JOURNAL OF CHEMOMETRICS
J. Chemometrics 2001; **15**: 497–509
DOI: 10.1002/cem.635

Calibration with absolute shrinkage

Henrik Öjelund*, Henrik Madsen and Poul Thyregod

Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

SUMMARY

In this paper, penalized regression using the L_1 norm on the estimated parameters is proposed for chemometric calibration. The algorithm is of the lasso type, introduced by Tibshirani in 1996 as a linear regression method with bound on the absolute length of the parameters, but a modification is suggested to cope with the singular design matrix most often seen in chemometric calibration. Furthermore, the proposed algorithm may be generalized to all convex norms like $\sum|\beta_j|^\gamma$ where $\gamma \geq 1$, i.e. a method that continuously varies from ridge regression to the lasso. The lasso is applied both directly as a calibration method and as a method to select important variables/wavelengths. It is demonstrated that the lasso algorithm, in general, leads to parameter estimates of which some are zero while others are quite large (compared to e.g. the traditional PLS or RR estimates). By using several benchmark data sets, it is shown that both the direct lasso method and the regression where the lasso acts as a wavelength selection method most often outperform the PLS and RR methods. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: lasso; variable selection; wavelength selection; NIR spectroscopy; linear regression

1. INTRODUCTION

In a chemometric calibration the number of parameters to be estimated normally exceeds the number of observations, since the number of available spectra is normally less than the number of measured wavelengths. In this situation the normal equations, corresponding to the OLS problem, have infinitely many solutions. Traditionally, this problem is handled by using regularization methods such as principal component regression (PCR), partial least squares (PLS), ridge regression (RR) or variable selection (VS) (see Reference [1] for an introduction). PLS and PCR decompose the explanatory data into a few latent factors which are then used as explanatory variables. RR obtains a unique solution by limiting the squared length of the estimated parameters. It is shown in Reference [2] that PCR, PLS and RR all have approximately the same performance and produce similar estimates. A characteristic property of these methods is that none of them will produce an estimate where any of the parameters are equal to zero. Furthermore, these three methods will lead to the same model, minimum length least squares (MLLS), in the extreme situation where all the latent factors are used (PLS and PCR) or where the squared length of the RR estimate is unlimited [1]. The last method, VS, finds a solution by selecting a set of explanatory variables of a specified size that minimizes the

* Correspondence to: Henrik Öjelund, Department of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.
Contract/grant sponsor: Danish Academy of Technical Sciences

Received 15 December 1998
Accepted 7 April 2000

standard least squares criterion. The problem of selecting the optimal set of explanatory variables is a 2^p problem, where p is the number of measured wavelengths. In many chemical calibration problems the number of wavelengths may be large (say thousands), which makes the optimal subset selection a very computationally demanding problem. This has spurred the development of a number of more or less heuristic selection methods such as forward selection or backward elimination. In recent years, optimization algorithms such as simulated annealing and genetic algorithms, and lately Bayesian variable selection with MCMC, have been applied to the selection of individual wavelengths [3–6]. These methods perform a stochastic search which tolerates temporary decreases in quality during an optimization. Search paths are neither predictable nor reproducible and there is no guarantee that the final set of wavelengths is the optimal one.

VS and RR can both be seen as regression methods which limit the length of the parameters [2]. The difference between the methods is in the metric used. For RR the length is bounded by the L_2 norm and for optimal VS the L_q norm where $q \to 0+$. In 1996, Tibshirani [7] introduced a new method, called *lasso*, for 'least absolute shrinkage and selection operator', which bounds the absolute length, or the L_1 norm, of the parameters. Thus the method is conceptually placed between RR and optimal VS. In 1995, however, Williams [8] applied the absolute penalty to neural networks. The absolute penalty was interpreted as a Laplace prior, distribution on the parameters.

By limiting the absolute length of the parameters, some of the parameters will become zero while others may become quite large. Osborne *et al.* [9] treated the lasso as a convex programming problem, which led to a very efficient algorithm to calculate the lasso estimate. Fu [10] describes a class of regression methods where the length of the parameters is described by convex norms. Furthermore, a new algorithm, called the shooting algorithm, was introduced to calculate the lasso estimate.

The rest of this paper is organized as follows. Section 2 starts by defining the lasso, and a geometrical interpretation of the method is given. The section is concluded with a discussion of the problem of estimating the standard errors of the parameters. In Section 3 it is shown that the lasso can be used as a variable selection method. The lasso has a hyperparameter λ which controls the absolute length of the parameters. How to select the value of λ using cross-validation is discussed in Section 4. In Section 5 a résumé is given of algorithms to calculate the lasso estimate. The lasso, used both directly for estimating the parameters and for obtaining a good subset of wavelengths, is thoroughly tested in Section 6 and compared with PLS and RR. The data sets used consist of both NIR spectra of wheat, gasoline and beer and UV-vis spectra of reaction products of ammonia and HOCl. Finally, Section 7 provides a short summary and suggests some directions for future research.

