Sparse regularization of NIR spectra using implicit spiking and derivative spectroscopy

R. S. Anderssen [a], **F. R. de Hoog** [a], **I. J. Wesley** [b], **A. Zwart** [a]

[a] *CSIRO Computational Informatics, GPO 664, Canberra, Australian Capital Territory, 2601*
[b] *Grain Growers Limited, PO Box 7, North Ryde, NSW 1670*
Email: Bob.Anderssen@csiro.au

Abstract: For the application of spectroscopic calibration and prediction (CAP), the data consists of measurements, for each sample in some representative set, of both the property of interest and a spectral encapsulation (e.g. for wheat samples, their glutenin contents and NIR spectra). The information in the spectra about the property is confounded by the other components in the samples (e.g. in wheat, the gliadin) since their proportional presences also change. Nevertheless, one has an *implicit spiking* situation in the sense that one knows the ordering of the proportional presence of the target property (e.g. the gliadin). Here, using glutenin and NIR spectra for wheat, the extent is examined to which the explicit sparse regularization protocol developed by Anderssen, de Hoog, Wesley and Zwart for milk powder samples spiked with casein can be applied in a CAP analysis of implicitly spiked data.

An often occurring situation in information recovery arises when the indirect measurements of the phenomenon of interest (an NIR spectrum of milk powder; an image recorded on a very high resolution CCD camera) contains two different phases: (i) the information which encapsulates the answer to the question under examination (the proportional presence in milk powder of the casein, the major protein component; a lower resolution image is all that is required) and (ii) a considerable amount of superfluous information, the presence of which compromises the reliability with which the question can be answered. In such situations, for the identification of the information that encapsulates the answer, *"sparse regularization"* and *"compressive sensing"* are widely utilized. From both theoretical and practical perspectives, though there is an overlap between these two methodologies, their basic *modus operandi* are different. The former is more suited to spectroscopic data applications, where the scientific basis for the structure within the data is known, while the latter plays a key role in image and data compression and sensor network applications, where often there is no rationale for the structure in the data other than the circumstances of the application.

In this paper, the focus is the application of sparse regularization to the analysis of near infrared (NIR) spectroscopic data. It has already been shown Anderssen et al. (2013) how the explicit spiking of biological data can be used to identify, in the corresponding NIR spectra, the wavelength bands associated with the spiking that are essentially independent of the other components in the sample. In particular, this was done using samples of the same milk powder explicitly spiked with known different amounts of casein and derivative spectroscopy to perform the sparse regularization. The goal in this paper is to give independent validation to the methodology developed for the milk powder situation Anderssen et al. (2013) and thereby establish that it extends to CAP data where the spiking is implicitly determined by the chosen samples. This is achieved by applying the derivative spectroscopic sparse regularization to the NIR spectra of individual wheat grains for which the different levels of the proteins gliadin and glutenin have been measured.

The essence of the explicit spiking methodology, on which the milk powder analysis was based, is the identification of the wavelengths at which the intensities of the spectra correlate strongly with the proportional presence of the target property. The underlying assumption/rationale is that they represent the locations in the spectra where the interaction of the target component with the other components in the sample are minimal.

The paper has been organized in the following manner. As motivation for the implicit spiking approach, the explicit spiking methodology is reviewed in Section 1. Relevant details about the structure of NIR spectra, and, in particular, the wheat spectra analysed, are discussed in Section 2. The results of the spiking analysis of NIR wheat spectra, for which the proportional presences of albumin, gliadin and glutenin were available, are presented and discussed in Section 3.

Keywords: *Calibration and prediction, spiked data, near infrared spectroscopy, single grain analysis*

1 Introduction

For the recovery of information about the relationship between structural features in measured spectra (e.g. NIR measurements of single wheat kernel samples) and the values of some associated target property (e.g. protein content in the wheat), a popular methodology is calibration and prediction (CAP). The goal of the *calibration* step is the identification of the spectral wavelength intervals where the most succinct information is located about the target property to be predicted, and the utilization of this information for the construction of a reliable and robust predictor for subsequent use in the *prediction* of the values of the target property for spectroscopic measurements of new samples.

Various algorithmic procedures are available for performing the calibration step computationally. They include partial least squares (PLS) (Phatak and de Hoog (2002); Kondylis and Whittaker (2013)), the support vector machine regularization, compressed sensing Candes and Wakin (2008) and derivative spectroscopy Anderssen and Hegland (2010). In all these procedures, the identification of the relationship is performed as a sparse regularization in that they implicitly identify within the given spectra the subset of wavelength intervals which are informative of the property of interest. The underlying relationship is determined by the matrix algebra that defines the procedure.

