Elsevier Editorial System(tm) for Journal of Food Engineering
Manuscript Draft

Manuscript Number: JFOODENG-D-08-00255R1

Title: A comparison of variate pre-selection methods for use in partial least squares regression: a case study on NIR spectroscopy applied to monitoring beer fermentation

Article Type: Research Article

Section/Category:

Keywords: NIR spectroscopy; brewing; PLS regression; variate selection; genetic algorithm

Corresponding Author: Dr. Henri Tapp,

Corresponding Author's Institution: Institute of Food Research

First Author: Georgina McLeod

Order of Authors: Georgina McLeod; Kirsty Clelland; Henri S Tapp; E K Kemsley; Reginald H Wilson; Graham Poulter; David Coombs; Christopher J Hewitt

Manuscript Region of Origin:

Abstract: This work investigates four methods of selecting variates from near-infrared (NIR) spectra for use in partial least squares (PLS) regression models to predict biomass and chemical changes during beer fermentation. The fermentation parameters studied were ethanol concentration, specific gravity (SG), optical density (OD) and dry cell weight (DCW). The four selection methods investigated were: Simple, where a fingerprint region is chosen manually; CovProc, a covariance procedure where variates are introduced based on the magnitude of the 1st PLS vector coefficients; CovProc-SavGo, a modification to CovProc where the window size of a Savitzky-Golay filter applied to the spectra is also optimised; and Genetic Algorithm (GA), where variates are selected based on the frequency of appearance in 8-variate multiple linear regression models found from repeated execution of the GA routine. The analysis found that all four methods produced good predictive models. The GA approach produced the lowest standard error in prediction (SEP) based on leave-one-out cross validation (LOO-CV), although this advantage was not

reflected in the standard error in validation values, SEV, where all four models performed comparably. From this work, we would recommend using the Simple approach if a suitable fingerprint region can be identified, and using CovProc otherwise.

Institute of Food Research

Institute of Food Research
Norwich Research Park
Colney
Norwich
NR4 7UA
UK

28 May 2008

Milan Houska
Editor
Journal of Food Engineering

Revised Manuscript: JFOODENG-D-08-00255

Dear Dr Houska
Please find attached revisions to the following manuscript, 'A comparison of variate pre-selection methods for use in partial least squares regression: a case study on NIR spectroscopy applied to monitoring beer fermentation'.

With this covering letter are the following new or modified files

1. reply_to_ref_comments.doc – our reply to the referees' comments
2. McLeod_JFoodEng_v2.doc – revised manuscript
3. figure1_RawSpectra.tif – modified figure 1
4. figure2_CovProc.tif – modified figure 2
5. figure3_CovProcSG.tif – modified figure 3
6. figure4_ga.tif – modified figure 4
7. Tables_v2.doc – modified document with duplicated figure captions removed

Yours sincerely

Dr Henri Tapp

28 May 2008

Authors' response

We would first like to thank the referees for their helpful comments. Our response is given below in italics.

Reviewers' comments:

Reviewer #1: This is a well-written paper containing interesting results which merit publication. However, it is necessary to clarify a number of points. These are given below.

1. Overall
Authors concluded that where a 'fingerprint' region can be identified, then it advisable to use the Simple method. In general, the Sample method is not better than the other procedures because where a fingerprint region is chosen manually. In terms of utility, the CovProc-SavGo and GA methods are considerably more laborious to implement than the Simple method. However, these performances were more insufficient than the expectation. Was the GA modeling appropriate?

We contend that the GA variate selection method was worthy of consideration. This approach uses the least a priori information during the selection phase. It was interesting to explore how this method compared with the more standard fingerprint region approach.

2. Page 3, line 24 - Page 4, line 9
In order to develop the on-line monitoring system there are too many problems. The sample pre-treatment might not be avoided.

We agree with referee's comments: there are some formidable technical challenges that would need to be addressed before a realistic on-line NIR sensor can be realised. Our manuscript clearly states that we are here only considering the performance of a near-line laboratory system.

3. 2.3 Chemometric analysis
This section contains the results (p.7, line 9-20; p.8, line 21-26).

P. 7 line 9… We have reworded to clearly describe methods rather than results. However we find nothing wrong on page 8: it describes the methodology associated with the implementation of the genetic algorithm rather than discussing any results, and therefore we have left this alone.

4. Figures
y-axis label?

We have made several changed to the figure & legends to improve clarity.

Reviewer #2: The authors in the manuscript had made an attempt to predict the fermentation parameters from information obtained using NIR spectroscopy. The four different modeling methods have been compared. The modeling methods involved multiple steps. From the results, the developed models appears to be predicting very well. It am eager to see the prediction capabilities of the developed models on a different system with a wide range of operating conditions in future.
I have few editorial comments.

Page 4, line 17, line 21 - both "h" and "hours" are used for hours. please use only one for consistency.

We have changed to h throughout

page 5, line 16 - "rpm" should be in place of "pm"

We have corrected this typo

page 10, line 22 - "weigh" should be in place of "weight"

We have reviewed the sentence. Here we do mean weight rather than weigh

Page 4, line 21 - what is vvm?

This stands for ‘ volume per volume per minute’. This is a standard abbreviation for this control parameter. We will leave it to Editorial discretion on whether this requires further expansion.

Marking in the figures are not clear. The markings could be lebeled for the clarification.
Figures numbers are not marked above the figures. There are four small figures in each figure. I would recommend to name the four figures as (a), (b), (c) and (d) to help explain these in the text.

We have made several changed to the figure & legends to improve clarity.

A comparison of variate pre-selection methods for use in partial least squares regression: a case study on NIR spectroscopy applied to monitoring beer fermentation.

Georgina M[c]Leod[1], Kirsty Clelland[1], Henri Tapp[2]*, E. Katherine Kemsley[2], Reginald H. Wilson[2], Graham Poulter[3], David Coombs[3] and Christopher J. Hewitt[4]

[1] Centre for Formulation Engineering, Biochemical Engineering, School of Engineering (Chemical Engineering), University of Birmingham, Edgbaston, B15 2TT, UK.

[2] Institute of Food Research, Norwich Research Park, Norwich NR4 7UA, UK.

[3] Specac Limited, River House, 97 Cray Avenue, Orpington, Kent, BR5 4HE, UK.