2. THE LASSO

2.1. Definition

Suppose that we have data $(\mathbf{x}^i, y_i)$, $i = 1, 2, \ldots, N$, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ are the explanatory variables and y_i are the response variables. The $N \times p$ design matrix is denoted by $\mathbf{X}$, where $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^N)^{\mathrm{T}}$. To give all the explanatory variables the same weight, the explanatory variables are standardized. The following description of the lasso has been adopted from Reference [7]. The lasso estimates $\hat{\alpha}$ and $\hat{\beta}$ are obtained as

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min_{(\alpha, \boldsymbol{\beta})} \left(\frac{1}{N} \sum_{i=1}^{N} (y_i - \alpha - \mathbf{x}^{i\mathrm{T}} \boldsymbol{\beta})^2 \right) \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t \tag{1}$$

where $t \geq 0$ is a hyperparameter. For all values of t the function is minimized with $\hat{\alpha} = \bar{y}$. The constrained minimization problem (1) may be transformed to an equivalent unconstrained problem

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min_{(\alpha,\boldsymbol{\beta})} \left(\frac{1}{N} \sum_{i=1}^{N} (y_i - \alpha - \mathbf{x}^{i^{\mathrm{T}}} \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{2}$$

where the Lagrange parameter λ is chosen in such a way that $\sum_{j=1}^{p} |\beta_j| \leq t$. Both of these equivalent problem formulations may be used to obtain an insight into the properties of the lasso.

2.2. *Geometric interpretation*

One of the attractive characteristics of the lasso is that it often produces estimates that are exactly zero. To obtain insight into why this happens, the contour of the function

$$f(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{x}^{i^{\mathrm{T}}} \boldsymbol{\beta})^2$$

is shown in Figure 1 as a simple example where the number of parameters p is two and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is singular (the response variable y is centered). Therefore infinitely many solutions to the normal equations corresponding to the OLS problem exist, and instead of an ellipsoid, the contour plot of f degenerates to parallel lines with the solutions to the normal equation as the centerline. In Figure 1 the solutions to the normal equation are marked with a dotted line. The constraint region for the lasso is a rotated square and for RR it is a circle, as illustrated in Figure 1. The solution to the optimization problem (1) is represented by the point where the contour first touches the constraint. For RR the probability that the contour will hit the constraint where any of the parameters are zero is very small. However, for the lasso there is a high probability that the contour and the constraint will intersect in one of the corners, i.e. some of the parameters are zero. When t increases, the lasso estimate will eventually be identical to one of the solutions of the normal equation, as seen in the example in Figure 1. It is also clear that for the singular problem the lasso solution of the normal equation will not, in general, be identical to the RR, PLS or PCR solution, namely MLLS. From studying Figure 1, it is seen that the lasso will not have a unique solution when some of the explanatory variables are identical, and the contour thus has an angle of 45 °.

2.3. *Standard errors of parameter estimates*

In Reference [7] it is shown that the lasso may be interpreted as a Bayesian method. However, it is not possible to express the posterior distribution of the parameters in a closed form, i.e. there does not exist an analytical expression of the normalization factor of the posterior distribution. If this had been the case, the most obvious way to calculate the covariance matrix of the lasso estimates would have been to calculate the second-order moment about the maximum posterior. In Reference [7], and later in Reference [9], approximations for estimating the covariance matrix in non-singular situations have been developed; but for singular design matrices, sampling methods such as bootstrapping or MCMC still have to be utilized. However, owing to the simulation time, sampling-based methods become infeasible when the number of explanatory variables is high, which is often the case in chemometrics.

3. VARIABLE SELECTION USING THE LASSO

If the Lagrange term $\lambda \sum_j |\beta_j|$ in (2) is replaced by $\lambda \sum_j |\beta_j|^{\gamma}$ where $\gamma \to 0+$, the minimization will be identical to optimal variable selection [2]. Unfortunately, this minimization is difficult to perform, because the function has up to $2^p - 1$ local minima. The smallest value of γ for which the Lagrange term $\sum_j |\beta_j|^{\gamma}$ is convex, i.e. the function has only one minimum, is one. Therefore an interesting

