# Calibration with Empirically Weighted Mean Subset

**HENRIK ÖJELUND,* HENRIK MADSEN, and POUL THYREGOD**

*Informatics and Mathematical Modelling, DTU, DK-2800, Lyngby, Denmark (H.O., H.M., P.T.); and Danfoss Analytical, Ellegårdvej 36, DK-6400, Sønderborg, Denmark (H.O.)*

**In this article a new calibration method called empirically weighted mean subset (EMS) is presented. The method is illustrated using spectral data. Using several near-infrared (NIR) benchmark data sets, EMS is compared to partial least-squares regression (PLS) and interval partial least-squares regression (iPLS). It is found that EMS improves on the prediction performance over PLS in terms of the mean squared errors and is more robust than iPLS. Furthermore, by investigating the estimated coefficient vector of EMS, knowledge about the important spectral regions can be gained. The EMS solution is obtained by calculating the weighted mean of all coefficient vectors for subsets of the same size. The weighting is proportional to $SS_{\gamma}^{-\omega}$, where $SS_{\gamma}$ is the residual sum of squares from a linear regression with subset $\gamma$ and $\omega$ is a weighting parameter estimated using cross-validation. This construction of the weighting implies that even if some coefficients will become numerically small, none will become exactly zero. An efficient algorithm has been implemented in MATLAB to calculate the EMS solution and the source code has been made available on the Internet.**

Index Headings: **NIR; Calibration; Best subset; Mean subset; Shrinkage; Regularization; Model averaging.**

## INTRODUCTION

In this paper we focus on prediction based on spectral measurements. The classical calibration methods for this type of data are partial least-squares regression (PLS) and principal component regression (PCR).[1] It has been demonstrated that these two methods, together with ridge regression (RR), create similar calibration models and have comparable prediction performance.[2,3] Moreover, all three methods use the whole spectrum for calibration. A different approach to obtain a calibration model is by variable selection. Most variable selection algorithms use the same criterion, i.e., the least squares criterion with constraint on the number of explanatory variables. The genetic algorithm used as a variable selection method has been compared to simulated annealing (SA) and stepwise variable selection methods.[4] However, in a later comparative study,[5] the authors of that study were unable to reproduce the result and, to overcome this, a different approach, to use SA, was suggested.

Another variable selection algorithm is the sequential replacement algorithm. This algorithm is described by Miller[6] and the algorithm can be described as the deterministic counterpart to the stochastic simulated annealing. The solution obtained from variable selection algorithms often yield predictions that are comparable to, and sometimes even better than, the prediction obtained by using the full-spectrum calibration methods.[7]

In the NIR range, specific chemical components often absorb light only in narrow spectral bands. Hence, methods that can discard noninformative spectral regions may show improved prediction performance. However, variable selection methods are sensitive to spectral noise because only a few wavelengths are used for prediction. Furthermore, there is a need for methods that can identify spectral regions and not just single wavelengths. Such methods may be used to develop new sensors based on optical bandpass filters or may be used to limit the required spectral scanning range. The direct way to develop an algorithm for this is to combine variable selection with PLS or PCR.[8–10] Nørgaard et al.[11] proposed a method where data from limited spectral intervals are selected and used for calibration with PLS. The method is called interval partial least squares (iPLS) and has primarily been presented as a graphical method to identify important spectral regions. For some data sets, iPLS has been found to improve on the prediction performance. Another approach is to calculate the mean subset under a Bayesian model assumption.[12] In this method the regression coefficient vectors of all subsets of the same size are weighted proportional to $SS_{\gamma}^{-\omega}$, where $\omega = n/2$. Here $SS_{\gamma}$ denotes the residual sum of squares from a linear regression using the subset of variables indexed by $\gamma$, and $n$ is the number of observations.

Using simulated data, it was shown that the mean subset has better prediction performance than the best subset. However, when the number of observations increases, the mean subset will in practice be identical to the best subset. In the present paper we instead estimate the weighting parameter $\omega$ from the calibration data using cross-validation, and therefore we call the method empirically weighted mean subset, or EMS for short. In this paper EMS is applied to spectral data only, but the method is applicable to all sorts of data (process data, chemical data, etc.) as opposed to, e.g., iPLS. In the Theory section the theory and the optimization criterion behind EMS are presented. An algorithm to efficiently calculate the EMS solution is developed in the Algorithm section. In the Experimental section, the method is compared to PLS and iPLS using benchmark data. The three benchmark data sets consist of NIR spectra of gasoline with corresponding octane number, NIR spectra of wheat with measured moisture and protein contents, and NIR spectra of beer with the corresponding concentration of the original extract. The paper is concluded with a discussion and some remarks.

## THEORY

Let $\mathbf{y}$ be the $(n \times 1)$ vector of observed values of the response variable (e.g., concentration) and $\mathbf{X}$ the $(n \times p)$ matrix including all potential explanatory variables (e.g., spectra). Here, $n$ denotes the number of observations and $p$ the number of explanatory variables. The linear model has the form:

Received 30 April 2001; accepted 4 March 2002.

* Author to whom correspondence should be sent.

0003-7028/02/5607-0887$2.00/0

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

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 $1 \leq j \leq p$. Define the cardinal function $q_\gamma = \Sigma_{i=1}^{p} i_j$. When a selection vector is used as a subscript to a matrix, this shall be understood to mean the matrix with $q_\gamma$ 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 zeros 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 $p = 5$ and $\gamma = (1, 0, 1, 1, 0)'$ and the coefficient vector $\hat{\boldsymbol{\beta}}_\gamma = (5, 6, 7)'$, then $\hat{\boldsymbol{\beta}}^{(\gamma)} = (5, 0, 6, 7, 0)'$. Furthermore, let $S_q$ be a set of $(p \times 1)$ selection vectors defined as

$$S_q = \{\gamma | \gamma = (i_1, i_2, \ldots, i_p), i_j \in \{0, 1\}, 1 \leq j \leq p, \\ q_\gamma = q\}$$

where $q$ is the number of selected variables. Hence, $S_q$ will be the set of all subsets where the number of variables is $q$.

