Chapter 3
Multivariate Nonparametric Regression

Charles Kooperberg and Michael LeBlanc

As in many areas of biostatistics, oncological problems often have multivariate predictors. While assuming a linear additive model is convenient and straightforward, it is often not satisfactory when the relation between the outcome measure and the predictors is either nonlinear or nonadditive. In addition, when the number of predictors becomes (much) larger than the number of independent observations, as is the case for many new genomic technologies, it is impossible to fit standard linear models. In this chapter, we provide a brief overview of some multivariate nonparametric methods, such as regression trees and splines, and we describe how those methods are related to traditional linear models. Variable selection (discussed in Chapter 2) is a critical ingredient of the nonparametric regression methods discussed here; being able to compute accurate prediction errors (Chapter 4) is of critical importance in nonparametric regression; when the number of predictors increases substantially, approaches such as bagging and boosting (Chapter 5) are often essential. There are close connections between the methods discussed in Chapter 5 and some of the methods discussed in Section 3.8.2. In this chapter, we will briefly revisit those topics, but we refer to the respective chapters for more details. Support vector machines (Chapter 6), which are not discussed in this chapter, offer another approach to nonparametric regression.

We start this chapter by discussing an example that we will use throughout the chapter. In Section 3.2 we discuss linear and additive models. In Section 3.3 we generalize these models by allowing for interaction effects. In Section 3.4 we discuss basis function expansions, which is a form in which many nonparametric regression methods, such as regression trees (Section 3.5), splines (Section 3.6) and logic regression (Section 3.7) can be written. In Section 3.8 we discuss the situation in which the predictor space is high dimensional. We conclude the chapter with discussing some issues pertinent to survival data (Section 3.9) and a brief general discussion (Section 3.10).

C. Kooperberg
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M3-A410 Seattle, WA 98109-1024, USA
email: clk@fhcrc.org

X. Li, R. Xu (eds.), *High-Dimensional Data Analysis in Cancer Research*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-0-387-69765-9_3,

3.1 An Example

We illustrate the methods in this chapter using data from patients diagnosed with multiple myeloma, a cancer of the plasma cells found in the bone marrow. The data were obtained from three consecutive clinical trials evaluating aggressive chemotherapy regiments in conjunction with autologous transplantation conducted at the Myeloma Institute for Research and Therapy, University of Arkansas for Medical Sciences (Barlogie et al., 2006). The outcome for patients with myeloma is known to be variable and is associated with clinical and laboratory measures (Greipp et al., 2005). In this data set, potential predictors include several laboratory measures measured at the baseline of the trials, age, gender, and genomic features, including a summary of cytogenetic abnormalities and approximately 350 single nucleotide polymorphisms (SNPs) for candidate genes representing functionally relevant polymorphisms playing a role in normal and abnormal cellular functions, inflammation, and immunity, as well as for some genes thought to be associated with differential clinical outcome response to chemotherapy.

In most of our analysis we analyze the binary outcome whether there was disease progression after 2 years, using the laboratory measures, age, and gender as predictors. In Sections 3.7 and 3.8 we also analyze the SNP data; in Section 3.9 we analyze time to progression and survival using a survival analysis approach.

3.2 Linear and Additive Models

Let Y be a numerical response, and let $\mathbf{x} = (x_1, \ldots, x_p)'$ be a set of predictors spanning a covariate space $\mathcal{X}$. We assume that the regression model Y takes the form of a generalized linear model

$$g(E(Y|\mathbf{x})) = \eta(\mathbf{x}), \tag{3.1}$$

where $g(\cdot)$ is some appropriate link function and

$$\eta(\mathbf{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i. \tag{3.2}$$

In this chapter, we mostly assume that Y is a continuous random variable and that $g(\cdot)$ is the identity function so that (3.1) is a linear regression model or that Y is a binary random variable and that $g(\cdot)$ is the logit function so that (3.1) is a logistic regression model, but most of the approaches that are discussed in this chapter are also applicable to other generalized linear models. In Section 3.9, we discuss some modifications that make these approaches applicable to survival data.

Estimation via the method of maximum likelihood (or least squares) is well established. Many nonparametric regression methods generalize the model in (3.2).

In particular, we can replace the linear functions x_i in (3.2) by smooth nonlinear functions $f_i(x_i)$. Now (3.2) becomes

$$\eta(\mathbf{x}) = \beta_0 + \sum_{i=1}^{k} f_i(x_i). \tag{3.3}$$

The functions $f_i(\cdot)$ are usually obtained by local linear regression (loess, e.g., Loader, 1999) or smoothing splines (e.g., Green and Silverman, 1994). The model (3.3) is known as a generalized additive model (Hastie and Tibshirani, 1990).

3.2.1 Example Revisited

Of the 778 subjects with complete covariate data in the multiple myeloma data, 171 subjects had progressed after 2 years while 570 subjects had not. Another 37 subjects were censored sufficiently early that we chose not to include them in our analysis to retain a binary regression strategy. These 37 subjects are included in the survival analysis (Section 3.9). We used nine predictors: age, gender, lactate dehydrogenase (ldh), C-reactive protein (crp), hemoglobin, albumin, serum β_2 microglobulin (b2m), creatnine, and anyca (an indicator of cytogenetic abnormality). The transformed values of ldh, crp, b2m, and creatnine on the logarithmic scale were used in the analysis. In a linear logistic regression model, anyca has a Z-statistic of 4.8 ($p = 10^{-6}$), log(b2m) has a Z-statistic of 2.7 ($p = 0.007$), and log(ldh), albumin, and gender are significant at levels between 0.02 and 0.04.

We then proceeded to fit a generalized additive model, using a smoothing spline to model each of the continuous predictors. We used the R-function `gam()`, which selects the smoothing parameter using generalized cross-validation, and provides approximate inference over the "significance" of the non-nonlinear components. Three predictors were deemed significantly nonlinear at $p = 0.05$: age, log(crp), and log(b2m), all at significance levels between 0.015 and 0.05. Note that these significance levels are approximate, and they should be treated with caution. The most interesting significant nonlinearity was probably in log(b2m). In Figure 3.1 we show the fitted component with a band of width twice the approximate standard errors. It appears that the effect of log(b2m) is only present when log(b2m) is above 1, which is approximately the median in our data set.

3.3 Interactions

Nonadditive regression models (models for $\eta(\mathbf{x})$ containing effects of interactions between predictors) occur frequently in oncology. Such models may be needed because additive models, as discussed above, may not provide an accurate fit to the data, but they may also be of interest to answer specific questions. For example, models containing interactions may be used to identify groups of patients that are

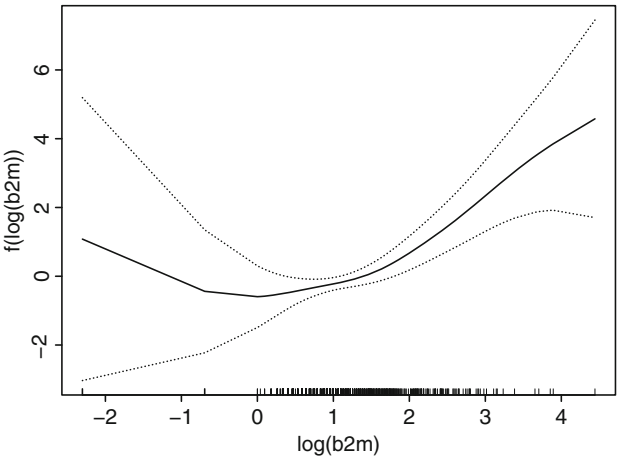

Fig. 3.1 The component of log(serum β_2 microglobulin) in the generalized additive model for progression after 2 years in the multiple myeloma data.

at especially high or low risk (e.g., LeBlanc et al., 2005), they may be of interest to identify subgroup effects in clinical trials (e.g., Singer, 2005), or to identify gene $\times$ environment interactions (e.g., Board on health sciences policy, 2002).