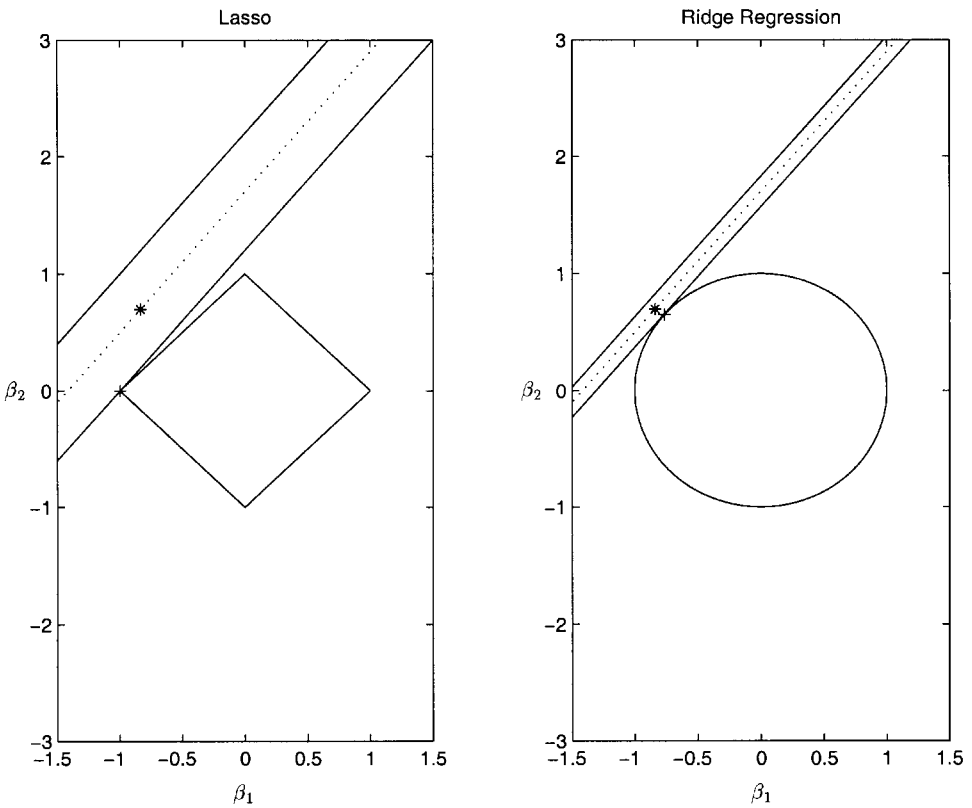

Figure 1. Illustration of how the solution of the minimization depends on the constraint. There are two parameters (β_1, β_2) to be estimated and the matrix $\mathbf{X}^T\mathbf{X}$ is singular. The MLLS solution is marked with an asterisk (*) and all the solutions to the normal equation are marked with a dotted line (…). The lines parallel to the solutions represent the contour of the linear regression function $f(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. The constraint $|\beta_1|^\gamma + |\beta_2|^\gamma = 1$ is drawn for $\gamma = \{1{\cdot}0, 2{\cdot}0\}$, i.e. the constraints for lasso and RR. The solution for each constraint is marked with a cross (+).

extension of lasso would be to use it as a variable selection method. Variable selection using lasso is a two-step procedure. In the first step, lasso is applied to identify which variables are important (non-zero parameter estimates), and in the second step the selected variables are used in ordinary linear regression. Expressed in detail, let the matrix $\tilde{\mathbf{X}}$ denote the matrix containing only those columns j of $\mathbf{X}$ for which the lasso estimate $\hat{\beta}_j \neq 0$. Then the variable selection estimate is defined as

$$\hat{\boldsymbol{\beta}}_{\text{vs}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{y} \tag{3}$$

The difference between lasso and lasso as a variable selection method is that in variable selection only the parameters that turn out to be zero are penalized (set to zero) and the others are estimated without any penalization. For the same value of λ the number of non-zero parameters will be the same for the two methods, but in practice, λ is selected as a larger value when lasso is used as a variable selection method. Hence models estimated with lasso used as a variable selection method will, in practice, have fewer non-zero parameters.

4. SELECTION OF HYPERPARAMETER

There are several methods to estimate the value of λ, such as bootstrapping or cross-validation. One common approach in chemometrics is to use full cross-validation [1] (N-fold cross-validation). This approach has been criticized as giving too large a model [11], and in Reference [12] it has been recommended to use fivefold or 10-fold cross-validation instead. Another advantage of fivefold or 10-fold cross-validation compared to full cross-validation is that these validation methods are less computationally demanding. For fivefold cross-validation the observations are first divided into five groups which are as equal as possible. Denoting these groups by L_1, …, L_5 and using an obvious notation, we define

$$L^{(v)} = L - L_v, \quad v = 1, \ldots, 5$$

where L is the entire data set. Now use the data $L^{(v)}$ to estimate the parameters and L_v to validate. Repeating this for $v = 1$, …, 5, the mean squared error of prediction becomes

$$\text{MSEP} = \frac{1}{N}\sum_{v=1}^{5} \sum_{(y_i,\mathbf{x}^i)\in L_v} (y_i - \mathbf{x}^i \hat{\boldsymbol{\beta}}^v)^2$$

where $\hat{\boldsymbol{\beta}}^v$ is the estimate found using the data $L^{(v)}$. It should be noted that the number of observations N in (2) and (4) is not fixed during the cross-validation but is equal to the number of observations in $L^{(v)}$. In the examples in Section 6 the fivefold cross-validation method has been used to estimate λ.

5. ALGORITHMS FOR THE LASSO

There are two different approaches to calculate the lasso estimates. The first approach directly solves problem (1) using quadratic programming, and the second reformulates the constrained minimization problem (1) to an unconstrained minimization problem (2).

The first algorithm, recommended by Tibshirani [7] and based on quadratic programming, starts by fixing $t \geq 0$ and then interprets problem (1) as a least squares problem with 2^p inequality constraints, corresponding to the 2^p different signs for the parameters. He found that by iteratively adding the necessary constraints until $\sum_j |\beta_j| \leq t$, the algorithm is quite efficient and only about $0{\cdot}5p$–$0{\cdot}75p$ iterations are needed. The drawback of this approach is that it only works when $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ has full rank. A more recent algorithm, called the shooting algorithm and introduced by Fu [10], is a more general algorithm and applicable for all convex penalties $\sum_j |\beta_j|^\gamma \leq t$ where $\gamma \geq 1$. However, the algorithm does not work when the design matrix is singular.

Very recently, a new algorithm to estimate the lasso by quadratic programming was suggested by Osborne *et al.* [9]. Compared to the algorithm suggested by Tibshirani, the new algorithm starts with the zero vector and adds variables iteratively instead of starting with the full OLS estimate and removing variables. Hence the algorithm works also in the singular case. In the present work this algorithm has only been used to validate the results. S-PLUS code to carry out the algorithms of Osborne *et al.* and Tibshirani may be found at lib.stat.cmu.edu/S.

