Improved predictions penalizing both slope and curvature in additive models

Magne Aldrin,
Norwegian Computing Center,
P.O.Box 114 Blindern, N-0314, Oslo, Norway
Tel: (+47 22 85 25 00)
email: magne.aldrin@nr.no

2nd August 2004

Abstract

A new method is proposed to estimate the nonlinear functions in an additive regression model. Usually, these functions are estimated by penalized least squares, penalizing the curvatures of the functions. The new method penalizes the slopes as well, which is the type of penalization used in ridge regression for linear models. Tuning (or smoothing) parameters are estimated by permuted leave-k-out cross-validation. The prediction performance of various methods is compared by a simulation experiment: penalizing both slope and curvature is either better than or as good as penalizing curvature only.

KEY WORDS: Penalized B-splines, Penalized least squares, Penalized likelihood, Ridge regression, Generalized additive models, Cross-validation

1 Introduction

We want to predict the outcome of a response variable y from the values of a p-dimensional predictor variable $\mathbf{x} = (x_1, ..., x_p)$, where the relationship can be assumed to follow on average the additive model

$$E(y|\mathbf{x}) \quad = \quad \mu(\mathbf{x}) = \beta_0 + s_1(x_1) + \cdots + s_p(x_p). \tag{1}$$

Here β_0 is an intercept, and $s_1(x_1), \cdots, s_p(x_p)$ are unknown, but smooth functions of the predictor variables. Inference is based on a training set of n observations $\{y_i, \mathrm{x}_i\}_1^n$, and all predictor variables are scaled to $[0, 1]$. It can be modelled by splines, and estimation is usually done by minimizing the penalized least squares criterion

$$\sum_{i=1}^{n} \{y_i - \tilde{\mu}(\mathbf{x}_i)\}^2 + \lambda \cdot \sum_{j=1}^{p} \int_0^1 \{\tilde{s}_j''(x)\}^2 dx \qquad\qquad \lambda \geq 0, \tag{2}$$

where $\tilde{\mu}$ and $\tilde{s}_j$ denote current estimates and $\tilde{s}_j''$ denotes the second derivative of the estimated function $\tilde{s}_j$. The optimal estimates according to the minimization of (2) or similar criterions will be denoted $\hat{\mu}$, $\hat{s}_j$ etcetera.

The curvatures of the functions are controlled by the tuning parameter λ. An "optimal" value of λ can be chosen for instance by cross-validation. Since λ is common to all functions, the predictor variables should be on a common scale, for instance between 0 and 1. Increasing the value of λ gives smoother functions and a smaller effective number of parameters, but worse fit to the data. When λ goes to infinity, the $\tilde{s}$-functions become linear, and the solution of (2) is then equivalent to the ordinary least squares (OLS) estimates of the linear model

$$\mu(\mathbf{x}) \quad = \quad \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_p \cdot x_p, \tag{3}$$

where the β_j's are regression coefficients.

It is well known that biased estimation methods can give considerably better predictions than OLS when there are few data, or when the predictor variables are highly collinear (see for instance Frank and Friedman, 1993). Biased estimation methods shrink the regression coefficients towards zero and yield estimates between those of the OLS and the pure intercept model estimated by the response mean ($\hat{\beta}_0 = \bar{y}$ and $\hat{\beta}_j = 0$ for $j > 0$). One of the most popular of these methods is ridge regression (Hoerl and Kennard, 1970). It can be formulated as a penalized regression problem where the regression coefficients are estimated by minimizing the penalized least squares criterion

$$\sum_{i=1}^{n} \{y_i - \tilde{\mu}(\mathbf{x}_i)\}^2 + \gamma \cdot \sum_{j=1}^{p} \tilde{\beta}_j^2, \qquad\qquad \gamma \geq 0. \tag{4}$$

To get the predictor variables on a comparable scale, they are usually scaled to have the same variance. Alternatively, the predictor variables are transformed to cover the range from 0 to 1. Then, since $s_j(x) = \beta_j \cdot x$ and $s_j'(x) = \beta_j$ under the linear model (3), the criterion (4) becomes

$$\sum_{i=1}^{n}\{y_i - \tilde{\mu}(\mathbf{x}_i)\}^2 + \gamma \cdot \sum_{j=1}^{p}\int_0^1 \{\tilde{s}_j'(x)\}^2 dx, \tag{5}$$

where the first derivatives are penalized instead of the second derivatives as in (2). The degree of

shrinking is controlled by the tuning parameter γ. When $\gamma = 0$, the method is equivalent to OLS with $p+1$ effective parameters, and when γ reaches infinity it is equivalent to the pure intercept model with 1 effective parameter.

Controlling the effective number of parameters by penalization is central to both the linear and nonlinear situations. I propose combining the two. Consider the additive model (1), but estimate the model by minimizing the penalized least squares criteria

$$\sum_{i=1}^{n}\{y_i - \tilde{\mu}(\mathbf{x}_i)\}^2 + \gamma \cdot \sum_{j=1}^{p}\int_0^1 \{\tilde{s}_j'(x)\}^2 dx + \lambda \cdot \sum_{j=1}^{p}\int_0^1 \{\tilde{s}_j''(x)\}^2 dx. \tag{6}$$

In this paper we investigate if the combined use of first and second order penalization can give better predictions than using either of the two penalizations on its own. It is difficult to find the solution that minimizes (6). Therefore, the actual implementation of the double penalization is based on the alternative formulation of generalized additive models by Marx and Eilers (1998) using penalized B-splines, also called P-splines. The model (1) is first approximated by a linear, but high-dimensional and thus still flexible model. Then the first and second order differences of adjacent estimated coefficients are penalized, which approximates the corresponding penalizations on the first and second order derivatives in (6). More details are given in section 3. Since the new method uses double penalization of both first and second order, I will call it DP12. Similarly, using the second order penalization only is called P2.

It is essential to use a precise procedure to estimate the two tuning parameters. I use a permuted leave-k-out cross-validation procedure (Shao, 1993), that has better properties than the ordinary leave-one-out cross-validation.

A simulation experiment with 16 real data sets has been carried out. DP12 is better than P2 when there are few training data, whereas the two methods have similar properties when there are many observations. DP12 also performs well compared to ridge regression. In situations with few observations (10, say, or

less), ridge regression may perform better than DP12. But when there are enough data to estimate the two tuning parameters (γ and λ) well, DP12 substantially outperforms ridge regression if the relationship between y and $\mathbf{x}$ is nonlinear.

