Prediction accuracy and stability of regression with optimal scaling transformations

Kooij, A.J. van der

Chapter 6

General Discussion

6.1. A short retrospect

In this monograph, the CATREG method for multiple regression with optimal scaling transformations and three major enhancements to the method were described. First, a remedy for the problem of local minima with monotonic transformations has been offered. This problem also occurs with other methods that are closely related to CATREG and has not been recognized in literature thus far. Second, a procedure for assessing the prediction accuracy of CATREG was incorporated into the method, estimating the expected prediction error using the .632 bootstrap, which was adapted to deal with transformed variables. And, finally, three regularization methods were implemented for improving prediction accuracy by stabilizing the parameter estimates and for predictor selection.

In the Introductory chapter (Chapter 1), the CATREG method was described in some detail and a dataset with nonlinear regression effects was used to graphically illustrate the basic elements of the method. The effect of the choice of the scaling level on the transformations was described and illustrated, as well as the optimal scaling and backfitting procedures.

With monotonic transformations, CATREG, and other methods for regression with transformations, sometimes ends up in a local minimum of the error function, resulting in a suboptimal solution. This problem was addressed in Chapter 2. The cause of ending up in a local minimum and a strategy to obtain the global minimum was described. This strategy, however, uses multiple systematic starts and involves complete enumeration, which is not feasible when the number of predictors is large. Several computationally less intensive strategies were discussed and results of evaluating the performance of these

strategies in a simulation study were given. Two strategies that work with a reduced number of multiple starts proved to give good results in reducing both the incidence and the severeness of local minima. One of these strategies uses a percentage of loss criterion, which has the advantage of flexibility. The other strategy, using a hierarchical approach, has the advantage of a smaller maximum number of starts, but is less flexible. Combining these two strategies, results in a flexible strategy that considerably reduces the incidence and severeness of local minima and considerably reduces the number of starts. Furthermore, the combined strategy has the additional advantage that the number of starts does not grow very fast with increasing number of predictor variables. In the simulation study also the effect of several data conditions on the incidence and severeness of local minima was investigated. It was found that local minima occur more frequently and are more severe with low to moderately low R^2 values, with higher number of categories and with higher multicollinearity.

In chapter 3, the prediction accuracy of CATREG was assessed by estimation of the expected prediction error with the .632 bootstrap. The complications that arised with the implementation of this form of the bootstrap for optimally transformed variables were described in detail. Part of the complication was due to the fact that in the .632 bootstrap literature at different places, different definitions are used for the leave-one-out bootstrap estimate of prediction error: three different definitions were found in Efron (1983); Efron and Tibshirani (1993), and Hastie and Tibshirani (1990). The prediction accuracy for all scaling levels was assessed and compared to five other regression methods that use transformations to fit nonlinear relations, and for which results of estimating expected prediction error with the .632 bootstrap have been reported in the literature (Hastie and Tibshirani 1990, sec. 10.3). In addition, a comparison with ACE (Breiman and Friedman 1985), a method closely related to CATREG, was made. For the assessment of prediction accuracy and the comparisons with other methods, a data set frequently used in the literature on nonlinear regression effects was used (the ozone data). It was concluded that, with the particular data set used, the CATREG method works as well or slightly better than the methods it was compared to. The effect on the prediction accuracy of reducing numeric predictors to categorical variables with less categories than the number of distinct values of the variable (binning) was also investigated, and found to be positive.

The study of the prediction accuracy of CATREG for the ordinal and monotonic scaling levels also provided a different view on local minima. With the Ozone data, the expected prediction error for local minimum solutions was found to be lower than for the global optimal solution. So, although the

global minimum solution minimizes the apparent prediction error, by definition, it may not do so with regard to the expected prediction error. The study of prediction accuracy confirmed the general conjecture that more restricted models lead to higher apparent error (again, by definition), but to lower expected error. However, when the relationship between the response and the predictors is severely nonlinear, more restrictive transformations may lead to reduced prediction accuracy. If and how much a too restrictive transformation will reduce the prediction accuracy, depends on the importance of the predictor and its relationship with the other predictors.

In linear regression, often regularization is applied to improve prediction accuracy by stabilizing the regression coefficients. At the same time, some regularization methods also accomplish predictor selection. Numerous regularization methods and algorithms exist, and three of them (Ridge Regression, the Lasso, and the Elastic Net) have been embedded in the CATREG method, enabling regularization for nonlinear regression problems (Chapter 4). Several rather complex and/or time consuming algorithms have been proposed in literature for the estimation of linear Lasso models, although a recent development provides an elegant and efficient procedure (LARS, Efron et al. 2004). All these algorithms are limited to select at most N predictors if $P > N$.

