A Computer Intensive Method for Choosing the Ridge Parameter

Karsten Luebke * Irina Czogiel Claus Weihs

February 2004

Universität Dortmund
Fachbereich Statistik

Abstract

In this paper we describe a computer intensive method to find the ridge parameter in a prediction oriented linear model. With the help of a factorial experimental design the method is tested and compared to a classical one.

Keywords
Ridge Regression, Experimental Design, Prediction model, Generalized Cross Validation

1 Introduction

The linear regression model is one of the most widely used statistical models. It is used to model relations of one or more dependent variables to one or more explanatory variables. Although the ordinary least squares estimator (OLS) is the uniformly best unbiased estimator for the regression vector when the errors are iid normally distributed, there are situations when there

*e-mail: luebke@statistik.uni-dortmund.de

are better estimators for the given problem – even if the errors are iid normal. This may happen because of numerical reasons in calculating the OLS estimator or in cases in which one is more interested in prediction so that a biased estimator may lead to less prediction error. One possibility to improve the OLS is to perform a ridge regression (RR). In a ridge regression an additional parameter, the ridge parameter, is used to control the bias of the regression towards the mean of the response variable. However, there exist a number of different methods for choosing the ridge parameter. The one we tried is a new computer intensive method which directly minimizes the prediction error.
In this paper we also tried to find factors or characteristics of the data which have an influence on the performance of a ridge regression. Using an experimental design to set up a simulation study is a possible tool to check the importance of potentially influencing factors.
This paper is organized as follows: In the next section the regression model and the ridge regression are introduced. Moreover in section 2 the new computer-intensive method we used is explained. Section 3 describes the simulation study and the experimental design. The results of the simulation study are given in section 4. In section 5 the new method is tested in a real life example. The paper is concluded with comments on the results found.

2 Regression Model and Ridge Regression

2.1 Linear Model

The standard basic multivariate linear model looks as follows:

$$y = 1_n\mu + X\beta + e, \tag{1}$$

where

- $y \in I\!R^n$ the data vector of the response variable.
- $\mu \in I\!R$ the mean of the response.
- $X \in I\!R^{n\times p}$ the data of the explanatory variables. For simplicity reasons it is assumed that X is of full column rank and mean centered.
- $\beta \in I\!R^p$ the unknown regression coefficient vector.

- $e \in I\!R^n$ the vector of the errors.

The ordinary least squares estimators are (if X is of full column rank and mean centered) :

- $\hat{\mu} = \bar{y}$.
- $\hat{\beta} = (X'X)^{-1}X'(y - 1_n\hat{\mu})$.

Since the matrix $X'X$ is inverted there may be numerical problems even if X is of full column rank, for example if the predictor variables are highly correlated.
It is interesting how the estimator will perform on n_0 future values X_0, y_0. The point prediction of the future response values is:

$$\hat{y}_0 = 1_{n_0}\hat{\mu}_y + X_0\hat{\beta}_{X,y}.$$

β and μ are estimated using the training set X, y. With a known test set X_0, y_0 the loss in n_0 (new) observations can be measured by

$$L = \frac{1}{n_0}\|(y_0 - \hat{y}_0)\Gamma^{-\frac{1}{2}}\|^2, \tag{2}$$

where Γ is a fixed weight. One possible choice is the variance of the response variable (Schmidli, 1995, p. 22). This was used here as the loss in y and y_0 is measured relative to the variance of the response.
The loss (2) is closely linked to the R^2 of the response variables (Schmidli, 1995, p. 23) where the R^2 is measured on the future values y_0:

$$R_0^2 = 1 - L. \tag{3}$$

Usually one is not only interested in the performance of the estimator for some observations but also in the 'general' or average performance. The corresponding mean loss (Mean Squared Error of Prediction) is defined as (Schmidli, 1995, p. 24):

$$\begin{aligned} MSEP &= \frac{1}{n_0}E_{y|X}E_{y_0|X_0}\|(y_0 - \hat{y}_0)\Gamma^{-\frac{1}{2}}\|^2 \\ &= \frac{1}{n_0}E_{y|X}E_{y_0|X_0}\|(y_0 - (1_{n_0}\hat{\mu} + X_0\hat{\beta}_{X,y}))\Gamma^{-\frac{1}{2}}\|^2. \end{aligned} \tag{4}$$

The $MSEP$ is a conditional expectation where the distribution is at least partially unknown, therefore it is necessary to estimate it. This can be done

for example by bootstrap or crossvalidation methods. In our problem, however, when $\hat{\beta}$ is not a linear function of X, y, one needs to generate different y, y_0 for the same X, X_0. To see this, note that the conditional expectation in the $MSEP$ depends on the conditional expectation of $\hat{\beta}_{X,y}$. As y itself is a function of X it follows that $E_{y|X}(\hat{\beta}_{X,1_n\mu+X\beta+e})$ only equals $\hat{\beta}_{X,1_n\mu+X\beta}$ if $\hat{\beta}$ is linear. To generate different y, y_0 two cases can be considered:

