Calibration with near-continuous spectral measurements

Henrik Aalborg Nielsen, Michael Rasmussen, and Henrik Madsen
Department of Mathematical Modelling
Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark

Abstract

In chemometrics traditional calibration in case of spectral measurements express a quantity of interest (e.g. a concentration) as a linear combination of the spectral measurements at a number of wavelengths. Often the spectral measurements are performed at a large number of wavelengths and in this case the number of coefficients in the linear combination is magnitudes larger than the number of observations. Traditional approaches to handling this problem includes principal components, partial least squares, ridge regression, LASSO, and other shrinkage methods. As a continuous-wavelength alternative we suggest replacing the linear combination by an integral over the range of the wavelength of a unknown coefficient-function multiplied by the spectral measurements. We then approximate the unknown function by a linear combination of some basis functions, e.g. B-splines. The method is illustrated by an example in which the octane number of gasoline is related to near infrared spectral measurements. The performance is found to be much better that for the traditional calibration methods.

1 Introduction

We consider the problem where a quantity characterizing a liquid, eg. the concentration of nitrate in waste water, is to be predicted from a spectrum measured on the liquid. To be able to perform such predictions we must first obtain reliable measurements of the quantity for a number of (well chosen) liquids, measure the spectrum for each liquid, and relate the measured quantities to the corresponding measured spectra. This process is known under the term *multivariate calibration*.

In chemometrics traditional multivariate calibration in case of spectral measurements express a quantity of interest (e.g. a concentration) as a linear combination of the spectral measurements at a number of wavelengths. Often the spectral measurements are performed at a large number of wavelengths and in this case the number of coefficients in the linear combination is magnitudes larger than the number of observations.

Traditional approaches to handling this problem includes principal components, partial least squares, ridge regression, LASSO, and other shrinkage methods. Variable selection methods have also been applied (Brown 1993, Osborne, Presnell & Turlach 2000). As a continuous-wavelength alternative the linear combination of the spectral values can be replaced with an integral over the range of the wavelengths of an unknown coefficient-function multiplied by the spectral measurements. The unknown function can then be approximated by a linear combination of some basis functions (e.g. B-splines). The problem then becomes a linear regression problem where the number of regressors depend on the number of basis functions and not the number of wavelengths.

To our knowledge the approach was first suggested by Hastie & Mallows (1993) who focused on smoothing splines for estimation of the coefficient-function. Similarly Goutis (1998) used smoothing splines to estimate a coefficient-function in the case where the predictive information is related to the second derivative of the spectrum. Marx & Eilers (1999) project the spectral measurements onto a moderate number of equally spaced B-spline bases. This approach is very similar to the approach presented here. However, the difference being that (i) we formulate the underlying model using an integral over the wavelengths, and (ii) we do not restrict the number of basis functions to be less than the number of observations. For near-continuous measurements (i) is largely a technicality which allow us, in a simple way, to study what happens if the predictive ability is related to derivatives of the spectra rather than the actual spectra.

In Section 2 the underlying model is described. Section 3 describes the approximations used. An application is presented in Section 4. Finally, in Section 5 some conclusions and discussions are listed, together with a short description of available computer programs which we are aware of. Figure 3 and 4 referred in Section 4 are placed after the list of references.

2 Model

The spectra are measured at a number of wavelengths $\lambda_j;\ j = 1, \ldots, m$. The measurements of the characteristic quantity is called $y_i;\ i = 1, \ldots, N$ and the measured spectrum corresponding to y_i is called $a_i(\lambda_j);\ j = 1, \ldots, m$. The traditional approaches to calibration focus on the model

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j a_i(\lambda_j) + e_i;\ i = 1, \ldots, N \tag{1}$$

where $e_i;\ i = 1, \ldots, N$ are the model errors which are assumed to be independently identical distributed (iid.) random variables, and $\beta_j;\ j = 0, \ldots, m$ are some coefficients which must be determined from data. Model (1) is a linear regression model. However,

