Chapter 4

Regularization with Ridge penalties, the Lasso, and the Elastic Net for Regression with Optimal Scaling Transformations

Regularized regression methods for linear regression have been developed the last few decades to overcome the flaws of ordinary least squares regression with regard to prediction accuracy. In this chapter, three of these methods (Ridge regression, the Lasso, and the Elastic Net) are incorporated into CATREG, an optimal scaling method for both linear and nonlinear transformation of variables in regression analysis. We show that the original CATREG algorithm provides a very simple and efficient way to compute the regression coefficients in the constrained models for Ridge gression, the Lasso, and the Elastic Net. The resulting procedures, subsumed under the term "regularized nonlinear regression" will be illustrated using the prostate cancer data, which have previously been analyzed in the regularization literature for linear regression. For model selection and the estimation of the prediction accuracy, we used the .632 bootstrap with the one-standard-error rule. We also show that the "CATREG-Lasso" with nominal transformations is equivalent to the

This chapter has been submitted for publication as Van der Kooij, A.J. & Meulman, J.J. (2006). Regularization with Ridge penalties, the Lasso, and the Elastic Net for Regression with Optimal Scaling Transformations.

recently developed methods “Group Lasso” and “Blockwise Sparse Regression” for nominal data using dummy variables. These methods as well as the “CATREG-Lasso” shrink nominal variables as a whole. A real data set is used to compare the results of shrinking nominal variables as a whole to the results of shrinking dummy variables, which boils down to shrinking the optimal quantifications of the categories. Finally, for a real data set three analytic model selection methods (AIC, BIC, and GCV) are compared to the nonparametric .632 bootstrap for model selection.

4.1. Introduction

Multiple regression is often used to estimate a model for predicting future responses, or to investigate the relationship between the response variable and the predictor variables. For the first goal the prediction accuracy of the model is important, for the second goal the complexity of the model is of more interest. Ordinary least squares (OLS) regression is known for often not performing well with respect to both prediction accuracy and model complexity. Several regularized regression methods were developed the last few decades to overcome these flaws of OLS regression, starting with Ridge regression (Hoerl and Kennard 1970a,b), followed by Bridge regression (Frank and Friedman 1993), the Garotte (Breiman 1995), and the Lasso (Tibshirani 1996), and more recently LARS (Efron, Hastie, Johnstone, and Tibshirani 2004), Pathseeker (Friedman and Popescu 2004), and the Elastic Net (Zou and Hastie 2005). In this chapter, we focus on Ridge regression, the Lasso, and the Elastic Net.

OLS regression may result in highly variable estimates of the regression coefficients in the presence of collinearity or when the number of predictors (P) is large relative to the number of observations (N). Ridge regression reduces this variability by shrinking the coefficients, resulting in more prediction accuracy at the cost of usually only a small increase of bias. In Ridge regression, the coefficients are shrunken towards zero, but will never become exactly zero. So, when the number of predictors is large, Ridge regression will not provide a sparse model that is easy to interpret. Subset selection, on the other hand, does provide interpretable models, but does not reduce the variability of the estimates of the coefficients. While not reducing the variability of the coefficient estimates of the selected variables, subset selection can reduce the variability of the prediction estimates, but not as much as Ridge regression or the Lasso. The Lasso was developed by Tibshirani (1996) to improve both predition accuracy and model interpretability by combining the nice features of Ridge regression and subset selection. The Lasso reduces the variability of the estimates by shrinking the coefficients and at the same time produces interpretable models by shrinking some coefficients to exacly zero. In Tibshirani (1996), it was demonstrated that in terms of prediction accuracy and interpretability, the Lasso outperforms Ridge regression and subset selection for data with a small to moderate number of moderate-sized effects; subset selection performs the best with a small number of large effects, and Ridge regression performs the best with a large number of small effects.

Recently, Zou and Hastie (2005) proposed the Elastic Net to overcome the limitations of the Lasso in some situations. The Elastic Net also combines shrinkage and variable selection, and in addition encourages grouping of variables: groups of highly correlated variables tend to be selected together, where the Lasso would only select one variable of the group. Also, in the case $P \gg N$, Lasso algorithms are limited because at most N variables can be selected. Zou and Hastie (2005) conjecture that, whenever Ridge regression improves on OLS, the Elastic Net will improve the Lasso.

Ridge regression, the Lasso, and the Elastic Net are regularization methods for linear models. In this chapter, we implement these three methods in CATREG, an algorithm that incorporates linear and nonlinear transformation of the variables. With respect to nonlinear transformations, CATREG transforms variables monotonically or non-monotonically, using either step functions or spline functions. Recently, a method was developed to simultaneously regularize and transform variables non-monotonically (Yuan and Lin (2006) and Kim et al. (2006)). This method expands the variables to blocks (dummy variables for categorical variables; basis functions for continuous variables) and applies the regularization to groups (blocks) of variables. We will show that with CATREG the same is achieved without the need to expand a variable to a group of dummy variables or basis functions.

4.2. Ridge penalties, the Lasso, and the Elastic Net for linear regression

The loss functions for Ridge regression, the Lasso, and the Elastic Net can be viewed as constrained versions of the ordinary least squares (OLS) regression loss function. In Ridge regression, the *sum of squares* of the coefficients is constrained as follows:

$$L^{\text{ridge}}(\beta_1, \ldots, \beta_P) = \|\mathbf{y} - \sum_{j=1}^{P} \beta_j \mathbf{x}_j\|^2, \text{ subject to } \sum_{j=1}^{P} \beta_j^2 \leq t_2, \tag{4.1}$$

with N the number of observations, P the number of predictor variables, $\beta_j, j = 1, \ldots, P$, the regression coefficients, and t_2 the Ridge tuning parameter, and where $\|\cdot\|^2$ denotes the squared Euclidean norm. The Lasso constrains the *sum of the absolute values* of the coefficients:

$$L^{\text{lasso}}(\beta_1, \ldots, \beta_P) = \|\mathbf{y} - \sum_{j=1}^{P} \beta_j \mathbf{x}_j\|^2, \text{ subject to } \sum_{j=1}^{P} |\beta_j| \leq t_1, \tag{4.2}$$

with t_1 the Lasso tuning parameter. Finally, the Elastic Net combines the Ridge regression and the Lasso constraints:

$$\begin{aligned} L^{\text{e-net}}(\beta_1, \ldots, \beta_P) &= \|\mathbf{y} - \sum_{j=1}^{P} \beta_j \mathbf{x}_j\|^2, \text{ subject to} \\ & \sum_{j=1}^{P} \beta_j^2 \leq t_2 \text{ and } \sum_{j=1}^{P} |\beta_j| \leq t_1. \end{aligned} \tag{4.3}$$

These constrained loss functions can also be written as penalized loss functions:

$$L^{\text{ridge}}(\beta_1, \ldots, \beta_P) = \|\mathbf{y} - \sum_{j=1}^{P} \beta_j \mathbf{x}_j\|^2 + \lambda_2 \sum_{j=1}^{P} \beta_j^2, \tag{4.4}$$

$$L^{\text{lasso}}(\beta_1, \ldots, \beta_P) = \|\mathbf{y} - \sum_{j=1}^{P} \beta_j \mathbf{x}_j\|^2 + \lambda_1 \sum_{j=1}^{P} \text{sign}(\beta_j)\beta_j, \tag{4.5}$$

$$\begin{aligned} L^{\text{e-net}}(\beta_1, \ldots, \beta_P) &= \|\mathbf{y} - \sum_{j=1}^{P} \beta_j \mathbf{x}_j\|^2 + \\ & \lambda_2 \sum_{j=1}^{P} \beta_j^2 + \lambda_1 \sum_{j=1}^{P} \text{sign}(\beta_j)\beta_j, \end{aligned} \tag{4.6}$$

with λ_2 the Ridge penalty parameter, penalizing the sum of the squared regression coefficients and λ_1 the Lasso penalty, penalizing the sum of the absolute values of the regression coefficients. In matrix notation, the penalized loss functions are written as

$$L^{\text{ridge}}(\beta_1, \ldots, \beta_P) = \|\mathbf{y} - \mathbf{Xb}\|^2 + \lambda_2 \mathbf{b}'\mathbf{b}, \tag{4.7}$$

$$L^{\text{lasso}}(\beta_1, \ldots, \beta_P) = \|\mathbf{y} - \mathbf{Xb}\|^2 + \lambda_1 \mathbf{w}'\mathbf{b}, \tag{4.8}$$

$$L^{\text{e-net}}(\beta_1, \ldots, \beta_P) = \|\mathbf{y} - \mathbf{Xb}\|^2 + \lambda_2 \mathbf{b}'\mathbf{b} + \lambda_1 \mathbf{w}'\mathbf{b}, \tag{4.9}$$

where the elements w_j of $\mathbf{w}$ are either $+1$ or -1, depending on the sign of the corresponding regression coefficient β_j.