1. Simulated data
   In a simulation study different e can be generated according to the error distribution.

2. Real life data
   The error distribution must be estimated by a bootstrap method (Stapleton, 1995, p. 225). Let

   $$\hat{e} = y - \hat{y}.$$

   This estimate of the error matrix builds the empirical error distribution. So let e^* be a random bootstrap sample of the empirical error distribution. A new bootstrap sample is:

   $$y^* = \hat{y} + e^*.$$

   By doing this it is possible to generate –for a given split into training and test data– different realizations of the response variables and estimate the conditional expectations in the $MSEP$.

2.2 Ridge Regression

Hoerl and Kennard (1970) propose the use of ridge regression to estimate β when the explanatory variables are highly correlated. Ridge regression is based on the James-Stein estimator and the basic idea is to reduce the variance by shrinking the estimator so that the MSE can be reduced. To achieve that in a ridge regression an additional parameter k is added to the OLS estimation problem:

$$\hat{\beta}(k) = (X'X + kI_p)^{-1}X'(y - 1_n\hat{\mu}), \quad k \geq 0. \tag{5}$$

If $k = 0$ the resulting estimator is the OLS estimator for β.
The procedure can be generalized to choose individual ridge parameters for

each predictor variable (Hoerl and Kennard, 1970).
There are a number of different methods for selecting the value of k. A widely used method is the generalized cross validation (gcv) criterion (Golub et al., 1979). The gcv criterion is a constant multiple of

$$V(k) = \frac{||(I-H)y||^2}{(trace(I-H))^2}, \tag{6}$$

with $H = X(X'X + kI)^{-1}X'$. The Minimization is often done by line search of k (Hawkins and Xin, 2002). The gcv estimate is based on the minimization of the PRESS criterion but transformed to be rotation-invariant (Golub et al., 1979). They also show, that it is equivalent to a weighted version of PRESS. Remember that PRESS is based on a leave-one-out cross validation. On the other hand Bunke and Droge (1984) show that a bootstrap estimator is preferable to a cross validation estimator for the prediction error, so that gcv might not choose the best k concerning prediction error.

2.3 Prediction Optimal Ridge Parameter

The loss (2) for a given training and test data in a ridge regression model can be calculated as:

$$L(k) = \frac{1}{n_0}\|(y_0 - (1_{n_0}\hat{\mu} + X_0(X'X + kI_p)^{-1}X'(y - 1_n\hat{\mu})))\Gamma^{-\frac{1}{2}}\|^2. \tag{7}$$

Including a ridge-parameter in the equation for the MSEP (4) the formula can be re-written as a function of k:

$$MSEP(k) = \frac{1}{n_0}E_{y|X}E_{y_0|X_0}\|(y_0-(1_{n_0}\hat{\mu}+X_0(X'X+kI_p)^{-1}X'(y-1_n\hat{\mu})))\Gamma^{-\frac{1}{2}}\|^2. \tag{8}$$

As the MSEP is estimated by a bootstrap method (see above) a computer intensive method for choosing k is developed quite easily. Because of the results of Bunke and Droge (1984) this method may result in a better prediction. It is summarized in Algorithm 1.

As the objective is to find a value k which minimizes (8) directly we call the new method PrK (Prediction K).
Direct minimization of MSEP as a function of a projection matrix turned out to be successful in a latent factor linear model (Luebke and Weihs, 2003). There the minimization is done by means of Simulated Annealing which is

Algorithm 1 Algorithm for selecting ridge parameter k

for $i = 1$ to n_{boot} **do**
  Generate Training (X, y) and Test data (X_0, y_0)
  $k_i = argminL(k)$, $L(k)$ from equation (7)
end for
$\hat{k} = T(k)$, with T being a statistic for the location parameter of $k_i, i = 1, \cdots, n_{boot}$, e.g. mean or median

not necessary here. One parameter (k) needs to be optimized and in our examples (7) has only got one minimum so there is no need to overcome local minima – a problem that is tackled by Simulated Annealing.

For the parameters of Algorithm 1 we used 20 splits in training and test data and 10 replications of the bootstrapped error, so $n_{boot} = 20 \cdot 10 = 200$. As the estimator for k we used the mean. The minimization is done by line search and all calculations are done by the statistical program R (Ihaka and Gentleman, 1996).

3 Design of Experiment of Simulation Study

In order to obtain most general results, important characteristics of the model are varied by using a 2^5 factorial experimental design. The design is a full factorial design so there were 32 different runs.