[4] Interdisciplinary Centre for Biological Engineering (ICBE), Department of Chemical Engineering, Loughborough University, Loughborough, Leicestershire LE11 3TU, UK

*Author for correspondence (Fax: +44 (0)1603 507723; E-mail: henri.tapp@bbsrc.ac.uk)

Keywords: NIR spectroscopy, brewing, PLS regression, variate selection, genetic algorithm

Abstract

This work investigates four methods of selecting variates from near-infrared (NIR) spectra for use in partial least squares (PLS) regression models to predict biomass and chemical changes during beer fermentation. The fermentation parameters studied were ethanol concentration, specific gravity (SG), optical density (OD) and dry cell weight (DCW). The four selection methods investigated were: Simple, where a fingerprint region is chosen manually; CovProc, a covariance procedure where variates are introduced based on the magnitude of the 1st PLS vector coefficients; CovProc-SavGo, a modification to CovProc where the window size of a Savitzky-Golay filter applied to the spectra is also optimised; and Genetic Algorithm (GA), where variates are selected based on the frequency of appearance in 8-variate multiple linear regression models found from repeated execution of the GA routine. The analysis found that all four methods produced good predictive models. The GA approach produced the lowest standard error in prediction (SEP) based on leave-

one-out cross validation (LOO-CV), although this advantage was not reflected in the standard error in validation values, SEV, where all four models performed comparably. From this work, we would recommend using the Simple approach if a suitable fingerprint region can be identified, and using CovProc otherwise.

1. Introduction

This paper describes an investigation into the use of NIR spectroscopy for monitoring beer fermentation. The data presented here are near-line measurements of the raw liquor. The near-line approach avoids some of the technical challenges that would need to be met by an on-line sensor, such as long-term stability and fouling, whilst allowing the potential of NIR for monitoring biomass (optical density (OD) and dry cell weight (DCW)) and composition (ethanol concentration and specific gravity (SG)) to be investigated. The aim is to establish the performance in terms of predicting biomass and composition that could ultimately be obtained from an on-line probe.

We present a comparison of four different methods of pre-selecting the variates for use in partial least squares (PLS) regression models for predicting the biomass and compositional parameters. The purpose of the analysis was to compare feature selection procedures that involved varying degrees of complexity, and evaluate the impact of this on their predictive ability.

1.1 Background

Measurements relating to biomass production, substrate consumption, and ethanol accumulation are not routinely carried out during the course of large-scale brewing fermentations. This is because traditional analytical techniques such as measurement of DCW and substrates by GC or HPLC are routinely performed off-line and require extensive sample preparation (Macaloney et al., 1996). Results are obtained too late for any meaningful changes to be made to the process (Hewitt and Nebe-Von-Caron, 2004), consequently brewers often control fermentations retrospectively, taking action only after the quality of the final product has not met expectations. This management could

be more proactive if the products of interest were more closely monitored and the smallest changes could be instantaneously detected.

Near infrared spectroscopy (NIR) is now recognised as a rapid (analysis within minutes) and non-destructive technique for the analysis of a range of substrates during pharmaceutical fermentation processes. A single spectral acquisition allows multiple component concentrations to be detected at a single point in time (Arnold et al., 2002a). The weakly absorbing nature of NIR allows high concentrations to be handled reproducibly with no sample preparation or dilution, making it a favourable technique for both immediate near-line and on-line analysis (Bakeev, 2003). Indeed NIR has been successfully used for the quantification of various substrates during bacterial fermentations (Arnold et al., 2002b) as well as mammalian and insect cell cultures (Arnold et al., 2002a; Harthun et al., 1998; Riley et al., 1997). However, these processes use defined well-mixed growth media, being both spatially and temporally homogeneous in nature, providing an environment that is relatively simple in spectroscopic terms and ideally suited to on-line NIR analysis. NIR has also been applied to both fungal and filamentous bacterial fermentations, where in contrast the mycelial broths are highly viscous, display non-newtonian behaviour and are chemically relatively undefined (Arnold et al., 2000; Vaidyanathan et al., 2003). In these cases, the complex and heterogeneous nature of the processes leads to probe fouling, and measurement at a single point is unrepresentative of the bulk; off-line NIR analysis of a number of samples taken from various positions within the fermentation vessel is more suitable. These more complex systems can also have a greater batch-to-batch variability, and therefore often require larger data sets (that is, more samples) to account for this variation (Vaidyanathan et al., 2003).

Similarly, there are three key problems associated with the successful NIR analysis of a typical brewing fermentation. First, no mechanical agitation is provided for large brewing vessels, although mechanical agitation is used in this study to improve mixing and reduce heterogeneity. Mixing relies solely on the evolution of carbon dioxide by the yeast cells, which is itself dependent on

metabolic activity. This creates spatial and temporal gradients with respect to cell and substrate concentration (Boswell et al., 2003), so analysis at a single point is not suitable. Second, pH is uncontrolled, and the natural decrease in pH that takes place during the brewing process leads to spectral base-line drift that needs to be taken into account for accurate models to be constructed. Third, brewing fermentations support relatively low cell concentrations (when compared to pharmaceutical processes) making detection and quantification potentially quite difficult by NIR (Arnold et al., 2000). In this work, we investigate the potential of NIR to be used as a quantitative tool to characterise two very different brewing processes, at laboratory scale (5L), for the production of two model beer types: Muntons pale ale and Grolsch lager.

2. Materials and methods

A brewing strain of *Saccharomyces cerevisiae* (NCYC 1324) obtained from the national collection of yeast cultures (Institute of Food Research, Norwich, UK) was maintained at 4°C and grown aerobically at 25°C on yeast extract malt extract agar (YM) prepared as per the manufacturers instructions (B.D. Ltd Oxford, UK). Inoculum for pitching was prepared by growing the culture aerobically at 25°C with 50ml YM broth (B.D. Ltd., Oxford, UK) per 250ml Erlenmeyer shake flask at 200 rpm on a rotary shaker for 13h. The pitching rate was 1.5×10^7 cells ml^{-1} wort. Two sets of five fermentations were carried out, the first with Munton's 'Hopped Light' pale ale wort (BRI, Surrey, UK) and the second with Grolsch lager wort (Coors brewery Ltd, Burton-upon-Trent, UK), both with an initial SG of 1.060. Prior to pitching, the wort was agitated at 200 rpm and aerated at 1 vvm for 2h until the dissolved oxygen tension (DOT) reached the 100% saturation level. After pitching no further agitation or aeration was carried out, except just before sampling, when the culture was gently roused (lightly agitated) to create a homogeneous distribution of cells and substrate within the bioreactor. The temperature was maintained at 12°C using an anti-freeze filled jacket coupled with a recirculating chiller unit (LTD, Grant Instruments, Cambridge, UK). Fermentations were operated in batch mode, for a duration of 168h, and samples were taken every 12h for measurement of biomass concentration, cell viability, ethanol concentration and SG.

All fermentations were carried out in a 5L nominal cylindrical glass bioreactor (162 mm diameter × 300 mm total height), with a working volume of 4L. The vessel was fitted with two 82 mm, six-bladed radial flow paddle type impellers, 80 mm apart, with the lower impeller situated 80 mm above the bottom of the vessel. The vessel was also fitted with three equally spaced baffles, width 15 mm and equipped for pH and DOT measurement, as well as temperature and impeller speed control. DOT and pH were uncontrolled and allowed to decrease naturally, from 100% to 0% and 5 to 3 respectively, as the fermentation progressed. Samples were analysed for biomass concentration, ethanol concentration and SG.

