CHAPTER 10
Additive and Interaction Splines

10.1 Variational problems with multiple smoothing parameters.

Let $\tilde{\mathcal{H}}$ be an r.k. space and let $\mathcal{H}$ be a (possibly proper) subspace of $\tilde{\mathcal{H}}$ of the form

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$$

where $\mathcal{H}_0$ is span $\{\phi_1, \ldots, \phi_M\}$ and $\mathcal{H}_1$ is the direct sum of p orthogonal subspaces $\mathcal{H}^1, \ldots, \mathcal{H}^p$,

$$\mathcal{H}_1 = \sum_{\beta=1}^{p} \oplus \mathcal{H}^\beta. \tag{10.1.1}$$

Suppose we wish to find $f \in \mathcal{H}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - L_i f)^2 + \lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} \|P^\beta f\|^2 \tag{10.1.2}$$

where P^β is the orthogonal projection in $\tilde{\mathcal{H}}$ onto $\mathcal{H}^\beta$ and $\theta_\beta \geq 0$. If $\theta_\beta = 0$, then the minimizer of (10.1.2) is taken to satisfy $\|P^\beta f\|^2 = 0$.

We can find the minimizer of (10.1.2) using the results of Chapter 1 by making some substitutions. Let the r.k. for $\mathcal{H}^\beta$ be $R_\beta(t, t')$. Then the r.k. for $\mathcal{H}_1$ with the squared norm $\|P_1 f\|^2_{\tilde{\mathcal{H}}} = \sum_{\beta=1}^{p} \|P^\beta f\|^2_{\tilde{\mathcal{H}}}$ is $R^1(t, t') = \sum_{\beta=1}^{p} R_\beta(t, t')$; this follows since the r.k. for a direct sum of orthogonal subspaces is the sum of the individual r.k.'s (see Aronszajn (1950)). If we change the squared norm on $\mathcal{H}_1$ from $\sum_{\beta=1}^{p} \|P^\beta f\|^2_{\tilde{\mathcal{H}}}$ to $\sum_{\beta=1}^{p} \theta_\beta^{-1} \|P^\beta f\|^2_{\tilde{\mathcal{H}}}$, then the r.k. for $\mathcal{H}_1$ changes from $\sum_{\beta=1}^{p} R_\beta(t, t')$ to $\sum_{\beta=1}^{p} \theta_\beta R_\beta(t, t')$. Using these facts and making the relevant substitutions in Chapter 1, it can be seen that the minimizer of (10.1.2) is of the form (1.3.8) with

$$\xi_i \equiv \sum_{\beta=1}^{p} P^\beta \xi_i$$

replaced everywhere by

$$\xi_i^\theta = \sum_{\beta=1}^{p} \theta_\beta P^\beta \xi_i \tag{10.1.3}$$

and that Σ in (1.3.9) is of the form

$$\Sigma = \theta_1\Sigma_1 + \theta_2\Sigma_2 + \ldots + \theta_p\Sigma_p, \tag{10.1.4}$$

where the ijth entry of Σ_β is

$$< P^\beta\xi_i, P^\beta\xi_j >_{\tilde{\mathcal{H}}} = L_{i(s)}\ L_{j(t)}R_\beta(s,t).$$

The minimizer of (10.1.2) is then given by

$$f_{\lambda,\theta} = \sum_{\nu=1}^{M} d_\nu\phi_\nu + \sum_{i=1}^{n} c_i \sum_{\beta=1}^{p} \theta_\beta P^\beta\xi_i \tag{10.1.5}$$

where the coefficient vectors c and d satisfy (1.3.16) and (1.3.17) with Σ given by (10.1.4). $I - A(\lambda)$ of (1.3.23) then becomes

$$I - A(\lambda,\theta) = n\lambda Q_2(\theta_1 Q_2'\Sigma_1 Q_2 + \ldots + \theta_p Q_2'\Sigma_p Q_2 + n\lambda I)^{-1}Q_2', \tag{10.1.6}$$