Regularization has the effect that the regression coefficients are shrunken, causing them to be more stable. The shrinkage is realized through adding a penalty parameter to the model, penalizing the (absolute) size of the regression coefficients. Because in linear regression the coefficients are estimated simultaneously, involving the predictor correlation matrix, estimation of the Lasso and the Elastic Net coefficients, both using an L_1 penalty, is not very straightforward. The backfitting procedure, however, does allow straightforward estimation of the shrunken coefficients. In the backfitting estimation procedure the correlation matrix is not involved, because the coefficients are estimated one at a time, while when estimating the coefficient for a particular predictor, the effect of the other predictors is removed from the response. When applying backfitting, the penalty term can be straightforwardly added to the equation for the estimation of a coefficient. Moreover, with the backfitting algorithm, more than N predictors can be selected. Linear regression can also be performed using backfitting; for unregularized regression this is of course an unnecessarily inefficient procedure. For regularized regression, however, it provides a straightforward way of estimating shrunken coefficients.

Starting with a zero penalty, resulting in the full model (i.e., including all predictors with unshrunken coefficients), and stepwise increase of the value of the penalty, the Ridge coefficients follow paths that go towards zero at different rates. The Lasso and Elastic Net paths go to exactly zero at different

rates and at the point where the coefficient for a predictor is shrunken to zero, the number of predictors included in the model changes and hence such points are called transition points. With linear regularized regression the Lasso and Elastic Net shrinkage paths are linear between the transition points and thus can be constructed by connecting the transition points. So, then only the models at the transition points need to be estimated. Finding the transition points in the same order of magnitude of computational effort as Ordinary Least Squared was achieved by the LARS method (Efron et al. 2004). The CATREG-Lasso for linear regression, using the backfitting approach, finds the transition points in a more greedy way, using the change in slope of the paths at a transition point, although considerably less greedy than having to estimate models for each value of the penalty. With the nonlinear CATREG-Lasso, however, the paths are not piecewise linear and models have to be estimated for each value of the penalty parameter (the number of models to estimate depends on the step size used in increasing the penalty). For selecting the optimal model (equivalent to selecting the optimal penalty value), traditionally model selection criterion like BIC, AIC, and GCV are used. Before the LARS algorithm was available, resampling methods like cross validation or the bootstrap would have been too time consuming. In Chapter 4, using model selection criteria for linear CATREG-Lasso is compared to model selection using the .632 bootstrap for a particular data set. The BIC, AIC, and GCV statistics require estimation of the degrees of freedom, which is not well studied yet for regression with nonlinear transformations. Comparing the results with model selection criteria to selection with the .632 bootstrap, suggests that the BIC, AIC, and GCV statistics might be applicable in nonlinear CATREG-Lasso as well.

Recently, Yuan and Lin (2006) developed Grouped Lasso, a Lasso method for analysis of variance and additive models, extended by Kim et al. (2006) for more general loss functions, including generalized linear models, called Blockwise Sparse Regression (BSR). These methods deal with categorical predictor variables by representing them by groups or blocks of dummies variables. The dummy variables are treated as a block by applying a norm restriction to their regression coefficients. This norm restriction is equivalent to the weighted standardization of nominal category quantifications in CATREG. So, for nominal predictors the Grouped Lasso and BSR are equivalent to the CATREG-Lasso. In the CATREG-Lasso however, nominal variables are straightforwardly transformed, without the need to expand them to groups of dummy variables. The Grouped Lasso and BSR also include nonmonotonic transformation of continuous variables by expanding them to a group of basis functions, such as polynomials. With the CATREG-Lasso, continuous

predictors can be straightforwardly transformed and shrunken as a whole by applying a nonmonotonic or monotonic spline transformation. When using dummies for nominal variables without applying a norm restriction to their regression coefficients, the categories of the variable are shrunken in stead of the variable as a whole. The effect of this form of shrinking was also illustrated.

Chapter 5 contains an application to a real life, new data set; it describes results of the CATREG-Lasso and the .632 bootstrap applied to a sample of patients suffering from bulimia nervosa in a varying degree. The data were collected by the Department of Psychosomatic Medicine and Psychotherapy, University of Freiburg, Germany, and include an expert severity rating of the disorder and a standard battery of outcome questionnaires containing expert and self report measures. The research goals were to select an optimal small subset of the outcome variables to predict the severity rating, to empirically determine a cut-off score between healthy/subclinical and clinical cases, and to determine the score region of the predictor variables that contributes to the prediction. In answering these questions, the researchers wished to take into account the non-interval scale of the variables and possible nonlinear relations between them. Their search for expertise on these matters led to a collaboration with the author of this monograph, who conducted the statistical analysis and wrote a major part of the statistical Analysis and Results sections of the resulting paper. The CATREG-Lasso was not applied here to select a shrunken model, but to select a subset of predictors, with unshrunken coefficients. Models including one predictor, two predictors, up to all but one predictors were determined from the results of a CATREG-Lasso analysis, with ordinal and monotonic spline transformations. Then the .632 bootstrap was applied to these models to select the optimal subset of predictors. Next, the transformations were studied, revealing that the selected predictors are informative for discriminating clinical cases from healthy/subclinical cases, and for discriminating between healthy and subclinical cases, but less informative for discrimination within the class of clinical cases. The lower accuracy of the predictors in discriminating cases in the clinical class was also observed when examining models including more or other predictors. Finally, to evaluate the diagnostic validity of the selected model, the cases were classified as clininal or healthy/subclinical using the predicted values resulting from the selected model. Using ROC curve analysis, cut-off values for the predicted values could be found for classifications with either very high sensitivity (true positive rate) and acceptable specificity (true negative rate), or with very high specificity and acceptable sensitivity, or with both high sensitivity and high specificity.