In the Appendix a new algorithm for calculating the lasso estimate that minimizes the unconstrained problem (2) is suggested. The algorithm is usable even in cases where the number of explanatory variables exceeds the number of observations. The developed algorithm basically works by minimizing the function using standard Newton–Raphson. Our main idea in the algorithm is an iterative quadratic approximation of the Lagrange term $\lambda \sum_j |\beta_j|$. Furthermore, it is possible to generalize the algorithm to all convex penalties, i.e. $\lambda \sum_j |\beta_j|^\gamma$ where $\gamma \geq 1$, although this

generalization has not been explored further in this work. A MATLAB implementation of the algorithm may be downloaded from www.imm.dtu.dk/~hoe. However, for $\gamma = 1$ the algorithm suggested by Osborne *et al.* yields exactly the same result but is found to be considerably faster.

6. EXAMPLES AND COMPARISONS

In this section the lasso is evaluated and compared to RR and PLS, both as a stand-alone and as a variable selection method. A common way to evaluate new methods is to use simulated data. This type of evaluation has been criticized to be unrepresentative, and instead the use of real-life data has been recommended [13]. On the other hand, there is a risk of developing a method that is optimized only for a specific set of data, so it is important that several different data sets are used in an evaluation. Therefore the considered methods are applied to five sets of data that originate from different applications. In the examples below the PLS algorithm is defined as in Reference [14], and RR as

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min_{(\alpha,\boldsymbol{\beta})} \left[\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_{\mathrm{RR}} \sum_{j=1}^{p} \beta_j^2 \right] \tag{4}$$

All the evaluated methods contain some kind of hyperparameter which controls how much the estimates are shrunk. For PLS the hyperparameter is the number of latent factors, for RR the hyperparameter λ_{RR} controls the quadratic length of the estimates, and for the lasso the hyperparameter λ controls the absolute length of the estimates. The estimation of these hyperparameters is done using fivefold cross-validation, and the hyperparameter with the smallest MSEP is selected as described in Section 4. It should be noted that when the lasso is used as a variable selection method, the number of selected variables may vary for each fold. However, this does not effect the selection of λ. In all the examples, both the response variable and the explanatory variables have been standardized. Hence MSEP may be interpreted as the fraction of unexplained variance.

6.1. *UV-VIS spectra of ammonia reaction products with HOCl*

The data used in the first example originate from UV-VIS measurements of ammonia reaction products with HOCl. The ammonia molecule itself does not absorb in the measured spectral range and therefore the reagents NaOH and HOCl have to be added. The number of spectra is $N = 30$ and the number of measured wavelengths is $p = 250$ (the spectra are measured in 1 nm intervals from 190 to 439 nm). The response variable is the concentration of ammonia. Some regions of the spectra are heavily influenced by stray light, which makes the absorbance for these regions non-linear with respect to the concentration [15]. The results are shown in Table I. In this example, all the methods work very well. The limited number of observations makes the calculation of MSEP rather sensitive to how the data are divided for the cross-validation. It was found that with another grouping of the data the lasso VS would perform slightly better.

6.2. *NIR spectra of wheat*

In this example, NIR spectra are used to predict the amounts of moisture and protein in wheat. The number of wheat samples is $N = 100$ and the number of measured wavelengths is $p = 701$ (the spectra are measured in 2 nm intervals from 1100 to 2500 nm). A description of the data can be found in Reference [16] and the data can be downloaded from ftp.clarkson.edu/pub/hopkepk/Chemdata/Kalivas/. The results in Table II show that the lasso and VS have similar performance compared to

Table I. Predicting the concentration of ammonia from UV-VIS spectra

Method	Hyperparameter	MSEP
Lasso	$\lambda = 0{\cdot}000750$, 21 parameters $\neq 0$	0·0016
VS with lasso	$\lambda = 0{\cdot}0024$, 13 parameters $\neq 0$	0·0025
PLS	14 components	0·0021
RR	$\lambda_{RR} = 0{\cdot}0024$	0·0021

Table II. Predicting the amount of moisture from NIR spectra of wheat

Method	Hyperparameter	MSEP
Lasso	$\lambda = 0{\cdot}000750$, 14 parameters $\neq 0$	0·024
VS with lasso	$\lambda = 0{\cdot}001$, 16 parameters $\neq 0$	0·024
PLS	6 components	0·026
RR	$\lambda_{RR} = 0{\cdot}0075$	0·025