2.1 Conventional analyses

Cell biomass was measured turbidimetrically by OD at 550 nm in a double beam spectrophotometer and by measurement of DCW (g L^{-1}). For the latter, cell samples were washed (centrifugation at 4500 rpm followed by resuspension in distilled water) then separated using 0.45μm pore size cellulose nitrate filters (Sartorius AG, Goettingen, Germany). Filters were dried at 105°C to constant weight. The SG of the supernatant was measured to 5 decimal places using a 10ml density bottle (VWR International Ltd) and standard laboratory scales. The concentration of ethanol was determined by headspace gas chromatography (Perkin Elmer Autosystem XL with HS40 headspace unit, Perkin Elmer, Beaconsfield, UK) using a ZB-wax column (30 m × 0.32 mm, 0.25μm film thickness, Phenomenex, USA), the output from which was fed to a flame ionisation detector (250°C).

2.2 NIR spectroscopy

All spectral acquisition was conducted near-line, using a desktop FT-IR spectrometer (Perkin Elmer, Beaconfield Bucks, UK) supplied by Specac (Orpington, Kent). A halogen light source was used, which allowed spectra to be collected over the NIR region ranging from 10000 – 4000 cm^{-1}. NIR spectra were acquired from all samples in transmission mode, using an omni cell (Specac,

Kent), with calcium fluoride windows and a 0.4 mm pathlength. 32 scans were co-added before Fourier transformation, and spectral resolution was 4 cm^{-1}. These acquisition parameters achieved a data collection time of 2.1 minutes per spectrum. For the Grolsch samples only, an additional set of spectra was collected using a 1 mm cuvette, 128 co-added scans and 16 cm^{-1} resolution. These parameters resulted in a data collection time of 3 minutes per spectrum. For this measurement protocol, the fourfold increase in co-averaging and accompanying four-fold reduction in resolution was chosen to accommodate the decreased signal caused by the longer path-length without excessively increasing the overall acquisition time. Prior to analysis, samples were degassed using mild sonication and equilibrated to room temperature. All single-beam spectra were converted to absorbance using a distilled water background, and truncated at 4124 cm^{-1}. Data interpolation resulted in 3001 and 1501 variates for the 4 cm^{-1} and 16 cm^{-1} resolution spectra respectively. In the following discussion, spectra collected on Grolsch samples using the 0.4 mm and 1 mm resolution protocols will be referred to as 'short-path Grolsch' and 'long-path Grolsch' respectively.

2.3 Chemometric analysis

The data were analysed using Matlab (The Mathworks Inc., Cambridge, UK). Predictive models were built using partial least square regression, PLS-R, (Martens and Naes, 1989). Each reference parameter was modelled separately, an approach called PLS-1 regression. Each dataset was pre-treated with a Savitzky-Golay filter (Press et al., 1992). Four methods of selecting subsets of variates to pass to the PLS-R procedure were then compared. All three datasets (Muntons, short-path Grolsch and long-path Grolsch) comprised measurements from five fermentation batches: data from four of these were used to build the PLS-R models, and data from the fifth was used as an independent test set. The model building step involved performing leave-one-out cross validation (LOO-CV) on sub-models using increasing numbers of PLS factors, up to a maximum of 15 factors. The optimum number of factors was determined using a modified Amemiya's prediction criterion APC (Norušis and SPSS Inc., 1990), given in Equation 1, where n_s is the number of observations, n_p is the number of PLS factors, and Q the correlation between the actual reference values and the

cross-validated predictions. The number of PLS factors that minimised the APC was chosen in the final model.

$$APC = \frac{(n_s + n_p)}{(n_s - n_p)}\left(1 - Q^2\right) \quad (1)$$

Savitzky-Golay filters are used to smooth and/or derivatise spectra using a local polynomial fit controlled by three parameters [N P D], where N is the number of neighbouring points (total filter window of 2N + 1 points), P is the order of the polynomial, and D is the derivative number (0 smoothed, 1 first derivative, etc). A preliminary examination of the spectra indicated that the presence of bio-material introduces a sloping baseline to the spectra. The two measures of biomass were therefore predicted using smoothed spectra only (D=0). The remaining two reference measures were predicted from 1st derivative spectra (D=1, which essentially removes the sloping baseline); the aim is that the subsequent predictive models should target chemical rather than environmental effects. Although derivative spectra can be calculated using a simple difference approach rather than a SavGo filter, in our present work, this would not lead to a fair comparison between spectra collected at different resolutions (or equivalently, different numbers of co-additions, and hence levels of noise). A second order polynomial was used throughout (P=2). One of the feature selection methods (CovProc-SavGo, described below) optimised N; the remaining three used values for N chosen from inspection of the smoothed spectra: N=8 for Muntons and short-path Grolsch, N=3 for long-path Grolsch.

The four selection methods will be referred to as: Simple, CovProc, CovProc-SavGo and GA, and are detailed as follows:

Simple. This involved defining two regions from a combination of previous experience and inspection of the raw spectra: a 'biomass region' to predict OD and DCW, and a 'chemical region' to predict ethanol and SG. The chemical region was between 4700 and 4200 cm^{-1}. The biomass region for Muntons and short-path Grolsch was between 10000 and 5500 cm^{-1}, and for long-path

Grolsch between 6200 and 5600 cm^{-1}. The reduction in this range was to avoid artefacts seen in the 1st derivative spectra.

CovProc. This involved evaluating many PLS-R models, each using increasing numbers of variates (Höskuldsson, 2001; Reinikainen and Höskuldsson, 2003). For ethanol and SG, the variates were chosen in descending order of the magnitude of their associated 1st PLS vector coefficients. For OD and DCW, the variates were chosen in order of increasing wavelength (reducing frequency). This was to encourage selection of variates associated with the sloping baseline shown in Figure 1. The two procedures were essentially equivalent for the Muntons and short-path Grolsch datasets, as the larger 1st PLS vector coefficients were associated with the shorter wavelengths. The subset of variates associated with the model with the overall minimum APC was chosen as the final model.

CovProc-SavGo. This involved repeated applications of the CovProc procedure on Savitzky-Golay filtered spectra that used different values for N. The variate subset and N value associated with the model with the lowest APC was chosen as the final model.