The next section gives an example that motivates and illustrates the use of DP12. The various regression methods and the cross-validation procedures are described in more detail in section 3. The simulation experiment is described in section 4, and finally I give some conclusions and indicate some possible extensions in section 5.

2 Illustrating example

The left-hand panel of Figure 1 is a scatter plot of the response variable *ozone* versus the predictor variable *temperature* with 330 observations. The data set was analyzed by Breiman and Friedman (1985) and by Hastie and Tibshirani (1990). The nonlinear relationship is clear, and a nonlinear smooth function seems appropriate. P2 and DP12 has been applied to these data, with tuning parameters chosen by cross-validation, leading to 5.0 and 4.9 effective parameters respectively (calculated as described in Appendix A). The plotted curve is that of DP12, but the P2-curve is almost identical, which is typical when there are many training data.

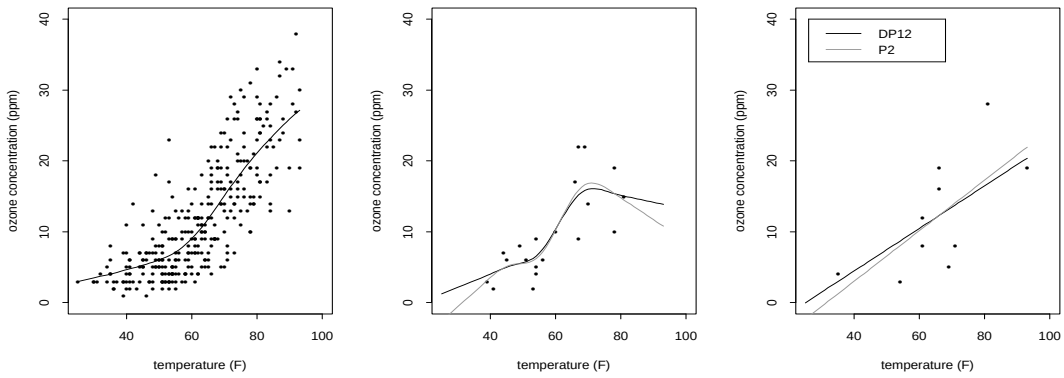

Figure 1: Ozone vs. temperature with regression functions by ordinary P-splines P2 and double P-splines DP12 (actually DP12seq21, see section 3.3).

Assume then that only 10 observations were available in the training data, as in the right-hand panel (drawn randomly from the full data sample). In this situation, a linear model seems appropriate. And indeed, the cross-validation procedure chose $\lambda = \infty$ for both P2 and DP12. Thus, here P2 is equivalent to

OLS with 2.0 effective parameters, whereas DP12 is equivalent to ridge regression with 1.8 effective parameters. The DP12-curve is flatter than the P2-curve.

In the middle panel, there are 20 observations. Both estimated curves are nonlinear, but the DP12-curve is again flatter than the P2-curve. The effective number of parameters is now 4.8 (P2) and 3.9 (DP12).

The above example illustrates how the "optimal" values of the tuning parameters γ and λ depend on the sample size of the training data. In addition, the "optimal" tuning parameters will depend on characteristics of the true relationship between y and x, such as the degree of nonlinearity and signal to noise ratio $[Var\{\mu(\mathbf{x})\}/Var\{y-\mu(\mathbf{x})\}]$.

3 Methods

3.1 Ridge regression and ordinary least squares

Let now y be a n-dimensional vector of response observations, $\mathbf{X}$ a nx$(p+1)$-dimensional design matrix with 1's in the first column and the predictor observations in the last p columns, and β the $(p+1)$-dimensional vector of regression coefficients. Further, let $\mathbf{I}^0$ be a $(p+1)$x$(p+1)$ diagonal matrix with 0 in the first diagonal element and 1 in the others. Then, assuming the linear model (3), the solution to the penalized least squares criteria (4) or (5) is given by the ridge regression estimate

$$\hat{\beta}^{RR} = (\mathbf{X}'\mathbf{X}+\gamma\mathbf{I}^0)^{-1}\mathbf{X}'\mathbf{y}, \tag{7}$$

where $'$ means transpose.

For practical purposes, the matrix $(\mathbf{X}'\mathbf{X}+\gamma\mathbf{I}^0)$ is considered as invertible if the ratio of the largest to the smallest eigenvalue is less than 10^{10}. If not, γ is increased until this is fulfilled.

3.2 Penalized B-splines - P-splines

Marx and Eilers (1998) proposed a variant of generalized additive models (Hastie and Tibshirani,1990) based on penalized B-splines. I give a short review below for p predictors. Eilers and Marx (1996) give a detailed description in the case with one predictor.

A B-spline of degree q consists of $(q+1)$ polynomial pieces, each of degree q, joined at q inner knots. At the joining points, derivatives up to order $(q-1)$ are continuous. The B-spline is positive on a domain spanned by $(q+2)$ knots, and

elsewhere 0. The left-hand panel in Figure 2 shows a B-spline of degree $q = 3$. The right-hand panel of the figure shows a set of $r = 13$ equidistant B-splines in the range from 0 to 1, constructed by dividing the interval from 0 to 1 into $(r-q) = 10$ equal intervals. At any given value between 0 and 1, $(q+1)$ B-splines are nonzero, and their sum is 1.

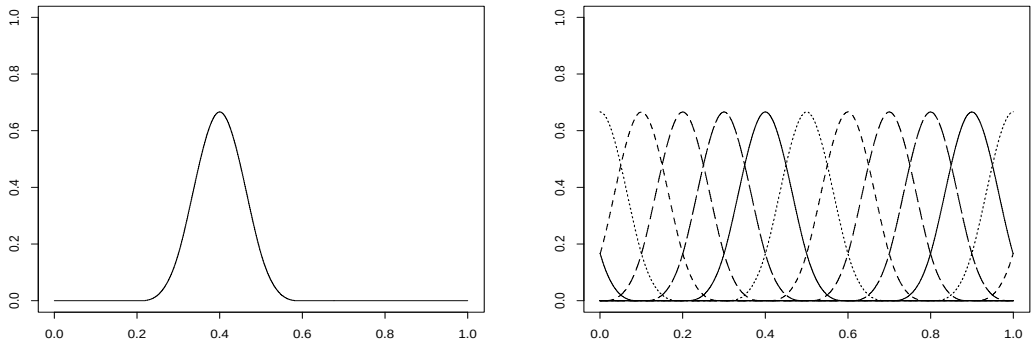

Figure 2: Left panel: One B-spline basis function. Right panel: A set of B-spline basis functions.