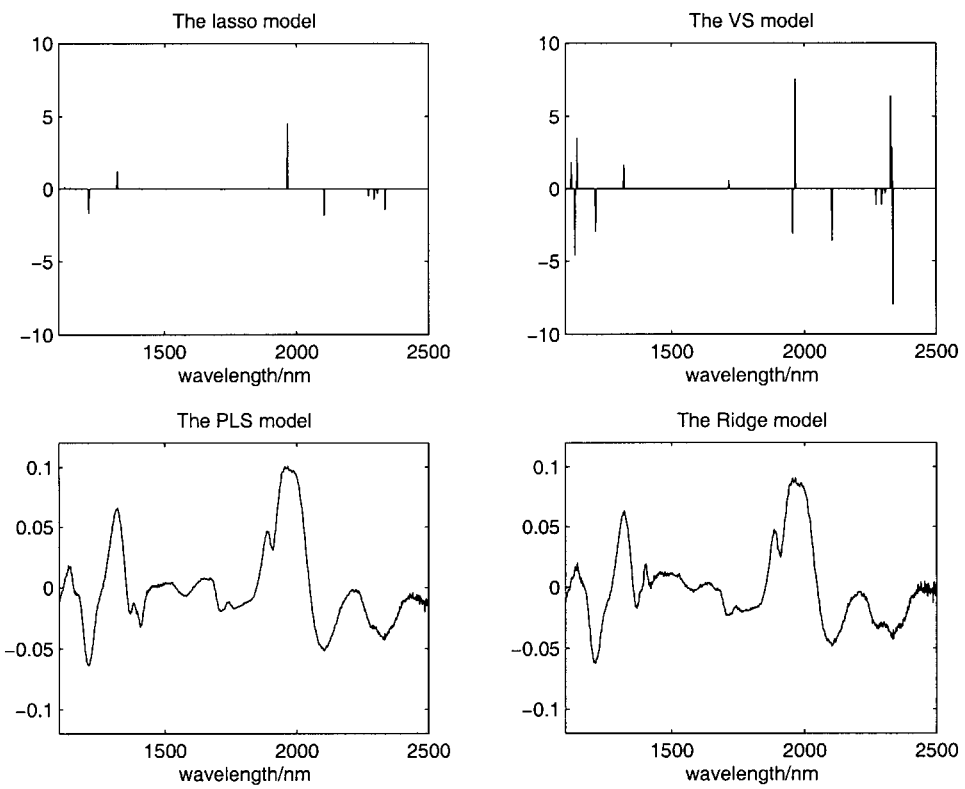

Figure 2. Estimated parameters ($\hat{\boldsymbol{\beta}}$) when the evaluated methods are applied to NIR spectra of wheat for predicting the amount of moisture. The parameters in the VS model consist of large parameters with different signs, indicating that the VS method captures the derivative of the spectrum. Also notice the similarity between the RR and PLS models.

Table III. Predicting the amount of protein from NIR spectra of wheat

Method	Hyperparameter	MSEP
Lasso	$\lambda = 1{\cdot}0 \times 10^{-4}$, 37 parameters $\neq 0$	0·084
VS with lasso	$\lambda = 3{\cdot}16 \times 10^{-4}$, 26 parameters $\neq 0$	0·085
PLS	19 components	0·098
RR	$\lambda_{RR} = 7{\cdot}0 \times 10^{-6}$	0·092

RR and PLS. An interesting point is that more parameters are set to zero in the lasso model than in the VS model, even if the value of λ for the lasso model is smaller. In Figure 2 the $\hat{\beta}$ estimates for the evaluated methods are shown. Notice the large neighbouring parameters with opposite signs in the VS model. This indicates that the VS model captures the derivative of the spectrum. The results of predicting the amount of protein in wheat are shown in Table III. The best result is again obtained with the lasso and VS methods. The RR and PLS models have similar performance and are almost equal to the MLLS solution.

6.3. *NIR spectra of beer*

The response variable is now the concentration of original extract in beer, which is closely related to the concentration of alcohol. The number of observations is $N = 60$ and the number of wavelengths is $p = 926$ (the spectra are measured in 2 nm intervals from 400 to 2250 nm). The results in Table IV show that the lasso and VS methods significantly outperform PLS and RR. By considering the parameter estimates in Figure 3, it is seen that the enhancement is achieved by neglecting a large part of the spectrum which is found to be non-informative. The lasso and VS models use only a few wavelengths, while the RR and PLS models, which are very similar, use the whole spectrum.

6.4. *NIR spectra of gasoline*

NIR spectra of $N = 60$ gasoline samples are measured in 2 nm intervals from 900 to 1700 nm ($p = 401$). The response variable is the octane number. As for the wheat data, a description of the data can be found in Reference [16] and the data can be downloaded from ftp.clarkson.edu/pub/hopkepk/Chemdata/Kalivas/. In this example the important information about the octane number is probably spread out over the whole spectrum. Hence a method like RR, implicitly assuming the parameters to be normally distributed, works well; see Table V.

6.5. *Discussion*

In all the examples the models obtained by the lasso, either used directly or used in conjunction with OLS as a variable selection method, performed as well as or better than models obtained by RR or

Table IV. Predicting the amount of original extract from NIR spectra of beer

Method	Hyperparameter	MSEP
Lasso	$\lambda = 0{\cdot}032$, 17 parameters $\neq 0$	0·010
VS with lasso	$\lambda = 0{\cdot}1$, 3 parameters $\neq 0$	0·008
PLS	16 components	0·042
RR	$\lambda_{RR} = 1{\cdot}0 \times 10^{-7}$ (MLLS)	0·042