We now introduce the empirically weighted mean subset method. The coefficient vector $\hat{\boldsymbol{\beta}}_{\mathrm{EMS}}^{(q)}$ is calculated as

$$\hat{\boldsymbol{\beta}}_{\mathrm{EMS}}^{(q)} = \sum_{\gamma \in S_q} w_\gamma \hat{\boldsymbol{\beta}}^{(\gamma)} \qquad (2)$$

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

$$w_\gamma \propto \mathrm{SS}_\gamma^{-\omega} \qquad (3)$$

where $\mathrm{SS}_\gamma = (\mathbf{y} - \mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)'(\mathbf{y} - \mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)$ and the two parameters controlling the complexity of the model are $\omega \geq 0$, which controls how the subsets are weighted, and $q$, which controls the size of the subsets. These parameters will be estimated using cross-validation. It is assumed that all matrices $\mathbf{X}_\gamma'\mathbf{X}_\gamma$ for which $\gamma \in S_q$ have full rank, so that necessarily $q \leq n$.

Depending of the value of $\omega$, EMS will take the shape of other methods. For $\omega = 1$, the subsets are weighted according to the sample variance. Conceptually this can be compared to weighted least-squares regression where the observations, instead of the parameter estimates, are weighted with the reciprocal variance of the observations instead of the estimated variance of the residuals. For $\omega = n/2$, it can be shown that under a particular Bayesian model, $\hat{\boldsymbol{\beta}}_{\mathrm{EMS}}^{(q)}$ will be equal to the posterior mean.[12] And for $\omega = \infty$, the method is identical to best subset regression. Hence, EMS changes smoothly from a full spectrum method to a variable selection method. Another appealing aspect of EMS is that the method is defined from a well-defined criterion and not just as an algorithm. This makes it easier to understand the properties of the method. Moreover, the weighting of the subsets does not depend on the scaling of the variables. Hence, prior standardization of the variables has no effect on the prediction performance. When the coefficient vector of EMS is calculated, it is necessary to calculate the regression coefficients of all subsets of size $q$. This is, of course, a computational challenge, and presently it is possible only to determine the exact solution for small values of $q$. In the next section, an efficient algorithm will be developed.

## ALGORITHM

The EMS method is computationally expensive because the regression coefficients for all subsets of size $q$ have to be calculated. For example, when the number of explanatory variables is $p = 1000$ and $q = 5$, the number of coefficient vectors to be calculated is 8.2503e + 12. Hence, the implemented algorithm is limited to the cases when $q \leq 4$. The algorithm has been written in C and interfaced to MATLAB using the MEX library. The source code can be down-loaded from `www.imm.dtu.dk/~hoe`.

A central part of the algorithm is the sweep operator.[13] The following description of the sweep operator has been adopted from Schatzoff et al.[14] A square matrix $\mathbf{A} = (a_{ij})$ is said to have been swept on the $r$th row and column (or $r$th pivotal element) when it has been transformed into a matrix $\mathbf{B} = (b_{ij})$ such that

$$\begin{aligned} b_{rr} &= 1/a_{rr} && \\ b_{ir} &= a_{ir}/a_{rr} && i \neq r \\ b_{rj} &= a_{rj}/a_{rr} && j \neq r \\ b_{ij} &= a_{ij} - a_{ir}a_{rj}/a_{rr} && i, j \neq r \end{aligned} \qquad (4)$$

If the square cross product matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{y} \\ \mathbf{y}'\mathbf{X} & \mathbf{y}'\mathbf{y} \end{pmatrix} \qquad (5)$$

is swept on the $r$th pivotal element, where $0 < r \leq p$, and transformed to $\mathbf{C}^* = (c_{ij}^*)$, the elements $c_{rp+1}^*$ will be the least-squares estimate of a linear model with only the $r$th variable. Furthermore, $c_{p+1p+1}^*$ will be the residual sum of squares, given the least-squares estimate, for the aforementioned model. If $\mathbf{C}^*$ is swept on the $s$th pivotal element, where $0 < s \leq p$ and $s \neq r$, and transformed to $\mathbf{C}^{**} = (c_{ij}^{**})$, the elements $c_{rp+1}^{**}$ and $c_{sp+1}^{**}$ will be the least-squares estimates of the corresponding two-variable linear model and $c_{p+1p+1}^{**}$ will be the residual sum of squares. Because the sweep operator is reversible and commutative, sweeping on the same pivotal element twice is equivalent to not having swept the matrix at all, and sweeping on the $r$th pivotal element and then the $s$th pivotal element is equivalent to sweeping the matrix in the opposite order. It has been shown[14] that it is possible to make the sweep operation more efficient by only working on the upper triangular part and introducing a parity vector which indicates whether a matrix has been swept an even or odd number of times on each pivotal position.

**Algorithm.** The following pseudo code illustrates how the algorithm works when $q \leq 3$.

```
for r:=1 to p-2
begin
  C1:=sweepA(C, r)
  ⋮
  ⋮ Include the regression coefficient
  ⋮ in the weighted summation for $q = 1$
  ⋮
  for s:=r+1 to p-1
  begin
    C2:=sweepB(C1, r, s)
    ⋮
    ⋮ Include the regression coefficients
    ⋮ in the weighted summation for $q = 2$
    ⋮
    for t:=s+1 to p
    begin
      C3:=sweepC(C2, r, s, t)
      ⋮
      ⋮ Include the regression coefficients
      ⋮ in the weighted summation for $q = 3$
      ⋮
    end
  end
end
```

The variable C denotes the cross product matrix defined in Eq. 5, while C1, C2, and C3 denote this matrix swept once, twice, and three times, respectively. The three different sweep functions sweepA(.), sweepB(.), and sweepC(.), sweep the matrix on position r, s, and t, respectively. The implemented sweep functions differ in the part of the matrix that is calculated. The function sweepA(.) calculates only rows ≥ r, function sweepB(.) calculates rows ≥ s and the $r$th row, and finally function sweepC(.) calculates only the regression coefficients at row r, s, and t and the residual sum of squares. In the sweep functions, the variables are tested not to be linear dependent. If this is the case, the coefficient vector of such a subset will not be included. When the number of variables is $p = 500$ and $q = 3$, the algorithm will take approximately 15 s on an HP 9000/785 server.