as measurement equipment get more advanced the spectra are measured at an increasing number m of wavelengths, so that each spectrum often can be considered known for every wavelength $\lambda \in [\underline{\lambda}, \overline{\lambda}]$. Therefore, the number of regressors m is often magnitudes larger than the number of observations N. Traditional approaches to handling this problem are mentioned in Section 1. We argue that, conceptually, it is more convenient to use a model which explicitly regard the spectra as functions $a_i(\lambda);\ i = 1, \ldots, N$ of the bandwidth λ. As a generalization of (1) it is convenient to replace the summation with an integral over the interval of wavelengths, i.e. to use the model

$$y_i = \beta_0 + \int_{\underline{\lambda}}^{\overline{\lambda}} \beta(\lambda) a_i(\lambda) d\lambda + e_i, \tag{2}$$

where the coefficient β_0 and the *function* $\beta(\cdot)$ must be determined from data, c.f. Section 3. It is interesting to note that if it is suspected that some predictive ability is related to the first- and second-order derivatives of the spectra rather than the spectra itself (2) can still be used if the range of wavelengths over which the spectra is measured is wide enough.

To see this consider the model

$$y_i = \beta_0 + \int_{\underline{\lambda}}^{\overline{\lambda}} \left(\phi_0(\lambda) a_i(\lambda) + \phi_1(\lambda) \frac{da_i}{d\lambda}(\lambda) + \phi_2(\lambda) \frac{d^2 a_i}{d\lambda^2}(\lambda) \right) d\lambda + e_i, \tag{3}$$

which take into account both the actual spectra and its first- and second order derivatives. Assuming that the derivatives exists, simple calculations (partial integration) show that (3) can be written

$$\begin{aligned} y_i &= \beta_0 \\ &+ \left(\phi_1(\overline{\lambda}) - \frac{d\phi_2}{d\lambda}(\overline{\lambda}) \right) a_i(\overline{\lambda}) - \left(\phi_1(\underline{\lambda}) - \frac{d\phi_2}{d\lambda}(\underline{\lambda}) \right) a_i(\underline{\lambda}) + \phi_2(\overline{\lambda}) \frac{da_i}{d\lambda}(\overline{\lambda}) - \phi_2(\underline{\lambda}) \frac{da_i}{d\lambda}(\underline{\lambda}) \\ &+ \int_{\underline{\lambda}}^{\overline{\lambda}} \left(\phi_0(\lambda) - \frac{d\phi_1}{d\lambda}(\lambda) + \frac{d^2\phi_2}{d\lambda^2}(\lambda) \right) a_i(\lambda) d\lambda + e_i. \end{aligned} \tag{4}$$

Given that the range of the wavelengths is so large that all important wavelengths are covered then $\phi_1(\overline{\lambda}) = \phi_1(\underline{\lambda}) = \phi_2(\overline{\lambda}) = \phi_2(\underline{\lambda}) = 0$. In this case the second line in (4) vanish and the term inside the parenthesis in the integral is a function of λ which can be handled by $\beta(\lambda)$ in (2). If not all important wavelengths are covered it is necessary to extent (2) with regression terms containing $a_i(\overline{\lambda})$, $a_i(\underline{\lambda})$, $\frac{da_i}{d\lambda}(\overline{\lambda})$, and $\frac{da_i}{d\lambda}(\underline{\lambda})$ in order to take first- and second-order derivatives of $a_i(\lambda)$ into account.

3 Approximations

To be able to determine the scalar β_0 and the function $\beta(\cdot)$ in (2) from data we approximate the function by a linear combination of a set of basis functions, such as B-spline

basis functions, natural spline basis functions, or wavelet basis functions (de Boor 1978, Bruce & Gao 1996).

$$\beta(\lambda) = \mathbf{B}'(\lambda)\boldsymbol{\theta}, \tag{5}$$

where $\mathbf{B}(\lambda) = [b_1(\lambda) \dots b_p(\lambda)]'$ are the basis functions and $\boldsymbol{\theta} = [\theta_1 \dots \theta_p]'$ are some coefficients to be determined from data. With (2) and (5) simple calculations show that

$$y_i = \beta_0 + \sum_{k=1}^{p} \theta_k x_{ki} + e_i, \tag{6}$$

where

$$x_{ki} = \int_{\underline{\lambda}}^{\overline{\lambda}} b_k(\lambda) a_i(\lambda) d\lambda;\ k = 1, \dots, p;\ i = 1, \dots, N, \tag{7}$$

does not depend on $\boldsymbol{\theta}$ and can be determined from the measurements of the spectra at the wavelengths $\lambda_j;\ j = 1, \dots, m$ by use of the trapezoid rule of integration.