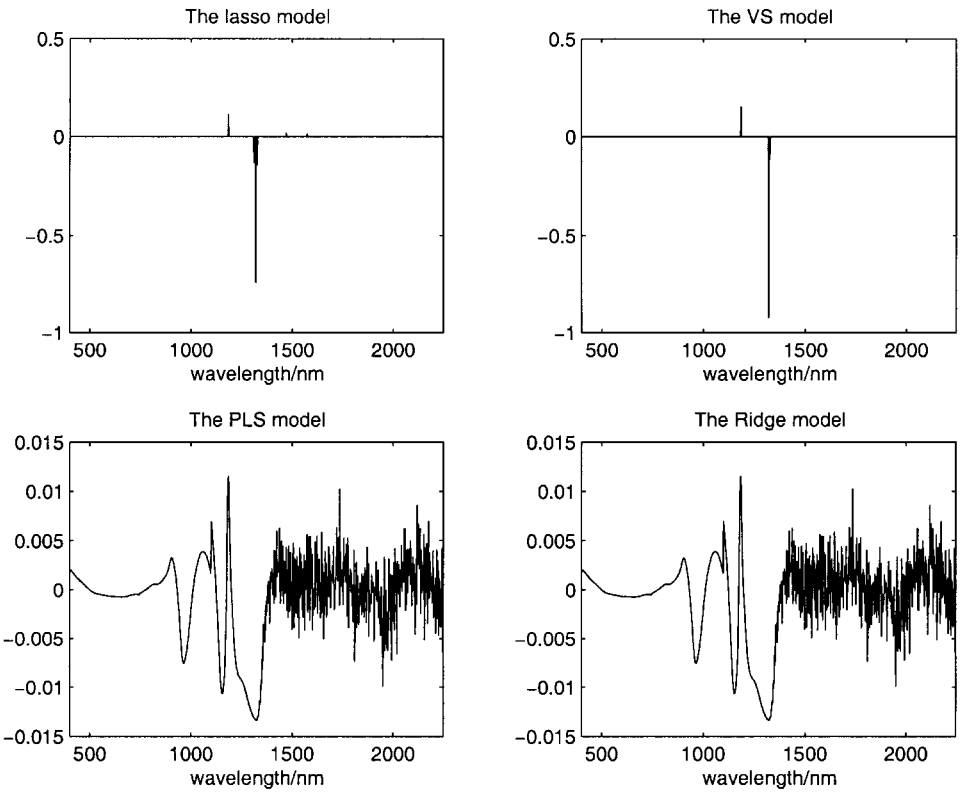

Figure 3. Estimated parameters ($\hat{\boldsymbol{\beta}}$ when the evaluated methods are applied to NIR spectra of beer for predicting the concentration of original extract. The VS and lasso methods put many of the parameters to zero, and by doing so, the variance of the prediction error is reduced significantly compared to the RR and PLS methods.

PLS. In one example the difference is substantial, namely the beer example. This is probably due to the fact that when the spectrum has non-informative variation in certain regions, the prediction can be improved by neglecting these regions in the model. Using a calibration method like the lasso that selects discrete wavelengths may help the researcher to make a meaningful physical interpretation of the model. For instance, in the beer example the lasso is able to locate two narrow regions of the spectrum that explain almost all the variation in the response variable.

Another interesting observation, which also corroborates the result in Reference [2], is that RR and PLS have very similar performances, and the estimated parameters are almost identical for these two methods. Some of the PLS models are rather large, but the number of parameters is chosen by fivefold

Table V. Predicting the octane number of gasoline from NIR spectra

Method	Hyperparameter	MSEP
Lasso	$\lambda = 0{\cdot}018$, 12 parameters $\neq 0$	0·020
VS with lasso	$\lambda = 0{\cdot}0751$, 9 parameters $\neq 0$	0·018
PLS	6 components	0·019
RR	$\lambda_{RR} = 0{\cdot}32$	0·018

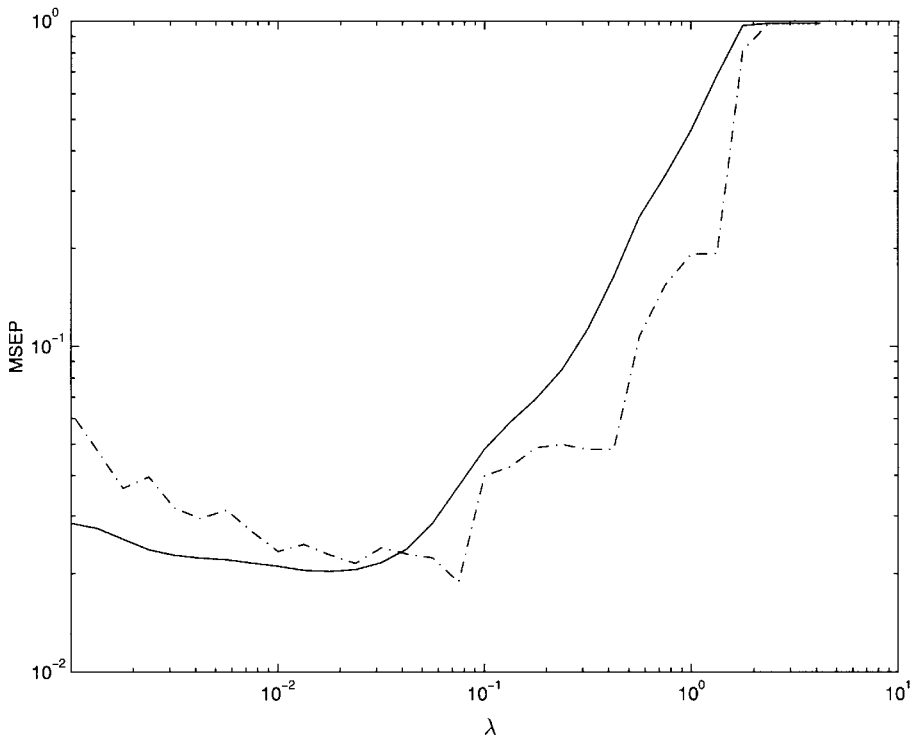

Figure 4. Estimated mean squared error of prediction as a function of λ for gasoline data (using cross-validation). The full curve shows MSEP for models estimated with the lasso, and the chain curve for the VS models. The explanatory variables included in the VS models are those with corresponding non-zero parameters in the lasso models.

cross-validation. Once the number of parameters has been fixed, the complete data set is used for the subsequent calibration, and both fewer and more parameters would result in an inferior PLS model. In Figure 4, MSEP as a function of λ is shown for the gasoline data. MSEP may be regarded as being the sum of a squared bias and a variance of the prediction. For models estimated with a small λ, the estimated parameters have small bias but high variance. For increasing λ, the bias increases and the variance decreases. The bias introduced by the lasso is caused both by removing parameters and by limiting the length of the remaining parameters. For the VS model, bias is only due to removed parameters. Hence for equal λ the VS model has less bias but higher variance than a model estimated with the lasso. In all the examples the selected values of λ for the VS models are consistently larger than those for the lasso models. One attractive property concerning the direct use of lasso is that the MSEP function of λ is a smooth function. This makes the estimation less sensitive to the selection of λ and hence more robust.