The algorithm described above can easily be extended to handle larger values of $q$. Today it is computationally feasible to calculate the coefficient vector of EMS when $p < 200$ and $q \leq 4$, or when $p < 1000$ and $q \leq 3$, say. An algorithm to handle $q \leq 4$ has been implemented and applied in the following sections.

## EXPERIMENTAL SETUP

Three data sets have been used to compare PLS and iPLS with the proposed EMS method. PLS and iPLS have been chosen for comparison as the former of these methods is implemented in several software packages and is routinely used for calibration and the latter addresses the same problem of identifying important spectral intervals. The data sets have been split into validation and calibration sets in the following manner: the observations are sorted with respect to the component for which a calibration model should be obtained, e.g., the octane number or the amount of protein. From the sorted data, observations $\{2, 5, \ldots, 3\,[(n + 1)/3] - 1\}$ are reserved for validation and the remaining observations for calibration, which means that one-third of the data is used for validation. This validation data will only be used to evaluate the calibrated models and not to estimate the hyper parameters.

All the hyper parameters, i.e., the number of latent variables (LV) in the PLS and iPLS methods, the size and placement of the interval in the iPLS method, and the size of the subsets (value of $q$) and the weighting of the subsets (value of $\omega$), are estimated using 5-fold cross-validation on the calibration data. By using 5-fold cross-validation, the computation as well as the risk of over fitting[15] is reduced compared to leave-one-out cross-validation. The cross-validation has been performed by dividing the sorted calibration data into five groups of equal size (or as close as possible) and sequentially using the data in four of the groups to predict in the fifth. Hence, one of the groups will include all the observations of the calibration data with the highest concentrations and one group all observations of the calibration data with the lowest concentrations. The validation data has been centered using the sample mean of the calibration data. To compare the methods, the following performance measures have been calculated: the root mean square error of prediction (RMSEP), the root mean square error of cross-validation (RMSECV), the root mean square error of calibration (RMSEC), and the percentage of explained variation of the prediction data ($R^2$). The values of the hyper parameters associated with the smallest RMSECV values have been selected.

**The Methods.** *Setup for PLS.* Two different criteria have been used to select the number of latent variables (LV) for the full spectrum PLS. In the first criterion, denoted A, the number of LV for which the lowest RMSECV value is obtained is selected. In the second criterion, denoted B, the number of LV is selected for which the RMSECV sequence has its first local minimum. In both criteria the limits on the number of LV have been set to minimum 3 and maximum 20.

*Setup for iPLS.* For interval partial least-squares regression,[11] the spectra are divided into a number of equally sized intervals. Initially, the spectra are not divided (only one interval) and iPLS is identical to PLS. In the next step, the spectra are divided in two intervals and PLS is applied separately in each interval. This has been repeated until the spectra are divided into 20 separate intervals. The total number of intervals for which PLS is applied then becomes $1 + 2 + \cdots + 19 + 20 = 210$. Two criteria have been used to select the interval and the number of LV. In the first criterion, denoted A, the interval and the associated number of PLS LV with the lowest RMSECV value is selected. In the second criterion, denoted B, the first local minimum principal, as described in the setup for PLS, has been applied to each of the 210 intervals. From these 210 local minima, the interval and the associated number of LV with the lowest RMSECV value are selected. In both criteria the limits have been set to minimum 3 LV and maximum 20 LV (or the number of variables in the interval when the number of variables is less than 20).

In Nørgaard et al.,[11] a more advanced iPLS method is described where spectral data from more than one interval is used in each PLS calibration. In this comparison, however, we will apply the simplest version of iPLS where spectral data from only one interval is used in each calibration.

*Setup for EMS.* The number of variables $q$, and the weighting parameter $\omega$, are estimated by calculating the $4 \times 31$ RMSECV values in the grid defined by $q = 1$,

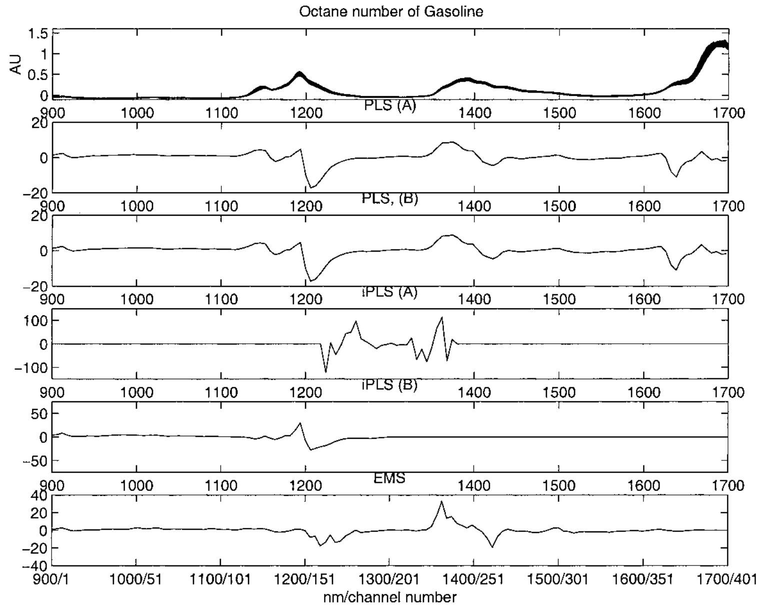

FIG. 1. Spectra of gasoline and the estimated coefficient vectors for predicting the octane number. For the EMS model, $\omega = 3.16$ and $q = 4$. PLS(A) and PLS(B) obtain identical models in this example.

2, 3, 4 and $\omega = 10^{-1.0}, 10^{-0.9}, \ldots, 10^{2.0}$. The grid point with the lowest RMSECV value estimates $q$ and $\omega$.