$$x_{ki} = \frac{1}{2} \sum_{j=1}^{m-1} (\lambda_{j+1} - \lambda_j) \left[b_k(\lambda_j) a_i(\lambda_j) + b_k(\lambda_{j+1}) a_i(\lambda_{j+1}) \right] \tag{8}$$

It is seen that, handled this way, the calibration problem is not dependent on m as long as the spectra is measured at fine enough intervals to allow the integrals in (7) to be evaluated with reasonable precision. Furthermore, although $m > N$, the number of basis functions can often be chosen so that $p < N$, whereby (6) becomes an ordinary regression problem. We may choose to use $N < p < m$, in this case principal components, partial least squares, ridge regression, LASSO, and other shrinkage methods may be applied.

As noted by Marx & Eilers (1999) the application of models like (6) regularize estimation as compared to models like (1). However, we argue that, depending on the spectra, the regressors in (6) may still be very collinear. Figure 1 shows a cubic B-spline basis with six equally spaced knots covering the interval 900 to 1700 nm, this results in $p = 8$. It is seen that the basis-functions are non-zero only for wavelengths around their maximum, this is the key feature by which (5) becomes a good approximation. However, if $a_i(\lambda)$ is constant across i for some wavelengths then the nature of the basis-functions may result in collinearity of the regressors (7). It is therefore suggested that instead of using model (6) directly the regressors are replaced by their principal components. If variable selection techniques are then applied to the principal components both problems where $p < N$ and $p \geq N$ can be handled.

The type of basis functions used influence the type of functions which can be approximated by (5). A B-spline basis of order n result in $\beta(\cdot)$ having continuous derivatives up to order n, i.e. a cubic B-spline basis is of order 2. This also hold for a natural spline basis, but here $\beta(\cdot)$ has the additional property that it is linear outside $[\underline{\lambda}, \overline{\lambda}]$. Opposed to this a wavelet basis can be used to approximate a function with sharp peaks.

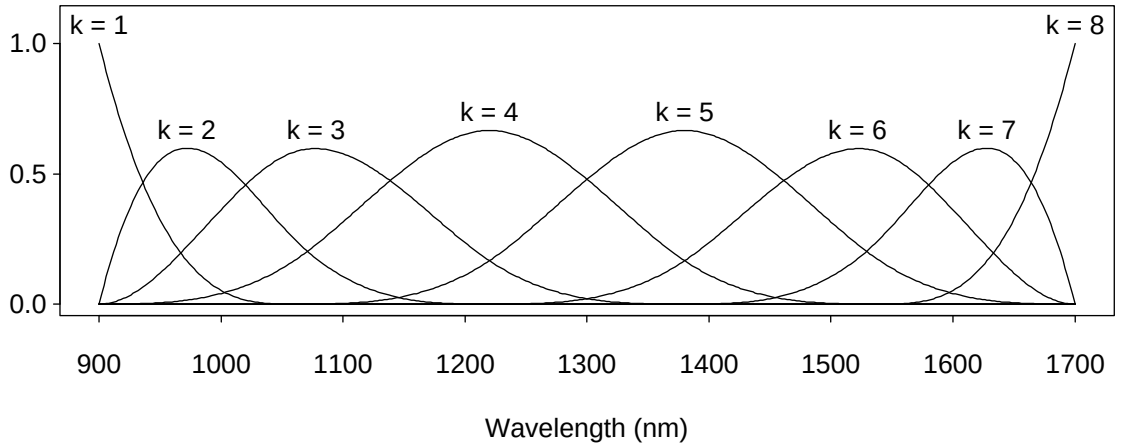

Figure 1: Cubic B-spline basis with six equally spaced knots (four internal) covering the interval 900 to 1700 nm.

4 Application to NIR spectra of gasoline

The method has been applied to predict the octane number based on near infrared (NIR) spectral measurements. The data set contains 60 gasoline samples with specified octane numbers. Samples were measured using diffuse reflectance (R) as $\log(1/R)$ from 900 to 1700 nm in 2 nm intervals. So we have $N = 60$ and $m = 401$. The data set was divided into five parts for use in a 5-fold cross-validation. To achieve this the octane numbers were sorted in ascending order and numbered successively from 1 to 5 in order to get five sets that cover approximately the same range. We follow Brown (1993, p. 42) and center all spectra, that is $\sum_i a_i(\lambda_j) = 0$; $j = 1, \ldots, m$. The octane numbers are also centered. The mean values of the calibration data is used to center the validation data.