In the following several sections, we will discuss general models for interactions in a regression context. There are, however, special cases in which dedicated methods are more appropriate. For example, if the goal is to only identify patients at especially high risk, we may not feel a need to model the risk (regression function) for patients at low risk accurately (LeBlanc et al., 2006). When we know that some predictors are independent of each other, as is sometimes the case for gene $\times$ environment interactions or for nested case–control studies within clinical trials, more efficient estimation algorithms are possible (Dai et al., 2008). We will not discuss these situations in this chapter.

The most straightforward interaction model is to include all linear interactions up to a particular level in model (3.2); for example, a model with two- and three-level interactions is

$$\eta(\mathbf{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{1 \le i < j \le k} \beta_{ij} x_i x_j + \sum_{1 \le i < j < l \le k} \beta_{ijl} x_i x_j x_l.$$

It is clear that with this approach the number of coefficients becomes very large quickly. The problems that this causes are even worse when we generalize the smooth model (3.3). This explosion of the size of the model is sometimes known as the "curse of dimensionality," and it can be formalized by establishing the convergence rates of parameters in such models under appropriate conditions (Stone, 1994). Instead we may want to include only those interactions that are really needed to accurately model the regression function $\eta(\mathbf{x})$. Often this is done using some form of stepwise regression. It turns out that approach can be generalized conveniently using a basis function approach.

3.4 Basis Function Expansions

The linear model (3.1) can also be used as the starting point for nonlinear, nonadditive, multivariate regression methods. Assume that the regression function $\eta(\mathbf{x})$ is in some p-dimensional linear space $\mathcal{B}(\mathcal{X})$, and let $B_1(\mathbf{x}),\ldots,B_p(\mathbf{x})$ be a basis for $\mathcal{B}(\mathcal{X})$. Then we can write

$$\eta(\mathbf{x}) = \sum_{i=1}^{p} \beta_i B_i(\mathbf{x}). \tag{3.4}$$

For a given set of basis functions $B_1(\cdot),\ldots,B_p(\cdot)$ estimation in (3.4) is a straightforward extension of (3.2).

Several nonparametric multivariate regression methodologies use a basis function approach, but rather than fixing the space $\mathcal{B}(\mathcal{X})$ these approaches select the space at the same time as the coefficients of the basis functions are estimated. Three of the methodologies that are discussed later in this chapter use this approach.

- Regression tree methods, such as classification and regression trees (CART, Breiman et al., 1984). The basis functions that are used for tree methods are indicator functions corresponding to rectangular regions of the predictor space. Tree methods are discussed in Section 3.5.
- Multivariate adaptive regression splines (MARS, Friedman, 1991) and related spline methods (e.g., Kooperberg et al., 1995; Stone et al., 1997). The basis functions that are used for MARS and related methods are piecewise polynomials (splines) and their tensor products. We discuss spline methods in Section 3.6.
- Logic regression (Ruczinski et al., 2003) is discussed in Section 3.7. The basis functions that are used for logic regression are Boolean combinations of binary predictors.

Stepwise regression methods provide useful tools for model selection using basis functions. As an example, suppose that we consider two linear spaces to model $\eta(\mathbf{x})$: a p-dimensional space $\mathcal{B}^p(\mathcal{X})$ that is a sub-space of a $(p+1)$-dimensional space $\mathcal{B}^{p+1}(\mathcal{X})$. After we fit model (3.4) using basis functions for the smaller space $\mathcal{B}^p(\mathcal{X})$ we can compute a score test (Rao statistic, Rao, 1973) to evaluate how much better $\eta(\mathbf{x})$ would be modeled if we would require that $\eta(\mathbf{x}) \in \mathcal{B}^{p+1}(\mathcal{X})$ instead. Similarly, after we fit model (3.4) using basis functions for the larger space $\mathcal{B}^{p+1}(\mathcal{X})$ we can compute a Wald statistic to evaluate how much worse $\eta(\mathbf{x})$ would be modeled if we would require that $\eta(\mathbf{x}) \in \mathcal{B}^p(\mathcal{X})$. If these would be prespecified spaces the score and Wald statistics could be compared with standard parametric distributions, similar to what is done in stepwise variable selection methods (see Chapter 2). The adaptivity of these approaches does typically require other approaches to obtain significant levels and prediction errors though (see Chapter 4).

We can generalize this stepwise procedure to an algorithm for stepwise model building, that is used both in tree and in spline methods.

1. Start with modeling $\eta(\mathbf{x}) \in \mathcal{B}_a^p$. A common situation is that $p = 1$ and $\mathcal{B}_a^1$ consists of only constant functions.
2. Stepwise addition: replace $\mathcal{B}_a^p$ by a $(p+1)$-dimensional space $\mathcal{B}_a^{p+1}$ of which $\mathcal{B}_a^p$ is a subspace by considering a (large) set of candidate spaces $\mathcal{B}_a^{p+1}$ that satisfy some method-dependent regularity conditions. Select the "best" $\mathcal{B}_a^{p+1}$ for example, by selecting the $\mathcal{B}_a^{p+1}$ corresponding to the largest score statistic.
3. Continue adding dimensions until either a prespecified dimension p^* is reached, or until the improvement in the fit between successive models becomes very small.
4. Set $\mathcal{B}_d^{p^*} = \mathcal{B}_a^{p^*}$.
5. Proceed with stepwise deletion: replace $\mathcal{B}_d^p$ by a $(p-1)$-dimensional subspace $\mathcal{B}_d^{p-1}$ that satisfies some method-dependent regularity conditions. Select the "best" $\mathcal{B}_d^{p-1}$, for example, by selecting the candidate corresponding to the smallest Wald statistic.
6. Continue until p reaches some minimum dimension (e.g., $p = 1$).
7. Out of all the linear spaces considered $\mathcal{B}_a^1, \ldots, \mathcal{B}_a^{p^*} = \mathcal{B}_d^{p^*}, \ldots, \mathcal{B}_d^1$, select one either using some penalized likelihood like the Akaike information criterion (Akaike, 1974) or the Bayesian information criterion (BIC, Schwarz, 1978), or an honest method to estimate the prediction error, such as cross-validation.

3.5 Regression Tree Models

3.5.1 Background

Regression and classification trees are primarily known for their easy-to-understand geometric representation. While a binary regression tree provides a simple description of groups of subjects, the model can also be cast in a regression spline form similar to the methods presented in Section 3.6. The CART algorithm (Breiman et al., 1984) is probably the best-known implementation of tree-based methods in the statistical literature and generally motivates the basics given in this section. There has also been extensive research of tree-structured methods in machine learning, for instance the C4.5 algorithm of Quinlan (1993). When extended to survival data, regression trees have found a significant following in medicine because the sequence of binary decisions leads to simple representation for prognostic groups of patients treated in a similar fashion. Most tree-based methods for survival data have adopted at least some aspects of the CART algorithm (Gordon and Olshen, 1985; Ciampi et al., 1986; Segal, 1988; LeBlanc and Crowley, 1993). Some recent examples in survival analysis using regression trees include Greipp et al. (2005), London et al. (2005), Farag et al. (2006), and Gimotty et al. (2007).