Minimization of (4.7) with respect to $\mathbf{b}$ has an analytic solution; the constrained coefficients in Ridge regression are obtained as

$$\mathbf{b}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}. \tag{4.10}$$

For the Lasso, however, minimization of the constrained loss function is more complicated. The regression coefficients are estimated as

$$\mathbf{b}^{\text{lasso}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - \frac{\lambda_1}{2}\mathbf{w}), \tag{4.11}$$

and this is a least squares problem with 2^P inequality constraints (there are 2^P possible sign patterns for the coefficients). In Tibshirani (1996), a quadratic programming algorithm is used to estimate the Lasso coefficients. This is a complex and computationally demanding procedure, and is hence not feasible for large values of P. Less complex and/or more efficients algorithms were developed by a.o. Fu (1998), Osborne, Presnell, and Turlach (2000), Perkins, Lacker, and Theiler (2003), Friedman and Popescu (2004), and Zhao and Yu (2004). The efficient LARS algorithm of Efron et al. (2004) finds the entire Lasso regularization paths with the computational effort of a single OLS fit, but the algorithm can not be applied when $P > N$, neither can the original Lasso algorithm of Tibshirani (1996) nor the “shooting” algorithm of Fu (1998). The algorithm of Osborne et al. (2000) is an improved quadratic programming algorithm that can handle $P > N$ predictors, but is still computationally demanding when P is large. The “Grafting” algorithm of Perkins et al. (2003), the “Pathseeker” algorithm of Friedman and Popescu (2004), and the “boosting” algorithm of Zhao and Yu (2004) are gradient descent algorithms, that can deal with $P > N$ predictors in a computationally less demanding way.

Zou and Hastie (2005) have proposed the Elastic Net and developed an algorithm, called LARS-EN, based on the efficient LARS algorithm, to overcome the Lasso limitations of selecting at most N predictors and of selecting only one predictor from a group of highly correlated predictors. For the Elastic Net the regression coefficients are estimated as

$$\mathbf{b}^{\text{e-net}} = (\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I})^{-1}(\mathbf{X}'\mathbf{y} - \frac{\lambda_1}{2}\mathbf{w}), \tag{4.12}$$

Minimization of this loss function is much like minimizing the Lasso loss function and the entire Elastic Net regularization paths can be estimated almost as efficiently as the Lasso paths with the LARS-EN algorithm (Zou and Hastie 2005).

All existing Ridge and Lasso algorithms and the LARS-EN algorithm are developed for linear regression. In the next section, we will show that Ridge regression, the Lasso, and the Elastic Net can easily be incorporated into the CATREG algorithm, resulting in a simple and efficient algorithm for linear regression as well as for nonlinear regression (to the extent one would regard the original CATREG algorithm to be simple and efficient). Also, with the CATREG regularization algorithm the Lasso can select more than N predictors in the $P \gg N$ case. In contrast to the existing Lasso algorithms, that find the Lasso coefficient paths in an iterative way (except for LARS-Lasso), CATREG-Lasso estimates the Lasso coefficient paths straightforwardly, but does so in the context of the CATREG backfitting algorithm, which is iterative.

4.3. Ridge penalties, the Lasso, and the Elastic Net with CATREG

The major motivation for the CATREG algorithm has been to include nonlinear transformations of both the predictors and the response variable in the regression model. Since the nonlinear transformations are not fixed, but have to be optimized, we have to update the coefficient and the transformation for one variable at a time. This is done by an algorithm that was first proposed in Kruskal (1965), subsequently applied in psychometrics in De Leeuw et al. (1976) and Gifi (1990) and in statistics in Breiman and Friedman (1985), Buja et al. (1989), and Hastie and Tibshirani (1990), labeled *backfitting* in Friedman and Stuetzle (1981). The CATREG loss function is written as

$$L(\varphi(\cdot); \beta_1, \ldots, \beta_P) = N^{-1} \|\varphi_r(\mathbf{y}) - \sum_{j=1}^{P} \beta_j \varphi_j(\mathbf{x}_j)\|^2, \tag{4.13}$$

where $\varphi_r(\mathbf{y})$ denotes the transformation of the response variable $\mathbf{y}$ and $\varphi_j(\mathbf{x}_j)$ the transformation for a predictor variable $\mathbf{x}_j$, with $j = 1, \ldots, P$. The loss function is minimized by iteratively estimating β_j and $\varphi_j(\mathbf{x}_j)$ for one variable at a time, keeping the estimates for the other variables $l \neq j$ fixed.

By rewriting (4.13), with fixed $\varphi_r(\mathbf{y})$, β_l, and $\varphi_l(\mathbf{x}_l)$ for all predictors $l \neq j$ as

$$L(\beta_j; \varphi_j) = N^{-1} \|\varphi_r(\mathbf{y}) - \sum_{l \neq j} \beta_l \varphi_l(\mathbf{x}_l) - \beta_j \varphi_j(\mathbf{x}_j)\|^2, \tag{4.14}$$

and setting partial derivatives in (4.14) with respect to β_j to zero, the updated estimate β_j^+ for the coefficient for variable j becomes

$$\beta_j^+ = N^{-1}(\varphi_j(\mathbf{x}_j))'(\varphi_r(\mathbf{y}) - \sum_{l \neq j} \beta_l \varphi_l(\mathbf{x}_l)). \tag{4.15}$$

At this point, we would estimate the optimal transformation $\varphi_j(\mathbf{x}_j)$, and move on to the next variable. After a loop over all variables, we obtain the transformation of the response variable $\varphi_r(\mathbf{y})$, and compute the squared multiple regression coefficient (R^2). The algorithm converges to a stationary point, and we continue the updating process until the difference in R^2 from one iteration to the next is below a preset convergence criterion.

To show how the three different regularization procedures are easily incorporated into the CATREG (backfitting) algorithm, we will only consider linear transformations in the next two subsections, so that $\varphi_r(\mathbf{y})$ and the $\varphi_j(\mathbf{x}_j)$ contain standardized scores, and only the regularized regression coefficients need to be estimated.

4.3.1 Updating the regularization regression coefficients in CATREG

Because we can remove the linear transformation from the loss function by applying the transformation before the minimization process, we here assume that $\mathbf{y}$ and $\mathbf{x}_j$ are standardized variables and use the loss functions given in (4.4), (4.5), and (4.6). These loss functions can be partitioned, analoguously to (4.13) and (4.14), fixing and separating all terms that apply to variables $l \neq j$ from the terms that involves only variable j. The CATREG versions of Ridge regression, Lasso, and Elastic Net are then written as

$$\begin{aligned}
L^{\text{ridge}}(\beta_j) &= \|\mathbf{y} - \sum_{l \neq j} \beta_l \mathbf{x}_l - \beta_j \mathbf{x}_j)\|^2 + \lambda_2 \beta_j^2 + \lambda_2 \sum_{l \neq j} \beta_l^2, && (4.16)\\
L^{\text{lasso}}(\beta_j) &= \|\mathbf{y} - \sum_{l \neq j} \beta_l \mathbf{x}_l - \beta_j \mathbf{x}_j)\|^2 + \lambda_1 w_j \beta_j + \lambda_1 \sum_{l \neq j} w_l \beta_l, && (4.17)\\
L^{\text{e-net}}(\beta_j) &= \|\mathbf{y} - \sum_{l \neq j} \beta_l \mathbf{x}_l - \beta_j \mathbf{x}_j)\|^2 + \lambda_2 \beta_j^2 + \lambda_1 w_j \beta_j + \\
&\quad \lambda_2 \sum_{l \neq j} \beta_l^2 + \lambda_1 \sum_{l \neq j} w_l \beta_l, && (4.18)
\end{aligned}$$