Here, we explore how the identification of the wavelength intervals can be performed in an explicit manner. It is based on the rationale that the methodology developed for the identification of the informative wavelength intervals in the NIR spectra of samples explicitly spiked with the target component, the property of which is required, can be extended to the analysis of NIR spectra where the levels of the spiking, though not performed in an explicit orchestrated manner, have been measured.

The essential rationale is that, with respect to the values of the chosen target property, the most important wavelength intervals are where the amplitudes of the spectra (or some (linear) transformation (such as a second order differentiation) of the spectra) correlate positively with the values.

An experimental approach, referred to as *explicit spiking* Anderssen et al. (2013), has been used to identify the wavelength intervals in the NIR spectra of milk powder, spiked with different known levels of casein, where there is a strong correlation of the amount of the added casein and the amplitudes of the spectra. The underlying rationale is the identification the wavelength intervals where the presence of the casein has minimal confounding with respect to the other components in the milk powder.

An illustration of the confounding that occurs in NIR spectra is given in Figure 1, where the second and fourth derivatives of the NIR spectra of the gliadin and glutenin from a wheat sample are plotted Wesley et al. (1999). Though it is not possible to obtain pure gliadin and glutenin samples and measure their NIR spectra, accurate estimates of their NIR spectra can be obtained using unmixing Wesley et al. (1999). That the extent of the confounding is very strong is reflected in the fact that even the fourth derivatives of the gliadin and glutenin NIR spectra are quite similar. This is not too surprising, since NIR spectra record the intensity of the vibrations of the molecular side chains with the confounding caused by the common side chains on the different components which form the gliadin and glutenin samples. However, as illustrated in Figure 1, the differences in molecular structure of the components and the way in which the components are arranged within the samples do produce minor differences in the structure of the NIR spectra, which their second and fourth derivatives highlight. For example, for both the second and fourth derivatives, differences can be seen in the wavelengths intervals 450-550nm, 1100-1300nm and 1600-1750nm. Interestingly, globally, the second and fourth derivatives look quite similar. Closer inspection shows however that there are wavelength intervals where the differences between the fourth derivative spectra are different from the differences in the second derivative spectra, with more pronounced differences in the second derivative spectra.

2 Explicit and implicit spiking analysis of NIR spectra

Spiking of data is an explicit process. However, when a selection of samples is chosen in a CAP to construct a predictor for some specific property of interest, the samples are in fact implicitly spiked. As already mentioned above, the potential to which this observation can be exploited using the procedure developed for explicit spiking is the focus of this paper.

2.1 Explicit Spiking

The explicit spiking procedure was developed for the CAP construction of a reliable and robust predictor of the proportional presence of casein in milk powder Anderssen et al. (2013). The *modus operandi* behind the

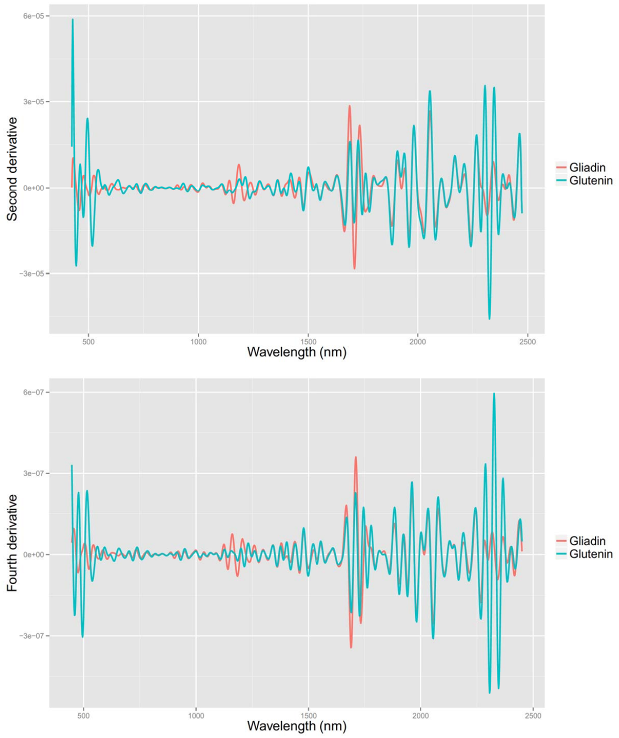

Figure 1. A comparison of the second (top) and fourth (bottom) derivatives of gliadin and glutenin NIR spectra, to illustrate the high level of confounding in their NIR spectra due to the similarity in their side chain structure and organization.

formulation of how the calibration should be performed was the need to identify the NIR wavelength intervals where the presence of the casein is not confounded by the presence of the other components in the milk powder.