3.5.2 Model Building

3.5.2.1 Model Basis Set as Partition Function

A tree model can be represented as a binary tree T, where the set of terminal nodes $\widetilde{T}$ corresponds to the partition of the covariate space $\mathcal{X}$ into a number of $M(\widetilde{T})$ disjoint subsets. A tree model can also be expressed by a basis function representation

$$\eta(\mathbf{x}) = \sum_{h \in \widetilde{T}} \eta_h B_h(x),$$

(compare with (3.4)) where $B_h(x) = I\{x \in R_h\}$, R_h is the region corresponding to a terminal node h, and η_h is a vector of parameters (e.g., a mean, a clinical response probability, or a higher-dimensional object such as a survival function $S(t|\eta(\mathbf{x}))$ corresponding to a terminal region. For instance, the survival function could be of semiparametric form $S_0(t)^{\exp(\eta(\mathbf{x}))}$ as in the proportional hazards model. We outline important components of algorithms used to construct regression trees, including specifying the types of partitions that are permitted; rules to prune the tree back; and methods to choose model or tree size.

3.5.2.2 Splitting or Basis Selection

Trees represent a sequence of splits of the data or predictor space where each split is induced by a rule of the form "$x \in S$" where $S \subset \mathcal{X}$. Typically, splits are dependent on a single covariate, so we may have $S = \{\mathbf{x} : x_j \leq c\}$ for an ordered predictor, or S is a subset $S \subset B = \{v_1, v_2, ..., v_r\}$ of the r values of x_j for categorical variables.

The tree model is grown in a forward stepwise fashion, similar to the stepwise algorithm described in Section 3.4. Starting with the entire data set and predictor space, each variable and potential split point is evaluated. The split point and variable that leads to the "best" split (as described below) is chosen. The data and the predictor space are partitioned into two groups. The same algorithm is then recursively applied to each of the resulting groups. Therefore, at any point on the regression tree, a split at a node h yields two nodes which can also be represented with the pair of basis functions

$$b^+_{h(j)}(\mathbf{x}) = I\{x_{h(j)} \in S_{h(j)}\} \text{ and } b^-_{h(j)}(\mathbf{x}) = I\{x_{h(j)} \notin S_{h(j)}\}.$$

Each step in the growing process geometrically replaces a current node h with a left and right daughter nodes $l(h)$ and $r(h)$ or equivalently a current basis function $B_h(\mathbf{x})$ for node h with the basis functions

$$B_{l(h)}(\mathbf{x}) = B_h(\mathbf{x}) b^+_{h(j)}(\mathbf{x}) \text{ and } B_{r(h)}(\mathbf{x}) = B_h(\mathbf{x}) b^-_{h(j)}(\mathbf{x}).$$

Most tree algorithms use error, likelihood, or partial likelihood (or score tests such as the logrank test) to select split points (or knots). The improvement for a split at node h into left and right daughter nodes can be represented by

$$G(h) = D(h) - [D(l(h)) + D(r(h))],$$

where $D(h)$ is the residual error at a node. For uncensored continuous response problems, $D(h)$ is typically the mean residual sum of squares or mean absolute error or for binary data it is typically binomial deviance. For survival data, it would be reasonable to use the deviance corresponding to the assumed survival model. For instance, the exponential model deviance for node h is

$$D(h) = \sum 2 \left[\delta_i \log \left(\frac{\delta_i}{\widehat{\lambda}_h t_i} \right) - (\delta_i - \widehat{\lambda}_h t_i) \right],$$

where $\delta_i = 1$ if the ith observation was a failure, and $\delta_i = 0$ if the observation was censored, and $\widehat{\lambda}_h$ is the maximum likelihood estimate of the hazard rate in node h (Davis and Anderson, 1989). Alternatively $G(h)$ can be an appropriate score test statistic, for example the logrank test statistic.

Typically a large tree is grown to avoid missing structure and then pruned back using a method described below.

3.5.3 Backwards Selection (Pruning)

Many stepwise regression methods utilize variations of backwards selection to select more simple models (see Section 3.4). The local nature of the tree-based methods leads to a fast backwards method, called cost complexity pruning in the CART algorithm, for evaluating all possible submodels. The cost-complexity objective function is defined as a penalized measure of fit

$$D_\alpha(T) = \sum_{h \in \widetilde{T}} D(t) + \alpha M(\widetilde{T}),$$

where α is a nonnegative complexity parameter, $D(h)$ is the estimated cost or impurity of a node, and $M(\widetilde{T})$ is the number of terminal nodes or constant regression regions R_h. Therefore, the cost-complexity measure controls the trade-off between the size or complexity of the tree, and how well the tree fits the data. Then, for any value of α the goal is to find $T(\alpha)$: the tree that minimizes $D_\alpha(T)$ among all pruned subtrees of T. The algorithm finds the sequence of optimally pruned subtrees by repeatedly deleting branches of the tree for which the average reduction in residual error per split in the branch is small. The process yields a nested sequence of optimal subtrees $T(\alpha) = T(\alpha_l) = T_l$ for $\alpha_l \le \alpha < \alpha_{l+1}$. The removal of a branch can again be viewed in regression context as replacing each of the basis functions corresponding to the pruned branch with the sum of the basis functions

$$B_l(\mathbf{x}) = \sum_{h \in Q_l} B_h(\mathbf{x})$$

where Q_l represent the nodes in a branch rooted at node l. The final tree size is selected by resampling (often K-fold cross-validation is used), although some difficulties arise for semiparametric survival regression models.

3.5.4 Example Revisited

Using the example data set and variables described earlier, we constructed a regression tree to characterize the probability of death or progression within 2 years of registration. Figure 3.2 show a large tree constructed on the available predictors. Below each terminal node is an estimate of the probability of progression or death. Since the tree likely over-fits the data, a pruned tree is selected using cost-complexity pruning and ten fold cross-validation of binomial deviance. The resulting tree model is presented in Figure 3.3; it includes just two splits on variables serum β_2

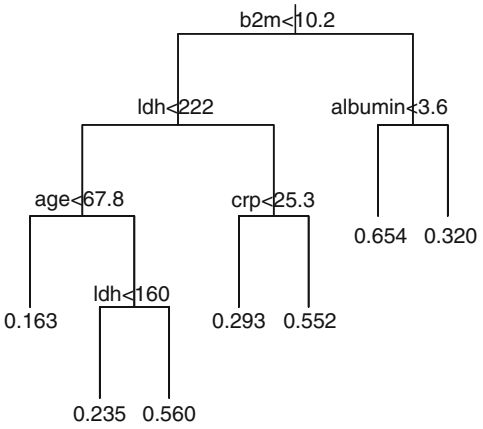

Fig. 3.2 An unpruned regression tree constructed to characterize 2-year progression probability for the multiple myeloma data.

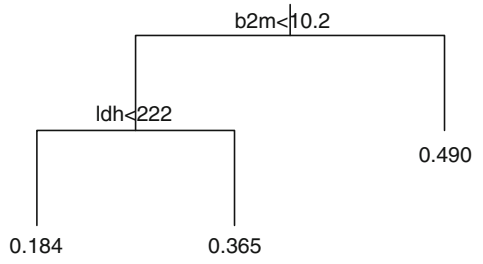

Fig. 3.3 A pruned regression tree constructed to characterize 2-year progression probability for the multiple myeloma data.

microglobulin and lactate dehydrogenase and identifies three outcome groups. Subjects with serum β_2 microglobulin ≥ 10.2 have the worst outcome with 49% having either progressed or died within 2 years.

3.5.5 Issues and Connections

An often cited limitation of regression trees is that they are piecewise constant functions when typically the underlying conditional distribution function of the outcome is a smooth function of the predictors. If interest is in studying groups of patients, this is not really a problem, other than the difficulty in specifying a specific fraction of patients to be indicated by the prognostic rule. However, for prediction applications the nonsmoothness does lead to reduced performance. Ensembles of trees, through boosting, bagging, and Random Forests (Freund and Schapire, 1996; Breiman, 1996; Friedman et al., 2000) have been used to circumvent this discreteness at the cost of losing the simple decision rules. Alternatively, spline methods such as HARE or MARS described in Section 3.6 can lead to substantially improved predictions.