GA. This involved three steps. First, a genetic algorithm GA (Goldberg, 1989; Kemsley et al., 2007; Mitchelle, 1998; Tapp et al 2003) was executed 1000 times. One execution of the GA is termed an *epoch*. The GA used the following settings: population size of 600; a fixed subset size of 8 variates; the top 50% retained for breeding; a mutation rate of 0.04; a fitness score based on the Q^2 from block validated multiple linear regression (MLR), using 7 randomly assigned validation blocks with new partitions in each epoch. Breeding likelihood was weighted strongly toward the most successful subsets. For each epoch, the termination criterion was either reaching 100 generations, or 30 generations without improvement. The best subset from each generation was retained in the following generation and each offspring within a generation had a unique subset of variates. Second, the 1000 × 8 variate identifiers were pooled and a histogram of the frequency of occurrence calculated. Starting with the most common variate, a peak picking procedure retained variates that had the highest occurrence within a 2 point neighbourhood (5 point window). Third, the best number of variates to be used was found by evaluating PLS models with increasing

numbers of variates, introduced in order of the peak-list. The subset of variates associated with the model with the overall minimum APC was chosen as the final model.

In summarising the performance of the variate subset selection methods, the standard error in prediction (SEP) and validation (SEV) were calculated as the root mean squared residuals from LOO-CV predictions and from predictions of the test set respectively. Similarly, Q and R are Pearson product moment correlation coefficients between actual reference values and LOO-CV predictions (Q), and test set predictions (R) respectively. The bias is the mean residual (predicted - actual reference values).

3. Results and discussion

Table 1 show the correlations between the four brewing parameters in the training sets for both Muntons and Grolsch experiments. As expected, the four parameters are highly correlated: OD and DCW should be correlated, since both are measures of biomass. The difference between original and present gravity is proportional to ethanol concentration, hence there should be a strong negative correlation between SG and ethanol. Increases in biomass should also be closely linked to increases in produced ethanol, and also to a reduction in SG due to the combined effect of increased lower-density ethanol and consumption of higher-density dissolved sugars. Table 1 also shows the mean and standard deviations in the parameter for the two experiments. The ethanol and SG values are similar in both experiments, and the biomass indicators slightly higher in the Grolsch experiment.

Figure 1 show the raw spectra over the range studied (10000 cm^{-1} to 4124 cm^{-1}) for the three analysed datasets. The biomass and chemical regions used in the Simple selection method are marked. Figure 1 also shows spectra collected at the longer path (1 mm path length, 16 cm^{-1} resolution, 128 co-added scans) on Grolsch supernatant. The presence of bio-matter causes an environmental effect on the NIR spectra, resulting in offsets and sloping baselines. This can be attributed to scattering. In contrast, ethanol and SG are related to specific chemical absorption peaks

(e.g. ethanol and sugars to the region 4200 ~ 4500cm^{-1} (C-H combination bands, and O-H stretch overtone); ethanol and water to the region 5100 ~ 5200 cm^{-1} (combination band of O-H stretch and deformation)). The relationships are both direct and indirect: for instance, higher sugar levels mean less water, and subsequently less absorption at the water peaks. A consequence of the strong water absorption at around 5300 cm^{-1} is very low signal levels in both sample and background spectra. Here, small differences are amplified during conversion to absorption units, making the data in this region appear somewhat unstable. However, given the strong correlations between the four parameters, and the indirect link with relative water absorption, this region may provide some useful information despite the relatively high noise levels. NIR tend to have smooth broad spectral features, which may be captured at lower spectral resolutions. An aim of the comparison between different resolution and path length protocols with similar acquisition times was to determine if the higher-resolution protocol would lead to any advantage in the predictive modelling.

Figures 2 to 4 show the variates selected by CovProc, CovProc-SavGo and GA methods respectively, superposed upon mean filtered spectra from the three datasets, offset for clarity. The SavGo filter settings and corresponding number of variates used in each model are given in Table 2. As a general observation, it can be seen that there is marked variation in the number of selected variates, both between the same parameter from different datasets and between related parameters (OD and DCW, ethanol and SG).

Figure 2 shows the CovProc results. Note first that here, as in all the selection methods, the data was not explicitly variance scaled. This was to weight preferentially the larger spectral features in an attempt to minimise the potential of incorporating coincidentally favourable noise into the model. Although this deemphasised the biomass influence, Figure 2 shows that applying CovProc to the derivatised spectra resulted in selection of the strong water absorption band around 5300 cm^{-1}, rather than, as might have been expected, the 4200 ~ 4500cm^{-1} region. This water region would normally be avoided; its selection is probably due to a combination of large noisy absorbance

values and an indirect link with the parameters as described above. In terms of utility, CovProc was considerably more laborious to implement than the Simple method, as it involved evaluating many more models. Because of the considerable computation time, CovProc was implemented by first introducing variates in coarse increment (around 25 variates) until all were included, and then successively refining both the variate range and incremental step size, based on plots of the variation in the minimum APC with number of variates used.

Figure 3 shows the results of the CovProc-SavGo selections. Compared to the CovProc results shown in Figure 2, there is some degree of agreement, in for example the variates selected for OD and DCW. There are also dramatic differences, such as between the variates used in predicting SG in the two Grolsch datasets. Differences in the chosen and optimised N values are also apparent, particularly in the long-path Grolsch dataset. In terms of utility, CovProc-SavGo was a step up in complexity to implement compared with CovProc. Although the general method was the same, there was the added step of searching and refining the N filter setting. Note also that CovProc-SavGo was implemented in a naïve manner, with the spectra filtered using N values determined in an outer loop, combined with an inner loop to investigate varying numbers of variates. A careful audit of the computational tasks may suggest opportunities for speeding up the process.

Figure 4 shows the results of the GA selections. This method had the least constraints regarding variate selection, although the choice of smoothed or derivatised spectra was predetermined. A popular application of GA is in the exploration, through feature selection, of multivariate datasets found in for example, metabolomics (Kemsley et al., 2007), and proteomics (Olias et al., 2006). In this study, therefore, it was anticipated that for ethanol and SG, the GA would favour variates associated with the chemical region used in the Simple method. However, this did not happen consistently. Nor is there much commonality between the variates chosen across the beer qualities. Despite the use of block validation in the GA selection process, which better guards against overfitting than LOO-CV, it is likely that some variates with a favourable noise structure were able

to out-compete generically useful features. Note also that in MLR, there is no issue of variance scaling; differences in the range of the variates are absorbed into the values of the regression coefficients. In terms of utility, GA was of comparable complexity to CovProc. The first stage in using the GA, although computationally intensive, was also wholly automated. Implementing the remaining two stages was comparatively straightforward.

Table 2 summarises the performance of the PLS-R analyses for the four selection methods investigated. They all performed reasonably well; compare the SEP and SEV values in Table 2 with the corresponding standard deviations given in Table 1. This performance is as expected, given that the feasibility of monitoring fermentation is well established.

Consider the summary statistics associated with model development (SEP and Q). Comparing Simple with CovProc, we find they performed similarly in terms of both Q and SEP. Comparing CovProc with CovProc-SavGo, we find that in most (11 /12) cases, CovProc-SavGo was better (larger Q, smaller SEP) than CovProc. This is to be expected considering the CovProc-SavGo procedure is an extension of CovProc. However, the improvements in both Q and SEP are marginal, and suggest the extra effort involved in the CovProc-SavGo procedure was not merited. With respect to the GA procedure, this was found not only to be consistently better (both Q and SEP) than the other three procedures, but often markedly so.