7. CONCLUSION

This paper demonstrates how the lasso method, which is a linear regression with L_1 bounds on the estimated parameters, can be used to solve chemometric problems.

It is shown that the lasso can be used both directly for calibration and as a variable selection method, i.e. as a method for finding a set of explanatory variables as a subset of the original set of explanatory variables. This subset can then be used in ordinary least squares regression.

By considering a number of benchmark cases, it is demonstrated that both ways of applying the lasso method create models that often are better than those found by the traditional methods PLS and RR. PLS and RR, on the other hand, are found to estimate the parameters almost identically. The minor extra computation needed to calculate the lasso VS model once a lasso model has been estimated makes the combination of methods attractive for calibration. Furthermore, it is demonstrated that the lasso leads to a calibration model where many of the parameters become zero, and in some situations this may help the researcher to make a meaningful physical interpretation of the model.

Tibshirani [7] discussed a taxonomy for what type of statistical method is adequate for a specific set of data. If a large number of explanatory variables have a small effect, RR is the best method, and if a small number of variables have a large effect, subset selection does best. However, if a moderate number of variables have a moderate effect, lasso, which conceptually is placed between subset selection and RR, is the preferred choice of method. The examples in Section 6 indicate that chemometric data often have this property.

A drawback of the lasso compared to PLS and RR is that it is more computationally demanding, but with the efficient algorithm developed by Osborne *et al.* [9], it is possible to perform the calibration on a standard PC.

The suggested algorithm to calculate the lasso may be generalized to all convex penalty terms like $\sum |\beta_j|^\gamma$ where $\gamma \geq 1$. In a paper in preparation (H. Öjelund *et al.*) we explore the properties of non-convex penalties where $\gamma < 1$. This kind of non-convex penalty may be used in a wide range of applications, e.g. automatic pruning of neural networks, modelling of time series, and identification of linear and non-linear dynamic models in general. We are also investigating the use of MCMC methods to estimate the uncertainty of the parameter estimates. If this leads to an efficient method to estimate the standard errors, chemometric calibration with the lasso will become even more interesting.

ACKNOWLEDGEMENTS

The authors thank L. Nørgaard at the Royal Veterinary and Agricultural University, Copenhagen, Denmark and Carlsberg A/S, Copenhagen, Denmark for permission to use the beer data set. We are also grateful to the reviewers and Cyril Goutte at IMM, DTU, Lyngby, Denmark for valuable comments. This work has been financially supported by the Danish Academy of Technical Sciences.

APPENDIX. AN ALGORITHM FOR CALCULATING THE LASSO

In this appendix an algorithm to solve problem (2) is suggested. The algorithm is a modified Newton–Raphson method where the Lagrange term $\lambda \sum_j |\beta_j|$ is iteratively approximated by a quadratic function in order to ensure differentiability. During the iterative minimization, some of the parameters will often become very small, and by introducing a threshold, typically 10^{-5} if the explanatory variables have been standardized, it is possible to put these equal to zero.

Denote the non-zero parameters after k iterations by $\tilde{\beta}$ and the corresponding explanatory variables by $\tilde{\mathbf{X}}^k$. Let the Lagrange term be approximated by $\lambda \sum_j (w_j^k + z_{j,j}^k \beta_j^2)$, where w_j^k and $z_{j,j}^k$ are defined as