where w_l and w_j are either $+1$ or -1 depending on the sign of the corresponding β_l and β_j. Because the effect of the $l \neq j$ other predictors has been removed from the response, the contribution of the jth predictor is corrected for the contribution of the $l \neq j$ predictors, and thus the estimate of the constrained regression coefficient for the jth predictor can simply be updated as

$$\begin{aligned} \beta_j^{+\text{ridge}} &= \beta_j^{+}/(1+\lambda_2), & (4.19)\\ \beta_j^{+\text{lasso}} &= (\beta_j^{+} - \frac{\lambda_1}{2}w_j)_+ \\ &= (\beta_j^{+} - \frac{\lambda_1}{2})_+ \text{ if } \beta_j^{+} > 0 \\ &= (\beta_j^{+} + \frac{\lambda_1}{2})_+ \text{ if } \beta_j^{+} < 0, & (4.20)\\ \beta_j^{+\text{e-net}} &= (\beta_j^{+} - \frac{\lambda}{2}w_j)_+ \\ &= \frac{(\beta_j^{+} - \frac{\lambda_1}{2}}{1+\lambda_2})_+ \text{ if } \beta_j^{+} > 0 \\ &= \frac{(\beta_j^{+} + \frac{\lambda_1}{2}}{1+\lambda_2})_+ \text{ if } \beta_j^{+} < 0. & (4.21) \end{aligned}$$

with β_j^{+}, equivalent to (4.15), defined as $\beta_j^{+} = N^{-1}\mathbf{x}_j'(\mathbf{y} - \sum_{l\neq j}\beta_l\mathbf{x}_l)$ and $(\cdot)_+$ denotes truncation at zero: when $\beta_j^{+} > 0$ and $\beta_j^{+\text{lasso}} < 0$, or when $\beta_j^{+} < 0$ and $\beta_j^{+\text{lasso}} > 0$, $\beta_j^{+\text{lasso}}$ is set to zero. The double amount of shrinkage in the estimation of the Elastic Net regression coefficients is corrected by rescaling the coefficients after convergence:

$$\beta_j^{+\text{e-net}} = \beta_j^{*\text{e-net}}(1+\lambda_2). \tag{4.22}$$

4.3.2 Paths for the coefficients

The varying size of the penalty parameter λ from ∞ to zero, determines a path for each of the regression coefficients. The Ridge and Lasso paths can be found by repeatedly applying the algorithm, starting with an initial value for λ high enough to exclude all predictors, and then gradually decrease λ to zero, at wich value the OLS estimates for the coefficients are obtained. For the Elastic Net, multiple paths for a predictor are created by gradually decreasing λ_1 for a (relatively small) number of fixed values of λ_2. The location on the Lasso paths where the models that contains k predictors (with k in between 1 and P, and indicating the number of predictors in the model) changes to a model that contains $k+1$ or $k-1$ predictors is called a *transition point*. To

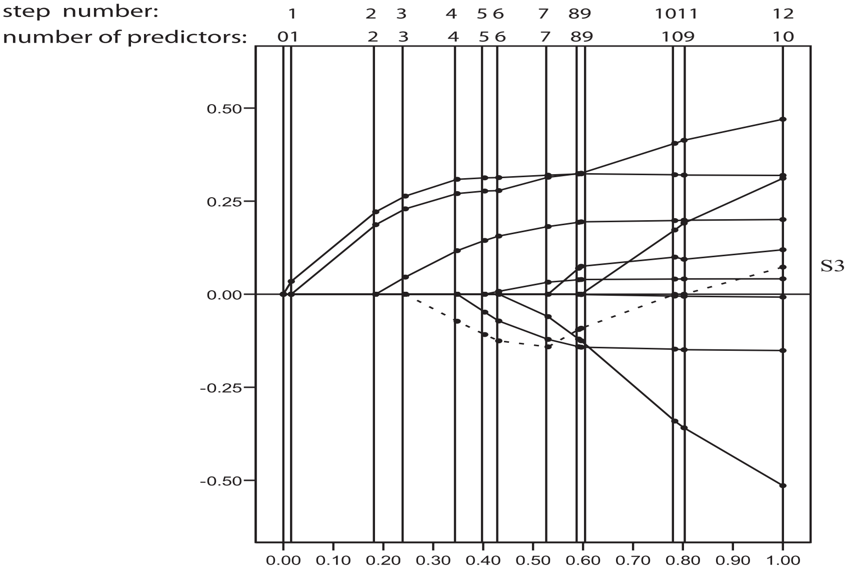

Figure 4.1. Lasso for diabetes data. The constrained coefficients on the vertical axis versus $s = \sum_{j=1}^{P} |\beta_j^{lasso}| / \sum_{j=1}^{P} |\beta_j^{ols}|$*. The vertical lines indicate the transition points; at the top the number of active predictors is given. The values of* λ_1 *at the transition points are 1.17, 1.10, 0.560, 0.400, 0.164, 0.111, 0.085, 0.028, 0.0087, 0.0072, 0.0032, 0.0020, and 0.*

ensure that all transition points are included in the path, a small enough step size should be used when λ is decreased. It may happen that there are more than P transition points because regression coefficients can cross the zero line.

The Lasso paths are illustrated in Figure 4.1 for the diabetes data (Efron et al. 2004 (downloaded from http://www-stat.stanford.edu/~hastie/Papers/LARS/diabetes.data). The data concern observations for 442 patients on ten baseline variables, and the outcome variable is a quantitative measure of disease progression one year after the baseline variables were measured. (For the Elastic Net, we would obtain multiple plots, one for each fixed value of λ_2.) The number of transition points for these data is $10 + 2 = 12$, because variable S3 (the dotted path) starts being active at step 3 with a negative coefficient, obtains a zero coefficient at step 11, and becomes active again after step 11 with a positive regression coefficient.

The paths for the Lasso and the Elastic Net are linear in between two transition points (Efron et al. 2004). Therefore, these paths can be found much more efficiently, because it suffices to find the transition points only. This is what the Lasso option of the LARS algorithm (Efron et al. 2004) achieves, by using correlations and equi-angular vectors. We have incorporated the LARS-Lasso approach into CATREG, but it turned out not to work when nonlinear transformations of the variables are called for. Here, the paths need to be found by repeatedly applying the algorithm with different values of λ_1. Although less efficient than the LARS-Lasso method, the CATREG-Lasso limits the number of repeats by using a new method for finding the transition points.

4.3.3 Finding the transition points in CATREG

We define a transition point as the point where the *slope* of the paths changes. The slope of piece k of the paths for Lasso models with k active predictors is given by

$$\gamma_{A_k}^{lasso} = -(\mathbf{X}'_{A_k}\mathbf{X}_{A_k})^{-1}\frac{1}{2}\mathbf{w}_{A_k}, \tag{4.23}$$

where A_k is the set of indices of the k predictors active in the model, and where the columns of $\mathbf{X}_{A_k}$ contain the standardized variables $\mathbf{x}_j$, for $j \in A_k$. Note that we need to compute the slopes of the path of the coefficients for *one* predictor only, because the transition points are the same for all paths. Using the slopes of the path for the coefficients of the predictor in A_1, we only need to invert $x'_{A_1}x_{A_1}$ to find the slope for the first piece of the paths. For the subsequent slopes, in stead of inverting the growing cross-product matrices $\mathbf{X}'_{A_k}\mathbf{X}_{A_k}$, we update $\mathbf{R}_{k-1}$ resulting from the Cholesky factorization of $(\mathbf{X}'_{A_{k-1}}\mathbf{X}_{A_{k-1}})$ in the previous step (Golub and Van Loan 1983). For the Elastic Net, the slopes for a fixed value of λ_2 are

$$\gamma_{A_k}^{e-net} = -(\mathbf{X}'_{A_k}\mathbf{X}_{A_k} + \lambda_2)^{-1}\frac{1}{2}\mathbf{w}_{A_k}(1+\lambda_2). \tag{4.24}$$

The intercepts for pieces of the path can be computed from the slopes as

$$\alpha_k = \beta_k - \lambda_k\gamma_k, \tag{4.25}$$

and the value of the Lasso penalty λ_1 at the point where transition takes place from a model with $k-1$ active variables to a model with k active variables, is the value where the the pieces $k-1$ and k of the path join:

$$\lambda_{1_k} = \frac{\alpha_{k-1} - \alpha_k}{\gamma_k - \gamma_{k-1}}. \tag{4.26}$$

Thus, to find the transition points, P solutions (or more than P if coeffients cross the zero line) have to be found: a solution for a model with only one active predictor, next a solution for a model with two active predictors, up to a solution with all P predictors active in the model. The solutions are found by decreasing the initial high value of λ_1 in rather big steps, and if this results in a model with more than one additional predictor compared to the model in the previous step (or more than one predictor less), the previous value of λ_1 is increased using a smaller stepsize, until a model is obtained with only one additional predictor compared to the number obtained in the previous step (or only one predictor less).