**The Data.** *Gasoline Data.* This data set was submitted by Kalivas[16] and proposed as a standard reference data set. It contains 60 gasoline samples with specified octane numbers. Samples were measured using diffuse reflectance as ($1/R$) from 900 to 1700 nm in 2 nm intervals. To limit the computation, we consider only every third data point, $p = 134$. All the spectra are shown in the upper graph in Fig. 1.

*Wheat Data.* This data set was also submitted by Kalivas.[16] It consists of 100 wheat samples with specified protein and moisture content. Samples were measured using diffuse reflectance as ($1/R$) from 1100 to 2500 nm in 2 nm intervals, but in this paper, only every fifth data point is considered, $p = 141$. All the spectra are shown in the upper graphs in Figs. 2 and 3. Calculating the difference of the spectra is a data pretreatment method often used on NIR data to depress noninformative variation. However, all pretreatment methods will also depress the informative variation to some extent. Depending on the type of calibration method used, different pretreatment methods will work well. To avoid this extra source of uncertainty in the calibration, we have chosen not to include any data pretreatment except for mean centering.

*Beer Data.* This data set has previously been used to demonstrate iPLS.[11] In that paper, it was found that by using only a small interval of the spectrum, the calibra-

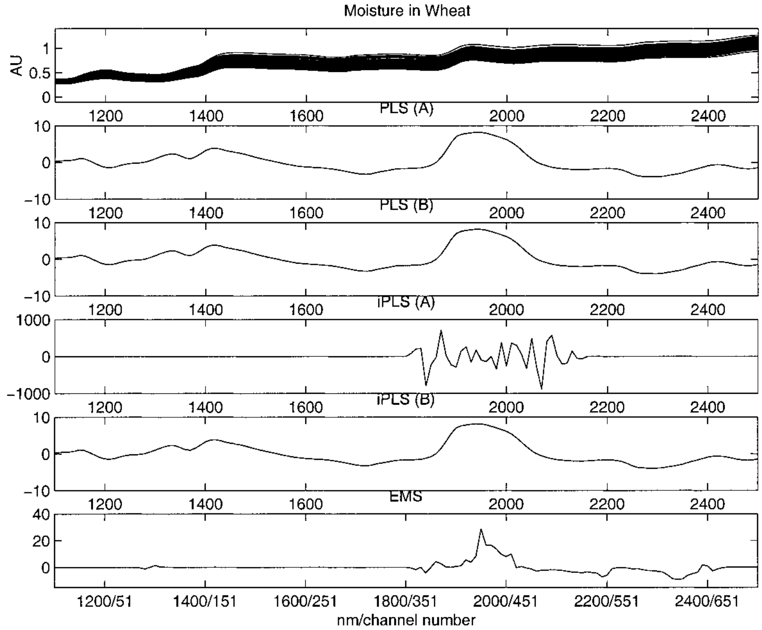

FIG. 2. NIR spectra of wheat and the estimated coefficient vectors for predicting the content of moisture. For the EMS model, $\omega = 5.01$ and $q = 2$. PLS(A), PLS(B), and iPLS(B) obtain identical models in this example.

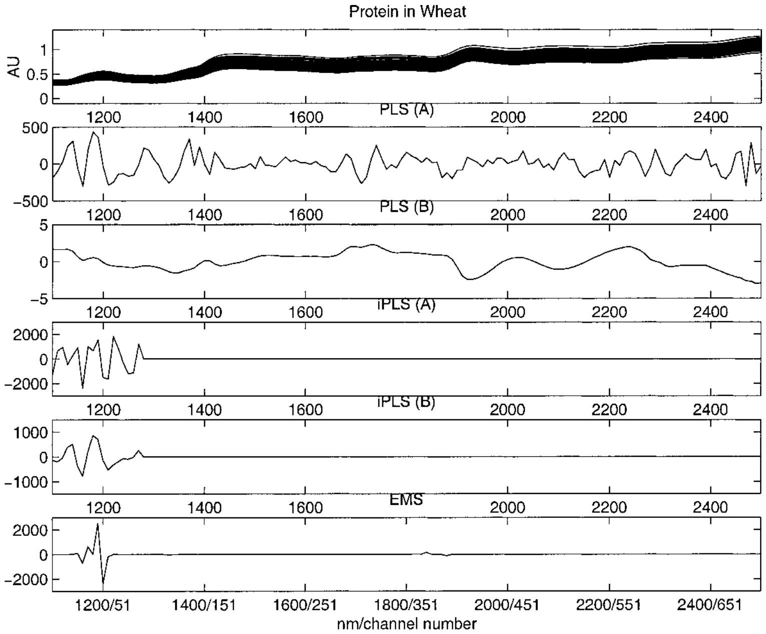

FIG. 3. NIR spectra of wheat and the estimated coefficient vectors for predicting the content of protein. For the EMS model, $\omega = 63.1$ and $q = 4$.

tion model could be enhanced in terms of squared prediction errors. The transmission spectra of beer were recorded at 2 nm intervals in the range from 400 to 2250 nm and were converted to absorbance spectra. In this paper the spectra have been subsampled and only every seventh data point is considered, $p = 133$. A total of 60 beer samples have been analyzed and the spectra are shown in the upper graph in Fig. 4.