6.2. TOPICS FOR FURTHER RESEARCH

First of all, the estimation of prediction accuracy of CATREG with the .632 bootstrap needs more extensive study, using simulated data with known error and transformation curves, varying the signal-to-noise ratio. Also, regularized CATREG needs to be investigated more extensively in a simulation study.

During the research described in this monograph, several interesting issues and ideas that deserve further examination arised, which will be outlined in the following subsections.

6.2.1 Transformations towards independence and regularization

Meulman and Van der Kooij (2000) conjectured that due to removing the variance in the response accounted for by the predictors not being estimated in the backfitting procedure, the predictor variables will be transformed towards independence. They conducted a modest simulation study, sampling a response variable and three predictors for 200 observations from a multivariate normal distribution and discretizing the variables by binning them into seven categories. The values for the multiple correlation R^2 were varied, as well as forms of nonlinearity and degrees of multicollinearity. The results of this simulation study supported this conjecture and indicated that the effect of optimal scaling transformations on the (in)dependence among the predictors depends on the size of R^2 and the smoothness of the transformation (the smoother the transformation, the smaller the effect). However, more definite evidence for the conjecture requires a larger simulation study, including more predictors. A new and closely related conjecture arose when examining the results of the CATREG-Ridge applied to the prostate cancer data (Chapter 4): with nonlinear optimal scaling transformations, CATREG-Ridge did not improve the prediction accuracy compared to regular CATREG. We could conjecture that the reason for this is that because the predictors were transformed towards indepence, regularization to deal with interdependence of the predictors was not warranted anymore in this particular case. So, it would be interesting to investigate if (and, if affirmed, to what extent) optimal scaling transformation is a form of regularization itself.

6.2.2 The efffect of shrinking on low-frequency categories

Optimal scaling sometimes results in a "degenerate" transformation for a variable when one of its categories has a very low frequency. Especially with

ordinal transformation, it can happen that the quantification results in a dichotomous variable, distinguishing only the low-frequency category from the other categories that all receive the same quantified value. In such a case, the regularization of dummy variables approach, described in Chapter 4, might be of interest, possibly resulting in the low-frequency category shrunken to zero.

6.2.3 Regularization in Discriminant Analysis

The CATREG regularization method can easily be extended to both linear and nonlinear regularized Discriminant Analysis (DA) for classifying cases into groups. CATREG with nominal scaling applied to a dependent categorical variable and linear transformations to continuous predictors is equivalent to linear DA (one-dimensional; only one discriminant function will result). By choosing nonlinear transformations, nonlinear DA is achieved. The adaptation of CATREG to CATDA does not concern the algorithm, but only the output: regression coefficients need to be converted to discriminant coeffients, which is straightforward because they are proportional to each other, and output specific to DA needs to be provided. Incorporating the *multiple nominal* scaling level in (regularized) CATREG will allow for (regularized) nonlinear multiple discriminant function analysis. Applying the multiple nominal scaling level to the dependent variable results in multiple sets of quantifications for the dependent variable — quantifications that are different for each dimension (function) — and multiple sets of discrimimant coefficients for the predictors.

6.2.4 CATREG with cluster restrictions

A method for combined regression and clustering could be incorporated in CATREG by applying cluster restrictions. For this purpose, the approach of the predecessor of CATREG, which was called MURALS, (Van der Kooij and Meulman 1997) is probably more suited. In this approach, the influence of the predictor variables is channelled through a compromise variable, which is as close as possible to the linear combination of the predictor variables and at the same time as close as possible to the dependent variable. This representation of multiple regression has been inspired by the generalized canonical analysis technique called OVERALS (Gifi 1990). In the OVERALS methodology, similaritiy between sets of variables is established by simultaneously comparing the linear combinations of the variables in each set to an unknown, compromise variable. When using this representation of regression, cluster

restrictions can easily be imposed on the compromise variable. Such a cluster restriction resembles much the cluster restriction applied in GROUPALS (Van Buuren and Heiser 1989), a method that combines nonlinear Principal Components Analysis with clustering. Also, this approach is similar in spirit to Projection Pursuit Regression (Friedman and Stuetzle 1981).