4.3.4 Including nonlinear transformations in the CATREG algorithm

When nonlinear transformations are called for, the estimation of the transformation functions $\varphi(\cdot)$ is not affected by the penalty terms that are added to the OLS loss function to attain the regularization. (How these transformatons are obtained, has been fully described in Chapter 2.) So, incorporating Ridge regression, the Lasso, and the Elastic Net in CATREG with nonlinear transformations, only requires the same slight adjustment of the OLS regression coefficient estimates as for linear regression with CATREG: (4.19), (4.20), and (4.21) also apply when nonlinear transformations are involved, the difference is only that now β_j^+ is defined as in (4.15). Thus, to the extent the CATREG algorithm is simple and efficient, we have a simple and efficient algorithm to estimate Ridge, Lasso and Elastic Net regression coefficients for *both* linear and nonlinear regression. In contrast to the Lasso paths resulting from regularized linear regression, the paths for regularized regression with nonlinear transformations are not piecewise linear. So, when including nonlinear transformations, the paths have to be found by repeatedly applying the algorithm, starting with a high penalty value and stepwise decreasing the value.

4.4. Selection of the optimal penalty parameter

Selecting the optimal value of the penalty parameter is equivalent to selecting the optimal value of the tuning parameter t in (4.1) - (4.3). We will use the .632 bootstrap method (Efron 1983), which is essentially a smoothed version of leave-one-out cross validation. The details of how to use the .632 bootstrap with CATREG are described in Van der Kooij and Meulman (2006a). Using the .632 bootstrap for model selection is time consuming, because it has to be repeated for a lot of models on the paths. Major advantages of the .632 bootstrap over other, analytic, methods to select λ, are that it does not require the estimation of the degrees of freedom involved, and that it also works when

$P \gg N$. (Some analytic selection methods are described in the Discussion section.)

4.4.1 Illustration

For an illustration we applied regularized CATREG[1] to the prostate cancer data from Stamey et al. (1989) (obtained from http://www-stat.stanford.edu/~tibs/ElemStatLearn/), consisting of 97 observations on eight predictors to predict (the log of) the prostate specific antigen measure. The predictors are (1) log(cancer volume) (lcavol), (2) log(prostate weight) (lweight), (3) age, (4) log(benign prostatic hyperplasia amount) (lbph), (5) seminal vesicle invasion (svi), (6) log(capsular penetration) (lcp), (7) Gleason score (gleason), and (8) percentage Gleason scores 4 or 5 (pgg45). The variables svi and gleason are categorical; all other variables are continuous. For svi and gleason, we fitted a nonmonotonic step function, and for lcavol, lweight, age, lbph, lcp, and pgg45, an optimal nonmonotonic spline transformation, using second degree polynomials with two interior knots. The response variable was linearly transformed to standard scores. We used the .632 bootstrap with 200 samples for model fitting and estimation of prediction accuracy for model selection.

In Table 4.1, estimates of the *generalization* error (the error when applying the selected model to a test set) are given, both for linear and nonlinear Ridge, Lasso, and Elastic Net models. For the estimation of the generalization error, the data set was divided into a training set ($N = 67$) and a test test ($N = 30$). The models were selected by applying the .632 bootstrap and the one-standard-error rule: the most parsimomuous model within one standard error of the minimum was selected[2]. It is clear that all regularization methods improve the prediction accuracy compared to no shrinking (OLS). The Elastic Net performs better than the Lasso, and the Lasso performs better than Ridge regression. When nonlinear transformations are included, the Elastic Net again performs best. Interestingly, here Ridge shrinking does not improve the prediction accuracy compared to OLS, so we could conjecture that nonlinear transformation is a form of shrinking in itself.

[1] Ridge regression, the Lasso, and the Elastic Net have actually been incorporated by adapting the version of CATREG that is available through SPSS (Meulman et al. 1999, 2004). In this version, the variables are assumed to be categorical (hence the name CATREG). However, a straightforward way is provided to allow continuous variables in the analysis by an internal procedure that digitizes continuous data by a linear transformation.

[2] The results for the linear models are very similar to the results reported in Zou and Hastie (2005), except for the Elastic Net. Zou and Hastie (2005) report the value .381 as the test mean squared error with optimal $\lambda_2 = 1000$. The CATREG-E-net mean squared error (.378) for $\lambda_2 = 1000$ is close, but the .632 bootstrap selects a different optimal value for λ_2.

Table 4.1. Generalization error for prostate cancer data.

Method	Model: λ / s	Test mean squared error (SE)	Selected predictors[1]
Linear			
OLS	0.00 / 1.00	0.589 (0.105)	all
Ridge	0.62 / 0.29	0.554 (0.066)	all
Lasso	0.36 / 0.37	0.505 (0.104)	1,2,5
Elastic Net	0.43 / 0.52 $\lambda_2 = 1$	0.441 (0.065)	1,2,4,5,6,8
Nonlinear			
OLS	0.00 / 1.00	0.472 (0.137)	all
Ridge	1.65 / 0.11	0.477 (0.084)	all
Lasso	0.40 / 0.31	0.411 (0.078)	1,2,5,8
Elastic Net	0.75 / 0.28 $\lambda_2 = 10$	0.348 (0.078)	1,2,5,6,7,8
Results Zou & Hastie (2005)			
OLS		0.586 (0.184)	all
Ridge	1.00 /	0.566 (0.188)	all
Lasso	/ 0.39	0.499 (0.161)	1,2,4,5,8
Elastic Net	/ 0.26 $\lambda_2 = 1000$	0.381 (0.105)	1,2,5,6,8

[1] The predictors are (1) log(cancer volume) (lcavol), (2) log(prostate weight) (lweight), (3) age, (4) log(benign prostatic hyperplasia amount) (lbph), (5) seminal vesicle invasion (svi), (6) log(capsular penetration) (lcp), (7) Gleason score (gleason), and (8) percentage Gleason scores 4 or 5 (pgg45).

Figure 4.2 displays the paths and the selected models for Ridge, Lasso, and Elastic Net regularization, both linear panels at the top) and nonlinear (panels at the bottom). We obtained the nonlinear regularization paths and the linear Ridge paths by stepwise decreasing the penalty. The paths for the linear Lasso and the linear Elastic Net were obtained by finding the transition points, and connecting them. For both linear and nonlinear regression, we determined the minimum by applying the .632 bootstrap for many different points on the path. Therefore, it may happen that the vertical line that

Running the cv.enet function in the Elastic Net R-package (Zou and Hastie, downloaded from http://cran.r-project.org/src/contrib/PACKAGES.html) with different values of λ_2 being $0, 1, 10, 100$ and 1000, several times reveals that sometimes 1000 is selected as the optimal value for λ_2 (applying the one-standard-error rule) but most other times the values 1 or 10 are selected.