Considering the summary statistics associated with the model validation (SEV and R) we find here that all four procedures perform similarly well, although these figures are based on quite small sample numbers (14 Muntons, 15 Grolsch). Comparing the correlation found during model development and validation (Q vs. R), we find some indication that the GA performed worse in validation than expected, which suggests some level of over-fitting at the model building stage. Comparing SEP with SEV confirms the indication of over-fitting in the GA procedure; we find that

the SEV values for all four Muntons ethanol models were poorer than expected from their SEP values. This may be due to the relatively high bias values found here.

Finally, we find that for all the variate selection methods studied, the long-path Grolsch predictions were in general poorer than might have been expected from their SEP values. This is also apparent when comparing the performance of the short- and long- path models in development and validation. In model development, the short- and long- path models performed comparably (both in terms of Q and SEP). In model validation, we find the short-path models had consistently better correlations and standard errors than their long-path counterparts did.

4. Conclusions

From the results of the analysis presented, the main conclusion is that where a 'fingerprint' region can be identified, then it advisable to use the Simple approach. There is a clear underlying mechanism (expressed by the Beer-Lambert law) that justifies the use of linear modelling, and this gives confidence behind the ability of models to generalise. An additional benefit of using a continuous spectral region is that other pre-treatments, such baseline correction, can be used instead of using derivatised spectra. In this study, the choice of the selection methods effectively imposed the use of Savitzky-Golay filters: as three of the methods picked variates from the whole spectral range, then the only remedy for suppressing the sloping background when analysing ethanol and SG was to use derivative spectra. Although filtering the spectra can, in principle, improve the predictive ability of the model, choosing the filter settings adds to the overall complexity of the modelling process. Here, we found optimising the filter window width only marginally improved the SEP values, and these improvements were not always transferred to the validation set.

Where a fingerprint region is not known then this analysis suggests using CovProc. Here its performance was found to be similar to the Simple approach, although the selected variates tended to arise from indirect relationships between the modelled parameter and spectral variations. This

would add doubt to the ability of the model to generalise. The method was found to be quite clumsy to implement although this may be remedied with more efficient PLS (e.g. SIMPLS; de Jong, 1993) and cross-validation routines.

We would not recommend using the CovProc-SavGo approach. This was considerably more involved to use than CovProc, while gaining little improvement in model performance. The work also demonstrates the benefit of setting aside data for validation. From the SEP values, the more involved procedures appeared better. Only by applying them to independent test data was their ability to generalize in a real-world situation revealed.

Finally, this work suggests that the short-path, high-resolution, low-averaging measurement protocol can offer real benefits in predictive performance, without any additional cost in spectral acquisition time. We conclude that it is comparatively better to collect data at a higher spectral resolution and shorter sample path length, forgoing some amount of noise improvement through signal-averaging.

Acknowledments.

The authors thank the BBSRC for funding this work and Coors brewery Ltd for providing the Grolsch lager wort.

References

Arnold, S.A., Crowley, J., Vaidyanathan, S., Matheson, L., Mohan, P., Hall, J.W., Harvey, L.M. & McNeil B. (2000). At-line monitoring of a submerged filamentous bacterial cultivation using near-infrared spectroscopy. *Enzyme and Microbial Technology*, 27(9) 691-697.

Arnold, S.A., Harvey, L.M., McNeil, B., & Hall J.W. (2002a). Employing near-infrared spectroscopic methods of analysis for fermentation monitoring and control; Part 1, method

development. *Biopharm International – The Applied Technologies of Biopharmaceutical Development*, 15(11), 26-32 November.

Arnold, S.A., Gaensakoo, R., Harvey, L.M. & McNeil, B. (2002b). Use of at-line and in-situ near-infrared spectroscopy to monitor biomass in an industrial fed-batch escherichia coli process. *Biotechnology and Bioengineering*, 80(4) 405-413.

Bakeev, K.A., (2003). Near-infrared spectroscopy as a process analytical tool. *Pharmaceutical Technology Europe*, September, 27-32.

Boswell, C.D., Varley, J., Boon, L., Hewitt, C.J. & Nienow, A.W. (2003). Studies on the impact of mixing in brewing fermentation: comparison of methods of effecting enhanced liquid circulation. *Food and Bioproducts Processing*, 81(C1), 33-39.

de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3) 251-263.

Harthun, S., Matischak, K. & Friedl, P. (1998). Simultaneous prediction of human antithrombin III and main metabolites in animal cell culture processes by near infrared spectroscopy. *Biotechnology Techniques*, 12(5), 393-397.

Goldberg, D.E. (1989). Genetic *Algorithms in Search, Optimization, and Machine Learning*. Addison Wesley Longman, Inc., Massachusetts.

Hewitt, C.J. & Nebe-Von-Caron G. (2004). The application of multi-parameter flow cytometry to monitor individual microbial cell physiological state. *Physiological Stress Responses in Bioprocesses – Advances in Biochemical Engineering / Biotechnology*, 89 197-223.

Höskuldsson, A. (2001). Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 55(1-2) 23-38.

Kemsley, E.K., Le Gall, G., Dainty, J.R., Watson, A.D., Harvey, L.J., Tapp, H.S. & Colquhoun I.J. (2007). Multivariate techniques and their application in nutrition: a metabolomics case study. *British Journal of Nutrition*, 98(1), 1-14.

Macaloney, G., Draper, I., Preston, J., Anderson, K.B., Rollins, M.J., Thompson, B.G., Hall, J.W. & McNeil, B. (1996). At-line control and fault analysis in an industrial high cell density Escherichia coli fermentation, using NIR spectroscopy. *Food and Bioproducts Processing*, 74(C4), 212-220.

Martens, H. & Naes, T. (1989). *Mutivariate Calibration*. Wiley, Chichester.

Mitchelle, M. (1998). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge Massachusetts.

Norušis, M.J. & SPSS Inc. (1990). *SPSS Base System User's Guide*. SPSS, SPSS Inc., Chicago.

Olias, R., Maldonado, B., Radreau, P., Le Gall, G., Mulholland, F., Colquhoun, I.J. and Kemsley, E.K. (2006). Sodium dodecyl sulphate-polyacrylamide gel electrophoresis of proteins in dry-cured hams: Data registration and multivariate analysis across multiple gels. *Electrophoresis*, 27(7), 1288-1299.

Press, W.H., Teukolosky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, (2nd ed., vol. 1). Cambridge University Press, Cambridge.

Reinikainen, S-P. & Höskuldsson A. (2003). COVPROC method: strategy in modeling dynamic systems. *Journal of Chemometrics*, 17(2) 130-139.