where $T'Q_2 = 0_{M\times(n-M)}$ as before, so that the GCV function $V(\lambda) = V(\lambda,\theta)$ of (4.3.1) becomes

$$V(\lambda,\theta) = \frac{z'(\theta_1\tilde{\Sigma}_1 + \ldots + \theta_p\tilde{\Sigma}_p + n\lambda I)^{-2}z}{[\mathrm{tr}(\theta_1\tilde{\Sigma}_1 + \ldots + \theta_p\tilde{\Sigma}_p + n\lambda I)^{-1}]^2} \tag{10.1.7}$$

where

$$z = Q_2'y,\ \ \tilde{\Sigma}_\beta = Q_2'\Sigma_\beta Q_2.$$

The problem of choosing λ and $\theta = (\theta_1,\ldots,\theta_p)'$ by GCV is then the problem of choosing λ and θ to minimize (10.1.7), where, of course, any (λ,θ) with the same values of $\lambda_r = \lambda/\theta_r$, $r = 1,\ldots,p$ are equivalent. Numerical algorithms for doing this will be discussed in Section 11.3. For future reference we note that

$$P^\beta f_{\lambda,\theta} = \theta_\beta \sum_{i=1}^{n} c_i P^\beta\xi_i \tag{10.1.8}$$

and so

$$\|P^\beta f_{\lambda,\theta}\|_{\tilde{\mathcal{H}}}^2 = \theta_\beta^2 c'\Sigma_\beta c \tag{10.1.9}$$

and

$$\sum_{i=1}^{n}(L_i P^\beta f_{\lambda,\theta})^2 = \theta_\beta^2\|\Sigma_\beta c\|^2. \tag{10.1.10}$$

Quantities (10.1.9) and (10.1.10) can be used to decide whether the βth components are significant.

10.2 Additive and interaction smoothing splines.

The additive (smoothing) splines and their generalizations, the interaction (smoothing) splines, can be put in the framework of Section 10.1, where $\tilde{\mathcal{H}}$ is the tensor product space $\otimes^d W_m$.

The additive splines are functions of d variables, which are a sum of d functions of one variable (main effects splines)

$$f(x_1, \ldots, x_d) = f_0 + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha),$$

the two factor interaction splines are of the form

$$f(x_1, \ldots, x_d) = f_0 + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, x_\beta),$$

and so forth, where certain side conditions on the f_α's, $f_{\alpha\beta}$'s etc., that guarantee uniqueness must hold. The additive spline models have become popular in the analysis of medical data and other contexts (see Stone (1985, 1986), Burman (1985), Friedman, Grosse, and Stuetzle, (1983), Hastie and Tibshirani (1986), Buja, Hastie, and Tibshirani (1989), and references cited therein). The interaction spline models have been discussed by Barry (1983,1986), Wahba (1986), Gu et al. (1989), Gu and Wahba (1991), and Chen (1987, 1989). These models, which in a sense generalize analysis of variance to function spaces, have strong potential for the empirical modeling of responses to economic and medical variables, given large data sets of responses with several independent variables, and represent a major advance over the usual multivariate parametric (mostly linear) models. They represent a nonparametric compromise in an attempt to overcome the "curse of dimensionality," since estimating a more general function $f(x_1, \ldots, x_d)$ will require truly large data sets for even moderate d.

To describe these splines, it will be convenient to endow $W_m[0,1]$ with a norm slightly different from the one given in Section 1.2.