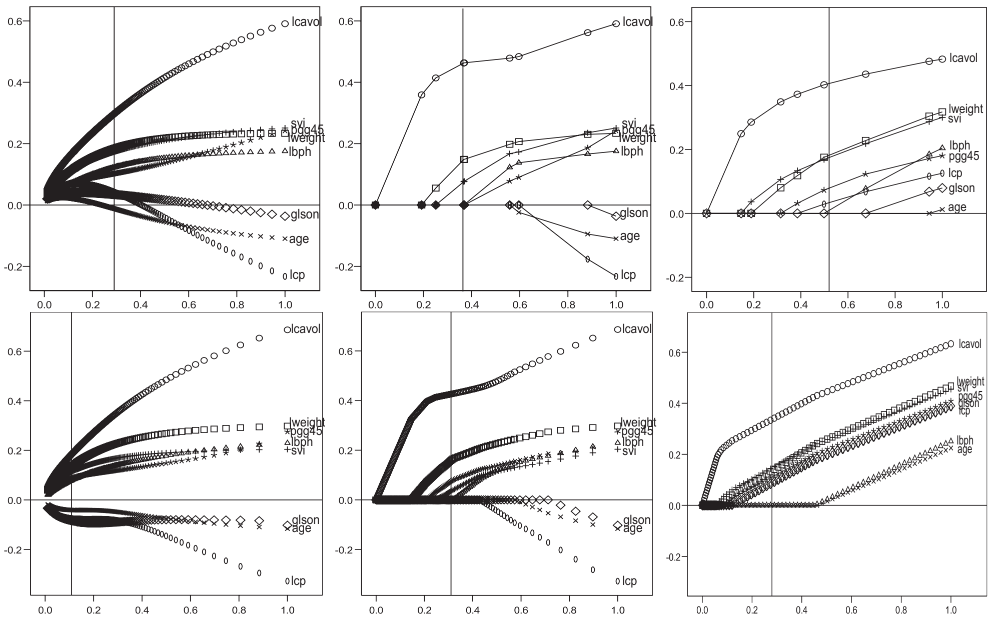

Figure 4.2. CATREG-Ridge (left), CATREG-Lasso (middle), and CATREG-E-net (right) for prostate cancer data (N = 67); linear (top) and nonlinear (bottom). The vertical line represents the model selected with the .632 bootstrap applying the one-standard-error rule.

indicates the best model does not coincide with a transition point (as for the linear Elastic Net). When we compare the linear with the nonlinear results, we note that for the nonlinear analysis the best model is always found for a larger penalty term (the vertical line is more to the left), but this does not imply that the model includes less predictor variables, as is most easily seen from Table 4.1.

Figures 4.3 and 4.4 display the transformations, for the full model (no shrinking) and for the Ridge, the Lasso, and the Elastic Net models selected with the .632 bootstrap applying the one-standard-error rule. These plots show that the transformations of the "stronger" predictors (lcavol, lweight, pgg45) that need large penalties to remove them from the model are rather unaffected by the various ways of regularization, while the transformations of the "weaker" predictors (age, gleason) do change considerably when regularization is applied, but are remarkably similar for the different regularization methods.

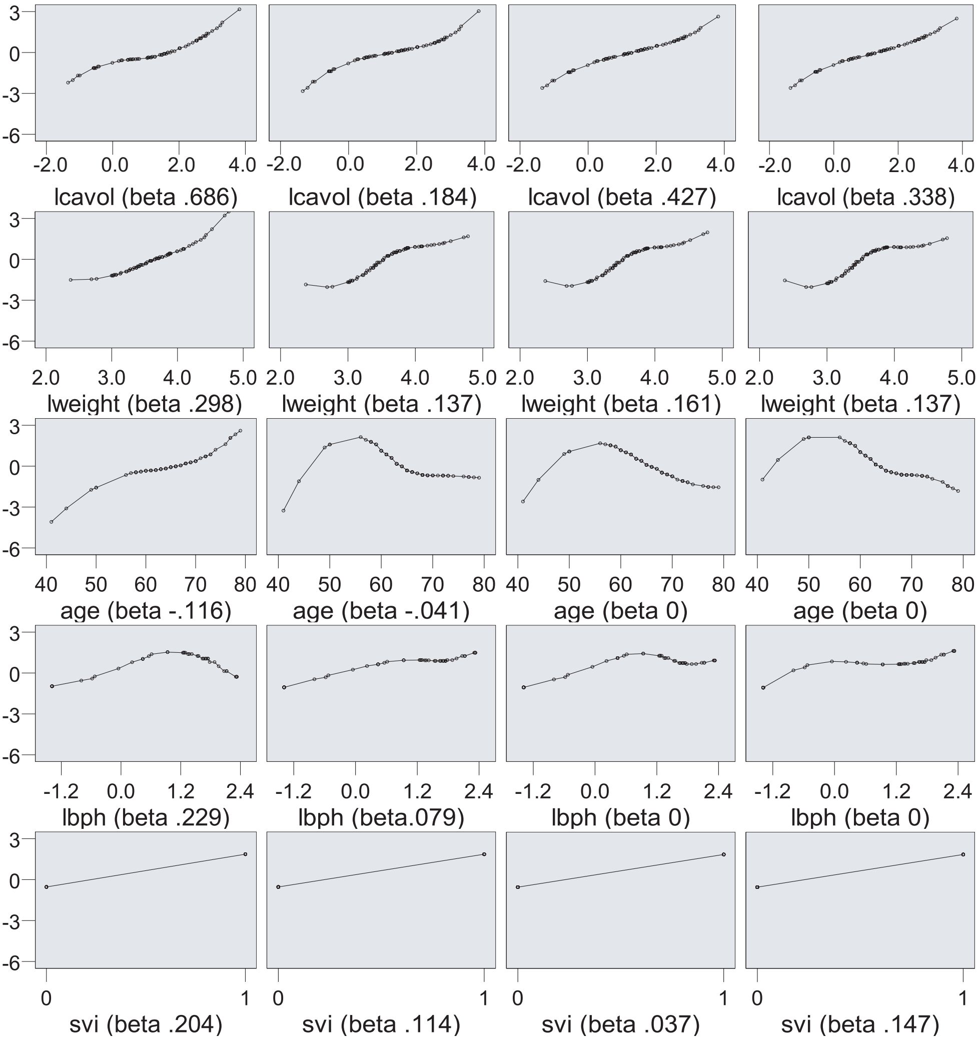

Figure 4.3. CATREG spline nominal transformations for prostate cancer data ($N = 67$): full model (1st column), Ridge (2nd column), Lasso (3rd column, and Elastic Net(4th column) models (to be continued in Figure 4.4).

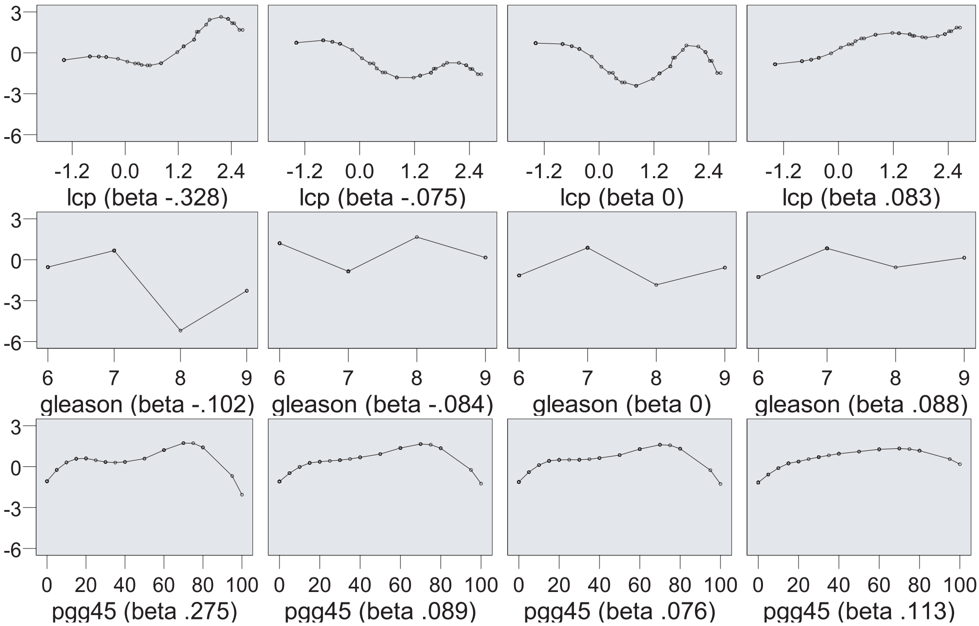

Figure 4.4. (Figure 4.3 continued).

4.5. Shrinking a nominal variable versus shrinking its categories

It is common practice to deal with a categorical variable in standard linear regression by replacing it by a set of dummy variables, where each dummy variable becomes a predictor variable, representing a category of the original nominal variable. As a result, applying a regularization procedure within such a treatment of a nominal variable, amounts to applying shrinking to the *categories* of the variable in stead of applying shrinking to the variable as a whole. Recently, this was remedied by a method called the "Grouped Lasso" in Yuan and Lin (2006) and "Blockwise Sparse Regression" (BSR) in Kim et al. (2006), generalizing the method proposed by Yuan and Lin (2006) for ANOVA and additive models to other loss functions. In this method, additional constraints are active in the regularization process, applied to the regression coefficients for dummy variables associated with the categories of the same variable. For nominal data this method is equivalent to CATREG. To show this equivalence, we need to go into some detail of the the CATREG algorithm.