Riley, M.R., Rhiel, M., Zhou, X., Arnold, M.A, & Maurhammer D.W. (1997). Simultaneous measurement of glucose and glutamine in insect cell culture media by near infrared spectroscopy. *Biotechnology and Bioengineering*, 55(1) 11-15.

Tapp H.S., Defernez M. & Kemsley E.K. (2003). FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils. *Journal of Agricultural and Food Chemistry* 51(21), 6110-6115.

Vaidyanathan S., White S., Harvey L.M. & McNeil B. (2003). Influence of morphology on the near-infrared spectra of mycelial biomass and its implication in bioprocess monitoring. *Biotechnology and Bioengineering*, 82(6), 715-724.

Table Captions

Table 1. Correlations between four brewing parameters and summary statistics for the Muntons pale ale and Grolsch lager fermentation experiments.

Table 2. Results of comparison between four variate selection methods prior to PLS regression: ntrain, size of training set; ntest, size of test set; savgo, Savitzky-Golay filter settings; nvars, number of variates used; npls, number of PLS factors used; Q, correlation with cross-validated predictions; sep, standard error in prediction; R, correlation with test-set predictions; sev, standard error in validation; bias, mean residual.

Figure Captions

Figure 1. Raw spectra, with ‘chemical’ and ‘biomass’ regions marked, for the three datasets studied: Muntons pale ale; Grolsch lager using 0.4 mm path length (short path); and Grolsch lager using 1 mm path length (long path). Also shown are spectra from Grolsch lager supernatant to demonstrate the influence of biomaterial on the spectra.

Figure 2. Variates selected by CovProc for the three datasets (Muntons, Grolsch short-path and Grolsch long-path) and four components studied, ethanol, OD SG and DCW

Figure 3. Variates selected by SavGo-CovProc for the three datasets (Muntons, Grolsch short-path and Grolsch long-path) and four components studied, ethanol, OD SG and DCW

Figure 4. Variates selected by the GA for the (Muntons, Grolsch short-path and Grolsch long-path) and four components studied, ethanol, OD SG and DCW

Tables

	Muntons				*Grolsch*			
Correlations	Ethanol	SG	OD	DCW	Ethanol	SG	OD	DCW
Ethanol	1.0000	-0.9071	0.8110	0.8339	1.0000	-0.9224	0.8278	0.7827
SG		1.0000	-0.8700	-0.9015		1.0000	-0.8448	-0.8085
OD			1.0000	0.9508			1.0000	0.8613
DCW				1.0000				1.0000
Mean	25.3	1.0236	14.70	2.79	25.3	1.0211	17.03	3.08
St. dev.	16.0	0.0190	7.35	1.41	16.4	0.0186	9.83	2.04

Table 1. Correlations between four brewing parameters and summary statistics for the Muntons pale ale and Grolsch lager fermentation experiments.

