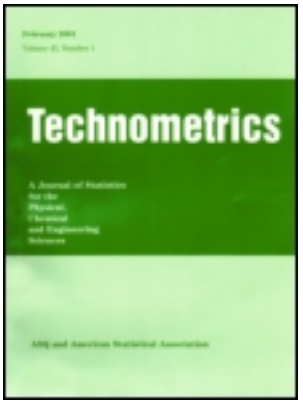

Technometrics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

Prediction Based on Mean Subset

H Ojelund[a], P J Brown[b], H Madsen[c] & P Thyregod[d]

[a] Informatics and Mathematical Modelling, The Technical University of Denmark, DK-2800 Lyngby, Denmark

[b] Institute of Mathematics and Statistics, University of Kent at Canterbury, Canterbury, Kent, CT2 7NF, UK

[c] Informatics and Mathematical Modelling, The Technical University of Denmark, DK-2800 Lyngby, Denmark

[d] Informatics and Mathematical Modelling, The Technical University of Denmark, DK-2800 Lyngby, Denmark

Published online: 01 Jan 2012.

To cite this article: H Ojelund, P J Brown, H Madsen & P Thyregod (2002) Prediction Based on Mean Subset, Technometrics, 44:4, 369-378, DOI: 10.1198/004017002188618563

To link to this article: http://dx.doi.org/10.1198/004017002188618563

Prediction Based on Mean Subset

H. ÖJELUND

Informatics and Mathematical Modelling
The Technical University of Denmark
DK-2800 Lyngby, Denmark
(hoe@imm.dtu.dk)

P. J. BROWN

Institute of Mathematics and Statistics
University of Kent at Canterbury
Canterbury, Kent, CT2 7NF, U.K.
(philip.J.Brown@ukc.ac.uk)

H. MADSEN

Informatics and Mathematical Modelling
The Technical University of Denmark
DK-2800 Lyngby, Denmark
(hm@imm.dtu.dk)

P. THYREGOD

Informatics and Mathematical Modelling
The Technical University of Denmark
DK-2800 Lyngby, Denmark
(pt@imm.dtu.dk)

Shrinkage methods have traditionally been applied in prediction problems. In this article we develop a shrinkage method (mean subset) that forms an average of regression coefficients from individual subsets of the explanatory variables. A Bayesian approach is taken to derive an expression of how the coefficient vectors from each subset should be weighted. It is not computationally feasible to calculate the mean subset coefficient vector for larger problems, and thus we suggest an algorithm to find an approximation to the mean subset coefficient vector. In a comprehensive Monte Carlo simulation study, it is found that the proposed mean subset method has superior prediction performance than prediction based on the best subset method, and in some settings also better than the ridge regression and lasso methods. The conclusions drawn from the Monte Carlo study is corroborated in an example in which prediction is made using spectroscopic data.

KEY WORDS: Bayesian variable selection; Best subset; Calibration; Garrote; Lasso; Model averaging; Shrinkage.

1. INTRODUCTION

In this article we focus on the problem of prediction, that is, the problem of finding a function of the predictor variables x_i that is in some sense a good predictor of the response variable y. Given a regression model, there is of course a superficial similarity between this problem of finding a predictor and the familiar problem of estimation, in the sense that for both problems a vector of regression coefficients is estimated. However, as Copas (1983) noted, the loss functions for the two problems are different.

Traditionally, prediction problems have been dealt with using shrinkage methods. In the Bayesian framework, shrinkage is an inherent property resulting from the choice of prior. The Stein estimator (James and Stein 1961) was the shrinkage method for which it was first proven that the mean squared prediction errors will decrease when a bias is introduced. Later, several other shrinkage schemes were proposed, the most well known being ridge regression (Hoerl and Kennard 1970).

The common procedure of selecting a subset of the available predictor variables and to estimate the regression coefficients on the subset by least squares can also be viewed as a shrinkage method. This somewhat extreme form of shrinkage involves the complete pull-back to 0 of a subset of coefficients. Variable selection and other more continuous shrinkage forms were derived and compared by Dempster (1973). That article's motivation was to compare more Bayesian versions of variable selection with continuous-shrinkage forms like ridge regression. In so doing, Dempster recognized that Bayesian versions of selection will typically take averages over different candidate subsets. This has the beneficial effect of both reducing overfitting (Copas 1983) and instability (Breiman 1996). A heuristic definition of an unstable shrinkage method is a method in which a small change in the data can lead to large changes in the sequence $\{\hat{\boldsymbol{\beta}}_\lambda\}$, where λ is a real parameter that indexes the amount of shrinkage Breiman (1996). Hence estimation of the shrinkage factor λ using cross-validation will be difficult when the method is unstable.

Two recently proposed methods, the garrote (Breiman 1995) and the lasso (Tibshirani 1996), try to combine variable selection and shrinkage. The motivation for the development of the garrote was the instability observed when a variable selection method is combined with cross-validation. It has been found (Vach, Sauerbrei, and Schumacher 2001) that the combination of selection and shrinkage makes the garrote and lasso methods more stable than regression with a subset of the variables, and better prediction models are obtained in general.

In fact, ridge regression itself offers a stable form of shrinkage and can be viewed as a weighted average of all least squares fitted subsets (see Leamer and Chamberlain 1976). Bayesian model averaging in regression has become widely advocated (see, e.g., Raftery, Madigan, and Hoeting 1997). In the context of the multivariate general linear model, it has been used for Bayesian variable selection and shown to be

TECHNOMETRICS, NOVEMBER 2002, VOL. 44, NO. 4
DOI 10.1198/004017002188618563

effective with spectroscopic data involving a large number of predictors (see Brown, Vannucci, and Fearn 1998).

In this article we develop a partially Bayes estimator that serves to form a weighted average of all subsets of a particular size. This has been developed independently but is similar in derivation to the REGF method of Dempster (1973). We apply cross-validation for choice of subset and develop fast search algorithms. Using Monte Carlo simulation, we compare the proposed mean subset method to the ridge regression, lasso, garrote, and best subset selection methods. We also apply it to a challenging spectroscopic example where the number of explanatory variables is much larger than the number of observations. In this application, the explanatory variables are discrete points recorded from a continuous curve, which makes them almost collinear. To limit the variance of the parameter estimates in this situation, it is very important to use some sort of shrinkage estimator.