Let

$$M_\nu f = \int_0^1 f^{(\nu)}(x)\,dx, \quad \nu = 0, 1, \ldots, m-1 \tag{10.2.1}$$

and note that

$$M_\nu f = f^{(\nu-1)}(1) - f^{(\nu-1)}(0), \quad \nu = 1, \ldots, m-1.$$

Let

$$\|f\|^2_{W_m} = \sum_{\nu=0}^{m-1} (M_\nu f)^2 + \int_0^1 (f^{(m)}(u))^2 du. \tag{10.2.2}$$

Let $k_l(x) = B_l(x)/l!$, where B_l is the lth Bernoulli polynomial (see Abramowitz and Stegun (1965)); we have $M_\nu k_l = \delta_{\nu-l}$ where $\delta_i = 1, i = 0$, and zero otherwise. With this norm, W_m can be decomposed as the direct sum of m

orthogonal one-dimensional subspaces $\{k_l\}$, $l = 0, 1, \ldots, m-1$, where $\{k_l\}$ is the one-dimensional subspace spanned by k_l, and $\mathcal{H}_*$, which is the subspace (orthogonal to $\oplus_l\{k_l\}$) satisfying $M_\nu f = 0$, $\nu = 0, 1, \ldots, m-1$. That is,

$$W_m = \{k_0\} \oplus \{k_1\} \oplus \ldots \oplus \{k_{m-1}\} \oplus \mathcal{H}_*.$$

This construction can be found in, e.g., Craven and Wahba (1979). Letting $\otimes^d W_m$ be the tensor product of W_m with itself d times, we have

$$\overset{d}{\otimes} W_m = \overset{d}{\otimes} [\{k_0\} \oplus \ldots \oplus \{k_{m-1}\} \oplus \mathcal{H}_*]$$

and $\otimes^d W_m$ may be decomposed into the direct sum of $(m+1)^d$ fundamental subspaces, each of the form

$$[\,] \otimes [\,] \otimes \ldots \otimes [\,] \text{ (d boxes)} \tag{10.2.3}$$

where each box ([ ]) is filled with either $\{k_l\}$ for some l, or $\mathcal{H}_*$. Additive and interaction spline models are obtained by letting $\mathcal{H}_0$ and the $\mathcal{H}^\beta$'s of Section 10.1 be direct sums of various of these $(m+1)^d$ fundamental subspaces ($\mathcal{H}_0$ must, of course, be finite-dimensional). To obtain (purely) additive spline models, one retains only those subspaces of the form (10.2.3) above whose elements have a dependency on at most one variable. This means that (at most) one box is filled with an entry other than $\{k_0\} \equiv \{1\}$.

The form of the induced norms on the various subspaces can most easily be seen by an example. Suppose $d = 4$ and consider, for example, the subspace

$$[\{k_l\}] \otimes [\mathcal{H}_*] \otimes [\mathcal{H}_*] \otimes [\{k_r\}],$$

which we will assign the index $l{**}r$. Then the squared norm of the projection of f in $\otimes^4 W_m$ onto this subspace is

$$\|P_{l{**}r} f\|^2 = \int_0^1 \int_0^1 \left[\frac{\partial^{2m}}{\partial x_2^m \partial x_3^m} M_{l(x_1)} M_{r(x_4)} f(x_1, x_2, x_3, x_4)\right]^2 dx_2\, dx_3,$$

where $M_{k(x_\alpha)}$ means M_k applied to what follows as a function of x_α.

The reproducing kernel (r.k.) for $\{k_l\}$ is $k_l(x)k_l(x')$ and the r.k. for $\mathcal{H}_*$ (found in Craven and Wahba (1979)) is $K(x, x')$ given by

$$K(x, x') = k_m(x)k_m(x') + (-1)^{m-1} k_{2m}([x - x']) \tag{10.2.4}$$

where $[\tau]$ is the fractional part of τ.