In part because of their nonsmoothness and the stepwise selection method, trees are subject to considerable variability. An important parameter to control variability is the minimum number of observations in a node (or uncensored observations in the case of censored survival data). This issue connects to the importance of avoiding placing knots in regression splines too close to the edge of the covariate distribution. Again, ensembles of trees have been used to reduce variability (sometimes dramatically) but again at the loss of the simple decision rule properties. Retaining decision rule but somewhat smoother methods have been proposed, such as rule induction via the PRIM method (Friedman and Fisher, 1999).

3.6 Spline Models

3.6.1 One Dimensional

Spline models are primarily used for the approximation of smooth univariate and multivariate functions. In univariate problems, splines are piecewise polynomial functions, that satisfy some regularity conditions. In particular, let $t_0 < t_1 < \cdots < t_K$ be a set of K knots. A function $f(x)$ is a cubic spline if in each of the intervals (t_{k-1}, t_k), $k = 1, \ldots, K$, the function $f(x)$ is a cubic polynomial, and it is twice differentiable everywhere. Different spline models may have boundary restrictions for $f(x)$ on the intervals $(-\infty, t_0]$ and $[t_K, \infty)$, but when there are no boundary conditions it is easy to see that these cubic spline functions form a linear space, with basis

$$1, x, x^2, x^3, (x - t_k)_+^3, \quad k = 0, \ldots, K, \tag{3.5}$$

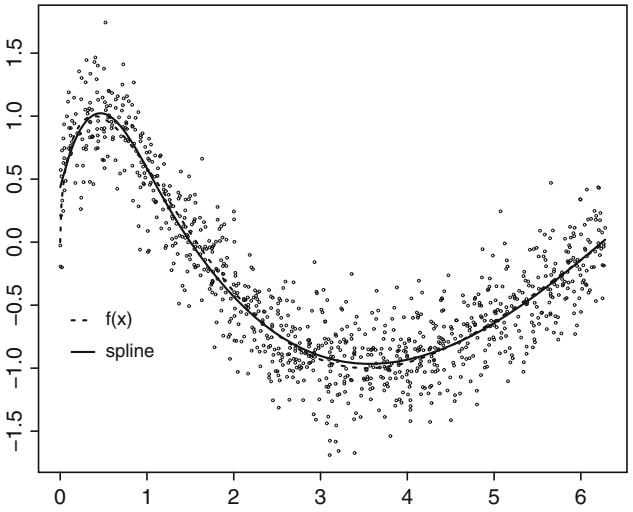

Fig. 3.4 Spline fit. The data were 1,000 i.i.d. samples generated from $Y = f(x) + \varepsilon$, $f(x) = \sin(\sqrt{2\pi x})$ where $x \sim \text{Unif}(0, 2\pi)$ and $\varepsilon \sim N(0, 0.25^2)$. The spline approximation only has a single knot at 1.13.

where $x_+ = x$ if $x > 0$ and 0 otherwise. Cubic spline functions can approximate functions very well, often with a small number of knots (Figure 3.4).

Similarly to cubic splines, a function $f(x)$ is a linear spline if it is continuous and linear on each of the intervals (t_{k-1}, t_k). A basis for linear spline functions is

$$1, x, (x - t_k)_+, \quad k = 0, \ldots, K. \tag{3.6}$$

Regression tree functions in one dimension can be seen as piecewise constant splines. Linear and piecewise constant splines are not as good as cubic splines in approximating smooth curves, but they are often easier to deal with algorithmically. As splines form a linear space, the spline model can be written in the form (3.4). Note that in most situations (3.5) and (3.6) are not the bases used for computations, as they are numerically very instable; instead usually a B-spline basis is used (de Boor, 1978).

Spline models naturally arise as solutions for some penalized regression problems. For example, based on regression data (Y_i, x_i), $i = 1, \ldots, n$, the solution of the minimization problem

$$\arg\min_{f(x)} \sum_i (Y_i - f(x_i))^2 + \lambda \int \left(\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2} \right)^2 \mathrm{d}x \tag{3.7}$$

is a (natural) cubic spline with knots at every unique data point x_i (Green and Silverman, 1994). In practice, having a model with so many knots causes problems in many nonlinear and high-dimensional problems. Instead, several other approaches use spline methods with fewer knots.

- Instead of using n knots, express $f(x)$ as a spline function with a fairly large number of knots, that is still much smaller than n, and then use a penalized optimization like (3.7) (O'Sullivan, 1988; Eilers and Marx, 1996). This approach works fairly well in more complicated one-dimensional problems, as well as for generalized additive models, in particular with automatic rules to select smoothing parameters.
- Use a much smaller number of pre-specified knots, and carry out estimation without penalty terms. The advantage is that the resulting problem is fully parametric, and that inference is thus well established. Estimation problems are often small (and easy). See Quantin et al. (1999) for an application in oncology. The disadvantage is that selection of the location of the knots can be arbitrary, and generalizations to nonadditive models are not immediate.
- A third alternative is to use a stepwise algorithm like the one described in Section 3.4 using knots and basis functions from (3.5). This approach was first used in univariate regression by Smith (1982) and is behind algorithms like MARS (Friedman, 1991) for linear regression, HARE (Kooperberg et al., 1995) for survival data, and Polyclass (Kooperberg et al., 1997) for logistic regression and classification. We will discuss those in more detail for multivariate models below. See Polesel et al. (2005) for an application in oncology.

3.6.2 Higher-Dimensional Models

The common approach to using regression splines in higher dimensions is to use basis functions that are tensor products of basis functions in one dimension. For example, if $B_1(\mathbf{x}) = g_1(x_k)$ and $B_2(\mathbf{x}) = g_2(x_l)$ are two basis functions that depend on a single predictor, then $B_3(\mathbf{x}) = g_1(x_k)g_2(x_l)$ is a tensor product basis function that depends on two predictors. For high-dimensional problems, it is common to consider only a few selected lower-order interactions. This has a variety of advantages: (1) lower-dimensional components are typically easier to interpret, interactions in models that do not contain the corresponding main effects are particularly difficult to interpret; (2) using all (higher order) tensor products of lower-order basis functions would yield an extremely large number of basis functions and may cause numerical instability, and (3) from a theoretical perspective, it has been established that spline functions have faster convergence rates if the largest order of interactions in models is small (Stone, 1994). The exact restrictions on when tensor product basis functions are allowed in spline models differs from one methodology to the other: for example, MARS (Friedman, 1991) has fewer restrictions than HARE (Kooperberg et al., 1995), Polyclass, and Polymars (Kooperberg et al., 1997). Here we will describe the Polyclass algorithm for logistic regression as an example.

Assume that we have an i.i.d. sample of size n with a binary response variable Y and a p-dimensional vector of predictors $\mathbf{x} = (x_1, \ldots x_p)'$. Polyclass uses linear splines, and uses interactions involving at most two predictors (although the generalization to higher-dimensional interactions is immediate). An allowable

linear space $\mathcal{B}(\S)$ can have basis functions 1, x_i, $(x_i - t_{k_i})_+$, $x_i x_j$, $(x_i - t_{k_i})_+ x_j$, and $(x_i - t_{k_i})_+ (x_j - t_{k_j})_+$, with $i \neq j \in \{1, \ldots, p\}$, where the t_{k_i} are knots in the range of x_i, with the additional conditions that

- $B(\mathbf{x}) = x_i x_j$ can only be in $\mathcal{B}(\S)$ if $B(\mathbf{x}) = x_i$ and $B(\mathbf{x}) = x_j$ are in $\mathcal{B}(\S)$;
- $B(\mathbf{x}) = (x_i - t_{k_i})_+$ can only be in $\mathcal{B}(\S)$ if $B(\mathbf{x}) = x_i$ is in $\mathcal{B}(\S)$;
- $B(\mathbf{x}) = (x_i - t_{k_i})_+ x_j$ can only be in $\mathcal{B}(\S)$ if $B(\mathbf{x}) = x_i x_j$ and $B(\mathbf{x}) = (x_i - t_{k_i})_+$ are in $\mathcal{B}(\S)$; and
- $B(\mathbf{x}) = (x_i - t_{k_i})_+ (x_j - t_{k_j})_+$ can only be in $\mathcal{B}(\S)$ if $B(\mathbf{x}) = x_i (x_j - t_{k_j})_+$ and $B(\mathbf{x}) = (x_i - t_{k_i})_+ x_j$ are in $\mathcal{B}(\S)$.