1.1 Introducing Mean Subset as a Shrinkage Method

Let $\boldsymbol{y} = \{y_1, \ldots, y_n\}'$ be the $(n \times 1)$ vector of observed values of the response variable and let $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}'$ be the $(n \times p)$ matrix including all potential explanatory variables where $\mathbf{x}_i = \{x_{i1}, \ldots, x_{ip}\}'$. The linear model has the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon}$ is a noise $(n \times 1)$ vector and $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of coefficients. A subset may be characterized by the p-dimensional *selection vector*, $\boldsymbol{\gamma} = (i_1, i_2, \ldots, i_p)'$, where $i_j \in \{0, 1\}$ and $1 \le j \le p$. Define the cardinal function, $q_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} i_j$. When a selection vector is used as a subscript to a matrix, this shall be understood to mean the matrix with $q_{\boldsymbol{\gamma}}$ columns corresponding to the nonzero elements of the selection vector and in that order. When a selection vector is used as a bracketed superscript for a vector, this shall be understood to mean the p-dimensional vector with 0s in all positions corresponding to the zero elements of the selection vector. The elements of the vector corresponding to the nonzero elements of the selection vector are equal to the elements of the vector with the same selection vector used as a subscript, and in that order. For example, if $p = 5$, $\boldsymbol{\gamma} = (1, 0, 1, 1, 0)'$, and the coefficient vector $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (5, 6, 7)'$, then $\hat{\boldsymbol{\beta}}^{(\gamma)} = (5, 0, 6, 7, 0)'$.

One commonly used approach for parameter shrinkage is best subset selection, in which the number of nonzero coefficients is controlled. The best subset method is described as follows. Let $\mathcal{S}_q$ be a set of $(p \times 1)$ selection vectors defined as

$$\mathcal{S}_q = \{\boldsymbol{\gamma} | \boldsymbol{\gamma} = (i_1, i_2, \ldots, i_p), i_j \in \{0, 1\}, 1 \le j \le p, q_{\boldsymbol{\gamma}} = q\},$$

where q is the number of selected variables. The *best subset* coefficient vector $\hat{\boldsymbol{\beta}}_{\mathrm{BS}}^{(q)}$ of size $q = 1, 2, \ldots, p$, is defined as the estimate $\hat{\boldsymbol{\beta}}^{(\gamma)}$ for which the selection vector $\boldsymbol{\gamma} \in \mathcal{S}_q$ minimizes the residual sum of squares,

$$\mathrm{SS}_{\boldsymbol{\gamma}} = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}_{\boldsymbol{\gamma}}\left(\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{X}_{\boldsymbol{\gamma}}\right)^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{y}, \tag{2}$$

and where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{y}$. It is assumed that all matrices $\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{X}_{\boldsymbol{\gamma}}$ for which $\boldsymbol{\gamma} \in \mathcal{S}_q$ have full rank, so that necessarily $q \le n$.

The mean subset coefficient vector $\hat{\boldsymbol{\beta}}_{\mathrm{MS}}^{(q)}$ is calculated as

$$\hat{\boldsymbol{\beta}}_{\mathrm{MS}}^{(q)} = \sum_{\boldsymbol{\gamma} \in \mathcal{S}_q} w_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}^{(\gamma)}, \tag{3}$$

where the q nonzero coefficient values of $\hat{\boldsymbol{\beta}}^{(\gamma)}$ are given by $(\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{y}$. The elements in the normalized weight sequence $\{w_{\boldsymbol{\gamma}} | \boldsymbol{\gamma} \in \mathcal{S}_q\}$ are

$$w_{\boldsymbol{\gamma}} \propto \mathrm{SS}_{\boldsymbol{\gamma}}^{-n/2}. \tag{4}$$

This weighting is motivated in the subsequent section. In calculating the mean subset coefficient vector no selection is involved, because the estimated coefficient vector $\hat{\boldsymbol{\beta}}_{\mathrm{MS}}^{(q)}$ is a weighted mean of all subsets. Furthermore, $\hat{\boldsymbol{\beta}}_{\mathrm{MS}}^{(q)}$ is an analytical function in $\boldsymbol{X}$ and $\boldsymbol{y}$, in contrast to $\hat{\boldsymbol{\beta}}_{\mathrm{BS}}^{(q)}$, and thus we expect the mean subset to be less sensitive to small variations in data. The mean subset may be seen as a shrinkage method in which q controls the amount of shrinkage.

1.2 Bayesian Motivation

It is possible to motivate (3) and (4) using Bayesian arguments. The following model largely follows the conjugate hierarchical mixture model suggested by George and McCulloch (1997). We start by assuming the standard linear model

$$f(\boldsymbol{y} | \boldsymbol{\beta}, \sigma) = N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}), \tag{5}$$

where σ is a positive scalar with the noninformative prior density proportional to $1/\sigma^2$ and the other variables satisfy the previously stated assumptions. The subsets of variables to be included are specified through the prior $\pi_q(\boldsymbol{\gamma})$. The prior is selected such that all index vectors with q 1s are assigned the same prior probability and all other index vectors are given zero prior probability. In practice, a priori q is unknown, and a cross-validation procedure is used to estimate q.

This particular prior distribution corresponds to the problem of finding a subset of q variables, as in best subset regression. Other prior probability assignments have been suggested by George and McCulloch (1997) and others since them. However, the prior probability assigned to models of different size has been rather arbitrary. The prior distribution suggested here allows one to look directly at pieces of the posterior that are individually quite sensible. Moreover, in the situation when there is an excess of explanatory variables in comparison to observations, it is not possible to work with the default or noninformative prior distribution, because the posterior will be improper when q is large.

Assume the following conditional prior for $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\bar{\gamma}} = 0 | \sigma, \boldsymbol{\gamma}) = N_q(0, \sigma^2 v_1 \boldsymbol{C}_{\boldsymbol{\gamma}}) \tag{6}$$

and

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\bar{\gamma}} \neq 0 | \sigma, \boldsymbol{\gamma}) = 0. \tag{7}$$

Here v_1 denotes an a prior-defined positive scalar and $\boldsymbol{C}_{\boldsymbol{\gamma}}$ denotes a suitably chosen positive definite $(q \times q)$ matrix.

An obvious prior setting for the correlation matrix $\boldsymbol{C}_\gamma$ is to replicate the correlation structure of the least squares estimates, that is,

$$\boldsymbol{C}_\gamma = \left(\boldsymbol{X}_\gamma'\boldsymbol{X}_\gamma\right)^{-1}.$$

This particular prior was used by Dempster (1973) and is sometimes called a g prior (see Zellner 1980). Integrating out both $\boldsymbol{\beta}$ and σ and letting $v_1 \to \infty$ yields

$$\pi(\boldsymbol{\gamma}|\boldsymbol{y}) \propto \mathrm{SS}_\gamma^{-n/2}\pi_q(\boldsymbol{\gamma}), \tag{8}$$

where SS_γ follows from (2). Hence $\pi(\boldsymbol{\gamma}|\boldsymbol{y}) = w_\gamma$ for $\boldsymbol{\gamma} \in \mathcal{S}_q$ and motivates using (4) to calculate the weights.