Since the r.k. for a tensor product of two r.k. spaces is the product of the two r.k.'s (see Aronszajn (1950) for a proof), the r.k. for this subspace, call it $K_{l{**}r}(x_1, x_2, x_3, x_4; x_1', x_2', x_3', x_4') = K_{l{**}r}(\mathbf{x}; \mathbf{x}')$, is

$$K_{l{**}r}(\mathbf{x}; \mathbf{x}') = k_l(x_1)k_l(x_1')K(x_2, x_2')K(x_3, x_3')k_r(x_4)k_r(x_4').$$

For more on the properties of tensor products of r.k. spaces, see Aronszajn (1950) and Weinert (1982). Tensor products of W_m were also studied by Mansfield (1972). If $L_i f = f(\mathbf{x}(i))$, where $\mathbf{x}(i)$ is the ith value of $\mathbf{x}$, then

$$(P_{l\cdots r}\xi_i)(\mathbf{x}) = K_{l\cdots r}(\mathbf{x}(i), \mathbf{x})$$

and

$$< P_{l\cdots r}\xi_i, P_{l\cdots r}\xi_j >= K_{l\cdots r}(\mathbf{x}(i), \mathbf{x}(j)).$$

In the purely additive model, $f(x_1, \ldots, x_d)$ is of the form

$$f(x_1, \ldots, x_d) = \mu + \sum_{\alpha=1}^{d} g_\alpha(x_\alpha) \tag{10.2.5}$$

where $g_\alpha \in \{k_1\} \oplus \ldots \oplus \{k_{m-1}\} \oplus \mathcal{H}_*$ and the penalty term in (10.1.2) is taken as

$$\lambda \sum_{\alpha=1}^{d} \theta_\alpha^{-1} \int_0^1 \left[\frac{\partial^m g_\alpha}{\partial x_\alpha^m}\right]^2 dx_\alpha. \tag{10.2.6}$$

To make the identifications with (10.2.3) and Section 10.1, for the purely additive spline model, $\mathcal{H}_0$ is the direct sum of the $M = 1+(m-1)d$ fundamental subspaces of the form (10.2.3) with $\{k_0\}$ in all the boxes except at most one, which contains some $\{k_l\}$ with $l > 0$. $\mathcal{H}_1 = \oplus_{\alpha=1}^{d} \mathcal{H}^\alpha$ where $\mathcal{H}^\alpha$ is of the form (10.2.3) with $\mathcal{H}_*$ in the αth box and $\{k_0\}$ in the other boxes.

If f of the form (10.2.5) is the additive spline minimizer of (10.1.2), with $L_i f = f(\mathbf{x}(i))$, then the g_α in (10.2.5) have a representation

$$g_\alpha(x_\alpha) = \sum_{\nu=1}^{m-1} d_{\nu\alpha} k_\nu(x_\alpha) + \theta_\alpha \sum_{i=1}^{n} c_i K(x_\alpha, x_\alpha(i))$$

where K is given by (10.2.4).

To discuss (two factor) interaction splines, it is convenient to consider the cases $m = 1$ and $m = 2, 3, \ldots$, separately. For $m = 1$, we have

$$\overset{d}{\otimes} W_m = \overset{d}{\otimes} [\{k_0\} \oplus \mathcal{H}_*].$$

In this case $\mathcal{H}_0$ consists of the single fundamental subspace $\otimes^d[\{k_0\}]$, there are d main effects subspaces, and there is one type of 2-factor interaction subspace, namely, one where the d boxes of (10.2.3) have $\mathcal{H}_*$ in two boxes and $\{k_0\}$ in the other $d-2$. For $m = 2, 3, \ldots$ we have two-factor interaction spaces that involve two $\{k_l\}$'s, with $l > 0$ (parametric-parametric). These may all be grouped in $\mathcal{H}_0$. Complicating matters, we may have interactions involving a $\{k_l\}$ and $\mathcal{H}_*$ (parametric-smooth) as well as two $\mathcal{H}_*$'s (smooth-smooth). For example, for $m = 2$ there are d (smooth) main effects subspaces, $d(d-1)$ fundamental subspaces with $\{k_1\} - \mathcal{H}_*$ interactions, and $d(d-1)/2$ subspaces with $\mathcal{H}_* - \mathcal{H}_*$ interactions. Similar calculations can be made for larger m and 3-factor and higher interactions. For $d = 4$, $m = 1$, we have $4 + 6 = 10$ main effects and