The algorithm then proceeds with the stepwise algorithm in Section 3.4. The final model is selected as the one that minimizes

$$\mathrm{AIC}_\alpha = -\widehat{\ell}(\mathcal{B}(\S); Y_i, \mathbf{x}_i, i = 1, \ldots, n) + \alpha p, \tag{3.8}$$

where $\widehat{\ell}(\mathcal{B}(\S); Y_i, \mathbf{x}_i, i = 1, \ldots, n)$ is the fitted log-likelihood for one of the models (of dimension p) that was considered, and α is a penalty parameter, or that maximizes the cross-validated likelihood.

3.6.3 Example Revisited

We applied the Polyclass methodology to the multiple myeloma data. The polyclass model with the default penalty parameter of $\alpha = \log n \approx 6.66$ (3.8) only involved the two predictors log(b2m) and anyca in a linear fashion:

$$\mathrm{logit}(P(\text{progression})) = 2.19 - 0.99 \log(\text{b2m}) - 0.50 \text{anyca}.$$

The model with $\alpha = 4$, while likely overfitting the data somewhat, is more interesting, as it also involves age, gender, log(ldh), log(creatinine), a knot in age, log(ldh), and log(b2m), and an interaction between age and gender. In Figure 3.5 we show a contour plot for the fitted 2-year progression probabilities as a function of creatinine and age, separately for men and women, when the other predictors are held at their median values. The figure indicates that while older ages lead to quite similar progression proportions, younger females tend to have higher risk than younger males.

3.7 Logic Regression

Logic regression is a generalized regression methodology that is particularly suited for situations in which (most) predictors are binary. Clearly this is the case when predictors are single nucleotide polymorphisms (SNPs), as is the case for the multiple myeloma data and many other oncological problems. The logic regression model is

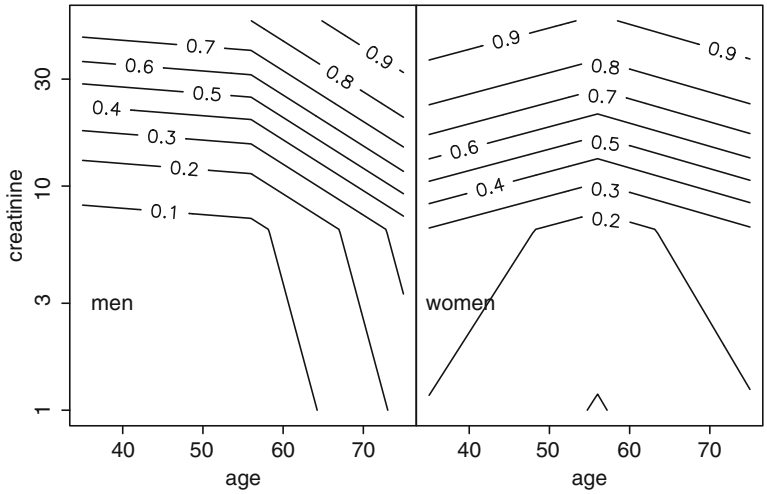

Fig. 3.5 Fitted 2-year progression probabilities for a Polyclass model selected with penalty $\alpha = 4$ as a function of creatinine and age, separately for men and women, when the other predictors are held at their median values.

$$\eta(\mathbf{x}) = \beta_0 + \sum_{i=1}^{m} \beta_i \mathrm{L}_i(\mathbf{x}). \tag{3.9}$$

Each of the L_i is a Boolean combination of binary predictors x_j, $j = 1, \ldots, J$ such as

$$\mathrm{L}_i = [(x_7 \text{ and } x_{13}^c) \text{ or } x_5],$$

where "1" equals "true," "0" equals "false," and c refers to the complement. Additional predictors Z or components to correct for population stratification can be included additively in model (3.9).

Logic regression is an adaptive algorithm which selects those logic terms L_i that minimize the residual sum of squares or maximize the log-likelihood corresponding to the model (3.9). Typically in logic regression the number of logic terms m is small (between 1 and 3), and the logic terms can be interpreted as "risk factors." The optimization of the logic regression model is carried out using a greedy stepwise algorithm or a stochastic simulated annealing algorithm.

For this simulated annealing algorithm it turns out to be very convenient to represent a logic expression $\mathrm{L}_i(\mathbf{x})$ in a logic tree form (Figure 3.6). During the simulated annealing algorithm, at each step one of the logic trees is replaced by another logic tree using one of the operations displayed in Figure 3.7. Based on the new tree the likelihood of $\eta(\mathbf{x})$ is evaluated. If the new model is an improvement over the existing model the new model is retained; if the old model was better the new model is retained with a probability that depends on the difference between the old and new log-likelihood and the stage of the algorithm: early on almost all new models are accepted, while toward the end of the algorithm only improved models are accepted.

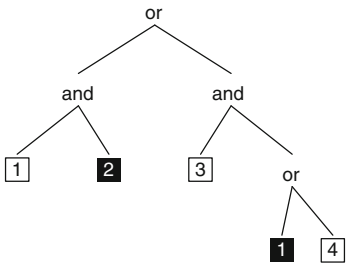

Fig. 3.6 A logic tree representation of the Boolean expression $(x_1 \wedge x_2^c) \vee (x_3 \wedge (x_1^c \vee x_4))$. Logic trees are evaluated from the bottom up; white numbers on a black background denote the complement.

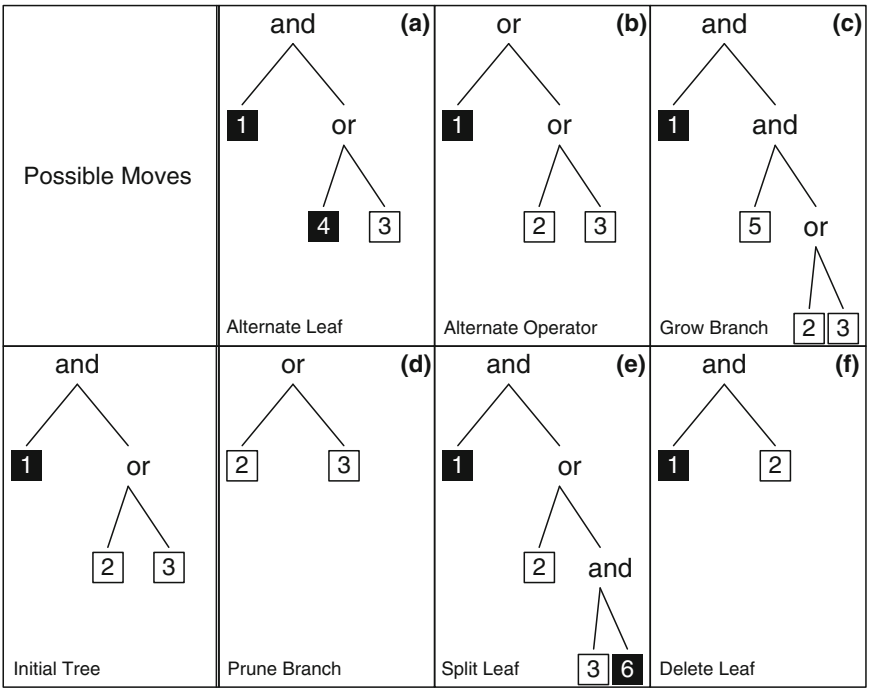

Fig. 3.7 Changes in logic regression trees considered during the simulated annealing algorithm.