2. COMPUTATIONAL ASPECTS AND FEASIBLE APPROXIMATIONS

We propose two methods based on a standard least squares exhaustive search to estimate the mean subset. The first method calculates the exact mean subset coefficient vector, and the second method approximates the mean subset coefficient vector using a weighted average of a subcollection of all possible models of size q. An efficient algorithm for finding the best subsets of all model sizes is called "regression by leaps and bounds" (Furnival and Wilson 1974). This algorithm avoids testing all variable combinations; nevertheless, it is able to guarantee that the k best subsets are found, where k is a prespecified positive integer. The algorithm is very efficient when k is small, but the performance degenerates quickly when k is increased. Today, it is possible to calculate the exact mean subset for all model sizes when the number of variables is less than, say, 30. If Moore's law, which states how the computer speed evolves over time, continues to hold, then this upper limit will be increased by one every $1\frac{1}{2}$ years. However, often it is interesting to find models with only a few variables, which makes an exhaustive search feasible for much larger problems. For example, it is possible to perform an exhaustive search for all models with three or fewer variables when the number of explanatory variables is 1,000.

An approximative exhaustive search is described as follows. Assume that we have obtained the k best subsets of size r with an exhaustive search. Each of these subsets is used as a starting point for forward selection, and one more variable is included in each subset. Next, these new subsets of size $r+1$ are used as initial subsets for variable exchange (Miller 1990). Variable exchange works by sequentially exchanging the variables in the subset. For example, assume a problem with 26 variables and denote the variables by the letters of the alphabet. If the initial subset is ABCD, then the following subsets are tested with variable exchange: ABCD, EBCD, FBCD, GBCD, …, ZBCD, AECD, AFCD, AGCD, …, AZCD, …, ABCZ. If the best of all the evaluated subsets is the initial subset, then the algorithm is stopped; otherwise, the procedure is repeated, with the best of the evaluated subsets as the new initial subset. The mean subset is approximated using the subsets evaluated during the forward selection and variable exchange procedures. Obviously, when k is increased, the approximation becomes better. Figure 1 describes the algorithm visually. The procedure is repeated when larger subsets are required.

The number of variables q is estimated using cross-validation. Breiman and Spector (1992) recommended using 5-fold or 10-fold cross-validation. An advantage of 5- or 10-fold cross-validation over leave-one-out cross-validation is that the former methods are less computationally demanding.

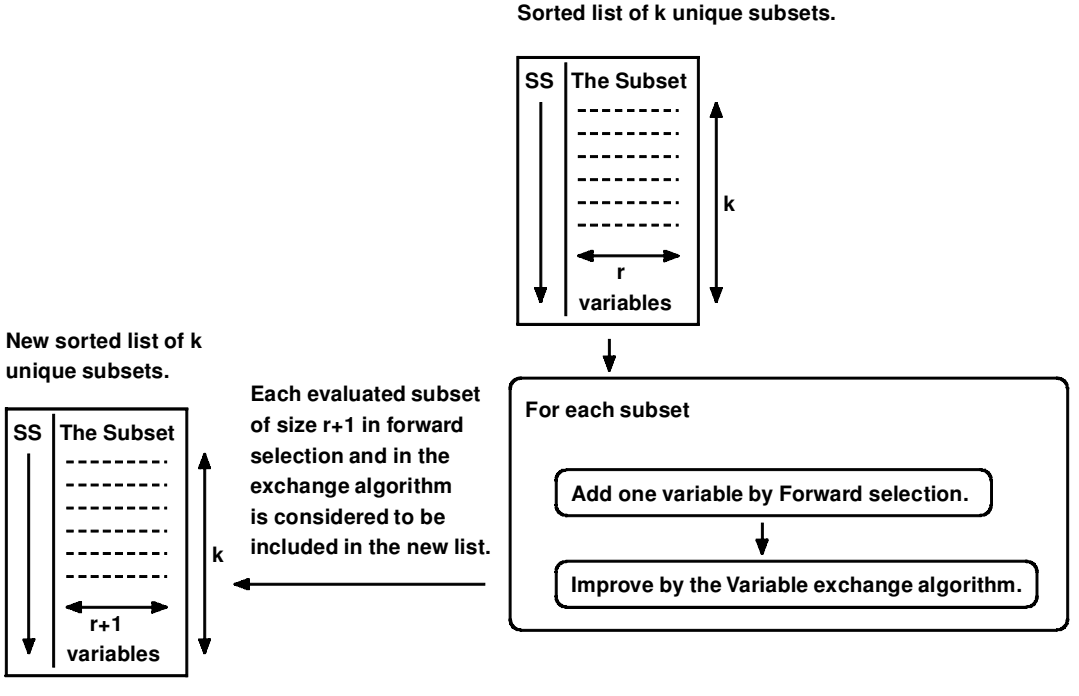

Figure 1. Algorithm for Finding the k Best Subsets of Size r + 1.

For 5-fold cross-validation, the observations are first divided into five equal-sized groups. Denoting these groups by $L_1, \ldots, L_5$ and using an obvious notation, define

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

where L is the entire dataset. Now use the data $L^{(v)}$ to estimate the parameters and L_v to validate. While repeating this for $v = 1, \ldots, 5$, the mean squared error of prediction (MSEP) becomes

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

where $\hat{\boldsymbol{\beta}}^v$ is the estimate found using the data $L^{(v)}$. The number of variables q is estimated by minimizing the MSEP value. In the following section, leave-one-out cross-validation is used to determine the shrinkage factor of ridge regression. Leave-one-out cross-validation is the same as n-fold cross-validation, where n is the number of observations.

3. MONTE CARLO SIMULATION STUDY

In this section prediction with the mean subset is compared to the best subset, ridge regression, garrote, and lasso methods through Monte Carlo simulations. The simulation study is constructed not to show the mean subset in a favorable light, but rather to demonstrate when the different shrinkage methods may be suitable. The competing shrinkage methods are defined as follows. The ridge regression estimate is obtained from

$$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'\boldsymbol{X} + k_R \boldsymbol{I})^{-1} \boldsymbol{X}' \boldsymbol{y}', \tag{9}$$

where k_R is the shrinkage factor.

The garrote starts with the ordinary least squares (OLS) estimates and shrinks them by nonnegative factors whose sum is constrained. For a given shrinkage factor $t \geq 0$, the garrote minimizes

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} c_j \beta_j^{\text{OLS}} x_{ij} \right)^2 \quad \text{subject to } \sum_{j=1}^{p} c_j \leq t, \quad c_j > 0. \tag{10}$$

The lasso estimate, $\hat{\boldsymbol{\beta}}_{\text{L}}$, is defined by

$$\hat{\boldsymbol{\beta}}_{\text{L}} = \arg\min_{(\boldsymbol{\beta})} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2 \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \tag{11}$$

where $t \geq 0$ is the shrinkage factor. In all of these shrinkage methods, the shrinkage factors are estimated by cross-validation.