The value x_j of a predictor variable is transformed into r new variables $b_l(x_j), l = 1, \ldots, r$, where $b_l(x_j)$ denotes the value at x_j of the lth B-spline. The jth smooth function s_j in (1) is then assumed to be satisfactorily represented by the linear combination

$$s_j(x_j) \quad = \quad \sum_{l=1}^{r} \alpha_{jl} b_l(x_j). \tag{8}$$

This is repeated for all p s-functions, with the same number r of B-splines used for each predictor variable. Then the full model becomes

$$\mu(\mathbf{x}) \quad = \quad \alpha_0 + \sum_{j=1}^{p} \sum_{l=1}^{r} \alpha_{jl} b_l(x_j), \tag{9}$$

now denoting the intercept by α_0.

We now have a flexible functional form of $\mu(\mathbf{x})$, but the model contains $r \cdot p + 1$ free or effective parameters, and when r is large, there are possibly too many which may lead to severe over fitting. To overcome this, Marx and Eilers (1998) introduce difference penalties on adjacent estimated α-coefficients. The 0th order difference of α_{jl} is the coefficient value itself, the first order difference is $(\alpha_{jl} -$

α_{jl-1}), and the second order difference is $\{(\alpha_{jl} - \alpha_{jl-1}) - (\alpha_{jl-1} - \alpha_{jl-2})\} = (\alpha_{jl} - 2\alpha_{jl-1} + \alpha_{jl-2})$. Higher order differences may be used as well, but here I will only consider differences up to order 2. The penalized fitting criterion is now

$$\sum_{i=1}^{n}\{y_i - \tilde{\mu}(\mathbf{x}_i)\}^2 \tag{10}$$
$$+\gamma \cdot \sum_{j=1}^{p}\sum_{l=2}^{r}(\tilde{\alpha}_{jl} - \tilde{\alpha}_{jl-1})^2 + \lambda \cdot \sum_{j=1}^{p}\sum_{l=3}^{r}(\tilde{\alpha}_{jl} - 2\tilde{\alpha}_{jl-1} + \tilde{\alpha}_{jl-2})^2.$$

The penalizations on the first and second order differences approximate the corresponding penalizations on $\tilde{s}'_j$ and $\tilde{s}''_j$ in (6) (see Eilers and Marx 1996). Using the same number r of B-splines for all predictors corresponds to scaling the predictor variables to cover the same range as assumed in (6).

This approach is called penalized B-splines or P-splines by Marx and Eilers (1998). They essentially consider penalizing with just one of the first or second (or higher) order differences. If $\gamma = 0$, (10) corresponds to P2, and thus gives ordinary (generalized) additive models. If $\lambda = \infty$, (10) corresponds to (5), and yields an alternative formulation of ridge regression. I denote this P1.

The vector $\hat{\alpha}$ of coefficients that minimizes (10) is rather simple to compute, and is given in Appendix A, which also includes a formula to compute the effective number of parameters. Table 1 shows the effective number of parameters for some special cases. In this paper, $r = 13$ for all methods.

γ	λ	Effective no. of parameters
0	0	$r \cdot p + 1$
0	∞	$p + 1$
∞	∞	1

Table 1: Effective number of parameters.

3.3 Double penalized B-splines - DP-splines

Double penalized B-splines were used by Eilers and Marx (2003), when the predictor variables were spectroscopic measurements ordered by wavelength. In contrast to the additive model I have in mind, their regression model was still linear, but the ordered regression coefficients varied smoothly from coefficient 1 to p. I shall investigate if the use of double penalization in the additive model (1) can give improved predictions. I will denote this method DP12, since it penalizes both the first and second order differences. The tuning parameters will be estimated by cross-validation, see next section. It is perhaps most natural to estimate both

tuning parameters simultaneously (called DP12sim). Another possibility is to estimate the tuning parameters sequentially. If we regard DP12 as a refinement of P2, we can first estimate λ by P2, and then estimate γ for fixed λ (DP12seq21). This is the method used in Figure 1. Or, conversely, first estimate γ by P1, and then estimate λ for fixed γ (DP12seq12). In a study of so called length-modified ridge regression (Aldrin, 1997), sequential estimation of two tuning parameters gave less prediction uncertainty than simultaneous estimation.

3.4 Predictions outside range of training data

Consider the challenging situation when predicting a new response for predictor variables outside the range of the training data. Within this range, $(q+1)$ B-splines are nonzero, but outside this range fewer than $(q+1)$ B-splines are nonzero, and at a distance $(2/(r-q))$ outside the range, all B-splines are 0. In other words, $s_j(x_j)$ is 0 if x_j is far enough from the range of the training data. This is an unwanted property that is not covered by Marx and Eilers (1998). To overcome this, I define $s_j(x_j)$ to be linear outside the assumed range $[0,1]$, with first derivatives $s_j^{'}(x_j) = s_j^{'}(0)$ if $x_j < 0$ and $s_j^{'}(x_j) = s_j^{'}(1)$ if $x_j > 1$.

Alternatively, extrapolation may also be handled by defining extra B-splines outside the range of the training data as in Currie, Durban and Eilers (2003), but I will not use this approach in the present paper.

3.5 Estimating tuning parameters by permuted leave-k-out cross-validation

Alternative methods for choosing tuning parameters includes Akaikes information criterion, (ordinary) cross-validation or generalized cross-validation, see for instance Eilers and Marx (1996) or Koenker and Mizera (2004), but here we restrict ourselves to (ordinary) cross-validation. The most popular variant is leave-one-out cross-validation, i.e. the tuning parameters are estimated by minimizing

$$\sum_{i=1}^{n} \{y_i - \hat{\mu}_{(-i)}(\mathbf{x}_i)\}^2, \tag{11}$$

where $\hat{\mu}_{(-i)}$ is the regression function estimated without the ith observation. In linear models, this can equivalently be calculated as

$$\sum_{i=1}^{n} \frac{\{y_i - \hat{\mu}(\mathbf{x}_i)\}^2}{(1-h_{ii})^2}, \tag{12}$$

i.e. without any need to recalculate $\hat{\mu}$ n times. Here h_{ii} is the ith diagonal element of the so called hat matrix $\hat{\mathbf{H}}$ (see Appendix A). This expression is used in Marx and Eilers (1998), since P-splines are linear in the $\mathbf{B}$-basis. But (12) does not cover extrapolation, i.e. $\mathbf{x}_i$ lying outside the range of $\mathbf{x}_{(-i)}$, as discussed in section 3.4. If we assume that (11) is calculated with extrapolation performed as described in the previous section, (12) will not be exactly equivalent to (11).