3.7.1 Example Revisited

We applied logic regression to the 348 SNPs of the multiple myeloma data, using again 2-year progression as the outcome. Each of the 348 SNPs was recoded as two binary predictors corresponding to a dominant and a recessive effect. In Figure 3.8

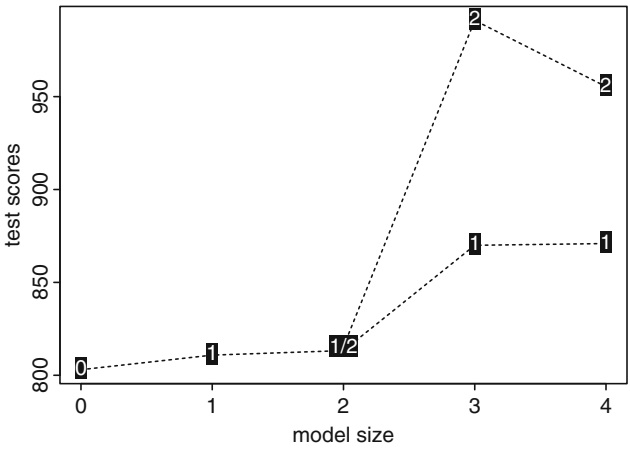

Fig. 3.8 Cross-validation (test set) deviance for the logic regression analysis of the multiple myeloma data. The white numbers in the black squares refer to the number of logic terms in the logic regression model, the model size refers to the total number of leaves in these models combines.

we show the test set deviance from tenfold cross-validation of the logic regression analysis of this data. We note from this figure that based purely on deviance, none of the models is better than the null-model. The model with two SNPs, however, has a deviance that is not much worse than the null-model, and may thus be of interest for further investigation. This model includes a logic regression term

$$\text{rs4148737D} \vee \text{rs1143627R}^c,$$

(rs1922242D was identical to rs4148737D) on this data. We will see the same SNPs appear in the analysis in the next section.

3.8 High-Dimensional Data

With the development of new genomic technologies, very high-dimensional data sets are now generated for oncological data. Data sets using gene expression data may have data on tens of thousands of genes (e.g., Rosenwald et al., 2002), data sets for whole genome association studies may have data on hundreds of thousands of SNPs (e.g., Easton et al., 2007; Yeager et al., 2007). The traditional statistical paradigm, where the number of cases n is much larger than the number of predictors p no longer holds in this situation. Typical statistical methods for this type of data involve substantial amounts of model selection, as well as shrinkage of the parameter estimates.

3.8.1 Variable Selection and Shrinkage

In moderate to high-dimensional predictor settings it is desirable to have parsimonious or sparse representations of prediction models. In the previous sections we have discussed stepwise basis function selection strategies. Alternatively, one can investigate smoother model selection methods.

3.8.2 LASSO and LARS

Consider the linear regression setting, where there are n independent observations $(y_i, x_{i1}, \ldots, x_{ik})$ of the response and k predictor variables. A technique proposed by Tibshirani (1996) introduces an L^1-penalty on the regression coefficients which leads to both shrinkage and variable selection called least absolute shrinkage and selection operator (LASSO). This is in contrast to ridge regression (Hoerl and Kennard, 1970) which minimizes the residual error subject to an L^2-penalty which does not lead to variable selection. The LASSO estimate $\widetilde{\beta} = (\tilde{\beta}_1, \ldots, \widetilde{\beta}_m)'$ is defined as the minimizer of

$$g(\beta) = \sum_{i=1}^{n} (y_i - \sum_k \beta_k x_{ik})^2 + \lambda_1 \sum |\beta_k|^1,$$

where λ_1 is a nonnegative penalty parameter. Often the response and predictors are standardized so that $\sum_i y_i = 0$ and $\sum_i x_{ik} = 0$ and $\sum_i x_{ik}^2 = 1$. This estimator has the attractive property that as λ_1 increases minimizing $g(\beta)$ with respect to β leads to some of the β_k set to zero and hence variable selection. For fixed λ_1, for optimization quadratic programming techniques or alternatives more efficient methods by Osborne et al. (2004) can be used. A related and highly efficient algorithm, the least angle regression algorithm (LARS, Efron et al., 2004), leads to efficient estimation and links forward stage-wise methods and LASSO. LASSO and LARS are discussed in more detail in Chapter 2.

LARS gives answers that are often close to LASSO; they are identical if the predictors are orthogonal. However, the estimation algorithm aligns closely with the forward stepwise model building strategies described in earlier sections. An outline of the algorithm is given below:

1. Start with $r = y$, $\widehat{\beta}_j = 0$, $j = 1, \ldots, p$. Assume that the x_j are standardized.
2. Find the predictor x_k that is most correlated with r.
3. Increase $\widehat{\beta}_k$ in the direction of $\text{sign}(\text{cor}(r, x_k))$ until another predictor x_j has equal correlation to r as it does with x_k. Put j in set of active predictors, S.
4. Move $(\widehat{\beta}_k : k \in S)$ in the joint least squares direction for $(x_k : k \in S)$ until yet another predictor has equal correlation with the current residual.
5. Repeat Step 4 until $\text{cor}(r, x_k) = 0$ for all k.

Note that the model can include at most $\min(p,n)$ variables. One strategy to alleviate this potential problem is the "elastic net" proposed by Zou and Hastie (2005). The elastic net can be expressed as an optimization problem with the objective function with both squared and absolute penalty terms

$$g(\beta) = \sum_{i=1}^{N}(y_i - \sum_k \beta_k x_{ik})^2 + \lambda_1 \sum |\beta_k|^1 + \lambda_2 \sum |\beta_k|^2.$$

Their simulations show that the elastic net method leads to grouping of variables where strongly correlated variables are either in or removed from the model as the penalty parameters λ_1 and λ_2 are increased.

Note, that in this section we have described these methods in terms of the original predictors x_k; we could generalize to sets of regression spline or regression tree basis functions, $B_j(\mathbf{x}), j = 1,\ldots,p$ as described in the previous section.

3.8.3 Dedicated Methods

While the methods described above directly lead to dimension reduction, there are a large number of other methods which can be viewed as two-stage procedures that at the first stage reduce the set of original variables x_i to a small number of combinations z_i and then at the second phase uses those combinations in further regression modeling. Many of the techniques can be viewed as generalizations or parallels to either principal components regression, which uses only the joint distribution of the x_i at the first stage, or partial least squares which constructs linear combinations of the predictors but also guides the selection by also using the outcome Y.

For instance, many gene expression modeling applications in oncology have used clustering of genes to derive predictor variables for association modeling. Jointly using outcome and expression was used by Hastie et al. (2001) and Dettling and Bühlmann (2002) and others. An important consideration when using both the joint distribution of outcome and predictors at the first stage is that appropriate assessment of prediction error and model fit is incorporated (for instance by cross-validation) and included in the modeling building.