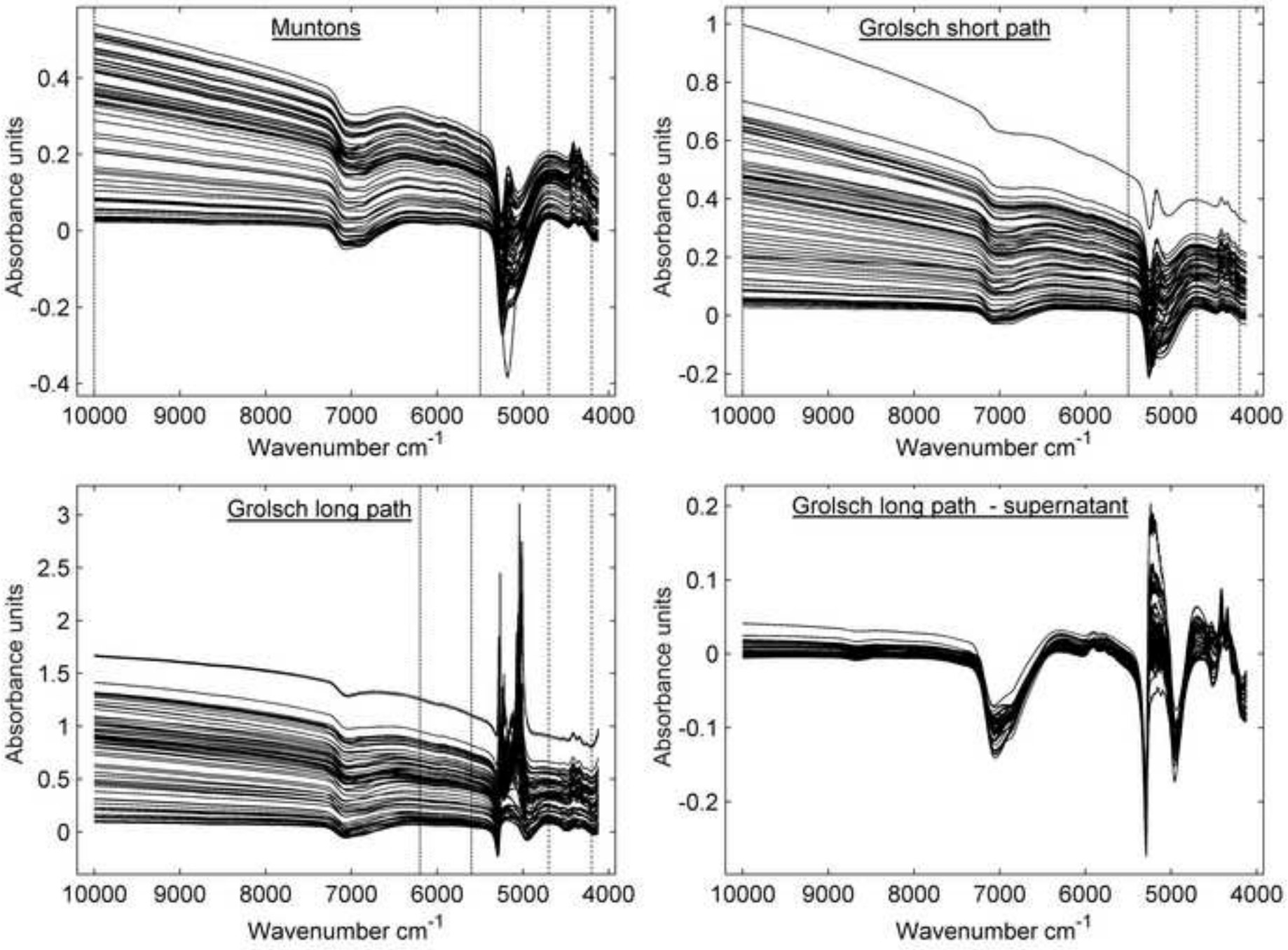
						Model				Test	
Dataset	*ntrain*	*ntest*	*Parameter*	*savgo*	*nvars*	*npls*	*Q*	*sep*	*R*	*sev*	*bias*
Simple											
Muntons	58	14	Ethanol	8 2 1	251	1	0.9730	3.59750	0.9631	6.0741	-3.7320
			SG	8 2 1	251	2	0.9682	0.00466	0.9830	0.0044	-0.0026
			OD	8 2 0	2251	3	0.9630	1.95160	0.9603	2.5536	-1.2805
			DCW	8 2 0	2251	3	0.9473	0.45773	0.9528	0.5412	-0.3623
Grolsch short	52	15	Ethanol	8 2 1	251	1	0.9763	3.62579	0.9899	2.4215	-1.1396
			SG	8 2 1	251	2	0.9732	0.00435	0.9865	0.0029	0.0002
			OD	8 2 0	2251	3	0.9453	2.82014	0.9836	2.7170	1.3747
			DCW	8 2 0	2251	4	0.9652	0.55195	0.9829	1.6225	1.3806
Grolsch long	52	15	Ethanol	3 2 1	126	8	0.9857	2.83197	0.9707	4.9569	0.8364
			SG	3 2 1	126	2	0.9695	0.00463	0.9809	0.0114	0.0109
			OD	3 2 0	151	6	0.9558	2.54779	0.9510	5.3305	3.1929
			DCW	3 2 0	151	9	0.9733	0.48670	0.9683	2.8844	2.0577
CovProc											
Muntons	58	14	Ethanol	8 2 1	15	3	0.9739	3.53900	0.9668	6.7907	-5.0432
			SG	8 2 1	20	3	0.9667	0.00477	0.9858	0.0042	-0.0028
			OD	8 2 0	1420	3	0.9664	1.86176	0.9743	2.1200	-1.1697
			DCW	8 2 0	403	7	0.9609	0.40157	0.9466	0.4348	-0.0947
Grolsch short	52	15	Ethanol	8 2 1	21	3	0.9727	3.89356	0.9902	2.2656	-0.7497
			SG	8 2 1	267	3	0.9678	0.00476	0.9796	0.0039	-0.0017
			OD	8 2 0	594	4	0.9583	2.46987	0.9743	3.2708	-0.7529
			DCW	8 2 0	1025	8	0.9762	0.46012	0.9791	1.5157	1.1538
Grolsch long	52	15	Ethanol	3 2 1	96	15	0.9688	4.37192	0.8521	12.9233	2.6519
			SG	3 2 1	1470	13	0.8825	0.00944	0.9713	0.0049	0.0024
			OD	3 2 0	41	3	0.9537	2.60012	0.9477	4.1738	0.8924
			DCW	3 2 0	1175	10	0.9784	0.43697	0.9186	3.2967	1.6873
CovProc-SavGo											
Muntons	58	14	Ethanol	9 2 1	13	3	0.9742	3.51739	0.9665	6.7405	-4.9621
			SG	10 2 1	11	6	0.9717	0.00441	0.9793	0.0049	-0.0030
			OD	1 2 0	1460	10	0.9770	1.54861	0.9851	1.5447	-0.7482
			DCW	13 2 0	509	9	0.9691	0.35689	0.9361	0.4802	-0.1192
Grolsch short	52	15	Ethanol	4 2 1	31	2	0.9772	3.55906	0.9914	2.5530	-1.3783
			SG	31 2 1	2938	5	0.9725	0.00440	0.9859	0.0037	0.0011
			OD	1 2 0	978	5	0.9614	2.39079	0.9710	3.3077	-0.4063
			DCW	3 2 0	1389	11	0.9807	0.41358	0.9873	1.6616	1.3651
Grolsch long	52	15	Ethanol	17 2 1	225	12	0.9816	3.23474	0.9855	3.7773	2.1974
			SG	23 2 1	184	9	0.9668	0.00484	0.9788	0.0040	0.0016
			OD	17 2 0	13	6	0.9647	2.27895	0.9398	4.3464	-0.5153
			DCW	32 2 0	899	9	0.9782	0.43904	0.8727	3.8248	1.6478
GA											
Muntons	58	14	Ethanol	8 2 1	33	6	0.9880	2.40733	0.9481	7.6944	-5.3049
			SG	8 2 1	60	7	0.9833	0.00340	0.9709	0.0061	-0.0035
			OD	8 2 0	29	11	0.9849	1.25510	0.9836	1.5901	-0.7141
			DCW	8 2 0	50	14	0.9855	0.24367	0.9637	0.3579	-0.0679
Grolsch short	52	15	Ethanol	8 2 1	42	8	0.9938	1.86680	0.9845	2.9653	0.7092
			SG	8 2 1	34	6	0.9884	0.00287	0.9849	0.0042	0.0021
			OD	8 2 0	27	8	0.9787	1.77644	0.9740	3.1932	1.1441
			DCW	8 2 0	40	11	0.9931	0.24842	0.9902	1.9002	1.5070
Grolsch long	52	15	Ethanol	3 2 1	18	8	0.9937	1.87907	0.9405	6.0098	-2.0720
			SG	3 2 1	7	7	0.9879	0.00293	0.8034	0.0127	0.0035
			OD	3 2 0	7	6	0.9752	1.91258	0.9049	5.9581	-2.2734
			DCW	3 2 0	11	11	0.9876	0.33119	0.9847	1.7829	1.3302

Table 2. Results of comparison between four variate selection methods prior to PLS regression: ntrain, size of training set; ntest, size of test set; savgo, Savitzky-Golay filter settings; nvars, number of variates used; npls, number of PLS factors used; Q, correlation with cross-validated predictions;

sep, standard error in prediction; R, correlation with test-set predictions; sev, standard error in validation; bias, mean residual.