2-factor interaction spaces. Fitting such a model with 10 smoothing parameters has proved to be difficult but not impossible in some examples we have tried. (See the discussion in Section 11.3.) However, $m = 1$ plots tend to be visually somewhat unpleasantly locally wiggly. This is not surprising since the unknown f is not even assumed to have continuous first derivatives. However, for $d = 4$, $m = 2$, there are 4 (smooth) main effects subspaces, 12 $\{k_1\} - \mathcal{H}_*$ interaction spaces and 6 $\mathcal{H}_* - \mathcal{H}_*$ interaction spaces. We would not recommend trying to estimate 22 smoothing parameters with the present technology.

To fit additive and interaction spline models then, strategies for model simplification (selection) and numerical methods for multiple smoothing parameters are needed. Gu (1988) has suggested that the $m = 1$ case be used as a screening device. If one fits an $m = 1$ model and decides that the $\alpha\beta$th interaction is not present, then one may feel confident in eliminating both types of $\alpha\beta$ interaction in an $m = 2$ model. Similarly, if smooth main effects can be deleted in an $m = 1$ model, it can be assumed they are not present in an $m = 2$ model. (Recall that $f \in W_m \Rightarrow f \in W_{m-1}$, etc.) In Section 11.3 we discuss numerical methods for finding the GCV estimates of multiple smoothing parameters that have been used successfully in some examples with p as large as 10. If the GCV estimate $\hat{\lambda}_\beta^{-1} = \hat{\theta}_\beta/\hat{\lambda}$ is zero, then $\|P^\beta f_{\hat{\lambda},\hat{\theta}}\|^2 = 0$, and the subspace $\mathcal{H}^\beta$ can be deleted from the model. However, the probability that $\hat{\lambda}_\beta^{-1}$ is greater than zero when the true f satisfies $\|P^\beta f\|^2 = 0$ may be fairly large. (Recall the numerical results for the null model in the hypothesis tests of Section 6.3.) Thus it is desirable to have further strategies for deleting component subspaces. Possible strategies are the following. Delete $\mathcal{H}^\beta$ if, say, the contribution of this subspace to the estimated signal is small as judged by the size of

$$\sum_{i=1}^{n} (L_i P^\beta f_{\lambda,\theta})^2 = \theta_\beta^2 \|\Sigma_\beta c\|^2. \tag{10.2.7}$$

Observing that

$$\begin{pmatrix} L_1 f_{\lambda,\theta} \\ \vdots \\ L_n f_{\lambda,\theta} \end{pmatrix} = Td + \sum_{\beta=1}^{p} \theta_\beta \Sigma_\beta c$$

one could consider deleting as many of the smaller terms as possible so that

$$\Big\| Td + \sum_{\substack{\text{terms} \\ \text{retained}}} \theta_\beta \Sigma_\beta c \Big\|^2 \geq .95 \Big\| Td + \sum_{\beta=1}^{p} \theta_\beta \Sigma_\beta c \Big\|^2$$

just holds. One could also compare (10.2.7) to an estimate of $\hat{\sigma}^2$ based on the most complex reasonable model. In principle, one could generalize the GCV and GML tests of Section 6.3, to test $H_0, f \in \mathcal{H}_0 \oplus \sum_{\beta=1}^{p-1} \mathcal{H}_\beta$ versus $f \in \mathcal{H}_0 \oplus \sum_{\beta=1}^{p} \mathcal{H}_\beta$. Here, the test statistic would be the ratio of V or M minimized over the larger model to V or M minimized over the smaller model. Unfortunately, the distribution of the test statistic under H_0 will contain the

nuisance parameters $\theta_1, \ldots, \theta_{p-1}$. In principle one could generate reference distributions by Monte Carlo methods, but simplifications would no doubt be necessary to make the problem tractable. This is an area of active research (Gu (1990), Chen (1989)).