For comparison reasons we have applied the well known methods, PCR, PLS, Ridge and LASSO, to the setup described above, i.e. using model (1) without β_0. The optimal model for the methods is chosen by minimizing the root mean squared error of prediction, (RMSEP), based on the 5-fold cross-validation. These values are summarized in Table 1. PCR, PLS and Ridge produce the best results. Finally, Figure 3 show the parameter estimates plotted against their corresponding wavelengths. The estimates are obtained using the tuning-parameters listed in Table 1 together with the full data set.

Method	Regularization parameter	RMSEP
PCR	no. of components = 13	0.2333
PLS	no. of components = 7	0.2327
Ridge	k = 0.002	0.2357
LASSO	$\sum(\lvert\beta\rvert)$=236.5	0.2779

Table 1: RMSEP-values for the regularization methods.

We now use the model defined by (6) and (8), with $b_1(\lambda), \ldots, b_k(\lambda)$ generated using a cubic B-spline basis with knots placed equidistantly over the range of wavelengths. If

we restrict the number of basis-functions to be less than the number of observations, N, we have a standard linear regression problem which can be solved using ordinary least squares. If we allow the number of basis-functions to exceed the number of observations we can apply PCR, PLS, Ridge or LASSO.

The same setup as mentioned earlier is used to find the best model. The approach is straightforward, find the RMSEP-values for a fixed regularization parameter and varying number of basis-functions, now fix the regularization parameter to another value and find new RMSEP-values. This produces a matrix of RMSEP-values; find the smallest RMSEP-value and the corresponding value for the regularization parameter and the number of basis-functions. For LASSO the optimum is $\sum(|\theta|)$=17.15 and 33 internal knots; Figure 2 indicates the curvature of the RMSEP-surface around the optimum.

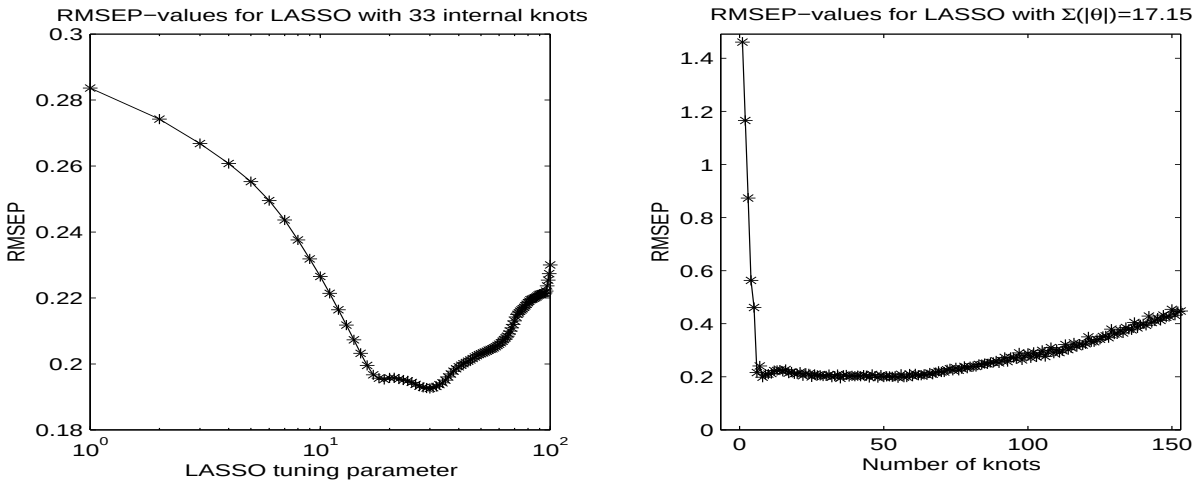

Figure 2: The RMSEP-values for LASSO for fixed number of internal knots (left) and fixed value for $\sum(|\theta|)$ (right).