The setup of the example is largely adopted from Breiman (1996), and the design makes it possible to investigate the influence of cross-validation to the prediction performance. The explanatory variables are sampled from a 0-mean, 20-variable multivariate normal with covariance matrix $\Omega_{ij} = \rho^{|i-j|}$. Three different values of ρ are tested—0, .45, and .9—which spans uncorrelated to highly correlated variables. For each correlation structure, five different coefficient vectors are used to generate the dependent data. The nonzero coefficients are in two clusters of adjacent variables with clusters centered at variables 5 and 15. The initial coefficient values for variables clustered around the variable 5 are given by

$$\beta_{5+j} = (h - |j|)^2, \quad |j| \leq h,$$

where h is a fixed integer controlling the cluster width. The cluster at variable 15 is generated in the same way. The influence of noise is studied by testing three levels of signal-to-noise (SN) ratio, SN $\in \{1, 5, 9\}$. To obtain the desired SN ratio, the coefficients are scaled so that $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}/n = \text{SN}$. The vector of dependent variables $\boldsymbol{y}$ is calculated from $\boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the number of observations is $n = 40$ and $\boldsymbol{\epsilon}$ is sampled from $\mathcal{N}_{40}(\boldsymbol{0}, \boldsymbol{I})$. In this study, the explanatory variables and response variable were centered before estimation.

The performance of the shrinkage methods is measured by calculating the mean model error (ME), defined as

$$\text{ME} = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \boldsymbol{\Omega} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + (\bar{y} - \bar{\boldsymbol{x}}' \hat{\boldsymbol{\beta}})^2, \tag{12}$$

where $\bar{y}$ and $\bar{\mathbf{x}}$ are the sample mean in each simulated dataset of the response variable and the explanatory variables.

Figures 2–4, show the average MEs for the methods. In the left-side graphs in each figure, the shrinkage factors are estimated using cross-validation. The Monte Carlo simulation has been repeated 2,000 times for each combination of h and ρ, and the estimated standard errors of the points in the graphs are less than .02. Fivefold cross-validation was used for the mean subset, lasso, garrote, and best subset, and leave-one-out cross-validation was used for ridge regression. This method of estimating the shrinkage factors was suggested by Breiman (1996). Whereas the leave-one-out estimate has lower bias, it is degraded by its higher variance. Hence, leave-one-out cross-validation may be used for stable methods like ridge regression, whereas 5-fold or 10-fold cross-validation with higher bias is suggested for unstable methods like best subset selection. In the right-side graphs in figures 2–4, the true data-generating model is assumed known (referred to as the *crystal ball*), and the value of the shrinkage factor is selected such that ME in (12) is minimized.

The graphs show that mean subset and ridge regression are complementary to one another. In cases with only a few nonzero coefficients, mean subset and best subset give good prediction, but in cases with many nonzero coefficients, ridge regression works best. When the shrinkage factor is estimated using cross-validation, mean subset consistently gives better prediction than best subset. The difference increases with an increasing number of nonzero coefficients and decreasing SN ratio. The graphs also reveal that the main reason for this difference in prediction performance is the instability of the best subset method. This is demonstrated by the degradation in prediction performance when q is estimated using cross-validation instead of the crystal ball. The graphs also show that the lasso is preferred over the garrote and that the difference between these two methods increases with increasing collinearity between the explanatory variables. Furthermore, the lasso is better or as good as mean subset, except when the underlying model is small and the SN ratio high.

4. PREDICTION OF MOISTURE AND PROTEIN CONTENT IN WHEAT

In this real data example, the objective is to predict the amount of moisture and protein in wheat using near-infrared

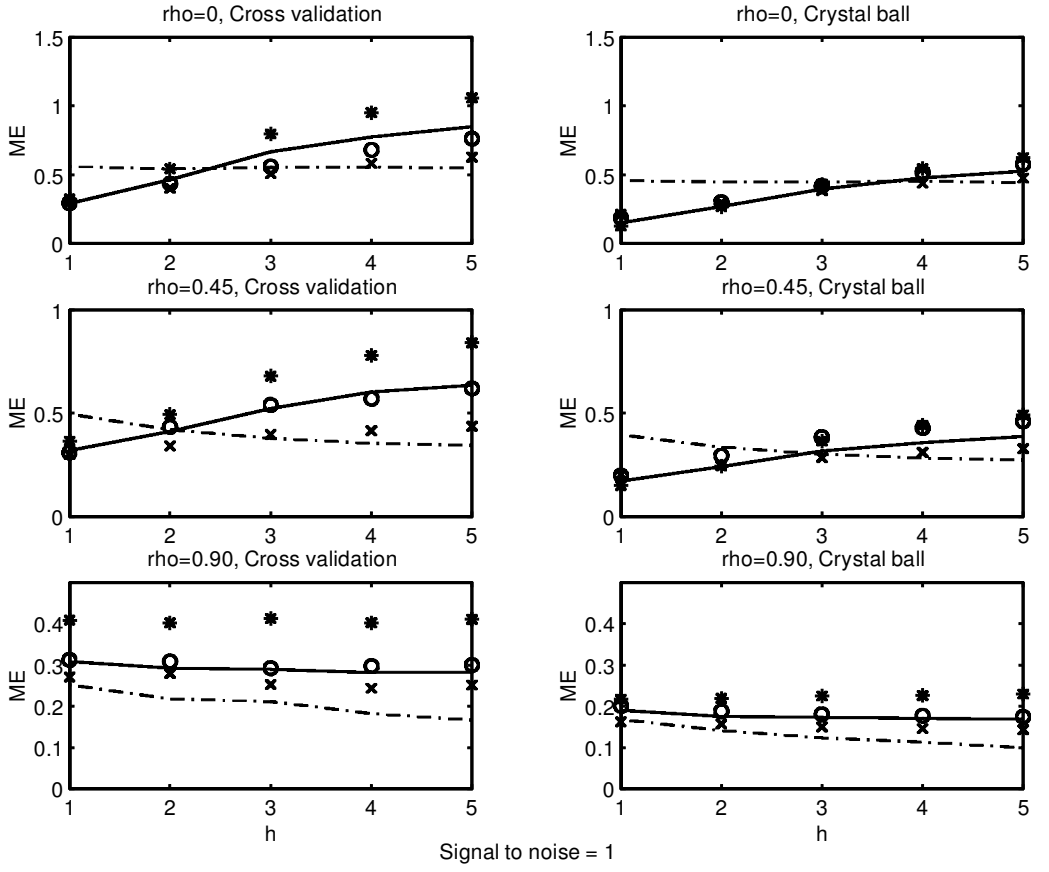

Figure 2. ME as a Function of Cluster Size h for SN Ratio SN = 1. Mean subset (full line); best subset (stars); garrote (circles); lasso (crosses); ridge regression (dashed and dotted line).