Kalivas[16] suggested partitions for validation and calibration of the wheat and gasoline data sets. We have chosen to split the data sets differently because, firstly, we are only interested in one validation set for each data set, and secondly, we would like to use the same partition strategy for all data sets. In Brenchley et al.,[17] a heuristic method to select spectral bands was proposed. In that article no improvement in prediction performance was found for the octane number and amount of moisture. However, some improvment was reported for the protein data. These results were also communicated in Kalivas.[16]

## RESULTS AND DISCUSSION

In Table I, the results using PLS, iPLS, and EMS are shown. It is found that EMS works as good as or better than PLS for all data sets and for all LV selection methods. With the exception of the protein data where

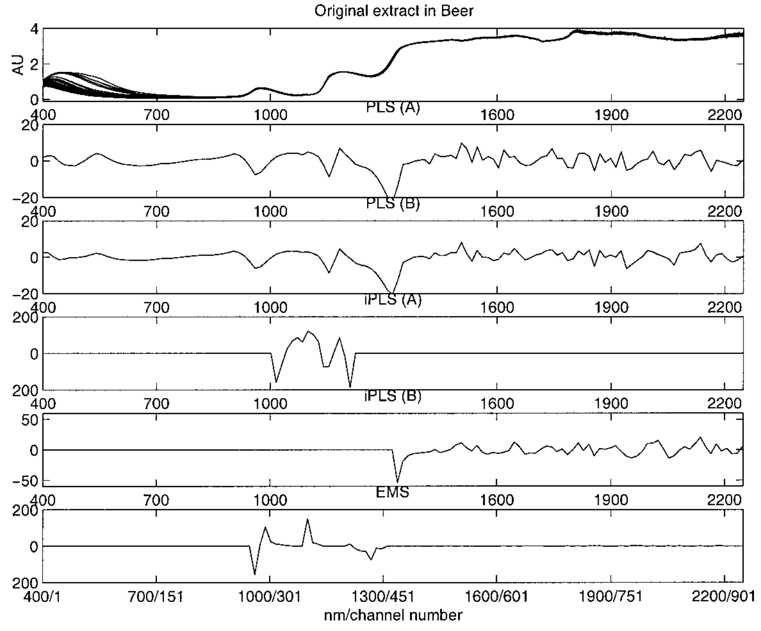

FIG. 4. NIR spectra of beer and the estimated coefficient vectors for predicting the concentration of original extract. For the EMS model, $\omega = 25.1$ and $q = 4$.

TABLE I. Comparison between PLS, iPLS, and EMS.

| Method | RMSECV | RMSEC | RMSEP ($R^2$) | LV/$q$ | Interval | ω |
|---|---|---|---|---|---|---|
| | | | Octane number of gasoline | | | |
| PLS(A) | 0.26 | 0.21 | 0.25 (0.971) | 4 | | |
| PLS(B) | 0.26 | 0.21 | 0.25 (0.971) | 4 | | |
| iPLS(A) | 0.21 | 0.18 | 0.20 (0.981) | 5 | 3 of 5 | |
| iPLS(B) | 0.21 | 0.19 | 0.19 (0.983) | 4 | 1 of 2 | |
| EMS | 0.27 | 0.18 | 0.19 (0.984) | 4 | | 3.16 |
| | | | Moisture in wheat | | | |
| PLS(A) | 0.24 | 0.23 | 0.21 (0.976) | 3 | | |
| PLS(B) | 0.24 | 0.23 | 0.21 (0.976) | 3 | | |
| iPLS(A) | 0.23 | 0.18 | 0.19 (0.980) | 11 | 3 of 4 | |
| iPLS(B) | 0.24 | 0.23 | 0.21 (0.976) | 3 | 1 of 1 | |
| EMS | 0.28 | 0.22 | 0.20 (0.979) | 2 | | 5.01 |
| | | | Protein in wheat | | | |
| PLS(A) | 0.41 | 0.16 | 0.63 (0.661) | 19 | | |
| PLS(B) | 0.65 | 0.41 | 0.78 (0.479) | 4 | | |
| iPLS(A) | 0.30 | 0.24 | 0.41 (0.853) | 9 | 1 of 8 | |
| iPLS(B) | 0.30 | 0.42 | 0.41 (0.853) | 5 | 1 of 8 | |
| EMS | 0.60 | 0.32 | 0.52 (0.766) | 4 | | 63.1 |
| | | | Org. extract in beer | | | |
| PLS(A) | 1.27 | 0.003 | 0.70 (0.917) | 20 | | |
| PLS(B) | 1.28 | 0.18 | 0.50 (0.959) | 7 | | |
| iPLS(A) | 0.13 | 0.14 | 0.22 (0.992) | 6 | 4 of 9 | |
| iPLS(B) | 0.42 | 0.51 | 1.34 (0.697) | 4 | 2 of 2 | |
| EMS | 0.18 | 0.10 | 0.18 (0.995) | 4 | | 25.1 |

iPLS(A) and iPLS(B) work slightly better than EMS, iPLS(A) and EMS seem to have similar prediction performance. For the gasoline and wheat examples, iPLS(A) and iPLS(B) derive to different models but with approximately the same prediction performance. For the beer data, however, iPLS(B) fails to find a suitable model. It is also seen that EMS can be used to classify the type of data. A small value of ω means that the individual subsets of variables should be weighted more equally and a large value means that the best subsets are more important. Hence, a large value of ω indicates that a narrow spectral interval is important and a small value indicates that the whole spectrum, or large parts, contain information about the component of interest. It is seen that when ω is small, PLS works well and only few LV are used and when ω is large, PLS does not work as well and more LV are included. The low RMSEC values in relation to the RMSECV values for PLS(A) method using the protein and beer data indicate overfitting. For the beer data, this can be seen in the poor prediction performance of the PLS(A) model. For the protein data, however, the ''overfitted'' PLS(A) model has better prediction performance than the PLS(B) model. This illustrates the difficulty in estimating the number of LV in PLS using cross-validation.