The so called V-fold cross-validation is a common alternative to leave-one-out cross-validation. It saves computer time, and several authors (e.g. Breiman and Spector 1992) have noted that it often yields better results than the leave-one-out variant. The training data are divided into V groups with approximately n/V observations in each group, and the response values in each group are predicted by the model estimated from the remaining $(V-1)$ groups. Denote the groups by $v, (v = 1, \ldots, V)$, and let the subscript $(-v)$ symbolize estimates without the vth group. The criterion to be minimized is then

$$\sum_{v=1}^{V} \sum_{i \in v} \{y_i - \hat{\mu}_{(-v)}(\mathbf{x}_i)\}^2. \tag{13}$$

In V-fold cross-validation about $k \approx n/V$ observations are left out for prediction, and the term leave-k-out may also be used to emphasize that more than one observation may be left out. In the context of variable subset selection in a linear model, Shao (1993) showed that when $n \to \infty$, the proportion to be left out should go to 1, the opposite of what happens in leave-one-out cross-validation! Thus V should decrease when n increases, whereas the number k of observations to be left out should increase faster than n.

There is no unique way to divide the training data into V groups. Assume that one divides the data into V groups by putting the first k observations in the first group, the next k observations in the second group etcetera. Then, if the data were permuted, the group assignment would be different, and the cross-validated sum of squares (13) would differ. An obvious extension of leave-k-out cross-validation is then to perform this several times for various permutations of the data, and taking the average over (13). In this way all response values y_i will be predicted by more than one estimated model, and the variance of the optimization criterion will be reduced. I will call this permuted leave-k-out cross-validation, see Shao (1993) and Pan (1999) who suggested a similar method called bootstrap-smoothed cross-validation.

4 Simulation experiment

4.1 Data

The experiment is based on 16 real data sets. I consider these 16 data sets as representative of a wide range of typical data sets and treat them as benchmarks in a comparative simulation study. In some of these, the relationship between the response and the predictor variables is nearly linear, whereas it is highly nonlinear in others. The number of predictor variables varies between 1 and 13. The basic action in the simulation experiment is to draw a training data set randomly from the full data set, and predict the remaining observations. The number of observations in the training data is varied between 10 and 320.

The 16 basic data sets are listed in Table 2. They have been selected according to the following criteria: They are real data sets with a continuous response, at least one continuous predictor variable, and at least 45 observations. The response variables are here consistently used on their original scale as they were published, even if they may have been transformed in the original analysis. To make the interpretation of the results easier, the data sets have been characterized as either nearly linear or nonlinear, according to the presence of a clear nonlinear relationship between the response and the predictor variables or not. If *all* the nonlinear methods give better predictions than the best linear method when the number of observations is at maximum, then the data set has been characterized as nonlinear, the remaining data sets are characterized as nearly linear. This division is rather rough, and for some of the data sets characterized as nearly linear, some (but not all) of the nonlinear methods still perform better than the linear methods when the number of observations is high. Within each group, the data sets are further sorted according to the number of predictor variables. Figure 3 shows scatter plots of the response versus each predictor in turn for each data set.

4.2 Design of the simulation experiment

For each data set, the following experiment is carried out: A training set of n observations is drawn without replacement from the original m observations. For each method (see Table 3), the prediction model, including tuning parameters, is estimated from the training data. Finally, the remaining $m - n$ observations are predicted, and the prediction error is calculated. This procedure is repeated, and an average prediction error is calculated for each method. The number n of observations in the training data varies through the values $10, 20, 40, 80, 160$ and 320, but of course only as far as $n < m$. 100 repetitions are carried out if n is 10 or 20, else otherwise the number of repetitions is 50. When the training