Using the spiking of milk powder with casein as the prototypical example, the basic steps in the explicit spiking procedure, in terms of utilizing the fourth derivative spectra to perform the wavelength interval identification, are:

(i) For different samples of the same milk powder, spike them with different amounts $m_j,\ j = 1,\ 2,\ \cdots,\ J,\ 0 = m_1 < m_2 < \cdots < m_J$ of casein.

(ii) Taking account of the amount of casein already in the milk powder, which can be independently measured, let the spiked samples be ordered in terms of their proportional casein content $a_1 < a_2 < \cdots < a_J$, where a_1 and a_J correspond, respectively, to the casein content of the unspiked milk powder and pure casein.

(iii) Record the NIR spectra of each sample and represent them as the row vectors

$$\mathbb{MP}_{a_j}^T = [\mathbb{MP}_{a_j}(\lambda_1)\ \ \mathbb{MP}_{a_j}(\lambda_2)\ \ \cdots\ \ \mathbb{MP}_{a_j}(\lambda_K)],$$

where the $\mathbb{MP}_{a_j}(\lambda_k)$ denote the values of the spectra at wavelength λ_k. The corresponding rectangular matrix array of row vectors $\mathbb{MP}_{a_j}^T$ will be denoted by

$$\mathbb{MP} = \begin{bmatrix} \mathbb{MP}_{a_1}(\lambda_1) & \mathbb{MP}_{a_1}(\lambda_2) & \cdots & \mathbb{MP}_{a_1}(\lambda_K) \\ \mathbb{MP}_{a_2}(\lambda_1) & \mathbb{MP}_{a_2}(\lambda_2) & \cdots & \mathbb{MP}_{a_2}(\lambda_K) \\ & \cdots\cdots\cdots & & \\ \mathbb{MP}_{a_J}(\lambda_1) & \mathbb{MP}_{a_J}(\lambda_1) & \cdots & \mathbb{MP}_{a_J}(\lambda_K) \end{bmatrix}.$$

(iv) Determine the fourth derivative of this set of spectra and denote them by $\mathbb{MP}^{(4)}$, with $\mathbb{MP}^{(4)}_{a_j}(\lambda_k)$ being the values of the fourth derivatives at wavelength λ_k. The opportunity that this set of spectra represents for identifying the appropriate intervals to be used as predictors of casein content is explained in Anderssen et al. (2013)..

(v) Assess the level of correlation between the a_j and the columns of $\mathbb{MP}^{(4)}$ with $\mathbf{MP}^{(4)}_{\lambda_k}$, $k = 1,\ 2,\ \cdots,\ K$, denoting the column vectors

$$[\mathbb{MP}^{(4)}_{a_1}(\lambda_k),\ \mathbb{MP}^{(4)}_{a_2}(\lambda_k),\ \cdots\ \mathbb{MP}^{(4)}_{a_J}(\lambda_k)]^T.$$

(a) mean center the columns in $\mathbb{MP}^{(4)}$, denoting the result matrix of columns as

$$\mathbf{S} = [\mathbf{s_1},\ \mathbf{s_2},\ \cdots\ \mathbf{s_K}],$$

(b) mean center of the column vector $[a_1,\ a_2,\ \cdots,\ a_J]^T$ to obtain $\mathbf{a}^*$,

(c) assess how closely some multiple of an $\mathbf{s_k}$ approximates $\mathbf{a}^*$ using the error measure

$$E_k = (\hat{a}_k\mathbf{s_k} - \mathbf{a}^*)^T(\hat{a}_k\mathbf{s_k} - \mathbf{a}^*), \quad \hat{a}_k = \frac{\mathbf{s_k}^T\mathbf{a}^*}{\|\mathbf{s_k}\|\|\mathbf{a}^*\|} \tag{1}$$

Fuller details about these steps along with various plots illustrating the individual steps can be found in Anderssen et al. (2013).

2.2 Implicit Spiking

The essence of the implicit spiking procedure is to apply steps (iii)-(v) to the NIR spectra and measured property for a representative set of (wheat) samples as if the measured property values corresponded to the a_j.

