}_j' \qquad (5)$$

where columns of $\mathbf{U}$ (the eigenvectors $\mathbf{u}_j$) comprise an orthonormal basis for $\Re^N$ and

$$\mathbf{D} = \mathbf{diag}\left(d_1, d_2, \ldots, d_N\right)$$

for eigenvalues of $\mathbf{S}_\lambda$

$$d_1 \geq d_2 \geq \cdots \geq d_N > 0$$

It turns out to be guaranteed that $d_1 = d_2 = 1$.

Eigen decompositions of the smoother matrix and $\boldsymbol{K}$

An eigenvalue for $\mathbf{K}$, say η, solves

$$\det(\mathbf{K} - \eta\mathbf{I}) = 0$$

Now

$$\det(\mathbf{K} - \eta\mathbf{I}) = \det\left(\frac{1}{\lambda}\left[(\mathbf{I} + \lambda\mathbf{K}) - (1 + \lambda\eta)\,\mathbf{I}\right]\right)$$

So $1 + \lambda\eta$ must be an eigenvalue of $\mathbf{I} + \lambda\mathbf{K}$ and $1/(1 + \lambda\eta)$ must be an eigenvalue of $\mathbf{S}_\lambda = (\mathbf{I} + \lambda\mathbf{K})^{-1}$. So for some j we must have

$$d_j = \frac{1}{1 + \lambda\eta}$$

Eigen decompositions and df

Observing that $1/\left(1+\lambda\eta\right)$ is decreasing in η, we may conclude that

$$d_j = \frac{1}{1+\lambda\eta_{N-j+1}} \tag{6}$$

for

$$\eta_1 \geq \eta_2 \geq \cdots \geq \eta_{N-2} \geq \eta_{N-1} = \eta_N = 0$$

the eigenvalues of $\mathbf{K}$ (that themselves *do not* depend upon λ). So in light of (3), (5), and (6), the smoothing effective degrees of freedom, $\mathrm{df}(\lambda)$, are

$$\mathrm{tr}\left(\mathbf{S}_\lambda\right) = \sum_{j=1}^{N} d_j = 2 + \sum_{j=1}^{N-2} \frac{1}{1+\lambda\eta_j}$$

which is clearly decreasing in λ (with minimum value 2 in light of the fact that $\mathbf{S}_\lambda$ has two eigenvalues that are 1).

Eigen vectors of the smoother matrix and $\boldsymbol{K}$

Further, consider $\mathbf{u}_j$, the eigenvector of $\mathbf{S}_\lambda$ corresponding to eigenvalue d_j. $\mathbf{S}_\lambda \mathbf{u}_j = d_j \mathbf{u}_j$ so that

$$\mathbf{u}_j = \mathbf{S}_\lambda^{-1} d_j \mathbf{u}_j = (\mathbf{I} + \lambda \mathbf{K})\, d_j \mathbf{u}_j$$

so that

$$\mathbf{u}_j = d_j \mathbf{u}_j + d_j \lambda \mathbf{K} \mathbf{u}_j$$

and thus

$$\mathbf{K}\mathbf{u}_j = \left(\frac{1 - d_j}{\lambda d_j} \right) \mathbf{u}_j = \eta_{N-j+1} \mathbf{u}_j$$

That is, $\mathbf{u}_j$ is an eigenvector of $\mathbf{K}$ corresponding to the $(N - j + 1)$st largest eigenvalue. That is, for all λ the eigenvectors of $\mathbf{S}_\lambda$ are eigenvectors of $\mathbf{K}$ and *thus do not depend upon* λ.

Shrinking of the prediction vector

Then, for any λ

$$\widehat{\mathbf{Y}}_{\lambda} = \mathbf{S}_{\lambda}\mathbf{Y} = \sum_{j=1}^{N} d_j \langle \mathbf{u}_j, \mathbf{Y} \rangle \mathbf{u}_j$$

$$= \langle \mathbf{u}_1, \mathbf{Y} \rangle \mathbf{u}_1 + \langle \mathbf{u}_2, \mathbf{Y} \rangle \mathbf{u}_2 + \sum_{j=3}^{N} \frac{\langle \mathbf{u}_j, \mathbf{Y} \rangle}{1 + \lambda \eta_{N-j+1}} \mathbf{u}_j \qquad (7)$$

and $\widehat{\mathbf{Y}}_{\lambda}$ is a shrunken version of $\mathbf{Y}$ that progresses from $\mathbf{Y}$ to the projection of $\mathbf{Y}$ onto the span of $\{\mathbf{u}_1, \mathbf{u}_2\}$ as λ runs from 0 to ∞. (It is possible to argue that the span of $\{\mathbf{u}_1, \mathbf{u}_2\}$ is the set of vectors of the form $c\mathbf{1} + d\mathbf{x}$, as is consistent with the original function optimization objective function.) The larger is λ, the more severe the shrinking overall. Further, the larger is j, the smaller is d_j and the more severe is the shrinking of $\mathbf{Y}$ in the $\mathbf{u}_j$ direction. (The unpenalized directions $\mathbf{u}_1$ and $\mathbf{u}_2$ have no associated shrinking.)

Shrinking of prediction and coefficient vectors

In the context of cubic smoothing splines, large j correspond to "wiggly" (as a functions of coordinate i or value of the input x_i) $\mathbf{u}_j$, and the prescription (7) calls for suppression of "wiggly" components of $\mathbf{Y}$.