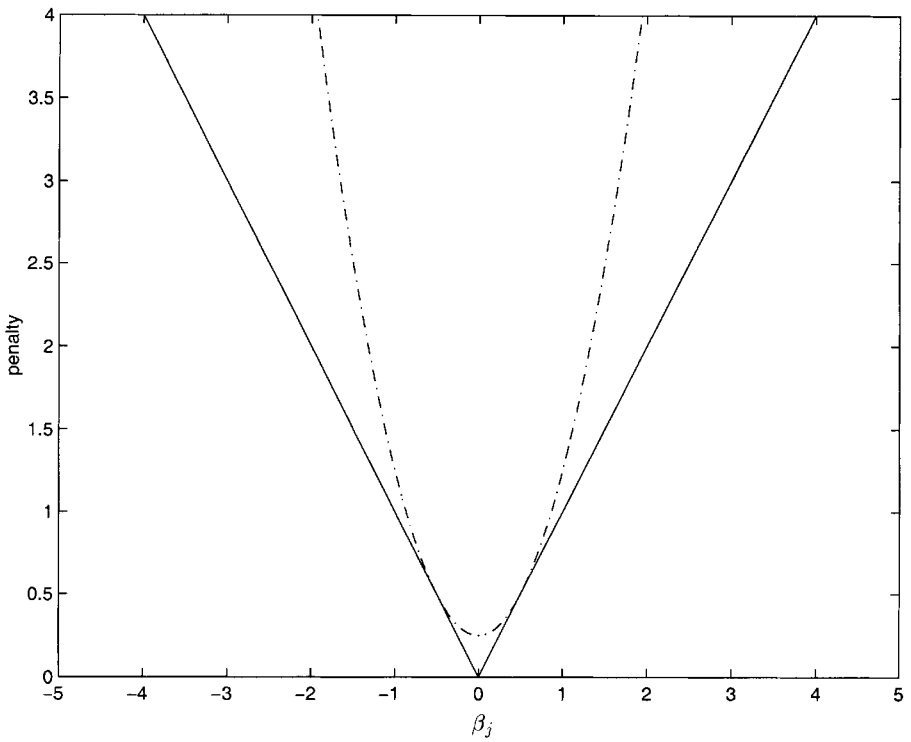

Figure 5. The proposed algorithm for calculating the lasso estimate works by iteratively approximating the Lagrange term with a sum of quadratic functions. In this figure, one quadratic function (chain curve) approximating $|\beta_j|$ is shown for $\tilde{\beta}_j^k = 0 \cdot 5$.

$$w_j^k = |\tilde{\beta}_j^k|/2 \tag{5}$$

$$z_{i,j}^k = \left\{ \begin{array}{ll} 1/2|\tilde{\beta}_j^k| & \text{if } i = j \\ 0 & \text{otherwise} \end{array} \right\}. \tag{6}$$

In Figure 5 the approximation for the jth Lagrange term is shown where $\tilde{\beta}_j^k = 0 \cdot 5$. The minimization may still be quite computationally demanding, and in an attempt to investigate why, let us consider the Newton–Raphson iterate

$$\tilde{\boldsymbol{\beta}}^{k+1} = \tilde{\boldsymbol{\beta}}^k - (\mathbf{G}^k)^{-1}\mathbf{g}^k$$

where $(\mathbf{G}^k)^{-1} = 2[(\tilde{\mathbf{X}}^k)^{\mathrm{t}}\tilde{\mathbf{X}}^k + \lambda \mathbf{Z}^k]$ is the Hessian and $\mathbf{g}^k = -2(\tilde{\mathbf{X}}^k)^{\mathrm{T}}(\mathbf{y} - \tilde{\mathbf{X}}^k\tilde{\boldsymbol{\beta}}^k) + \lambda \mathrm{sign}(\tilde{\boldsymbol{\beta}}^k)$ is the gradient. If the number of parameters is large, the inversion of the Hessian is very computationally demanding. However, if the number of observations N is less than the number of parameters p, it is possible to speed up this inversion by using the *matrix inversion lemma*

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1}\mathbf{DA}^{-1}$$

with $(\mathbf{G}^k)^{-1} = (\mathbf{A} + \mathbf{B}\,\mathbf{C}\,\mathbf{D})^{-1}$, $\mathbf{A} = 2\mathbf{Z}^k$, $\mathbf{B} = 2(\tilde{\mathbf{X}}^k)^{\mathrm{T}}$, $\mathbf{D} = \tilde{\mathbf{X}}^k$ and $\mathbf{C}$ the identity matrix of size N. With this reformulation it is possible to reduce the size of the matrix inverse to $N \times N$ instead of $p \times p$.

The selection of the starting point is important to get stable convergence. We have found that the MLLS estimate is a good choice of starting point, and MLLS has been used as such in all the examples. The MLLS estimate is easily obtained from

$$\tilde{\boldsymbol{\beta}}^1 = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

where $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}$ is the Moore–Penrose inverse. The developed algorithm is quite general and works for all convex Lagrange terms, i.e. $\lambda \sum_j |\beta_j|^\gamma$ where $\gamma \geq 1$, i.e. a method that continuously varies from RR to lasso. In a typical chemometric calibration the algorithm converges within thousands to hundreds of thousands of iterations. The time the algorithm needs to converge depends on the data and the value of λ.

REFERENCES

1. Brown PJ. *Measurement, Regression, and Calibration*. Oxford Science Publications: Oxford, 1993.
2. Frank IE, Friedman JH. *Technometrics* 1993; **35**: 109–135.
3. Leardi R, Boggia R, Terrile M. *J. Chemometrics* 1992; **6**: 267–281.
4. Jouan-Rimbaud D, Massart DL, Leardi R, De Noord OE. *Anal. Chem.* 1995; **67**: 4295–4301.
5. Hörchner U, Kalivas JH. *J. Chemometrics* 1995; **9**: 283–308.
6. Brown PJ, Vannucci M, Fearn T. *J. Chemometrics* 1998; **12**: 173–182.
7. Tibshirani R. *R. Statist. Soc. B* 1996; **58**: 267–288.
8. Williams PM. *Neural Comput.* 1995; **7**: 117–143.
9. Osborne MR, Presnell B, Turlach BA. *J. Comput. Graph. Statist.* in press.
10. Fu W. *J. Comput. Graph. Statist.* 1998; **7**: 397–416.
11. Shao J. *J. Am. Statist. Assoc.* 1993; **88**: 486–494.
12. Breiman L, Spector P. *Int. Statist. Rev.* 1992; **60**: 291–319.
13. Wold S. *Chemolab* 1995; **30**: 109–115.
14. Wold S, Martens H, Wold H. In *Matrix Pencils*, Ruhe A, Kågström B (eds). Springer: Heidelberg, 1983; 286–293.
15. Owen T. Hewlett Packard Publ. 12-5965-5123E, 1996.
16. Kalivas JH. *Chemolab* 1997; **37**: 255–259.