3.8.4 Example Revisited

We applied a generalization of the LARS regression method appropriate for binary data (Park and Hastie, 2006) to the 2-year progression-free survival outcome, and the multiple myeloma SNP data. Each of the SNPs was coded in dominant and recessive form. In the Figure 3.9, we show the first few steps of the coefficient path. Three SNPs appear to enter the model early, "rs1143627R," "rs2756109D," and "rs703842R." Note that the SNPs that were selected by logic regression entered

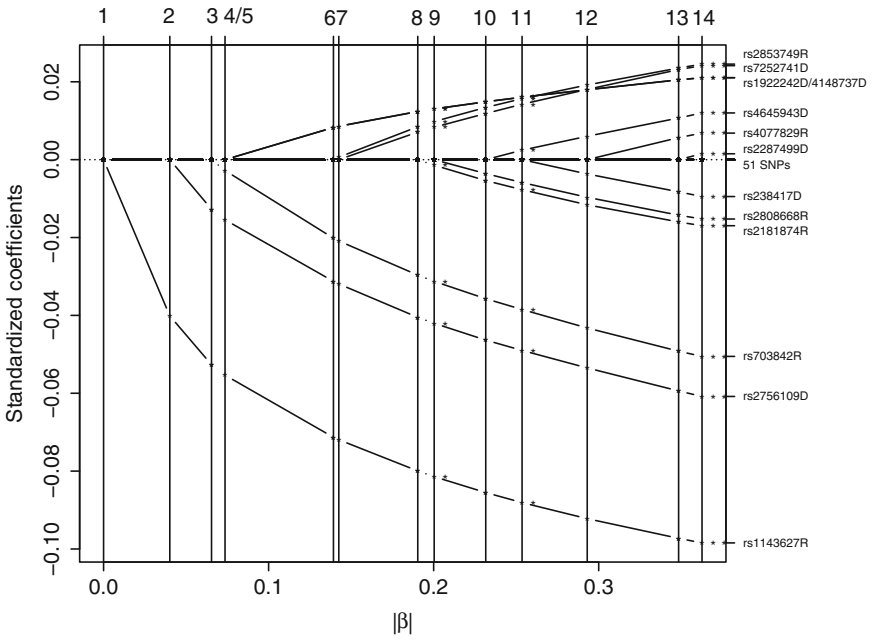

Fig. 3.9 Coefficient path for myeloma SNP data.

the model as the first, fourth, and fifth SNP. Cross-validating the model building process leads to the conclusion that the cross-validated estimates of deviance are relatively flat with respect to model complexity and then start to increase for models with larger numbers of predictors. Therefore, there is not strong evidence that the combination of SNPs are significantly associated with disease progression.

Often there is interest in assessing if genomic information adds to prediction beyond traditional laboratory measures. This can be easily incorporated by adjusting for known myeloma clinical variables then fitting SNP data using the Park and Hastie algorithm. This was done for the above example and while not unexpected given the earlier analysis, it suggested no additional impact with SNP data on prediction over the laboratory variables previously described.

3.9 Survival Data

An important goal in survival regression analysis is to determine how the distribution of survival times depends on the predictors. A complication in analyzing survival data in the context of oncology trials is that typically not all patients have died (or progressed) by the time the analysis for the study is completed. Those patients alive at the time of analysis are called censored.

We denote the true survival time as a positive random variable T, whose distribution may depend on a set of predictors $\mathbf{x} = (x_1, \ldots, x_k)'$. Often it can be assumed that the censoring mechanism is independent which facilitates likelihood construction and inference. Let the observed data be denoted by $(T_i, \delta_i, \mathbf{x}_i),\ i = 1, \ldots, n$.

While one can express the conditional survival distribution using an accelerated failure time specification which links the $\log(T)$ to a linear model of the predictors, $\log(T) = a + b'X + e$, hazard function modeling is most often used. The conditional hazard function is defined as

$$\lambda(t|\mathbf{x}) = \lim_{\Delta \to 0} \frac{P(t \le T < t + \Delta | t \le T; \mathbf{x})}{\Delta}.$$

Here, we limit discussion to predictors which are real values measured at baseline; in some survival settings they may represent time-dependent functions, $x_1(t), \ldots, x_k(t)$, as well. For instance, they may be measures of health status of the patient evolving over time. The (conditional) hazard function can be interpreted as the probability that someone dies in the next time interval of infinitesimal length Δ, given that he is alive at time t. It is convenient to specify models on the logarithm scale so we denote the logarithm of the hazard function as

$$\alpha(t|\mathbf{x}) = \log \lambda(t|\mathbf{x}).$$

If one assumes an additive model on the log scale,

$$\alpha(t|\mathbf{x}) = f(t) + \eta(\mathbf{x})$$

implies a proportional hazards assumption which is a focus of the model of Cox (1972), which also assumes the baseline hazard function to be an unspecified nonparametric function. Estimation in that case utilizes the partial likelihood. Note that $\eta(\mathbf{x})$ can represent a simple linear model or more flexible models depending on a regression spline basis described in earlier sections. For instance, let $B_1(\mathbf{x}), \ldots, B_p(\mathbf{x})$ be a basis for $\mathcal{B}(\mathcal{X})$. Then we can write

$$\eta(\mathbf{x}) = \sum_{i=1}^{p} \beta_i B_i(\mathbf{x}). \tag{3.10}$$

Within the proportional hazards class, tree-based or logic regression models can be used to characterize the basis section used in the above expression.

However, regression models can be more general. For instance, the HARE model (Kooperberg et al., 1995), can link both time and the predictors using a model specified as

$$\alpha(t|\mathbf{x}) = \sum_{i} \beta_i B_i(t|\mathbf{x}). \tag{3.11}$$

The basis functions $B_i(t|\mathbf{x})$ in HARE can depend solely on time or a predictor or both on time and a predictor which allows specification on nonproportional hazards

models. The basis functions are selected with an algorithm similar to the Polyclass algorithm in Section 3.6.2.

Modeling the full survival distribution is slightly more general than modeling within the proportional hazards framework. But there are also disadvantages: coefficients in model (3.10) are interpretable as log-relative risk estimates, while the nonproportionality in (3.11) removes this interpretation. Computationally the partial likelihood computations for (3.10) are much easier than the full likelihood computations for (3.11), as these later require integrating the conditional survival function for every unique set of covariates $\mathbf{x}$ which, except for piecewise linear splines, becomes very demanding.

We end this section by noting that a simple transformation of the survival times may facilitate modeling. Suppose that T is a continuous random variable having distribution function F. Then $U = F(T)$ has a standard uniform distribution and $\log(U) = \Lambda(T)$, where Λ represents the cumulative hazard function, has a standard exponential distribution and thus a constant hazard function. In the context of hazard function modeling with HARE, the regression model applied to survival times transformed by the marginal cumulative hazard function tends to require fewer knots applied to the time variable allowing more focus on the impact of predictors on the (transformed) outcome. The overall transformation applied to the data can be semiparametric, for instance using a regression spline model for the hazard function (the HEFT method of Kooperberg et al., 1995) or non-parametric using the empirical cumulative hazard function estimate. This transformation can facilitate the use of other flexible regression procedures utilizing exponential model likelihood, which typically allows for much faster computation than partial likelihood. For instance, after transforming the survival times by the cumulative hazard transformation, the survival times may be sufficiently well approximated by an exponential distribution, so that a regression tree program based on the exponential likelihood may perform well.

3.9.1 Example Revisited

The multiple myeloma data set included both overall survival and progression-free survival endpoint data. In this analysis, we consider all 778 subjects with complete covariate data. The HARE analysis of the time to progression is very similar to the Polyclass analysis presented in Section 3.6.3. The analysis of the survival time turned out more interesting, as it depended on serum β_2 microglobulin, anyca, log(ldh), and age, and included a nonproportionality component for log(ldh). In Figure 3.10, we show the fitted hazard function for a person of age 56, with a log(b2m) of 1, no anyca, and log(ldh) values of 4, 5, and 6, which roughly correspond to the 25th, 50th, and 75th percentile of the log(ldh).