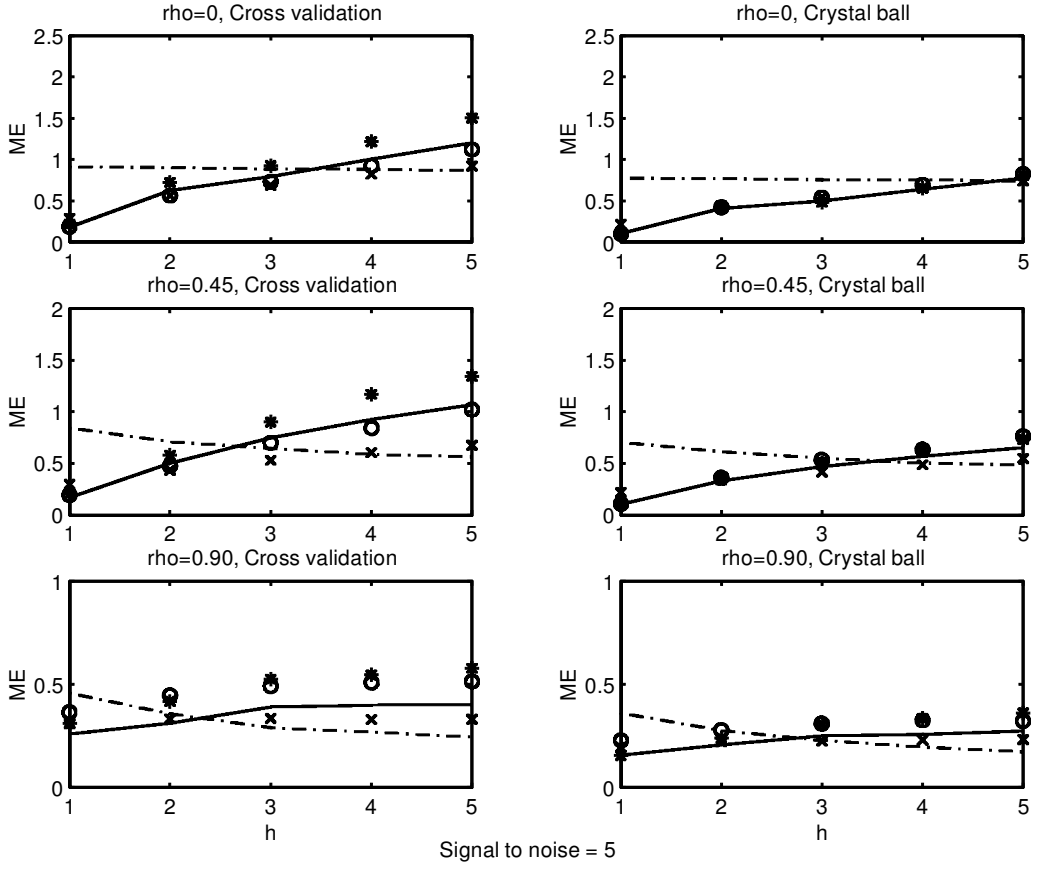

Figure 3. ME as a Function of Cluster Size h for SN Ratio SN = 5. Mean subset (full line); best subset (stars); garrote (circles); lasso (crosses); ridge regression (dashed and dotted line).

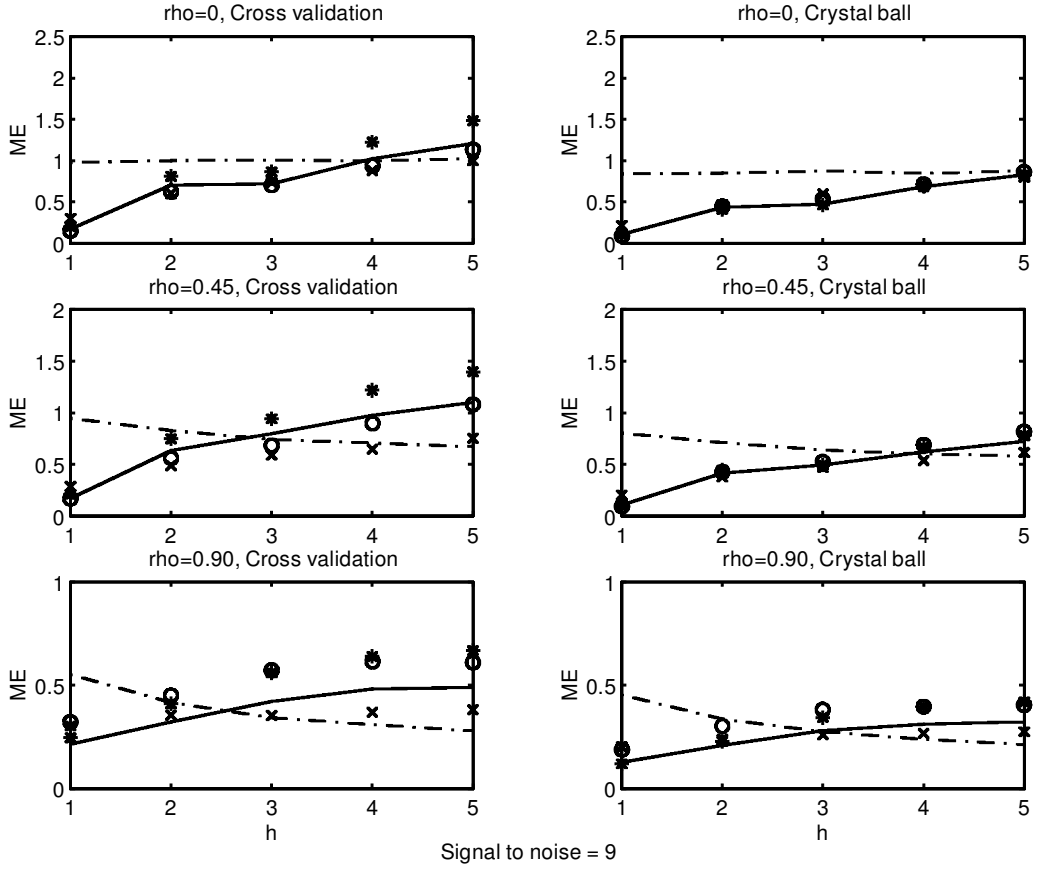

Figure 4. ME as a Function of Cluster Size h for SN Ratio SN = 9. Mean subset (full line); best subset (stars); garrote (circles); lasso (crosses); ridge regression (dashed and dotted line).