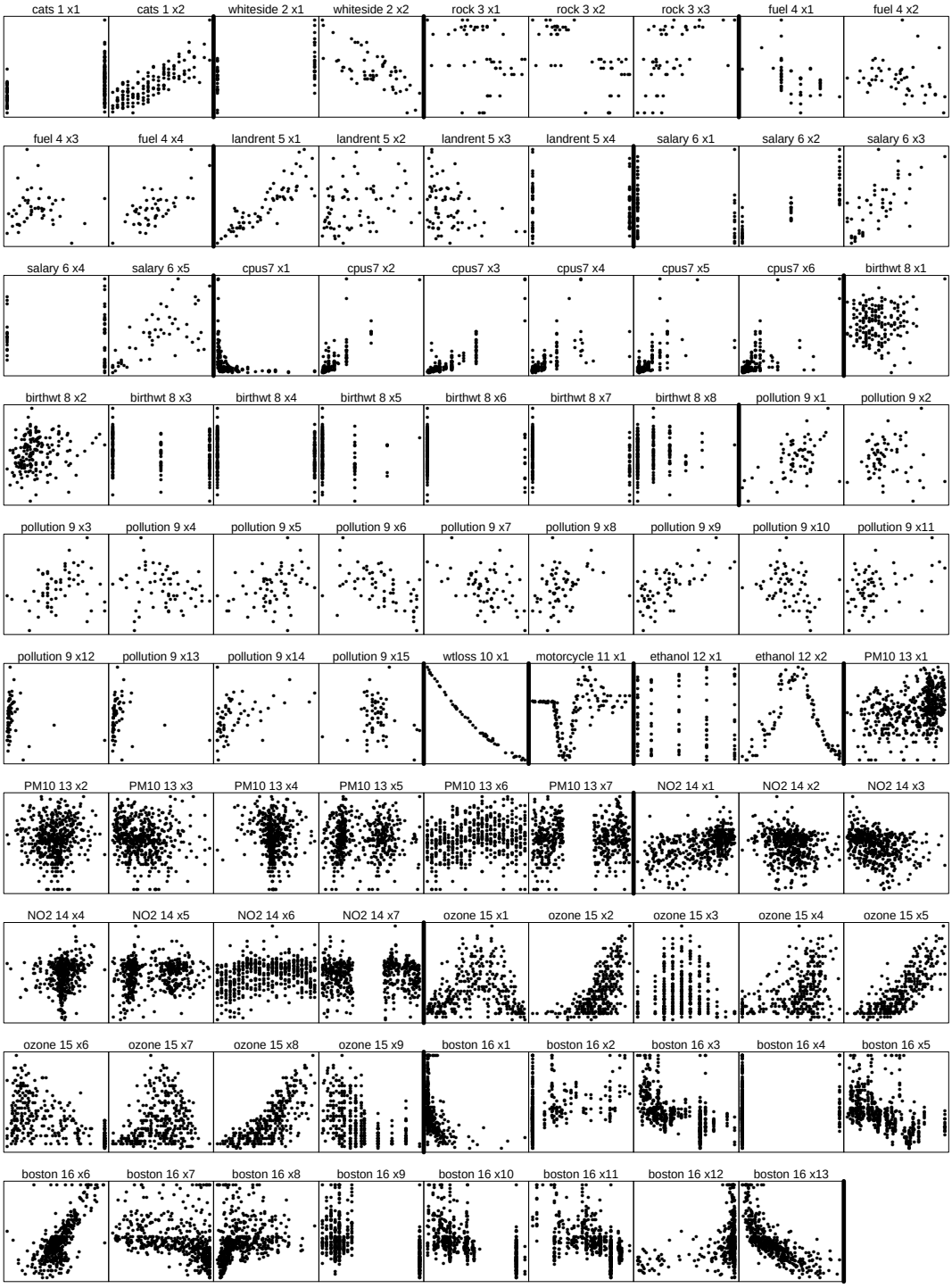

Figure 3: Scatter plots of data sets, with responses on the y-axes and predictor variables on the x-axes. The headings give name and number of data sets and the number of the predictor variables.

No.	Name	Source	Electronic source	m	p	
1	cats	Venables and Ripley (1999)	S-plus MASS	144	2	nearly linear
2	whiteside	Venables and Ripley (1999)	S-plus MASS	56	2	nearly linear
3	rock	Venables and Ripley (1999)	S-plus MASS	48	3	nearly linear
4	fuel	Weisberg	StatLib	48	4	nearly linear
5	landrent	Weisberg (1985)	StatLib	67	4	nearly linear
6	salary	Weisberg (1985)	StatLib	52	5	nearly linear
7	cpus	Venables and Ripley (1999)	S-plus MASS	52	7	nearly linear
8	birthwt	Venables and Ripley (1999)	S-plus MASS	189	8	nearly linear
9	pollution	McDonald and Schwing (1973)	StatLib	60	15	nearly linear
10	wtloss	Venables and Ripley (1999)	S-plus MASS	52	1	nonlinear
11	motorcycle	Hardle (1990)		133	1	nonlinear
12	ethanol	S-plus	S-plus	88	2	nonlinear
13	PM_{10}	this paper, see Appendix B	StatLib	500	7	nonlinear
14	NO_2	this paper, see Appendix B	StatLib	500	7	nonlinear
15	ozone	Breiman and Friedman (1985)	BLSS	330	9	nonlinear
16	boston	Breiman and Friedman (1985)	StatLib	506	13	nonlinear

Table 2: Data sets. MASS is a library in S-plus. BLSS: Abraham and Rizzardi (1988). StatLib: http://lib.stat.cmu.edu .

data are drawn randomly, it may happen that one or more predictor variables in a training data set has no variation, i.e. all observations have the same value. Then the training data set is rejected, and a new one is drawn.

Name	Explanation
Mean	Model with intercept only
OLS	Ordinary least squares
RR	Ridge regression, standardized by equal variance
P1	γ estimated and $\lambda = 0$
P2	λ estimated and $\gamma = 0$
DP12sim	γ and λ estimated simultaneously
DP12seq21	λ estimated first by P2, then γ
DP12seq12	γ estimated first by P1, then λ

Table 3: Methods.

The tuning parameters are estimated by permuted leave-k-out cross-validation. The tuning parameters are varied over the values $10^{-4}, 10^{-3}, \ldots, 10^{4}$, and the values that minimize (13) are chosen. The proportion of data to be left out ($1/V$) varies from 1/5 for small n to 1/3 for large n. The number of permutations is 5 for small n and 3 for large n. The detailed values of V, k and number of permutations are given in Table 4.

n	V	k	Number of permutations
10	5	2	5
20	4	5	5
40	3	13	5
80	3	27	4
160	3	59	3
320	3	107	3

Table 4: Number of observations in training data (n), number of groups (V), approximate number of observations left out (k) and number of permutations in permuted leave-k-out cross-validation.

In addition, two other variants of cross-validation are tried out for some methods: i) the approximate leave-one-out cross-validation based on (12), where extrapolation is not being handled properly; ii) the leave-k-out cross-validation without permutations.

The root mean squared error of predictions is chosen as the principal measure of quality of predictions. For simulation number $s, s = 1, \ldots, S$, let D_s denote the $m - n$ observations to be predicted. Further, let M denote a specific method and let $\hat{\mu}^M_{(-D_s)}(\mathbf{x}_i)$ denote the prediction by method M for the ith observation, based on the training data without the observations in D_s. The root mean squared error of predictions is then calculated as

$$RMSE^M = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\frac{1}{m-n}\sum_{i\in D_s}\{y_i - \hat{\mu}^M_{(-D_s)}(\mathbf{x}_i)\}^2}. \quad (14)$$

In most comparisons, a reference method R will be chosen, and the RMSE log ratio $log(RMSE^M/RMSE^R) = log(RMSE^M) - log(RMSE^R)$ will be presented. This quantity is positive if the reference method R is better than method M and vice versa. A difference of 0.10, say, means that one method has an approximately 10% smaller RMSE than the other.

4.3 Results

The aim of this section is first to choose the best cross-validation procedure (leave-one-out, leave-k-out or permuted leave-k-out), the best variant of ridge regression (RR or P1) and the best variant of DP12 (DP12sim, DP12seq12 or DP12seq21). Then, the selected methods will be compared.