The original CATREG approach was designed to deal with categorical data (and thus with nonlinear transformations). For a categorical variable, a transformation is written as $\mathbf{G}_j\mathbf{v}_j$, the product of an $N \times K_j$ indicator matrix $\mathbf{G}_j$ and a K_j vector of category quantifications $\mathbf{v}_j$, where K_j indicates the number of categories of variable j. In the indicator matrix $\mathbf{G}_j$, the elements are coded in the following way: object i obtains a one in row i of column k_j if observation i is in category k_j of variable j, and obtains a zero otherwise. So, a column $\mathbf{g}_{k_j}$ of $\mathbf{G}_j$ is a dummy variable for category k of variable j. The inner product $\mathbf{G}_j'\mathbf{G}_j$ is a diagonal matrix $\mathbf{D}_j$, containing the marginal frequencies of the categories for variable j on its main diagonal.

At this point, we assume (without loss of generality) that all predictor variables are categorical. Then, choosing a nonmonotonic step function to transform the jth predictor variable, the optimal quantifications are found as the averages of the scores of the objects that have scored in a particular category of the jth predictor on the response variable, *corrected for the contribution of the other predictor variables on the response variable.* Thus, if $\mathbf{y}$ denotes the standardized response, the category quantifications (the level values of the step function) for predictor variable j are written as

$$\tilde{\mathbf{v}}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'(\mathbf{y} - \sum_{l \neq j} \beta_l \mathbf{G}_l \mathbf{v}_l). \tag{4.27}$$

Then weighted normalization is applied to $\tilde{\mathbf{v}}_j$:

$$\mathbf{v}_j^+ = N^{1/2}\tilde{\mathbf{v}}_j(\tilde{\mathbf{v}}_j'\mathbf{G}_j'\mathbf{G}_j\tilde{\mathbf{v}}_j)^{-1/2}, \tag{4.28}$$

to render the transformed variable $\varphi_j(\mathbf{x}_j) = \mathbf{G}_j\mathbf{v}_j$ to be standardized, and the update for the regression coefficient is estimated as

$$\beta_j^+ \ = N^{-1}(\tilde{\mathbf{v}}_j'\mathbf{G}_j'\mathbf{G}_j\mathbf{v}_j^+). \tag{4.29}$$

The estimate (4.29) is equivalent to

$$\beta_j^+ = N^{-1}(\tilde{\mathbf{v}}_j'\mathbf{G}_j'\mathbf{G}_j\tilde{\mathbf{v}}_j)^{1/2}. \tag{4.30}$$

(Some simple matrix algebra shows that Equation (4.29) is equivalent to equation(4.15), which, writing a transformed variable as the product of the indicator matrix and the vector of category quantifications, is $\beta_j^+ = N^{-1}(\mathbf{G}_j\mathbf{v}_j^+)'$ $(\mathbf{G}_r\mathbf{v}_r^+ - \sum_{l \neq j}\mathbf{G}_l\mathbf{v}_l^+\beta_l)$.)

Because the dummy variables associated with the categories of a particular variable are uncorrelated, the unstandardized regression coefficients for linear regression on dummy variables are obtained as

$$a_{k_j} = (\mathbf{x}_{k_j}'\mathbf{x}_{k_j})^{-1}\mathbf{x}_{k_j}'(\mathbf{y} - (a_0 + \sum_{l \neq j}\sum_{k_l=1}^{K_l-1} a_{k_l}\mathbf{x}_{k_l})), \tag{4.31}$$

with a_{k_j} the unstandardized regression coefficient for the k'th category of variable j and $\mathbf{x}_{k_j}$ the k'th dummy variable (k'th column of $\mathbf{G}_j$). Equation (4.31) corresponds to Equation (4.27) written for a single category, because $\mathbf{x}_{k_j}$ is equal to column $\mathbf{g}_{k_j}$ of $\mathbf{G}_j$:

$$\tilde{v}_{k_j} = d_{k_j}^{-1}\mathbf{g}'_{k_j}(\mathbf{y} - \sum_{l \neq j} \beta_l \mathbf{G}_l \mathbf{v}_l). \tag{4.32}$$

So, by collecting the a_{k_j}'s in the vector $\mathbf{a}_j$ and normalizing $\mathbf{a}_j$ as is done in Equation (4.28) for v_j, the regression coefficients from linear regression on dummy variables yields the category quantifications obtained for CATREG with nominal transformations, and the regression coefficient for the variable as a whole can be computed as in (4.29) or (4.30). (NB: one of the columns of $\mathbf{G}_j$ is redundant because each column of $\mathbf{G}_j$ can be perfectly predicted from the other columns; this gives trouble in standard regression where $\mathbf{a}$ is computed as $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Then for each predictor one dummy/category has to be omitted. To obtain the CATREG nominal quantifications from the a_{k_j}'s resulting from linear regression on dummies, the omitted category is included with a regression coefficient of zero. It does not matter which dummy/category is omitted; omitting different categories results in different unstandardized $\mathbf{a}_j$'s, but standardized $\mathbf{a}_j$'s are always the same.)

The Grouped Lasso method of Yuan and Lin (2006) and the Blockwise Sparse Regression (BSR) method of Kim et al. (2006) treat the dummy variables for a predictor as a group/block by applying a norm restriction to the regression coefficients for the dummy variables in a group/block. As was shown in the previous paragraph, this restriction is equivalent to the weighted normalization of the CATREG nominal category quantifications $\mathbf{v}_j$. In the Grouped Lasso and BSR approach a continuous predictor is represented by a group/block of basis funtions, such as polynomials. In the CATREG approach, continuous predictors can be smoothly transformed and shrunken as a whole by applying nonmonotonic or monotonic spline transformations.

As an illustration, we use the breast cancer recurrence data (M. Zwitter and M. Soklic, University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia; available at http://www.ics.uci.edu/~mlearn/MLRepository.html). The response is a binary variable coded 0 for no-recurrence-events (201 cases) and 1 for recurrence-events (85 cases); deleting the nine cases with missing values, the 0/1 frequencies are 196/81. The predictor variables are described in Table 4.2. Kim et al. 2006 use these data to illustrate their BSR procedure, applying logistic regression with each categorical variable expanded to a block of dummy variables, and a numerical variable expanded to a block of tranformations up to third-order polynomial. In the analysis with CATREG, *Table 4.2. breast cancer recurrence data.*

Predictor	Number of categories / values		Measurement level
menopause	3		categorical
node-caps	2		categorical
breast	2		categorical
breast-quad	5		categorical
irradiat	2		categorical
age		6	numerical
tumor-size		11	numerical
inv-nodes		7	numerical
deg-malig		3	numerical

we used nonmonotonic step functions for all predictor variables since the numerical variables have only a limited number of values. The predictor variable tumor-size is an exception, for which we optimized a nonmonotic spline function, of degree two with one interior knot. Figure 4.5 displays the paths for the CATREG-Lasso and for the linear Lasso on dummy variables (for the latter, the regression coefficients for the variables were computed from the regression coefficients for the categories as explained above). The paths are rather similar; the main difference being that with linear Lasso on dummy variables, the variable "node-caps" (+) becomes active earlier than with CATREG-Lasso and the variable "age" (□) later.

To compare the CATREG-Lasso results to the logistic-regression-BSR misclassification results for this data set as reported in Kim et al. 2006, the predicted classification variable is computed by recoding the predicted value variable to 0 if its value is closer to the lowest value of the standardized dependent variable than to the highest value and 1 otherwise. For model selection and estimation of the prediction error, the .632 bootstrap was applied to the total data set (following as closely as possible the procedure of Kim et al. 2006, who used 10 repetitions of 10-fold cross validation). The results (given in Table 4.3) are very similar to the results of the BSR: the CATREG-Lasso and logistic-regression-BSR estimates of the misclassification rates (MCR) are very close. With CATREG and the .632 bootstrap, dummies-Lasso performs somewhat better than CATREG-Lasso, while Kim et al. 2006 found that logistic-regression-BSR is somewhat better than dummies-Lasso.