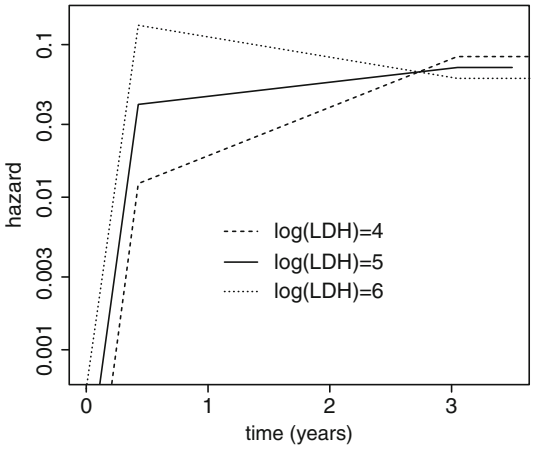

Fig. 3.10 Fitted hazard functions for the HARE analysis of the multiple myeloma data.

3.10 Discussion

Many choices exist for flexible regression modeling of patient data from oncology studies. Selection of appropriate methods, of course, depends on the goals in the particular analysis. For instance, it could be best to characterize the risk of progression as a smooth function of a single important prognostic variable or to develop a more general risk models using multiple predictors and variable selection. Adaptive regression spline methods such as HARE are well suited to such problems. Alternatively, one may want to characterize groups of patients or subjects, or identify interactions of binary predictor variables. Tree-based methods or logic regression are two tools useful for such problems.

A common aspect of cancer data is that the strength of associations between predictors and patient outcome is quite weak as demonstrated with the myeloma data. While sometimes it is useful to slightly overfit the data to suggest models that may be worth investigating further, in general we should prevent selecting regression models that are not supported by the data. Therefore, using methods to obtain honest prediction error to help avoid over-fitting is critical.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Barlogie, B., Tricot, G., Rasmussen, E., Anaissie, E., van Rhee, F., Zangari, M., Fassas, A., Hollmig, K., Pineda-Roman, M., Shaughnessy, J., Epstein, J., and Crowley, J. (2006). Total therapy 2 without thalidomide in comparison with total therapy 1: role of intensified induction and posttransplantation consolidation therapies. *Blood*, 107:2633–2638.

Board on health sciences policy. (2002). *Cancer and the Environment: Gene–Environment Interaction.* Institute of Medicine, National Academy Press, Washington D. C.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.
Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.
Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Pacific Grove, California.
Ciampi, A., Thiffault, J., Nakache, J. P., and Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition. *Computational Statistics and Data Analysis*, 4:185–204.
Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34:187–220.
Dai, J. Y., LeBlanc, M., and Kooperberg, C. (2008). Semiparametric estimation exploiting covariate independence in two-phase randomized trials. *Biometrics*, May 12. [Epub ahead of print].
Davis, R. B. and Anderson, J. R. (1989). Exponential survival trees. *Statistics in Medicine*, 8:947–961.
Dettling, M. and Bühlmann, P. (2002). Supervised clustering of genes. *Genome Biology*, 3:69.1–69.15.
Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447:1087–1093.
Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32:407–499.
Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11:89–121.
Farag, S., Archer, K. K., Mrózek, K., Ruppert, A. S., Carroll, A. J., Vardiman, J. W., Pettenati, J., Baer, M. R., Qumsiyeh, M. B., Koduru, P. R., et al. (2006). Pretreatment cytogenetics add to other prognostic factors predicting complete remission and long-term outcome in patients 60 years of age or older with acute myeloid leukemia: results from Cancer and Leukemia Group B 8461. *Blood*, Jul. 1; 108(1):63–73.
Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Tirteenth International Conference*, pp. 148–156. Morgan Kauffman, San Francisco.
Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, 19:1–141.
Friedman, J. H. and Fisher, N. I. (1999). Bump-hunting for high dimensional data. *Statistics and Computation*, 9:123–143.
Friedman, J. H., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *The Annals of Statistics*, 38:337–407.
Gimotty, P. A., Elder, D. E., Fraker, D. L., Botbyl, J., Sellers, K., Elenitsas, R., Ming, M. E., Schuchter, L., Spitz, F. R., Czerniecki, B. J., and Guerry, D. (2007). Identification of high-risk patients among those diagnosed with thin cutaneous melanomas. *Journal of Clinical Oncology*, 25:1129–1134.
Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports*, 69:1065–1069.
Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman and Hall, London.
Greipp, P. R., San Miguel, J., Durie, B. G., Crowley, J. J., Barlogie, B., Bladé, J., Boccadoro, M., Child, J. A., Avet-Loiseau, H., Kyle, R. A., et al. (2005). International staging system for multiple myeloma. *Journal of Clinical Oncology*, 23:3412–3420.
Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.
Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. (2001). Supervised harvesting of regression trees. *Genome Biology*, 2:3.1–3.12.
Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association*, 90:78–94.

Kooperberg, C., Bose, S., and Stone, C. J. (1997). Polychotomous regression. *Journal of the American Statistical Association*, 92:117–127.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88:457–467.

LeBlanc, M., Moon, J., and Crowley, J. (2005). Adaptive risk group refinement. *Biometrics*, 61:370–378.

LeBlanc, M., Moon, J., and Kooperberg, C. (2006). Extreme regression. *Biostatistics*, 13:106–122.

Loader, C. (1999). *Local Regression and Likelihood*. Springer-Verlag, New York.

London, W. B., Castleberry, R. P., Matthay, K. K., Look, A. T., Seeger, R. C., Shimada, H., Thorner, P., Broderu, G., Maris, J. M., Reynolds, C. P., and Cohn, S. L. (2005). Evidence for an age cutoff greater than 365 days for neuroblastoma risk group stratification in the Children's Oncology Group. *Journal of Clinical Oncology*, 23:6459–6465.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2004). On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9:363–379.

Park, M. Y. and Hastie, T. (2006). L_1 regularization path models for generalized linear models. *Journal of the Royal Statistical Society B*, page in press.

Polesel, J., Dal Maso, L., Bagnardi, V., Zucchetto, A., Zambon, A., Levi, F., La Vecchia, C., and Franeschi, S. (2005). Estimating dose-response relationship between ethanol and risk of cancer using regression spline models. *International Journal of Cancer*, 114:836–841.

Quantin, C., Abrahamowicz, M., Moreau, T., Bartlett, G., MacKenzie, T., Tazi, M. A., Lalonde, L., and Faivre, J. (1999). Variation over time of the effects of prognostic factors in a population-based study of colon cancer: Comparison of statistical models. *American Journal of Epidemiology*, 150:1188–1200.

Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*. Morgan Kaufman, San Francisco, CA.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley, New York.

Rosenwald, A., Wright, G., Chan, W., Connors, J., Campo, D., Fisher, R., Gascoyne, R., Muller-Hermelink, H., Smeland, E., Staudt, L., et al. (2002). Molecular diagnosis and clinical outcome prediction in diffuse large B-cell lymphoma. *New England Journal of Medicine*, 346:1937–1947.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics*, 12:475–511.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, 44:35–47.

Singer, E. (2005). Personalized medicine prompts push to redesign clinical trials. *Nature*, 452:462.

Smith, P. L. (1982). Curve fitting and modeling with splines using statistical variable selection methods. Technical report, NASA, Langley Research Center, Hampla, Virginia.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *The Annals of Statistics*, 22:118–184.

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*, 25:1371–1470.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288.

Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P., Wacholder, S., Minichiello, M. J., Fearnhead, P., Yu, K., Chatterjee, N., et al. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics*, 39:645–649.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320.