- For the data matrix X we vary the number of training observations (rowx) and the number of variables (colx).

- Numerical stability, as one of the main reasons to use a ridge regression model is tested by means of the degree of multicollinearity of the matrix X (Belsley et al., 1980, p. 86). Multicollinearity may happen when the explanatory variables are correlated. As a measure of the degree of multicollinearity the condition number (kond) is used (Belsley et al., 1980, p. 104). To achieve different degrees of multicollinearity we use the Cholesky decomposition $U'U = \Psi$ of a correlation matrix Ψ where the off-diagonal elements of Ψ are identical. Then X is transformed $X = XU$ (Frank, 1989).

- The influence of error variances (sn) is assessed by different variances of e varying the signal-to-noise ratio (Frank and Friedman, 1993).

- The elements of the regression vector β are either equal to one or unequal ($[\beta_j = j^2]_{j=1}^p$, compare Frank and Friedman (1993)).

The chosen values of the characteristics of model (1) are shown in Table 1. The low level (-1) is used for the situation in which calculating $\hat{\beta}$ should be easier whereas the high level (+1) is used for the more complicated situation.

Table 1: Design of experiment of simulation study

Characteristic	*Realization -1*	*Realization +1*
rowx	250	50
colx	5	25
kond	50	5000000
sn	3	2
eq	$[\beta_j = 1]_{j=1}^p$	$[\beta_j = j^2]_{j=1}^p$

4 Results

Each run of the factorial experimental design is repeated 50 times, so there were 160 runs altogether. The regression parameter and the ridge parameter are estimated by the training data and were then tested on 500 observations of validation data not used for estimation. As a result the logarithmed relative (relative to the true model) prediction error is calculated:

$$logrel(method) = log(\frac{(y - \hat{y}_{method})'(y - \hat{y}_{method})}{(y - \hat{y}_{true})'(y - \hat{y}_{true})}). \tag{9}$$

In order to analyze the influence of the possible influencing factors, a regression of the logrel (9) on the coded factors is performed. As the regression is based on the coded factors the variances are equal, so the values of the estimators can be compared directly. The results of the regression coefficients of the various methods (OLS, gcv, PrK) together with the associated p-values are shown in Table 2. Table 2 reveals that only the three characteristics of the data matrix X (colx, rowx, kond) are significant for the relative performance of the different (ridge) regression methods.

Table 2: Regression of the *logrel* on the coded factors

	ols		gcv		PrK	
	est.	p-value	est.	p-value	est.	p-value
(Intercept)	0.240	0.000	0.022	0.000	0.017	0.000
sn	0.000	1.000	−0.001	0.453	−0.001	0.050
colx	0.178	0.000	0.004	0.000	0.001	0.074
rowx	0.177	0.000	0.014	0.000	0.010	0.000
kond	0.000	1.000	−0.010	0.000	−0.003	0.000
eq	−0.000	1.000	0.002	0.112	0.001	0.021

Table 3: Regression of k on the coded factors

	gcv		PrK	
	est.	p-value	est.	p-value
(Intercept)	13.172	0.000	27.493	0.000
sn	2.124	0.000	5.894	0.000
colx	10.357	0.000	17.534	0.000
rowx	−4.331	0.000	−10.489	0.000
kond	−10.451	0.000	2.517	0.000
eq	−0.394	0.390	−0.020	0.946

The computer intensive method PrK performs better than gcv in the overall mean (Intercept). This does not indicate that PrK is outperforming gcv in general – as we used simulated data. The magnitude of the influence on the performance of the factors is quite similar for gcv and PrK.
The factors that influence the regression vector (eq) and the signal-to-noise have no significant influence on the relative performance of the methods.
The overall MSEP of the three methods is illustrated in Figure 1. It can be seen in Figure 1 that both ridge regression methods are more stable than OLS in the complicated situations.

Another point of interest was whether there are factors in the simulation study which influence the magnitude of k. The results of a regression of k on the coded factors are given in Table 3. In general PrK chooses a greater k then gcv. A rather surprising result is that gcv chooses a smaller k in situations