The RMSEP-values are listed in Table 2. It is seen that LASSO performs best and that all methods are superior to the methods listed in Table 1. Comparing the best results between Table 1 and 2 results in a 17% reduction in RMSEP-values when using spline functions. The simple OLS-solution results in a 12% reduction.

Method	Regularization parameter	No. of internal knots	RMSEP
OLS		4	0.2053
PCR	no. of components = 7	4	0.2001
PLS	no. of components = 7	4	0.2012
Ridge	k=0.0008	4	0.2002
LASSO	$\sum(\|\theta\|)$=17.15	33	0.1926

Table 2: RMSEP-values for the regularization methods combined with the spline-method.

Figure 4 show the estimates of $\beta(\lambda) = \mathbf{B}'(\lambda)\boldsymbol{\theta}$. The estimates are obtained using the tuning-parameters listed in Table 2 together with the full data set. For the OLS-solution curves indicating two times the pointwise standard error are also shown (obtained by

disregarding that the number of internal knots are selected by use of cross-validation). For all but LASSO the estimates are quite similar. Comparing the standard error bands of the OLS-solution with the LASSO-solution reveals that LASSO selects basis-functions corresponding to wavelengths for which the OLS-solution is significantly different from zero.

5 Conclusion and discussion

An approach to multivariate calibration in which the model is formulated using an integral of an unknown coefficient-function multiplied with the measured spectrum is presented. The unknown function is approximated as a linear combination of some basis functions, whereby estimation is made feasible. A key feature of the method is that the dimension of the resulting model is not dependent on the number of wavelengths at which the spectral measurements are performed.

When the number of basis functions is low, standard linear regression techniques can be applied. However, some of the regressors may be collinear when the number of basis functions increase. In this case standard shrinkage methods used in multivariate calibration can be applied. This also opens the possibility of using more basis functions than the number of observations.

In an example with 60 near infrared spectra of gasoline, which are used to predict octane numbers, cubic B-spline bases with knots placed equidistantly are used. Compared to standard methods the spline-based methods performs 12% to 17% better in terms of the root mean square of five-fold cross-validated prediction errors (RMSEP). Even the simple linear regression model obtained when using eight basis functions results in a 12% reduction in RMSEP compared to the standard methods.

When the number of basis functions are low the knot placement may have large influence; it may move the valleys and peaks (Hastie & Tibshirani 1990, pp. 251-254). To avoid this the smoothing splines solution used by Hastie & Mallows (1993) and Goutis (1998) may be applied. The P-spline approach by Marx & Eilers (1999) provides a mix between these two approaches. All these approaches result in estimates of the coefficient-function which have approximately the same degree of smoothness for all wavelengths for which the spectral measurements are performed. There is no reason to believe that this is desirable.

The B-spline/LASSO approach is one solution to the problem just outlined. Another solution would be to use wavelet basis functions together with LASSO. Since wavelets cover a large range of scales and positions, they may be more appropriate than B-splines.

As yet another solution an adaptive knot-placement procedure could be applied together with standard linear regression. It is however not clear how to construct such a procedure.

For people using the traditional multivariate calibration techniques the main problem of applying the techniques presented here is the generation of spline bases. In S-PLUS (www.splus.mathsoft.com) and R (www.r-project.org) these can be generated with the build-in functions `bs` (B-splines) or `ns` (natural splines). In Matlab (www.mathworks.com) we use `bsplval.m` by Dr. Graeme A. Chandler, Mathematics Department, The University of Queensland, Australia. A ZIP-archive containing this function can be downloaded as www.maths.uq.edu.au/∼gac/mn309/mfilez.zip and in www.maths.uq.oz.au/∼gac/mn309/bspl.html examples of how to apply it can be found.

References

Brown, P. J. (1993), *Measurement, Regression, and Calibration*, Clarendon Press, Oxford. (ISBN 0198522452).

Bruce, A. & Gao, H.-Y. (1996), *Applied Wavelet Analysis With S-Plus*, Springer-Verlag, Berlin/New York.

de Boor, C. (1978), *A Practical Guide to Splines*, Springer Verlag, Berlin.

Goutis, C. (1998), 'Second-derivative functional regression with applications to near infra-red spectroscopy', *Journal of the Royal Statistical Society, Series B, Methodological* **60**, 103–114.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall, London/New York.