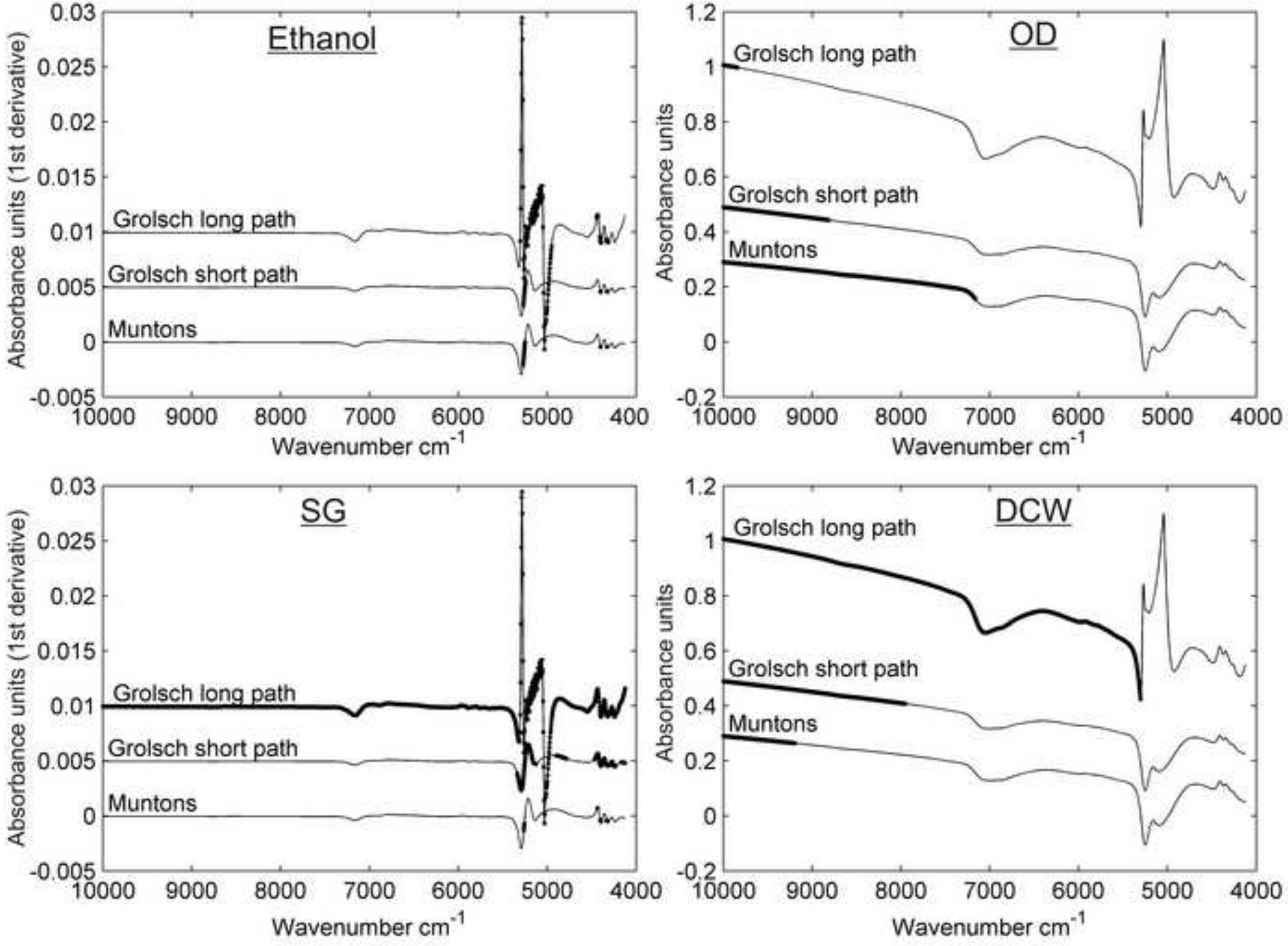
Figure

Click here to download high resolution image



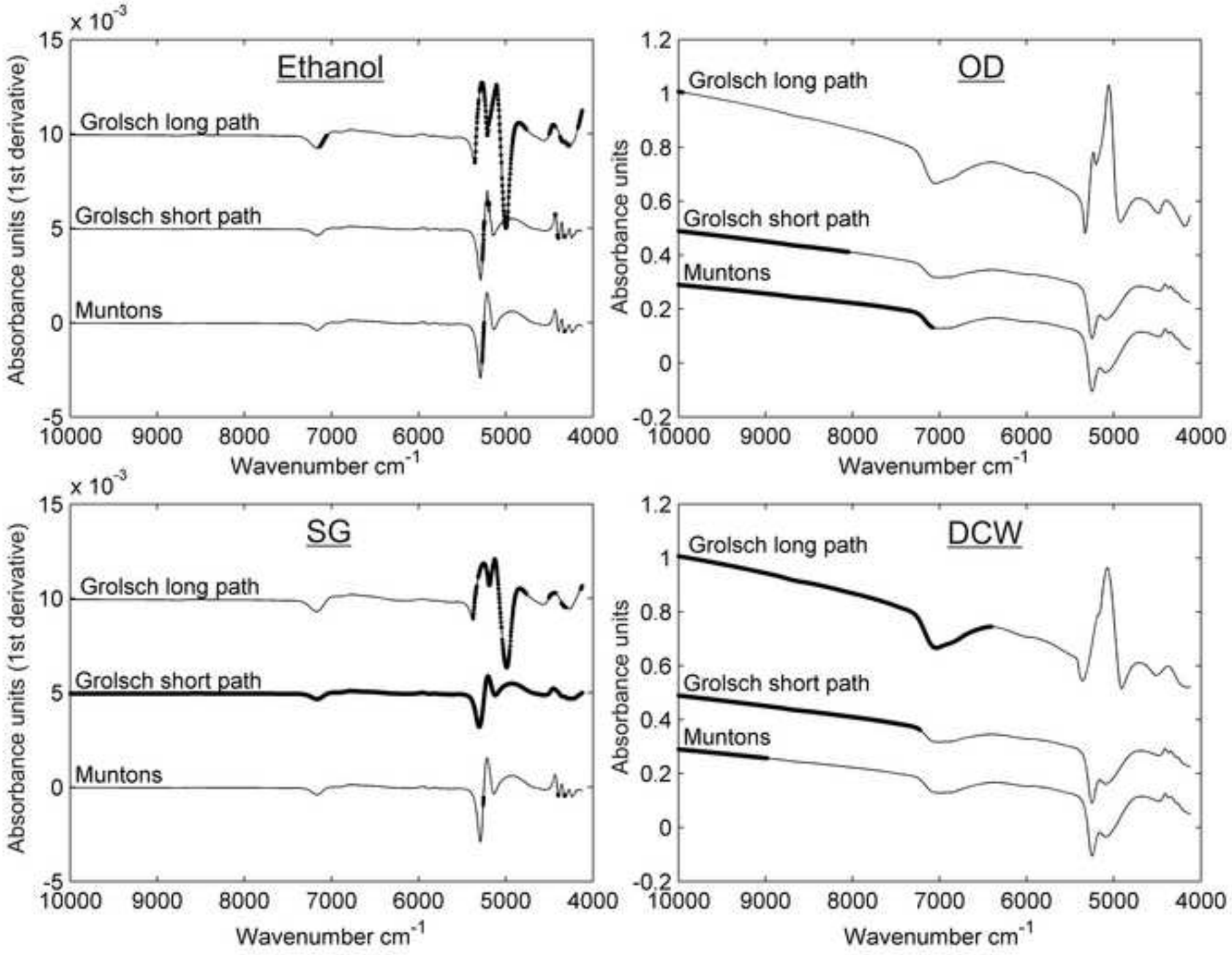
Figure

Click here to download high resolution image



Figure

Click here to download high resolution image



Figure

Click here to download high resolution image