Figure 4 show box plots of various RMSE log ratios. Each box plot is based on the results from 65 combinations of data set and sample size n. The left-hand

panel compares leave-one-out cross-validation to leave-k-out cross-validation by $\log(RMSE^{l-o-o}/RMSE^{l-k-o})$. The box plots tend to lie above 0, which means that leave-k-out cross-validation tends to be the best method of the two. This is confirmed by the mean values, given at the top of the plot. The mean value for RR is 0.01. For the nonlinear methods the difference is larger, up to 3%. In the middle panel, leave-one-out is compared to permuted leave-k-out, and the differences are up to 7% in favour of permuted leave-k-out. Finally, in the right-hand panel, non-permuted leave-k-out is compared to permuted leave-k-out. Here, permuted leave-k-out is the best one, up to 3% better than non-permuted leave-k-out. From now on, I will consider only results using the permuted leave-k-out cross-validation.

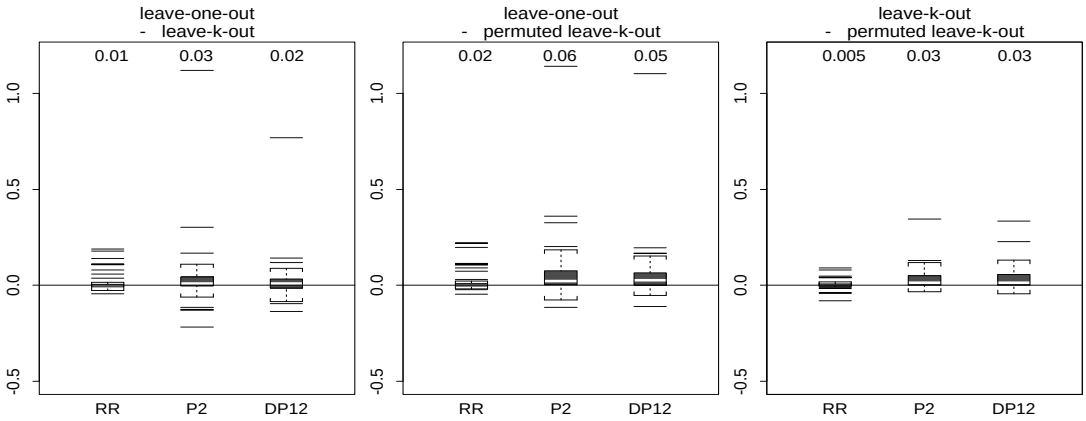

Figure 4: Boxplots of RMSE log ratios for various methods and variants of cross-Validation. Mean values are printed above each boxplot.

Consider then the two variants of ridge regression, as shown in Figure 5. As expected, the difference is rather small, on average only 0.3%. Since P1 is slightly better than RR, I used P1 in the rest of the paper.

Then the variants of DP12 is compared in Figure 6. We see that the sequential DP12seq12, i.e. with γ estimated first, performs rather poorly in some situations, which happen to be when the response-predictor relationship is both highly non-linear and non-monotone as in the motorcycle and ethanol data sets. Thus, this variant is not a good choice as a general method. The simultaneous DP12sim can both perform slightly better and slightly worse than DP12seq21, and is on average only 1% worse. However, DP12sim is computationally slower than DP12seq21, since in the estimation of tuning parameters it needs a simultaneous grid search instead of two single searches. Therefore, DP12seq21 is chosen as the preferred variant of DP12.

We are now ready to compare our main methods. The leftmost panel of Figure 7 shows box plots of the RMSE log ratios for the various methods compared to

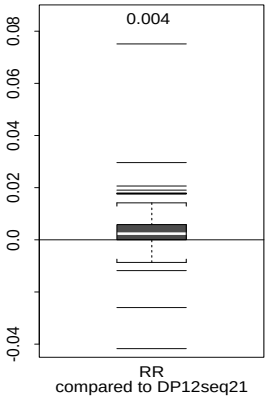

Figure 5: Box plots of RMSE log ratios for ordinary ridge regression (RR) compared to P1 (reference). The mean value is printed above the box plot.

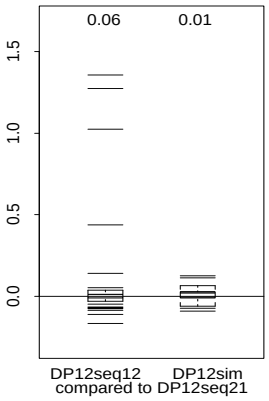

Figure 6: Box plots of RMSE log ratios of various methods compared to DP12seq21 (reference). Some values for the "Mean" and OLS methods lie above the upper limit of the plot. Mean values are printed above each box plot.

DP12seq21. First of all, no method is the best in all situations. However, note that all box plots tend to lie above the zero line, which means that DP12seq21 is the best overall method. Further, the other methods are never very much better than DP12seq21, whereas DP12seq21 sometimes clearly outperforms the competitors. That DP12seq21 is better than "Mean" and OLS is not surprising. Compared to P1, we see that DP12seq21 sometimes is clearly better. DP12seq21 can also often be clearly better than P2, which typical happens when there are few observations (see Figure 8 below).

To investigate this further, the 65 combinations of data set and sample size n are divided into two groups: The first group consists of those 32 combinations where the best of the linear methods is better than the best of the nonlinear methods. Most of these combinations are within the data sets that are characterized as nearly linear, but some are within the nonlinear data sets with small n. The second group consists of the remaining 33 combinations where the best of the nonlinear methods is better than the best of the linear methods. The results are shown in the middle and rightmost panels of Figure 7. Not surprisingly, most of the combinations where DP12seq21 is clearly better than P2 are within the group that are in favour of the linear methods. However, there are still some combinations in the other group where DP12seq21 is considerably better than P2, which means that there are situations where P12seq21 outperforms both P1 and P2.

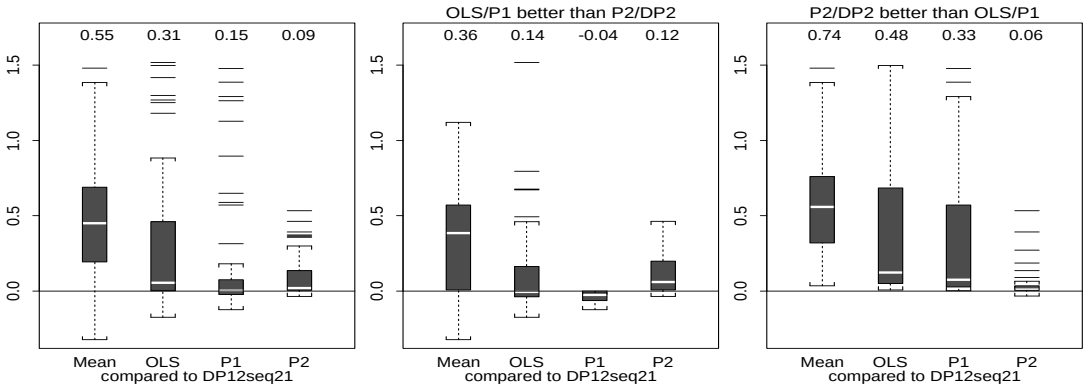

Figure 7: Box plots of RMSE log ratios of various methods compared to DP12seq21 (reference), for all combinations (leftmost panel), combinations in favour of linear methods (middle panel) and combinations in favour of nonlinear methods (rightmost panel). Some values for the "Mean" and OLS methods lie above the upper limit of the plot. Mean values are printed above each box plot.