Figures 4.6 and 4.7 displays the transformations for the full model and the models selected with the .632 bootstrap (using 200 samples) for the CATREG-

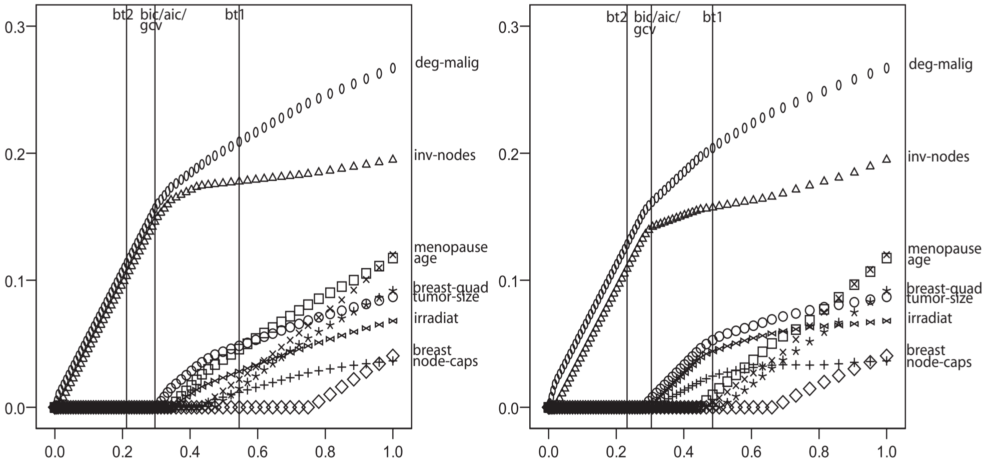

Figure 4.5. The CATREG-Lasso with nonmonotonic transformations (on the left; spline function for "tumor-size", step function for the other predictors) and the linear Lasso on dummy variables (on the right) for breast cancer recurrence data ($N = 277$). The vertical line represents the model selected with the .632 *bootstrap applying the one-standard-error rule.*

Table 4.3. Expected prediction error and misclassification rate for breast cancer recurrence data.

	Model:	Expected prediction	Expected mis-
	λ / s	error (SE)	class. rate (SE)
CATREG (full model)	0.00/1.00	0.1904 (0.0094)	0.2585 (0.0039)
CATREG-Lasso	0.15/0.55	0.1807 (0.0080)	0.2547 (0.0038)
Dummies-Lasso	0.15/0.47	0.1786 (0.0078)	0.2482 (0.0050)
Results reported in		Expected Logistic	
Kim et al. 2006		loss (SE)	
BSR		0.6917 (0.0015)	0.2578 (0.0028)
Dummies-Lasso		0.6964 (0.0016)	0.2646 (0.0028)

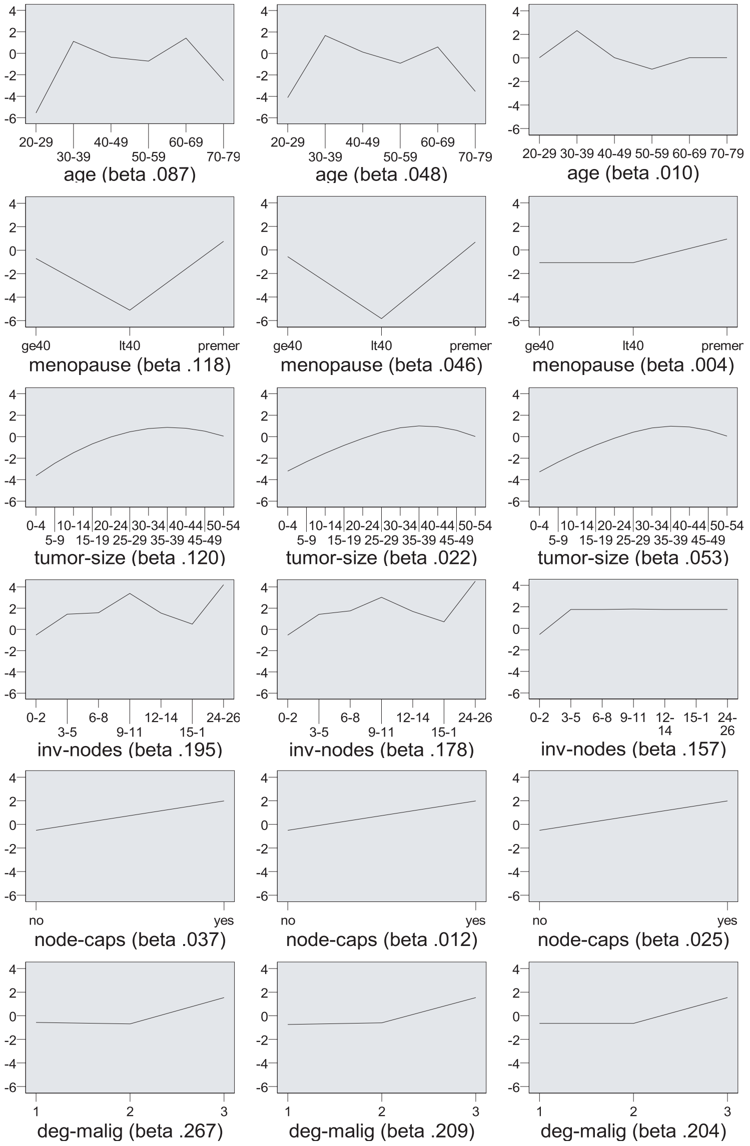

Figure 4.6. Transformations breast cancer recurrence data (N = 277): CATREG full model (left), CATREG-Lasso .632 bootstrap model (middle), and dummies-Lasso .632 bootstrap model (right) (to be continued in Figure 4.7).

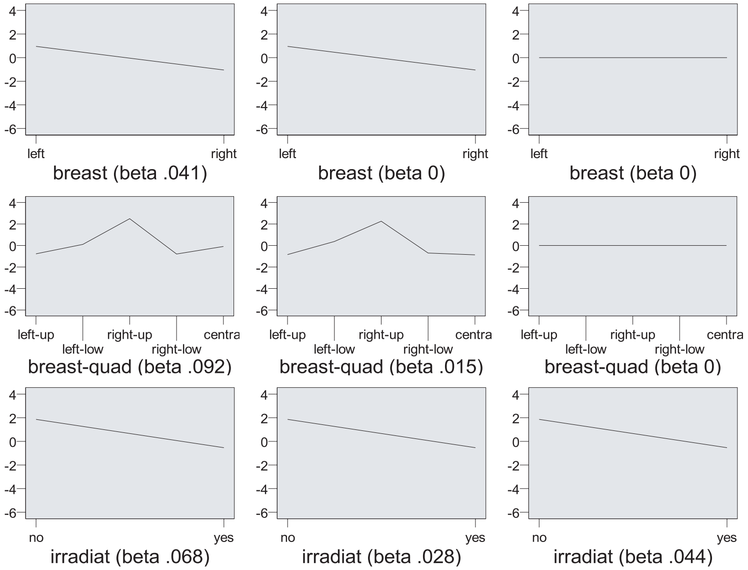

Figure 4.7. (Figure 4.6 continued).

Lasso and the dummies-Lasso (for the latter, the regression coefficients and category quantifications are computed as explained above). Comparing the transformations for the CATREG-Lasso and the dummies-Lasso, we notice that the dummies-Lasso transforms the variables “menopause” and “inv-nodes” to variables with two categories, contrasting the categories “ge40” and “lt40” to “premenopause”, and category “0–2” of “inv-nodes” to all higher categories. By allowing for shrinkage of category quantifications to zero, the dummies-Lasso results in dichotome variables. (In Van der Kooij and Meulman 2006a results are presented that suggest that binning a continuous variable might be beneficial for prediction accuracy.)

4.6. Discussion

We have shown that regularization methods such as Ridge regression, the Lasso, and the Elastic Net, can easily be incorporated into the CATREG algorithm for regression with nonlinear transformations, resulting in a simple and efficient way to estimate the constrained regression coefficients. At the same time, because we can apply CATREG to fixed linear transformations, we also have an algorithm for regularized linear regression that is an attractive alternative for the algorithms that have been proposed in the literature thus far.