Now, however, one is constrained by the available samples, not the chosen levels of the spiking, and, thereby, by the level of the implicit spiking that these samples have for the chosen property under investigation. Consequently, the motivation for this paper is an examination of the potential for this strategy to highlight appropriate wavelength intervals in NIR wheat spectra using the NIR spectra for the individual wheat kernels and their laboratory-determined albumin, gliadin and glutenin contents as published in Wesley et al. (2008). The second derivatives of the corresponding NIR spectra are plotted in Figure 2. In order to obtain the accurate NIR measurements utilized to obtain the derivatives in Figure 2, the NIR instrument used had two detectors with a silicon detector recording in the wavelength range 400-1100nm and a lead sulphide in the range 1100-2500nm. Consequently, the gap, centered a 1100nm, is where the join between these two separate measured spectra occurs. It is clearly visible in Figure 2 with a gap occurring around the 1100nm wavelength because centered moving averages have been used to perform the numerical differentiation in the two separate regions, utilizing the numerical differentiation methodology developed in Anderssen and de Hoog (1984); Anderssen et al. (1998).

From a vibrational spectroscopic analysis perspective, there are four separate regions of interest, which will be treated separately in the deliberations below: the visible 400-800nm; the third NIR overtone 800-1100nm; the second 1100-1800nm; the first and combination 1800-2500nm. In the sequel, attention will mainly focus on the second NIR overtone region.

As well as computing the values of E_k as a function of the wavelength values λ_k, the goal is to have a quick procedure for initially assessing, in a given situation, its potential utility. This is done in two separate ways:

(i) **The Rug.** In Figure 2, for the gliadin measurements, the values of E_k have been plotted as a red-white *rug* with white identifying the smallest values of E_k and red the highest. The role of the rug is to give a quick identification of the wavelength intervals in the spectra for which the values of the errors E_k are smallest. With respect to the property of interest (gliadin in Figure 2, and albumin in Figure 3), the errors are smallest where the confounding is minimal for the vibrational response of that property with the vibrational response of the other components in the (wheat) samples.

(ii) **The E_k Histogram.** In Figure 4, the histogram of the errors E_k are plotted, in order to give a size distribution summary of the errors as recorded in the rug in the corresponding spectral (or derivative spectral) plot. They allow the relative significance of the red-white pattern in a rug to be assessed and

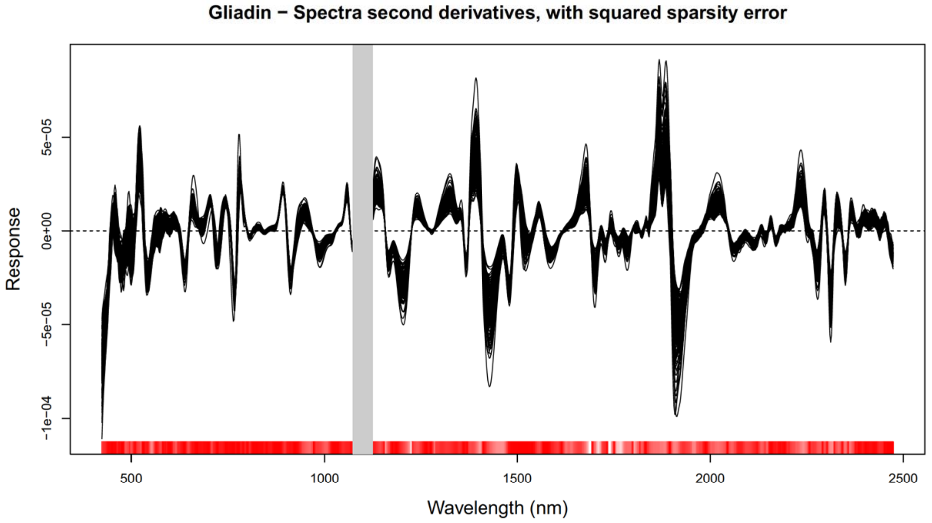

Figure 2. Plot of the second derivatives of the wheat NIR spectra along with the rug for Gliadin and the cross-over point at 1100nm clearly identifiable. The gap occurs because centered moving averages have been used to perform the numerical differentiation.