(NIR) spectra (Kalivas 1997). Determining the amount of moisture and protein in wheat normally involves costly and time-consuming laboratory experiments. Hence it is interesting to investigate the predictability of these factors using cheap and quickly obtainable NIR spectra. This example is also interesting because of the high number of regressors $p = 700$ (recorded from 1,100–2,500 nm in 2-nm intervals), compared to the relatively low number of observations, $n = 100$. Consequently, the problem is severely indeterminate, and it is necessary to use a shrinkage method to obtain a unique solution to the least squares problem. Hence shrinkage methods that depend on a unique least squares estimate under the full model, such as the garrote, will not work.

The spectra have baseline variation caused by the light-scattering effects of particles of different sizes and shapes. To depress this noninformative variation, the difference spectra are calculated, that is,

$$\boldsymbol{X} = \begin{bmatrix} s_{1,2} & \cdots & s_{1,701} \\ \vdots & \ddots & \vdots \\ s_{100,2} & \cdots & s_{100,701} \end{bmatrix} - \begin{bmatrix} s_{1,1} & \cdots & s_{1,700} \\ \vdots & \ddots & \vdots \\ s_{100,1} & \cdots & s_{100,700} \end{bmatrix},$$

where $s_{i,j}$ is the measured reflectance at wavelength number j of spectrum i. Figure 5 shows three typical difference spectra of wheat. As in the previous example, the data are centered before calibration.

To allow comparison of the methods, the data are split into a validation set (34 observations) and a calibration set (66 observations). The calibration set is used to estimate the shrinkage factors by cross-validation. To make the calculation computationally feasible, the best subset and mean subset for $q > 3$ are based on the approximation described in Section 2. Furthermore, the number of subsets used for approximating the mean subset is $k = 1{,}000$ for all values of $q > 3$. Tables 1 and 2 give the R^2 values for prediction of protein and moisture using the validation data.

The tables show that the mean subset method is best for predicting protein but that the results are more alike for moisture, with the ridge regression method slightly better. In an attempt to explain why ridge regression is better in predicting the amount of moisture, the difference spectra were smoothed before variable selection. It was then found that when the spectra were smoothed using a local polynomial of order two and a bandwidth of about 70 nm, a prediction performance of $R^2 = .972$ was obtainable with a single-variable best subset model. This indicates that the information for moisture is spread over several highly correlated variables and explains why the ridge regression method performs better than the subset selection

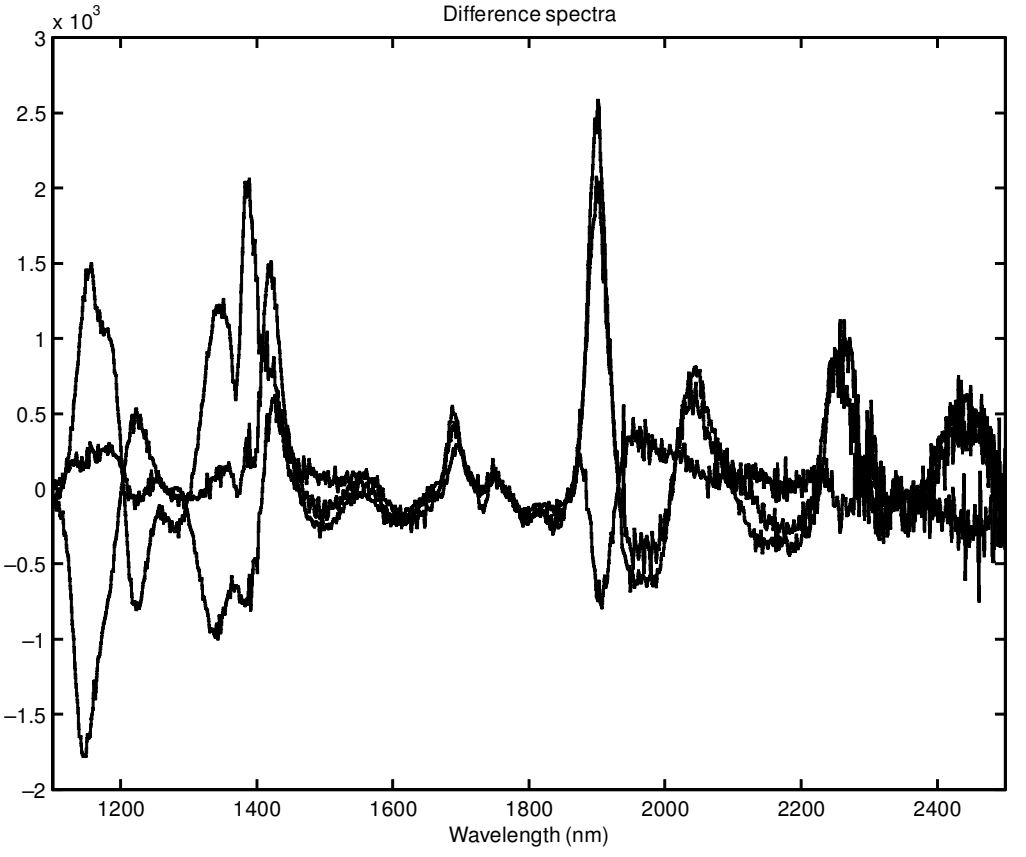

Figure 5. Three of the Difference Spectra. The number of regressors is 700.

methods. Figure 6 shows the estimated coefficient vectors for the four methods.

Comparing the amount of shrinkage in Table 1 and 2 reveals that all methods shrink the coefficient vectors less for the protein data than for the moisture data. Furthermore, for the protein data, the mean subset method has better prediction performance than the other methods. Figure 7 shows the coefficient vectors for predicting protein. An important difference between the ridge regression and variable selection methods, is that the later give a more clear-cut interpretation of the data. For instance, studying the coefficient for mean subset shows that spectral data above 1,800 nm shows no useful relation to protein. By removing this noninformative data, a more robust calibration model can be obtained. In general, chemical substances absorb radiation only in limited spectral regions. Hence, mean subset may be used to identify these important spectral regions.

Figure 8 plots the R^2 value of the validation data as a function of the number of variables. The figure clearly shows that the mean subset is much more stable than the best subset. Notice that the best subset model of size three has been found by exhaustive search and not by the approximative exchange algorithm. This corroborates the result of the Monte Carlo simulation study and explains why the difference in prediction performance between the best subset and mean subset methods is greater when cross-validation is used instead of the crystal ball.