TABLE II. Length of estimated coefficient vectors.

| | PLS(A) | PLS(B) | iPLS(A) | iPLS(B) | EMS |
|---|---|---|---|---|---|
| | | | Octane number of gasoline | | |
| 2-norm | 42.2 | 42.2 | 258 | 64.1 | 60.9 |
| 1-norm | 288 | 288 | 930 | 318 | 354 |
| | | | Moisture in wheat | | |
| 2-norm | 34.9 | 34.9 | 1990 | 34.9 | 51.0 |
| 1-norm | 319 | 319 | 9140 | 319 | 269 |
| | | | Protein in wheat | | |
| 2-norm | 1620 | 14.3 | 5220 | 1750 | 3630 |
| 1-norm | 14 800 | 141 | 12 000 | 5870 | 7020 |
| | | | Org. extract in beer | | |
| 2-norm | 52.3 | 46.0 | 360 | 81.9 | 257 |
| 1-norm | 420 | 343 | 1200 | 456 | 680 |

In Figs. 1–4, the estimated coefficient vectors for all methods are shown. It is seen that when ω is small (octane and moisture), several of the coefficients of the EMS estimate are important, and when ω is large (protein and beer) few coefficients are important for the prediction. A feature of EMS is that the important regions of the spectrum can be identified. This is remarkable because, in contrast to iPLS, no relations between neighboring variables are explicitly modeled, i.e., EMS obtains the same model if the data has been permuted. In Table II, the 2-norm of the estimated coefficient vectors for the evaluated methods are given. The 2-norm is calculated from

$$\|\boldsymbol{\beta}\| = \sqrt{\sum_{j=1}^{p} \beta_j^2} \qquad (6)$$

In order to better reflect the visual impression in Figs. 1–4, the 1-norm is also given in Table II. In Fig. 5, the RMSECV values calculated to estimate ω and $q$ are shown. For the gasoline and moisture data, it seems to be sufficient to only consider models with $q = 2$, i.e., weighting of simple linear models with only two parameters. But for the protein and beer data, models with $q = 4$ are the best and the graphs indicate that even models where $q > 4$ should be considered.

In Fig. 6, the gasoline data has been used to demonstrate the influence of ω. The coefficient vectors of EMS for $q = 2$ and increasing values of ω are shown. It can be seen that the important regions of the spectrum smoothly appear when ω increases. EMS can therefore be used as a tool to identify important spectral regions. It should be noted that even if the numerical value of

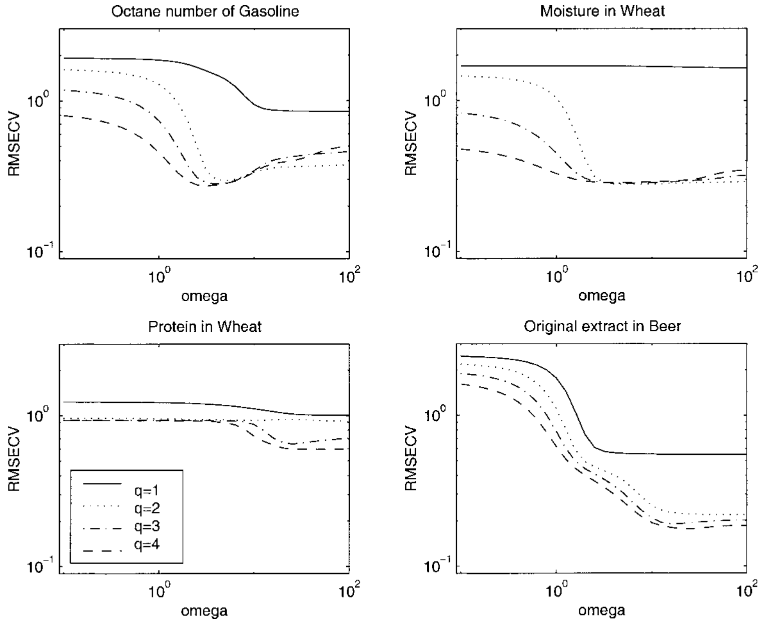

FIG. 5. Results from cross-validation used for estimating ω and $q$.

some of the coefficients becomes small for increasing ω, none of them will become exactly zero.

One important property of a calibration method is robustness to noise in data. To investigate the robustness of the considered methods, a resampling scheme is developed. The resampling has been performed by resampling in the calibration data and keeping the validation data intact. The number of observations resampled is equal to the original size of the calibration data. After a set of observations has been obtained by resampling, the new set of observations is sorted with respect to the quantity of interest, e.g., the octane number or amount of protein. Notice that in the new resampled calibration set, several of the original observations will be duplicated and others will be left out. After the resampled calibration set has been ordered, it is used as a new set of calibration data and the cross-validation procedures described above are commenced. The resampling procedure has been repeated 200 times, and for each repetition, one RMSEP value for each calibration method is obtained. In Figs. 7–10 histograms of these RMSEP values are shown.

A good calibration method should result in low and stable RMSEP values. Here, stable means that the RMSEP values should not vary too much from resampling to resampling. The first observation is that the gasoline data is easy to model and all calibration methods work fairly well. Since full spectrum PLS is a submethod of iPLS, one could be tempted to expect that iPLS will

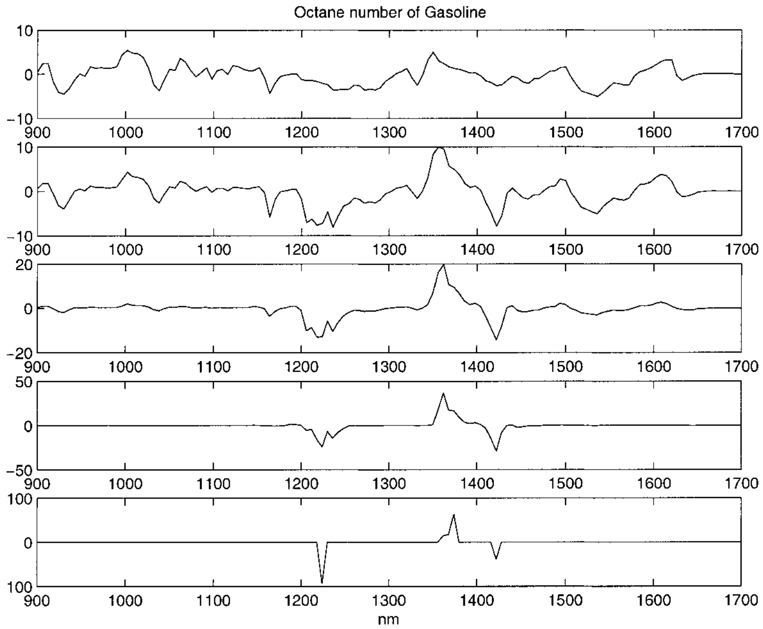

FIG. 6. Estimated coefficient vectors for the gasoline data for $q$ = 2 and ω = (0.1, 1.26, 2.0, 4.0, 100). In the top graph, ω = 0.1 and in the bottom graph ω = 100.