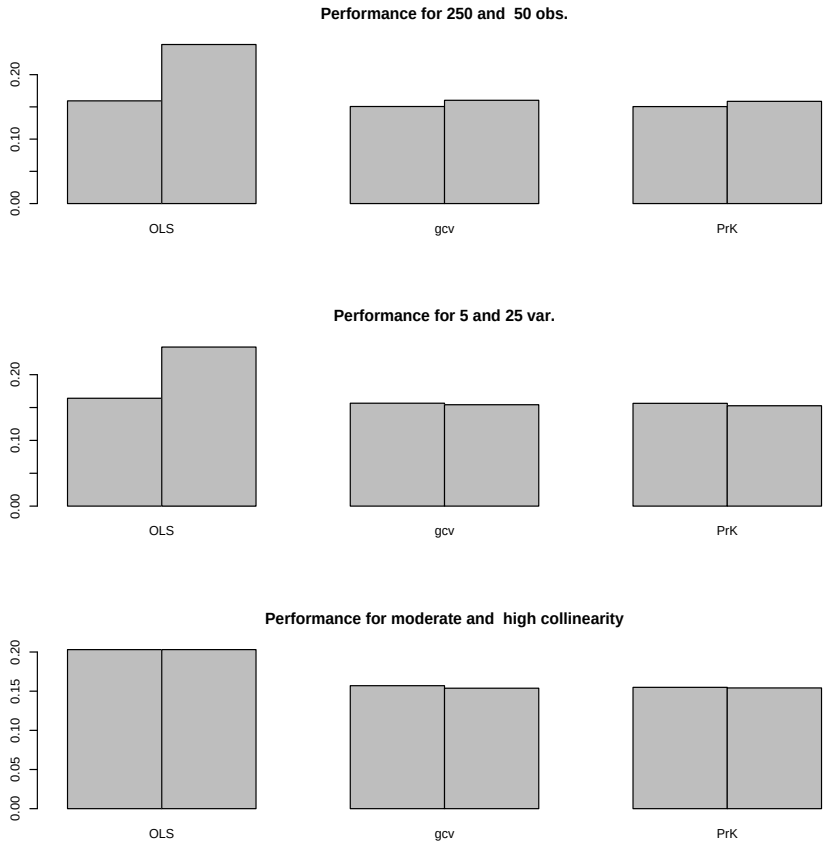

Figure 1: Overall MSEP for different methods

with a high condition number. This is caused by outliers for choosing k. The median of k in situations with low collinearity is smaller than the median in situations with high collinearity.

5 Example: Pollution Data

We also applied the new PrK method to real data considered in the literature. A description of the data can be found in McDonald and Schwing (1973).

Table 4: Comparison of MSEP in real life data

ols	gcv	PrK
0.784	0.659	0.489

The data is available from Statlib[1]. In this data the age adjusted mortality rate is linear related to some weather, socioeconomic and pollution variables. There are 60 observations and 15 explanatory variables. As the explanatory variables are correlated it is a common example for ridge-regression.
In order to compare the OLS, gcv and PrK we used an e0 bootstrap estimator. For the e0 bootstrap estimator, the training data set consists of n cases sampled with replacement from a size n sample. Cases not found in the training set form the test set. The e0 estimator was calculated 80 times. As there are some outliers in the MSEP we show the median MSEP in Table 4. For this data the new method clearly outperforms the gcv method for finding the ridge parameter (the differences in the mean of the e0 bootstrap estimators are even larger). This superiority may be caused by the fact that there are few training observations compared to the number of variables. Shao (1993) shows that in this situation the prediction ability of a model chosen by leave-one-out cross validation is not optimal.

6 Conclusion

The superiority concerning prediction power of a ridge regression over the ordinary least squares estimator was confirmed. We also found in our simulation study, that the new method is outperforming the gcv method slightly. This was confirmed by a real-life data example.
In the experimental design some factors could be identified that have a significant influence on the performance of estimators as well on the magnitude of k in the regression model.

[1]URL: http://lib.stat.cmu.edu/

Acknowledgment

This work has been supported by the Collaborative Research Center 'Reduction of Complexity for Multivariate Data Structures' of the German Research Foundation (DFG).

References

David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression diagnostics*. Wiley, 1980.

Olaf Bunke and Bernd T. W. Droge. Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Annals of Statistics*, 12(4):1400–1424, 1984.

Ildiko E. Frank. Comparative monte carlo study of biased regression techniques. Technical Report 105, Department of Statistics, Stanford University, 1989.

Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):199–209, 1993.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Douglas M. Hawkins and Xiangrong Xin. A faster algorithm for ridge regression of reduced rank data. *Computational Statistics & Data Analysis*, 40: 253–262, 2002.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

Karsten Luebke and Claus Weihs. Prediction optimal data analysis by means of stochastic search. In Martin Schader, Wolfgang Gaul, and Maurizio Vichi, editors, *Between Data Science and Applied Data Analysis*, pages 305–312. Springer, 2003.

Gary C. McDonald and Richard C. Schwing. Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15:463–481, 1973.

Heinz Schmidli. *Reduced Rank Regression*. Physica Verlag, 1995.

Jun Shao. Linear model selection by cross-validation. *Journal of the american statistical assosiation*, 88(422):486–494, 1993.

James H. Stapleton. *Linear Statistical Models*. Wiley, 1995.