Hastie, T. & Mallows, C. (1993), 'Comment on "A statistical view of some chemometrics regression tools"', *Technometrics* **35**, 140–143.

Marx, B. D. & Eilers, P. H. C. (1999), 'Generalized linear regression on sampled signals and curves: A P-spline approach', *Technometrics* **41**(1), 1–13.

Osborne, M. R., Presnell, B. & Turlach, B. A. (2000), 'On the LASSO and its dual', *Journal of Computational and Graphical Statistics* **9**(2), 319–337.

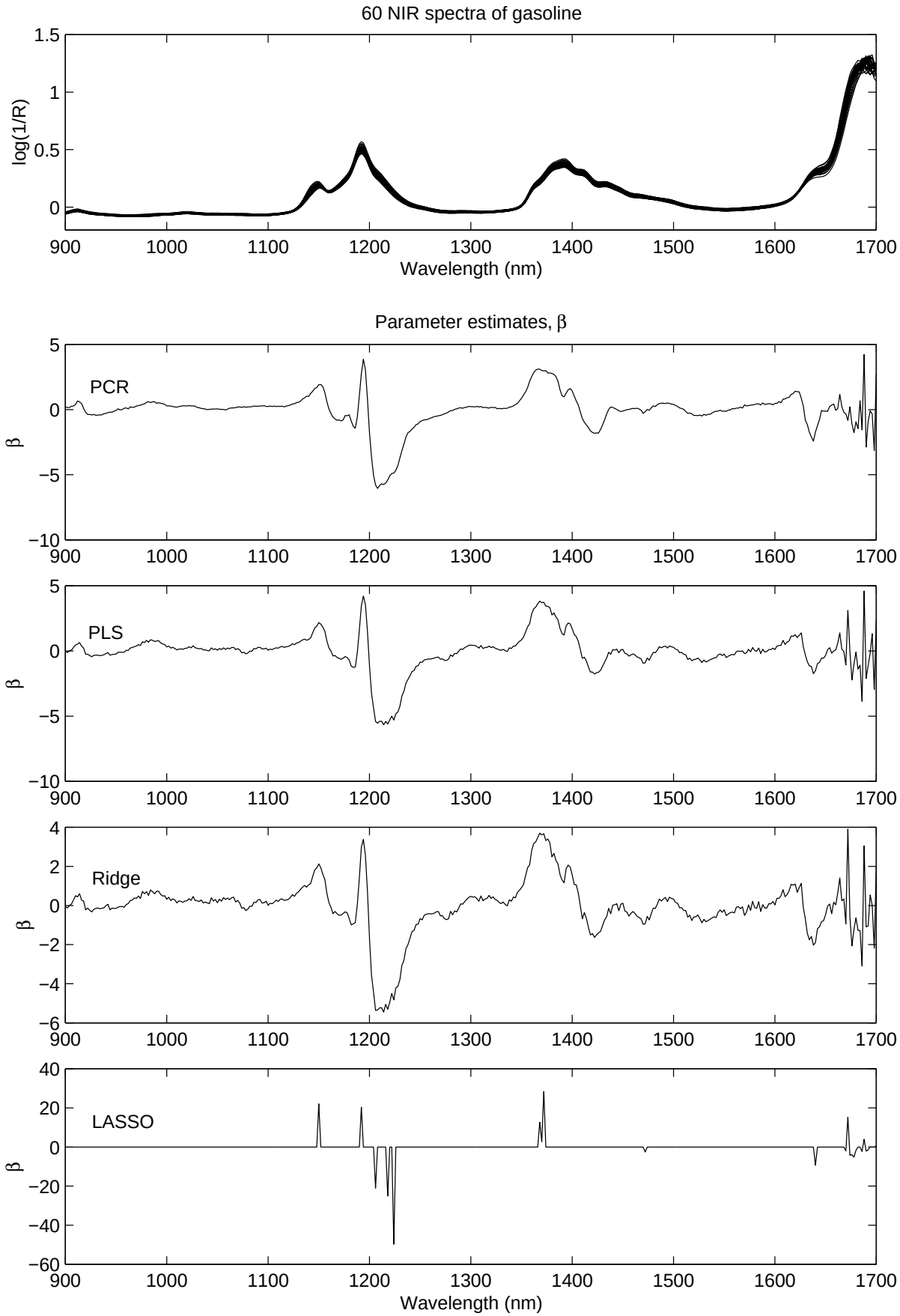

Figure 3: The 60 NIR spectra, together with the parameter estimates for PCR, PLS, Ridge and LASSO.