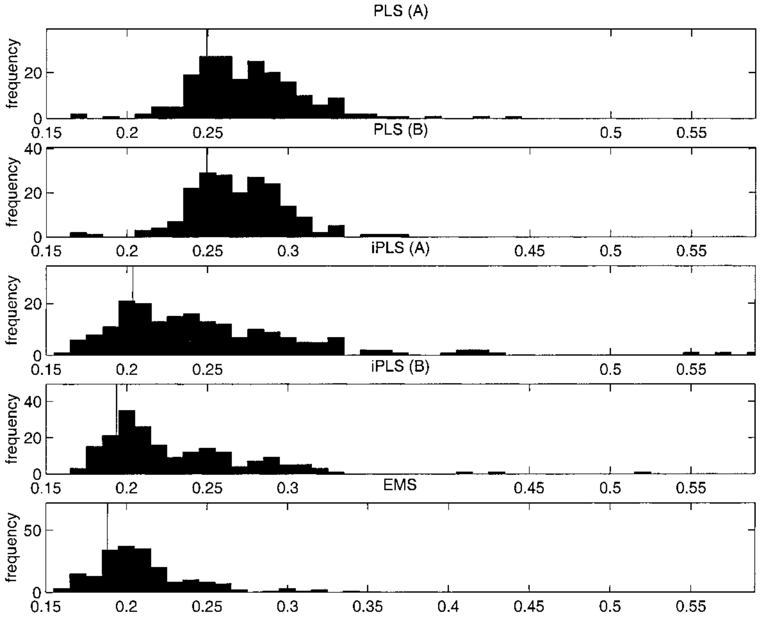

FIG. 7. Histograms of the RMSEP values obtained by resampling of the gasoline data. The line marks the RMSEP value obtained when the original calibration data is used.

always perform as good as or better than PLS. However, this is not the case, as can be seen in Figs. 7 and 8. In these two sets of data, large parts of the spectrum are important for the prediction. The selection of interval procedure in iPLS is therefore unnecessary and adds only uncertainty to the estimated coefficient vector. For the gasoline and moisture data, EMS shows both a stable and good prediction performance and does not encounter the increased variability found for iPLS. The situation is different for the protein data, where iPLS is found to work as stable and slightly better than EMS. This improved performance is probably due to the selected interval efficiently capturing the important variation, or/and $q \leq 4$ is too limiting for the EMS method. Depending on the type of method used to select the number of LV, PLS will either have large variation or poor prediction performance for the protein data.

In the final example, the beer data, EMS and iPLS(A) are found to be the best and most robust prediction methods and none of the PLS methods are found to perform well.

The figures also show the difficulty in selecting the number of LV for PLS (or iPLS). In Fig. 8, iPLS(B) is found to be more robust than iPLS(A), but in Fig. 10, the opposite holds.

## CONCLUSION

The most important aspect of the empirically weighted mean subset (EMS) method is the insight to the problem

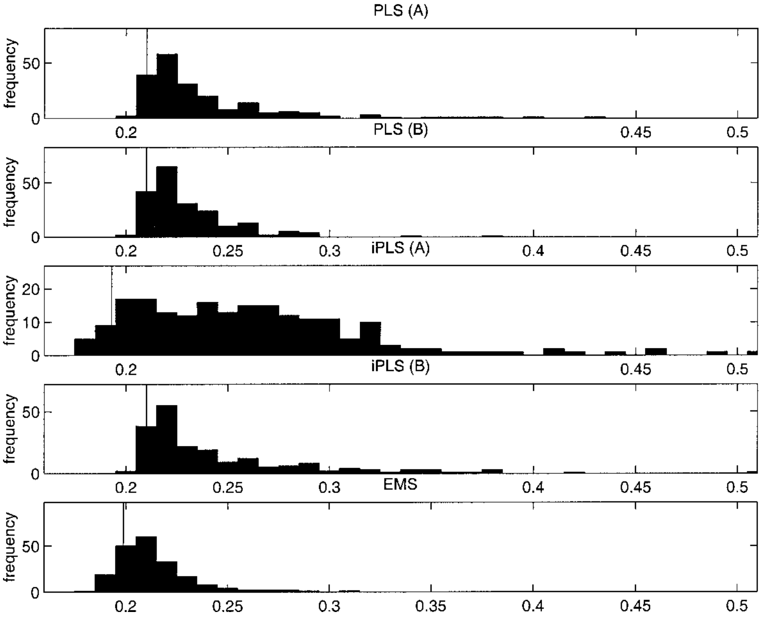

FIG. 8. Histograms of the RMSEP values obtained by resampling of the moisture data. The line marks the RMSEP value obtained when the original calibration data is used.

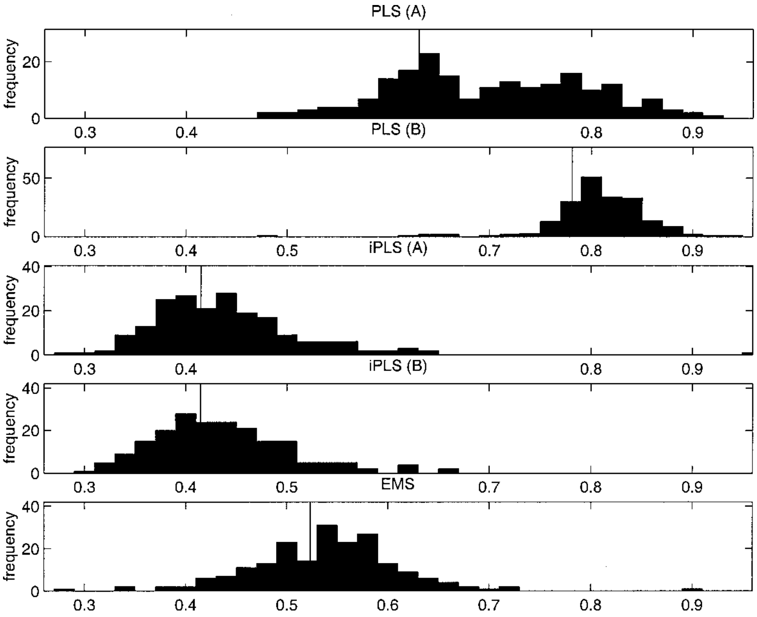

FIG. 9. Histograms of the RMSEP values obtained by resampling of the protein data. The line marks the RMSEP value obtained when the original calibration data is used.