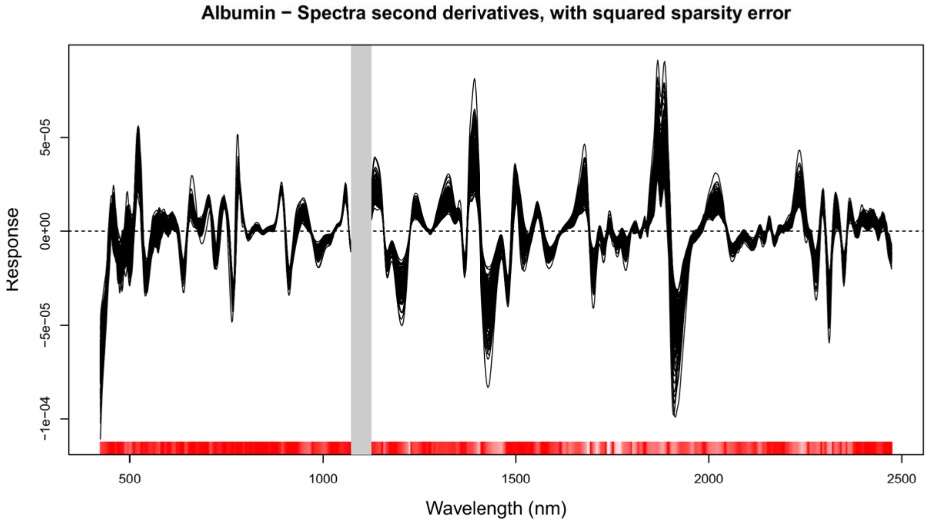

Figure 3. Plot of the second derivatives of the wheat NIR spectra along with the rug for Albumin. The second derivatives are the same as in Figure 2, but the rug is different as the values of E_k that determine the red-white colouring in the rug have been computed with respect to the Albumin content.

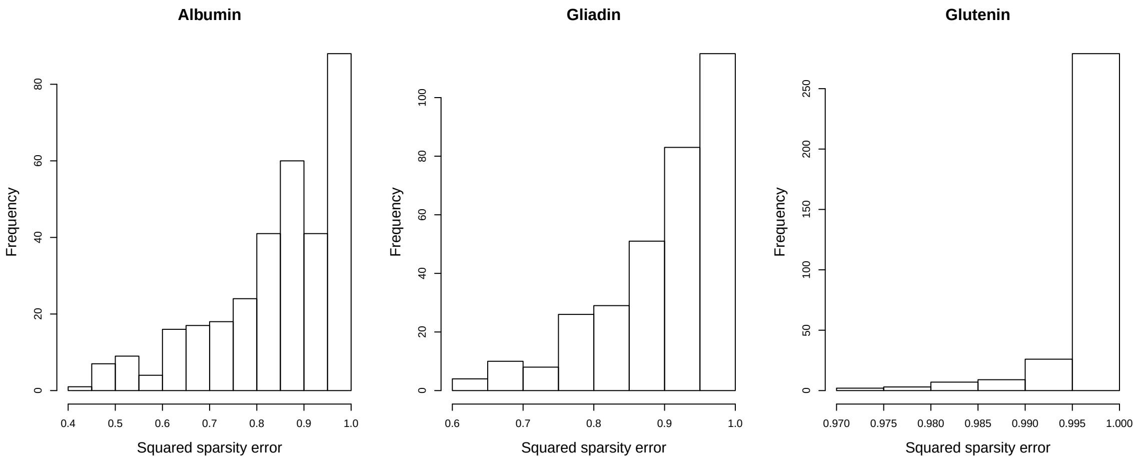

Figure 4. The E_k-Histograms for the second overtone region which gives a basis for the comparison of the significance of the pattern in the albumin, gliadin and glutenin rugs with respect to the second derivatives of the NIR wheat spectra.