Further, since $\widehat{\mathbf{Y}}_\lambda = \mathbf{H}\widehat{\boldsymbol{\beta}}_\lambda$ and $\mathbf{H}$ is nonsingular, as λ runs from 0 to ∞, $\widehat{\boldsymbol{\beta}}_\lambda$ runs from $\mathbf{H}^{-1}\mathbf{Y}$ to $\mathbf{H}^{-1}\left(\langle \mathbf{u}_1, \mathbf{Y}\rangle \mathbf{u}_1 + \langle \mathbf{u}_2, \mathbf{Y}\rangle \mathbf{u}_2\right)$. And there is "shrinking" enforced on $\widehat{\boldsymbol{\beta}}_\lambda$ in the sense that the quadratic form $\widehat{\boldsymbol{\beta}}_\lambda' \boldsymbol{\Omega} \widehat{\boldsymbol{\beta}}_\lambda$ must be non-increasing in λ. (If not, the fact that $\left\|\mathbf{Y} - \widehat{\mathbf{Y}}_\lambda\right\|^2$ increases in λ would produce a contradiction.)

Eigen decomposition of *K* and penalization

Now large j (indexing *late/small* eigenvalues of $\mathbf{S}_\lambda$) correspond to *early/large* eigenvalues of the smoothing spline penalty matrix $\mathbf{K}$. Letting $\mathbf{u}_j^* = \mathbf{u}_{N-j+1}$ so that

$$\mathbf{U}^* = (\mathbf{u}_N, \mathbf{u}_{N-1}, \ldots, \mathbf{u}_1) = (\mathbf{u}_1^*, \mathbf{u}_2^*, \ldots, \mathbf{u}_N^*)$$

the eigen decomposition of $\mathbf{K}$ is

$$\mathbf{K} = \mathbf{U}^* \mathbf{diag}\left(\eta_1, \eta_2, \ldots, \eta_N\right) \mathbf{U}^{*\prime}$$

and the criterion

$$\underset{\mathbf{v} \in \Re^N}{\text{minimize}} \left((\mathbf{Y} - \mathbf{v})' (\mathbf{Y} - \mathbf{v}) + \lambda \mathbf{v}' \mathbf{K} \mathbf{v} \right)$$

can be written as

$$\underset{\mathbf{v} \in \Re^N}{\text{minimize}} \left((\mathbf{Y} - \mathbf{v})' (\mathbf{Y} - \mathbf{v}) + \lambda \mathbf{v}' \mathbf{U}^* \mathbf{diag}\left(\eta_1, \eta_2, \ldots, \eta_N\right) \mathbf{U}^{*\prime} \mathbf{v} \right)$$

Eigen decomposition and penalization cont.

This criterion is then

$$\underset{\mathbf{v} \in \Re^N}{\text{minimize}} \left((\mathbf{Y} - \mathbf{v})' (\mathbf{Y} - \mathbf{v}) + \lambda \sum_{j=1}^{N-2} \eta_j \langle \mathbf{u}_j^*, \mathbf{v} \rangle^2 \right) \tag{8}$$

(since $\eta_{N-1} = \eta_N = 0$) and we see that eigenvalues of $\mathbf{K}$ function as penalty coefficients applied to the N orthogonal components of $\mathbf{v} = \sum_{j=1}^N \left\langle \mathbf{u}_j^*, \mathbf{v} \right\rangle \mathbf{u}_j^*$ in the choice of optimizing $\mathbf{v}$. From this point of view, the $\mathbf{u}_j$ (or $\mathbf{u}_j^*$) provide the natural alternative (to the columns of $\mathbf{H}$) basis (for $\Re^N$) for representing or approximating $\mathbf{Y}$, and

$$\widehat{\mathbf{Y}}_\lambda = \langle \mathbf{u}_1, \mathbf{Y} \rangle \mathbf{u}_1 + \langle \mathbf{u}_2, \mathbf{Y} \rangle \mathbf{u}_2 + \sum_{j=3}^{N} \frac{\langle \mathbf{u}_j, \mathbf{Y} \rangle}{1 + \lambda \eta_{N-j+1}} \mathbf{u}_j$$

provides an explicit form for the optimizing smoothed vector of responses.

Orthonormal bases and penalization

Here $\mathbf{K}$ has a specific meaning derived from the $\mathbf{H}$ and $\boldsymbol{\Omega}$ matrices connected specifically with smoothing splines *and* the particular values of x in the training data set. But an interesting possibility brought up by the development is that of forgetting the origins (from $\mathbf{K}$) of the η_j and $\mathbf{u}_j$ and beginning with any interesting/intuitively appealing orthonormal basis $\{\mathbf{u}_j\}$ and set of non-negative penalties $\{\eta_j\}$ for use in (8). Working backwards one is then led to a corresponding smoothed vector of responses and its "smoothing matrix". Slightly more detail on this line of argument is provided in Section 5.3.

Equivalent kernels

It is worth remarking that since $\widehat{\mathbf{Y}}_\lambda = \mathbf{S}_\lambda \mathbf{Y}$, the rows of $\mathbf{S}_\lambda$ provide weights to be applied to the elements of $\mathbf{Y}$ to produce predictions/ smoothed values corresponding to $\mathbf{Y}$. These can for each i be thought of as defining a corresponding "equivalent kernel" (for an appropriate "kernel-weighted average" of the training output values). (See Figure 5.8 of HTF in this regard.)