Figure 8 shows the RMSE log ratios in detail for selected methods, compared to DP12seq21. If some values are outside the plot borders, they are set at the border.

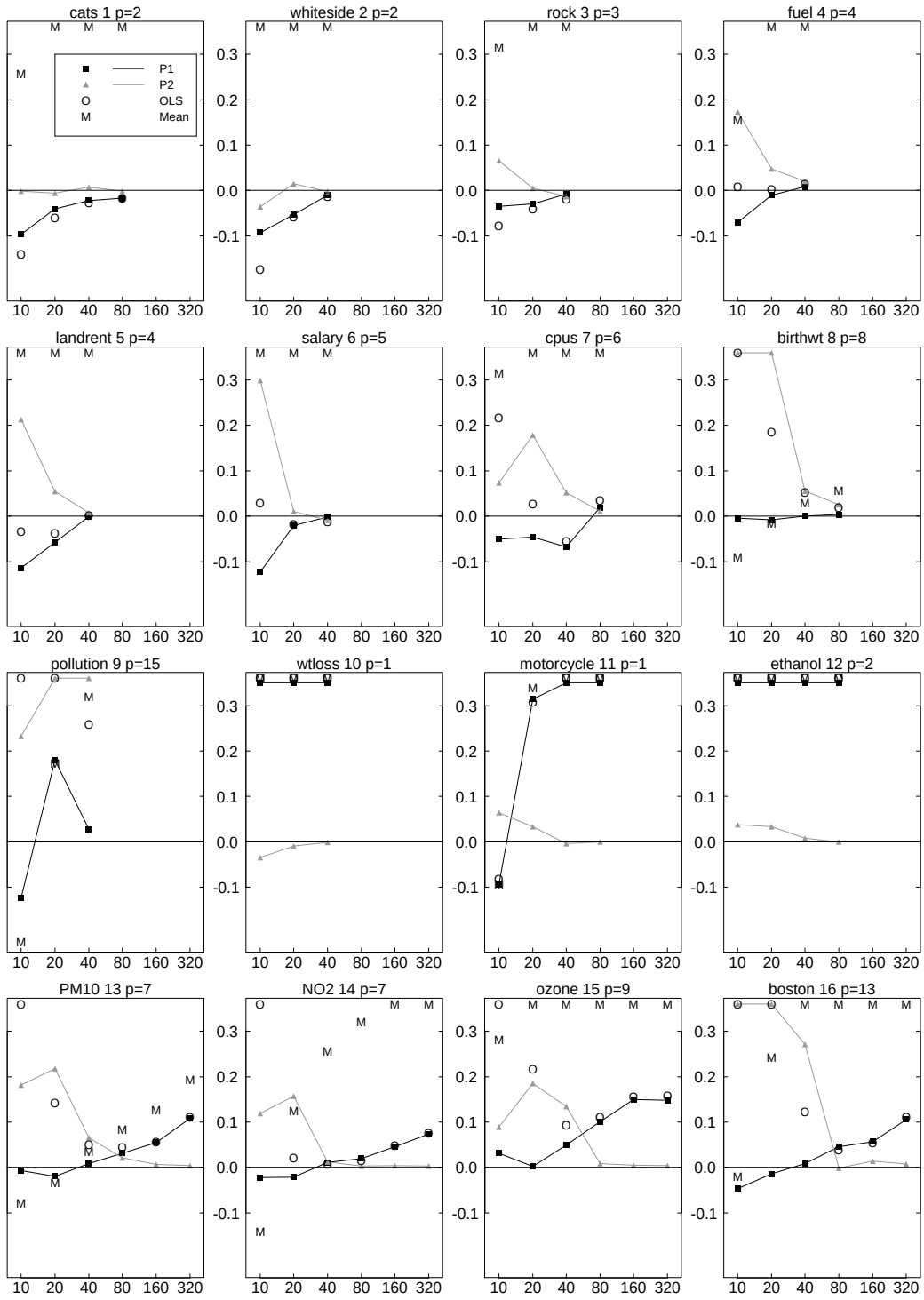

Figure 8: RMSE log ratios of various methods compared to DP12seq21 (reference). Values above upper border or below lower border are plotted at the borders.

For the nearly linear data sets ($1-9$), the tendency is that P1 is better than DP12seq21 when there are few observations in the training data, and that the two methods become almost equal when more data are available. For the nonlinear data sets ($10-16$) the picture is naturally more in favour of the nonlinear method DP12seq21, and P1 is clearly outperformed when there are many data. The comparison between P2 and DP12seq21 is very much in favour of the latter. When there are few data available for estimation, DP12seq21 is

usually considerably better for both the nearly linear and the nonlinear data sets, whereas the two methods become more similar when the number of observations increases.

5 Discussion

The results of the previous section show that penalizing both slope and curvature (DP12 or actually DP12seq21) is a very good strategy that outperforms in general the ordinary approach in additive models where only curvature is penalized (P2). The gain can be large when there are few data, whereas the potential loss seems to be small for large training data sets. Compared to the linear ridge regression, DP12 can be worse in situations with few data. On the other hand, DP12 performs substantially better than ridge regression in situations with a nonlinear response-predictor relationship and enough training data. Thus, DP12 seems to be a good candidate as an overall prediction method. DP12 fits into the general framework of Wahba (1990).

The permuted leave-k-out cross-validation procedure is more precise than the usual leave-one-out and the non-permuted leave-k-out procedures.

DP12 will also be useful in situations where one is interested in the regression curve $\mu(\mathbf{x})$ rather than predicting new response values. The prediction error $y - \hat{\mu}(\mathbf{x})$ can be written as $\{y - \mu(\mathbf{x})\} + \{\mu(\mathbf{x}) - \hat{\mu}(\mathbf{x})\}$, i.e. it includes the model error $\hat{\mu}(\mathbf{x}) - \mu(\mathbf{x})$ as well as the unpredictable quantity $y - \mu(\mathbf{x})$. Therefore, if one method yields smaller prediction errors than another method, the improvement in the model error will be even larger.