5. SUMMARY

In this article we have addressed the problem of using subset selection as a shrinkage method. It is known that using best subset selection as a shrinkage method is an unstable approach, because a small change in the data may lead to large changes in the selected explanatory variables. To avoid this

Table 1. Prediction of Protein Content

Method	R^2	*Shrinkage factor*
Mean subset	.834	$q = 5$
Best subset	.777	$q = 5$
Ridge regression	.793	$k_R = 1.80e-7$
Lasso	.790	$t = 16{,}926$

Table 2. Prediction of Moisture Content

Method	R^2	*Shrinkage factor*
Mean subset	.960	$q = 4$
Best subset	.955	$q = 2$
Ridge regression	.976	$k_R = 2.97e-6$
Lasso	.960	$t = 5{,}967$

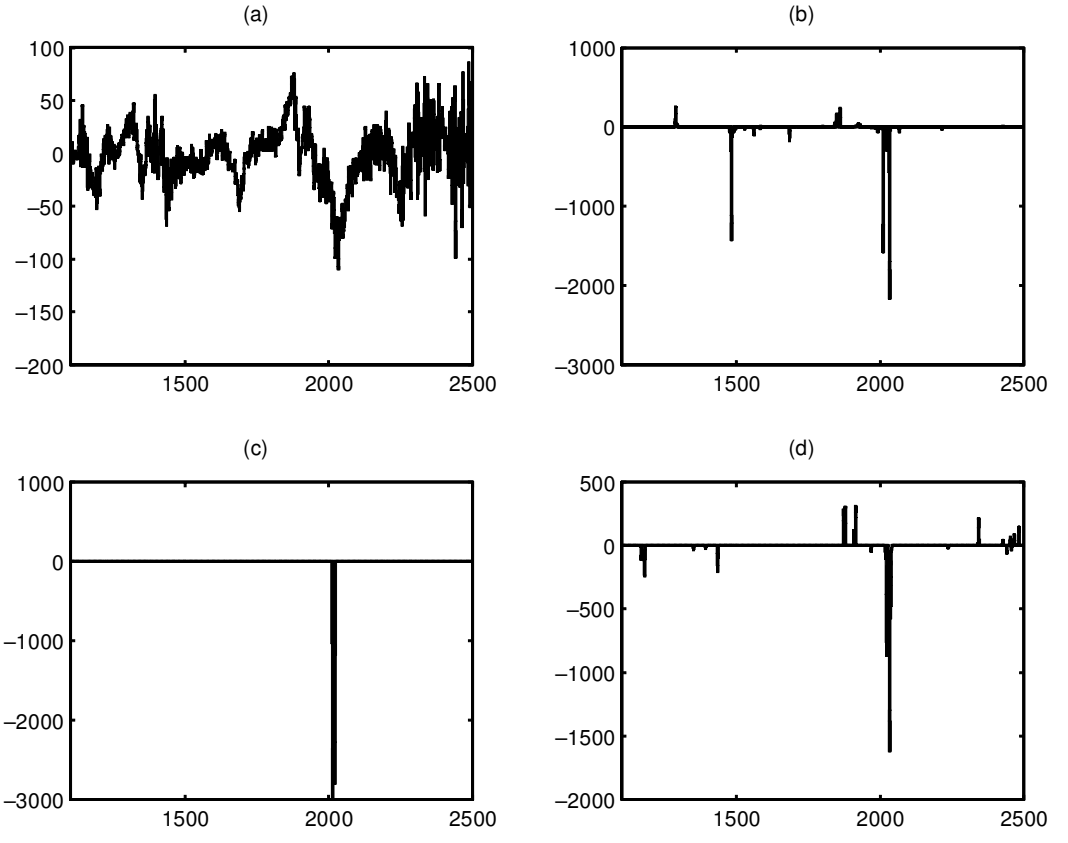

Figure 6. Coefficient Vector for Prediction of Moisture. (a) Ridge; (b) mean subset; (c) best subset; (d) lasso.

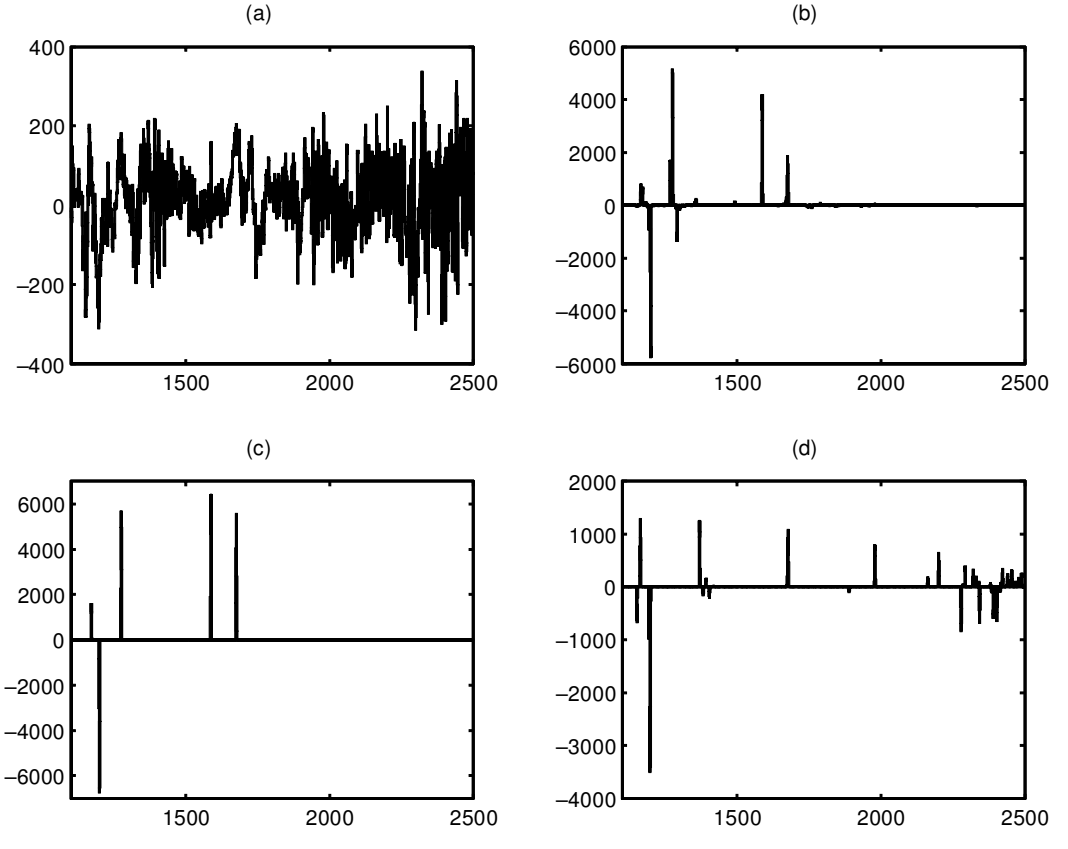

Figure 7. Coefficient Vector for Prediction of Protein. (a) Ridge; (b) mean subset; (c) best subset; (d) lasso.