In the context of regularized analysis, there are two goals: model selection and assessment of the selected model. To achieve these goals, the best approach is a *three-way* data split, dividing the data into a training set, a validation set, and a test set. The training set is used for model fitting and the prediction error for model selection is estimated using the validation test. In the end, the prediction error for the selected model (the generalization error) is estimated by applying the model to the test set. When there are not enough data for a three-way split, the data set is divided into two parts, a training and a test set, and the validation step is approximated either analytically, with Generalized Cross Validation (GCV; Golub, Heath, and Wahba (1979)), AIC, or BIC, or by using a resampling technique, such as cross validation or bootstrapping, on the training set. Throughout this chapter, we have used the .632 bootstrap as the tool for estimating prediction error for selection of the optimal penalty parameter. Here, we will discuss the analytic model selection methods mentioned above.

Defining rss(λ) as the residual sum of squares for the constrained fit, $df(\lambda)$ as the effective number of degrees of freedom, and estimating σ^2 as $\text{rss}^{\text{ols}}/(N - df^{\text{ols}})$, the GCV, AIC, and BIC statistics are written as

$$\text{GCV}(\lambda) = \frac{\text{rss}(\lambda)}{N(1 - \frac{df(\lambda)}{N})^2}, \tag{4.33}$$

$$\text{AIC}(\lambda) = \frac{\text{rss}(\lambda)}{N\sigma^2} + \frac{2}{N} df(\lambda), \tag{4.34}$$

$$\text{BIC}(\lambda) = \frac{\text{rss}(\lambda)}{N\sigma^2} + \frac{\log(N)}{N} df(\lambda), \tag{4.35}$$

and the optimal value of λ is the value minimizing the function. For Ridge regression, the effective number of degrees of freedom is

$$df^{\text{ridge}}(\lambda_2) = \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}' = \sum_{j=1}^{P} \frac{d_j}{d_j + \lambda_2}, \tag{4.36}$$

with d_j the jth eigenvalues of the matrix $\mathbf{X}'\mathbf{X}$. The effective number of degrees of freedom for the Elastic Net is

$$df^{\text{e-net}}(\lambda_2) = \text{tr}(\mathbf{X}_A(\mathbf{X}'_A\mathbf{X}_A + \lambda_2\mathbf{I})^{-1}\mathbf{X}'_A = \sum_{j=1}^{card(A)} \frac{d_j}{d_j + \lambda_2}, \tag{4.37}$$

where A denotes the active set of predictors and card(A) the number of predictors in this set. Zou, Hastie, and Tibshirani (2004) show that the number of effective degrees of freedom for the Lasso is very well approximated by the number of variables in the constrained model, $k(\lambda_1)$. Furthermore, Zou et al. (2004) show that the AIC and BIC statistics for Lasso models with k active predictors are minimal for the model at the point where transition from k to $k + 1$ or $k - 1$ active predictors takes place. Following their argument, this would also be true for the GCV statistic. So, to select the optimal λ_1 with GCV, AIC, and BIC, only the transition points need to be taken into account. We applied the GCV, AIC, and BIC to the analysis of the prostate cancer data, and we found that for Ridge regression and the Lasso, the .632 bootstrap with the one-standard-error rule chose more parsimonous models than GCV and AIC, while BIC was in between. For the Elastic Net, the results for the Boostrap and BIC were very close, and AIC and GCV were less parsimonous again.

If we wish to apply GCV, AIC and BIC for Ridge regression, the Lasso and the Elastic Net with CATREG, we need to adjust the degrees of freedom. For the Lasso, in stead of the number of predictors in the restricted model, the sum of the degrees of freedom for each predictor in the restricted model is used. With CATREG, the degrees of freedom depends on the transformation that is chosen. For a spline transformation, the number of degress of freedom amounts to the number of interior knots plus the number of the degrees of the splines, minus the number of spline coefficients that became zero. For the nominal and ordinal step functions, the number of degrees of freedom amounts to the number of distinct category quantifications minus one. For Ridge regression and the Elastic Net the effective degrees of freedom is computed as in Equations (4.36) and (4.37), replacing a variable $\mathbf{x}_j$ with $\varphi_j(\mathbf{x}_j)$.

Application to the prostate cancer data showed that for Ridge regression selection results are very similar to the linear analysis, with GCV and AIC much less parsimonous than BIC and the .632 bootstrap. For the Lasso, compared to the linear analysis, the role of BIC and the .632 bootstrap was reversed, and for the Elastic Net, AIC en BIC are more conservative than the .632 bootstrap and GCV. If we can conclude anything from the analysis of this single data set, it would be that the .632 bootstrap behaves neither extremely conservative (like the BIC), nor extremely liberal (like GCV). It remains true, of course, that the .632 bootstrap is much more time consuming, but since in a lot of interesting applications (such as in genomics, transcriptomics, proteomics, and metabolomics) the number of variables is much larger than the number of observations, we would need to apply a nonparametric method such as the .632 bootstrap in any case.

As mentioned before, the Ridge paths are not piecewise linear and neither are the Lasso paths when applying nonlinear transformations. (However, for the particular data sets that were analyzed in this chapter and for several other real data sets, we observed only slight non-linearities in the Lasso-paths.) So, in these cases, the paths have to be constructed by computing solutions for many values of the penalty parameter (as was done in our applications). This also applies to the linear Lasso and the Elastic Net when model selection is done with the .632 bootstrap. However, a plot of the .632 bootstrap estimates of the expected prediction error as a function of the model complexity usually shows a regular curve: we obtain the highest error estimates for the highest values of the penalty parameter and the values of the error estimates decrease with decreasing values of the penalty parameter, until we reach the minimum. From that point, the error estimates increase again until we reach the point for the zero penalty term. Thus, application of the .632 bootstrap can be made much less time-consuming by performing the CATREG analysis twice. In the first analysis, we use a rather big step size for the penalty parameter to obtain a region that contains the minimum. In the second analysis, we use a much smaller stepsize, but now only for a small range of penalty values in the region obtained in the first run to determine the minimum.

Finally, we would like to emphasize that CATREG has a unique approach to regularization in regression compared to the other approaches proposed in the literature. The CATREG approach deals with each predictor variable separately, isolating the estimation of the regression weights β_j from the estimation of the optimal transformation of each predictor variable (it is crucial to note that optimal scaling includes transformation of a continuous variable to standard scores). Contrasting the CATREG approach to nominal variables with the approach that obtains weights a_{k_j} for dummy predictors, CATREG attaches category quantifications $\mathbf{v}_j$ to the dummy variables in $\mathbf{G}_j$, applies normalization to obtain standardized transformed variables $\phi_j(\mathbf{x}_j) = \mathbf{G}_j\mathbf{v}_j$, $(\mathbf{G}_j\mathbf{v}_j)'\mathbf{G}_j\mathbf{v}_j = N$, and can next separate the estimation of the β_j. At this point, it is important to realize that in the process to obtain optimal quantifications, we do not need to use the indicator matrices themselves in the computations. The indicator matrices $\mathbf{G}_j$ (that are extremely sparse) are only used in the equations. In a computer program, the matrix multiplications can be replaced by simple additions.

In the statistical literature that uses backfitting to obtain optimal transformations in regression, regression weights are never explicitly computed, since they are subsumed in the transformed variables $\phi_j(\mathbf{x}_j)$. In CATREG, the transformed variables are always standardized, $\phi_j(\mathbf{x}_j)'\phi_j(\mathbf{x}_j) = N$, and in this way the β_j can be identified. Backfitting has been used by, among others, Friedman and Stuetzle (1981), Breiman and Friedman (1985), Buja et al. (1989), and Hastie and Tibshirani (1990) to incorporate smooth transformations in regression. Step functions to deal with categorical data were never considered in their context. If transformed variables absorbe the weights, it is not obvious how to impose regularization constraints on the regression weights. Also, backfitting is very inefficient when linear regression is concerned. This might be an explanation of the fact that the backfitting algorithm has not been proposed before to impose simple constraints on the estimates for the weights to regularize regression. We call these constraints simple since we do not have to pay attention to the other predictor variables (their contributions have been removed from the prediction). The separation between weights and quantifications/transformations in CATREG showed the way for the easy implementation of the three regularization penalties in regression proposed in this chapter.