provided by the method. For some data sets, EMS is able to identify narrow spectral intervals that explain most of the variation of the response data. Furthermore, by using several reference data sets, it is found that EMS has as good or better prediction performance than PLS in terms of the mean squared prediction errors. Another interesting property of EMS is that it changes smoothly from a full spectrum method to a variable selection method depending on the value of ω. Moreover, the method is scale independent, which makes the data pretreatment easier. In comparison to iPLS, where explicit intervals of the spectrum are tested, EMS does not depend on any prior assumptions about neighboring wavelengths. It is found that the selection of intervals makes iPLS less robust to variation in data compared to EMS. To facilitate other researchers using EMS, the source code has been made available on the Internet. In future work, a penalty on the squared length of the parameters could be included, and thereby a method that smoothly changes from variable selection to ridge regression is obtained. The combination of the ability to identify important spectral regions, as seen for the EMS method, with a penalty on the squared length, as in ridge regression, might yield a method with even better prediction performance.

## ACKNOWLEDGMENTS

The authors thank L. Nørgaard at the Royal Veterinary and Agricultural University, Copenhagen, Denmark, and Carlsberg A/S, Copenha-

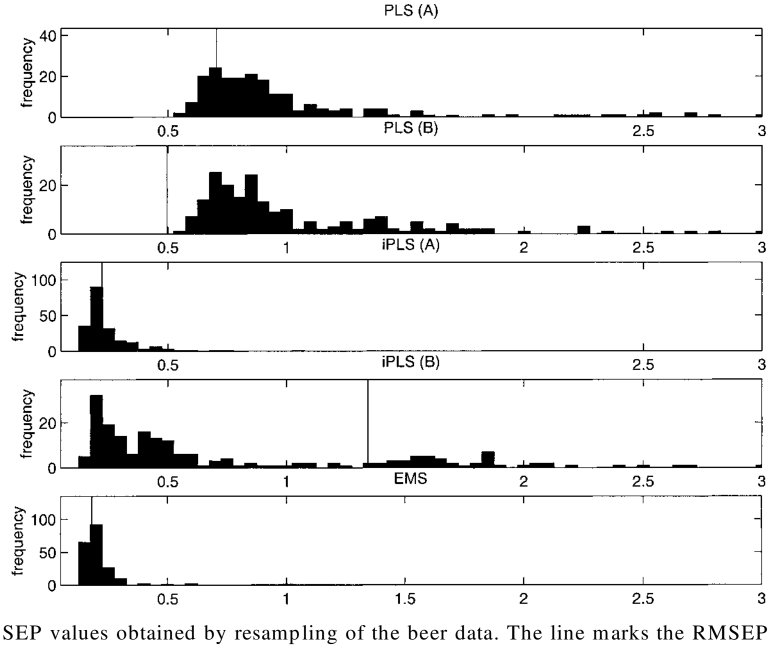

FIG. 10. Histograms of the RMSEP values obtained by resampling of the beer data. The line marks the RMSEP value obtained when the original calibration data is used.

gen, Denmark, for permission to use the beer data set. The work was done in the Centre for Water Quality Sensors, which is supported by the Danish Agency for Development of Trade and Industry. In addition, we are grateful to the reviewer for helpful comments.

1. H. Martens and T. Naes, *Multivariate Calibration* (John Wiley and Sons, New York, 1989).
2. I. E. Frank and J. H. Friedman, Technometrics **35,** 109 (1993).
3. H. Öjelund, H. Madsen, and P. Thyregod, J. Chemom. **15,** 497 (2001).
4. C. B. Lucasius, M. L. M. Beckers, and G. Kateman, Anal. Chim. Acta **286,** 135 (1994).
5. U. Hörchner and J. H. Kalivas, Anal. Chim. Acta **311,** 1 (1995).
6. A. J. Miller, *Subset Selection in Regression* (Chapman and Hall, Ltd., London, 1990).
7. V. Centner, J. Verdú-Andrés, B. Walczak, D. Jouan-Rimbaud, F. Despagne, L. Pasti, R. Poppi, D. Massart, and O. E. de Noord, Appl. Spectrosc. **54,** 608 (2000).
8. A. S. Bangalore, R. E. Shaffer, and G. W. Small, Anal. Chem. **68,** 4200 (1996).
9. D. J. Rimbaud, B. Walczak, D. L. Massart, I. R. Last, and K. A. Prebble, Anal. Chim. Acta **304,** 185 (1995).
10. F. Navaroo-Villoslada, L. V. Pérez-Arribas, M. E. León-González, and L. M. Polo-Díez, Anal. Chim. Acta **313,** 93 (1995).
11. L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, Appl. Spectrosc. **54,** 413 (2000).
12. H. Öjelund, P. J. Brown, H. Madsen, and P. Thyregod, Technometrics, paper submitted (2001).
13. A. E. Beaton, Research Bulletin RB-64-51 (Educational Testing Service, Princeton, New Jersey, 1964).
14. M. Schatzoff, R. Tsao, and S. Fienberg, Technometrics **10,** 769 (1968).
15. L. Breiman and P. Spector, International Statistical Review **60,** 291 (1992).
16. J. H. Kalivas, Chemom. Intell. Lab. Syst. **37,** 255 (1997).
17. J. M. Brenchley, U. Hörchner, and J. H. Kalivas, Appl. Spectrosc. **51,** 689 (1997).