When the aim is to estimate the effect of each predictor variable on the response one should be more careful, since the estimates will be biased towards 0, i.e. no effect. However, sometimes the main focus is in estimating the effect of *one* or a few predictor variables of interest, and controlling for the effect of the other predictor variables. Then it may be useful to drop the penalization on the first difference for that predictor variable(s) of interest, and keep penalization for the others.

Finally, DP12 may certainly be applied to non-Gaussian responses as in general-

ized additive models, by replacing the sum of squares by the log likelihood, in the penalized fitting criterion as well as in the cross-validation. However, estimating the tuning parameters may be more difficult for instance for binary responses, if the number of observations is moderate. On the other hand, estimating the s-functions may also be more difficult, thus increasing the need for penalizing the first difference. The potential benefit of using DP12 in generalized additive models has to be studied further.

Acknowledgements

This work was partially sponsored by the Norwegian Research Council, project 154079/420. The author thanks Arnoldo Frigessi and David Hirst for helpful comments.

A Details on P-Splines

The vector $\hat{\alpha}$ of coefficients that minimizes (10) is rather simple to compute. First, the original design matrix $\mathbf{X}$ is transformed to a new nx$(1 + r \cdot p)$ design matrix $\mathbf{B} = (\mathbf{1}|\mathbf{B}_1|\cdots|\mathbf{B}_p)$, where each $\mathbf{B}_j$ is a nxr matrix with the values of $b_l(x_j)$ in the lth column.

Then define the three matrices $\mathbf{D}^d$, $d = 0, 1, 2$ of dimension $(r-d)$xr: $\mathbf{D}^0$ is the identity matrix; $\mathbf{D}^1$ has the value -1 in element (l, l), the value 1 in element $(l, l+1)$ for $l = 1, \ldots, r-1$, and is 0 elsewhere; $\mathbf{D}^2$ is 1 in element (l, l), -2 in element $(l, l+1)$ and 1 in element $(l, l+2)$ for $l = 1, \ldots, r-1$, and 0 elsewhere. Then the three matrices $\mathbf{P}^d$, $d = 0, 1, 2$, are defined by blockdiag$(0, \mathbf{D}^{d'}\mathbf{D}^d, \ldots, \mathbf{D}^{d'}\mathbf{D}^d)$, where the $\mathbf{D}^d\mathbf{D}^d$ is repeated p times.

Marx and Eilers (1998) notice that there is no unique solution to (10) because $\mathbf{B}$ is singular, since the columns of each $\mathbf{B}_j$ sum to 1. To overcome this, they add a small penalization to the 0th order difference, i.e. the term $\delta \cdot \sum_{j=1}^{p} \sum_{l=1}^{p} (\tilde{\alpha}_{jl})^2$ is added to (10), where δ is a small positive constant. The solution to this slightly modified version of (10) is then

$$\hat{\alpha} \;=\; (\mathbf{B}'\mathbf{B} + \delta\mathbf{P}^0 + \gamma\mathbf{P}^1 + \lambda\mathbf{P}^2)^{-1}\mathbf{B}'\mathbf{y}. \tag{15}$$

The role of δ is only to ensure that the solution is unique, i.e. to ensure that $(\mathbf{B}'\mathbf{B} + \delta\mathbf{P}^0 + \gamma\mathbf{P}^1 + \lambda\mathbf{P}^2)$ is invertible. This matrix is considered as invertible if the ratio of the largest and smallest eigenvalue is less than 10^{10}. If not, δ is increased until this is fulfilled. The minimum value of δ is 10^{-4}.

The effective number of parameters in the model may be computed as the trace of the so called hat matrix defined by

$$\hat{\mathbf{H}} \;=\; \mathbf{B}(\mathbf{B}'\mathbf{B} + \delta\mathbf{P}^0 + \gamma\mathbf{P}^1 + \lambda\mathbf{P}^2)^{-1}\mathbf{B}'. \tag{16}$$

B Description of two air pollution data sets

Two data sets used in the simulation experiment are unpublished, but are available at StatLib (http://lib.stat.cmu/edu). They have been collected by The Norwegian Public Roads Administration, and originate in a study where air pollution at a road is related to traffic volume

and meteorological variables. The response variables consist of hourly values of the logarithm of the concentration of PM_{10} (particles) in data set 13 and of NO_2 in data set 14, measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. The predictor variables are the logarithm of the number of cars per hour, temperature 2 meter above ground (degree C), wind speed (meters/second), the temperature difference between 25 and 2 meters above ground (degree C), wind direction (degrees between 0 and 360), hour of day and day number from October 1. 2001. The observations in the two data sets are taken at different time points, so their $\mathbf{X}$-matrices are different.

References

Aldrin, M (1997). Length-modified ridge regression. *Computational Statistics & Data Analysis*, **25**, 377-398.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of American Statistical Association*, **80**, 580-619.

Abrahams, D. and Rizzardi, F (1988). *BLSS - The Berkeley interactive statistical system.* New York: W. W. Norton.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review*, **60**, 291-319.

Currie, I., Durban, M. and Eilers, P. (2003). Using P-splines to extrapolate two-dimensional Poisson data. Paper presented at the 18th International Workshop on Statistical Modelling, Leuven, Belgium.

Eilers, P. H. C. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, **66**, 159-174.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89-121.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109-147.

Härdle, W. (1990). *Applied nonparametric regression.* Cambridge: Cambridge University Press.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models.* London: Chapman and Hall.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **8**, 27-51.

Koenker, R. and Mizera, I. (2004). Penalized triograms: total variation regularization for bivariate smoothing. *Journal of the Royal Statistical Society, Series B*, **66**, 145-164.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, **28**, 193-209.

McDonald, G.C and Schwing, R. C. (1973). Instabilities of regression estimates relation air pollution to mortality. *Technometrics*, **15**, 463-482.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of American Statistical Association*, **88**, 486-494.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied statistics with S-PLUS - third edition.* New York: Springer.

Weisberg, S. (1985). *Applied Linear Regression - second edition.* New York: Wiley.

Wahba, G. (1990). *Spline Models for Oservational Data.* Philadelphia: SIAM.