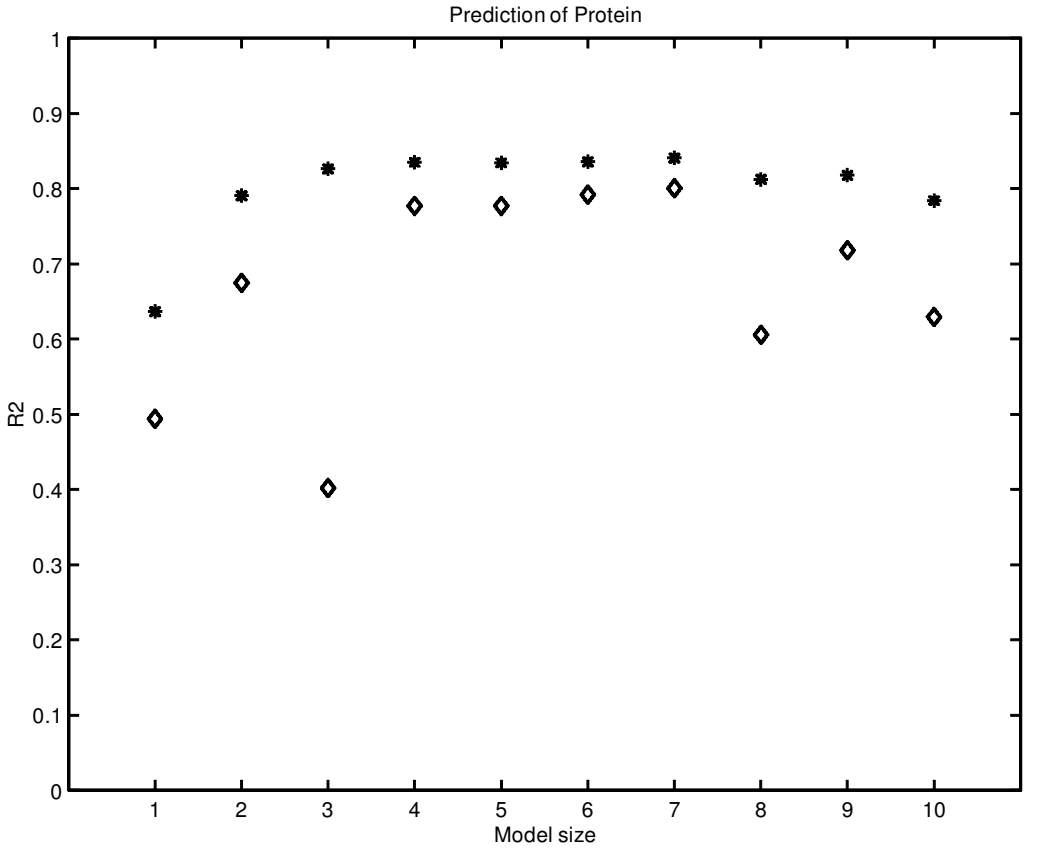

Figure 8. The Amount of Explained Variation of the Validation Data as a Function of the Number of Selected Variables. The results plotted with stars and diamonds are the mean subset and best subset.

problem, we suggest instead using the *mean subset* for prediction. This method is motivated using Bayesian arguments. We described a numerical method for finding the mean subset based on an approximating exhaustive search.

The results from the Monte Carlo simulation study, summarized in Table 3, show that mean subset works well when the underlying model is small and the SN ratio is high. The study also showed that the garrote is dominated by the lasso and best subset is dominated by mean subset.

Finally, in an example that used NIR spectra of wheat to predict the amount of moisture and protein, the mean subset method was found to produce both simple and competitive prediction models. The promising result of using mean subset instead of best subset (i.e., integration instead of maximization) suggests that the lasso method also could be enhanced by interpreting the penalty function as an a priori distribution and thereby calculate the expectation. This is a subject for future work.

Table 3. Classification of When the Different Shrinkage Methods Work Well

Model size\SN	*Low*	*Medium*	*High*
Small	Lasso	Lasso	Mean subset
Medium	Ridge regression	Ridge regression	Lasso
Large	Ridge regression	Ridge regression	Ridge regression

ACKNOWLEDGMENTS

The authors would particularly like to thank the associate editor and the referees, whose comments led to significant improvements in this article. The first author wishes to acknowledge that this work was partially supported by the Danish Academy of Technical Sciences.

[Received February 2001. Revised November 2001.]

REFERENCES

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

——— (1996), "Heuristics of Instability and Stabilization in Model Selection," *The Annals of Statistics*, 24, 2350–2383.

Breiman, L., and Spector, P. (1992), "Submodel Selection and Evaluation in Regression—The x-Random Case," *International Statistical Review*, 60, 291–319.

Brown, P. J., Vannucci, M., and Fearn, T. (1998), "Multivariate Bayesian Variable Selection and Prediction," *Journal of the Royal Statistical Society*, Ser. B, 60, 627–641.

Copas, J. B. (1983), "Regression, Prediction and Shrinkage," *Journal of Royal Statistical Society*, 45, 311–354.

Dempster, A. P. (1973), "Alternatives to Least Squares in Multiple Regression," in *Multivariate Statistical Analysis*, eds. D. Kabe and R. P. Gupta, Amsterdam: North-Holland, 25–40.

Furnival, G. M., and Wilson, R. W. (1974), "Regression by Leaps and Bounds," *Technometrics*, 16, 499–511.

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistical Sinica*, 7, 339–373.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

James, W., and Stein, C. (1961), "Estimation With Quadratic Loss," in *Proceedings of the 4th Berkeley Symposium*, Vol. 1, pp. 361–379.

Kalivas, J. H. (1997), "Two Data Sets of Near Infrared Spectra," *Chemometrics and Intelligent Laboratory Systems*, 37, 255–259.

Leamer, E. E., and Chamberlain, G. (1976), A Bayesian Interpretation of Pretesting. *Journal of the Royal Statistical Society*, Ser. B, 38, 85–94.

Miller, A. J. (1990), *Subset Selection in Regression*, Monographs on Statistics and Applied Probability, Vol. 40, London: Chapman and Hall.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Vach, K., Sauerbrei, W., and Schumacher, M. (2001), "Variable Selection and Shrinkage: Comparison of Some Approaches," *Statistica Neerlandica*, 55, 53–75.

Zellner, A. (ed.) (1980), *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, Amsterdam: North-Holland.