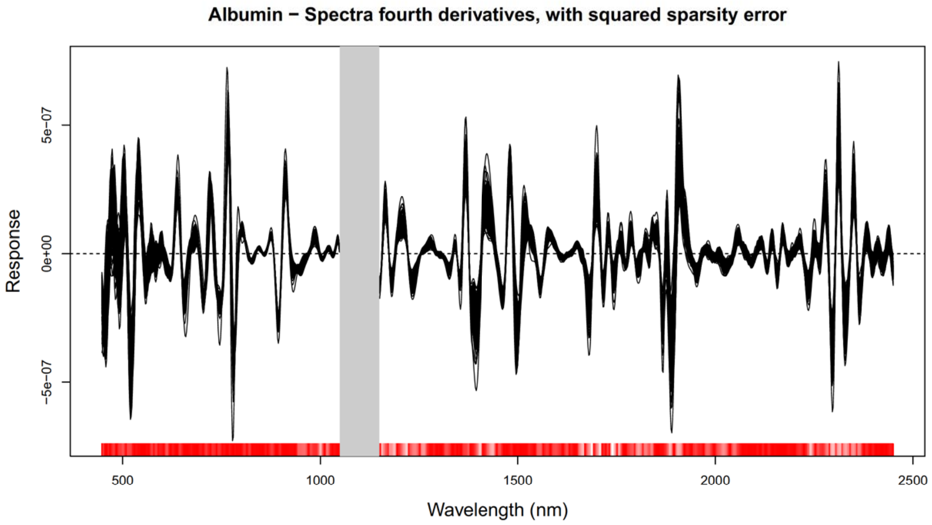

Figure 5. The albumin rug for the fourth derivative of the NIR wheat spectra.

> thereby yield a basis for comparing the pattern in a rug in the different spectral regions for a particular derivative, for comparing the pattern in the rugs of different derivatives for the same property and comparing patterns in the rugs for different properties.

Consequently, in a given situation, one uses the rug to identify the wavelength regions were the values of E_k are small and the corresponding E_k-histogram to assess the relative significance of the smallest values, with the significance being minimal if the smallest values of the E_k are not sufficiently small. For example, on the basis of a comparison of the three histograms in Figure 4, one concludes that the rug patterns for albumin and gliadin are significant, with that for albumin being the better of the two, whereas the rug pattern for glutenin is not, since the glutenin values for E_k indicate that the correlation is very poor implying a high level of compounding at all wavelengths.

A comparison of the E_k-histograms yields different insight about the situation. For example, a comparison of the E_k-histograms of Figure 4 leads to the following conclusion. There is a much stronger correlation of the albumin values with the structure in the histograms than that for the gliadin which is clearly better than the glutenin. In addition, the very high peak on the right for the E_k-Histogram for glutenin implies that there is only a marginal association of the glutenin values with the ordering of the curves in the second derivatives of the NIR spectra.

A validation of the approach adopted here relates to the fact that $E_k = 1 - r_k^2$, where r_k^2 corresponds to the Pearson correlation between the amplitudes of the response at wavelength k and the values of the property being investigated.

For comparison with Figure 2, the fourth derivative of the NIR wheat spectra along with the albumin rug is plotted in Figure 5. It shows that the rugs are similar, yet not identical. It represents confirmation that the proposed wavelength identification procedure is robust and consistent.

3 Conclusions

The above results establish that the methodology developed for explicit spiking can be successfully utilized in the analysis of implicitly spiked NIR spectra. It is anticipated that this will have wider application in the spectroscopic analysis.

References

Anderssen, B., F. de Hoog, and M. Hegland (1998). A stable finite difference ansatz for higher order differentiation of non-exact data. *Bull. Austral. Math. Soc. 58*, 223–232.

Anderssen, R. S. and F. R. de Hoog (1984). Finite-difference methods for the numerical differentiation of non-exact data. *Computing 33*, 259–267.

Anderssen, R. S., F. R. de Hoog, I. J. Wesley, and A. Zwart (2013, July). How much of an nir spectrum is useful? sparse regularization - let the data decide! In S. McCue and A. J. Roberts (Eds.), *Proceedings of the 16th Biennial Computational Techniques and Applications Conference, CTAC-2012*, Volume (in review) of *ANZIAM J.* url http://anziamj.austms.org.au/ojs/index.php/ANZIAMJ/article/view/3909 [July 10, 2011].

Anderssen, R. S. and M. Hegland (2010,). Derivative Spectroscopy – An enhanced role for numerical differentiation. *J. Integ. Eqn. Appl. 22*(3), 355–367.

Candes, E. J. and M. B. Wakin (2008). An introduction to compressive sampling. *IEEE Signal Processing Mag. 25*, 21–30.

Kondylis, A. and J. Whittaker (2013). Feature selection for functional PLS. *Cheom. Intel. Lab. Syst. 121*, 82–89.

Phatak, A. and F. de Hoog (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *J. Chemometrics 16*.

Wesley, I., S. Uthayakumaran, R. Anderssen, G. Cornish, F. Bekes, B. Osborne, and J. Skerritt (1999). A curve-fitting approach to the near infrared reflectance measurement of wheat flour proteins which influence dough quality. *JNIRS 7*, 229–240.

Wesley, I. J., B. G. Osborne, O. Larroque, and F. Bekes (2008). Measurement of the protein composition of single wheat kernels using near infrared spectroscopy. *JNIRS 16*, 505–516.