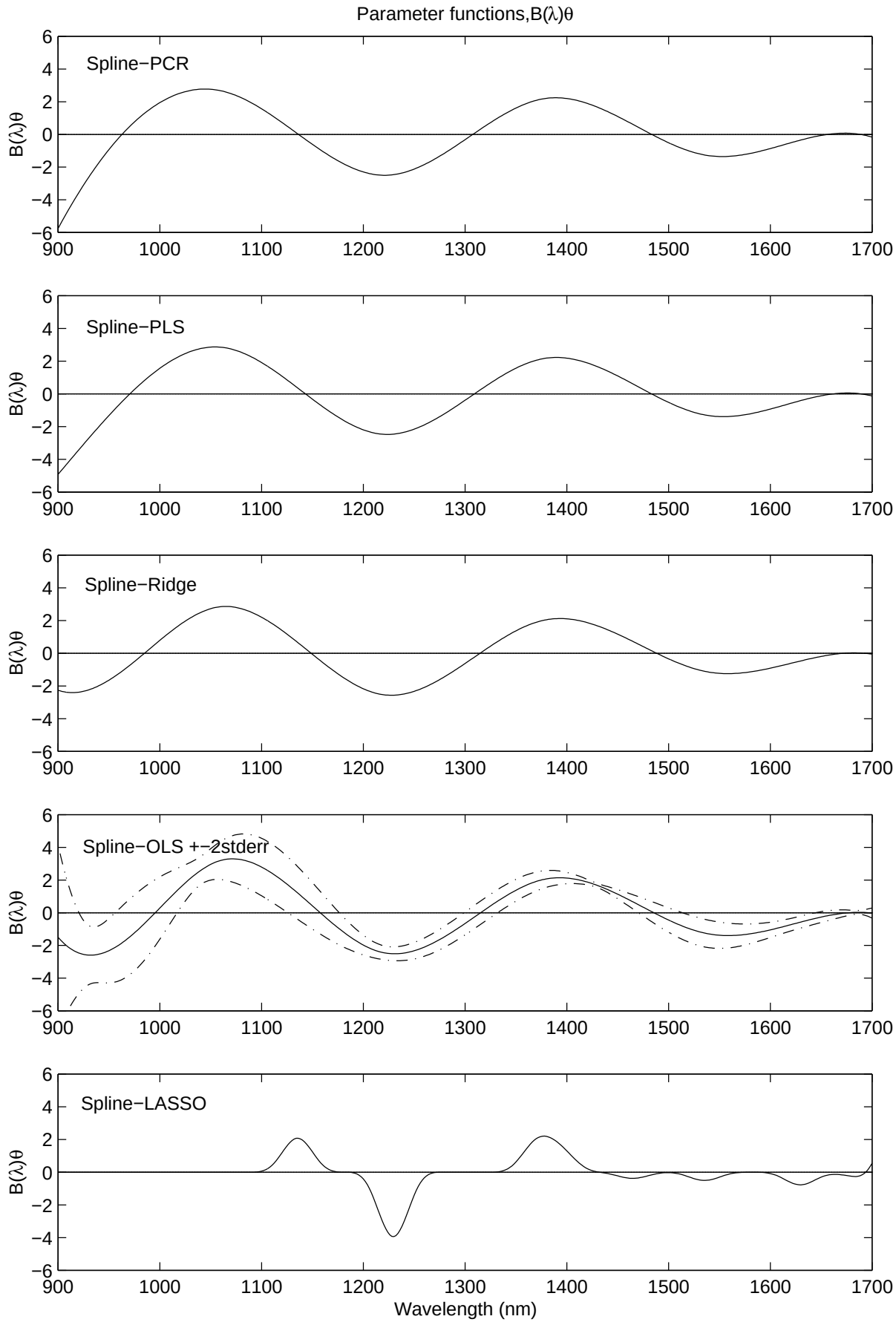

Figure 4: Estimated coefficient-functions using PCR, PLS, Ridge, OLS, and LASSO together with spline bases.