A Bayesian framework for near infrared calibration

T. Fearn,[a] D. Pérez-Marín,[b] A. Garrido-Varo[b] and J. E. Guerrero-Ginel[b]

[a]*University College London, London WC1E 6BT, United Kingdom.*
E-mail: tom@stats.ucl.ac.uk
[b]*University of Córdoba, 14014 Córdoba, Spain*

Introduction

Most approaches to NIR calibration involve a regression of reference values on spectral data. This direct modeling approach has many advantages, but it has one property that is not always an advantage: the predictions are biased towards the mean of the distribution of the reference values in the training set. In the alternative framework described here, a model for the dependence of spectral data on reference values is combined, using Bayes' Theorem, with a prior probability distribution describing the population of samples to be predicted. This framework allows the reproduction of standard results, but also provides the flexibility to incorporate non-standard prior distributions for the reference values. It also lends itself to the use of a variety of methods for modeling the dependence of spectral data on reference values, including one local-type approach that dispenses with the regression model altogether and which gave very good results for an application to the composition of animal feed samples. The Bayesian methodology and its application to the feed samples have both been described in detail elsewhere.[1] Here only the main findings are reported.

Materials and methods

A total of 7523 commercial feed samples, for which reference data on ingredient composition were provided by the manufacturer, were ground to 1mm and NIR spectra (1100–2500 nm) were measured on a FOSS NIRSystems 5000 instrument. The SNV-treated first derivative spectra were reduced to scores on 30 PCs. A training set of 7423 and a validation set of 100 samples were used to study calibrations for wheat and sunflower meal contents.

The equation at the heart of the approach to calibration is Bayes theorem

$$p(y_i|x) = p(x|y_i)p(y_i)/p(x) \quad \text{for } i = 1, \ldots, I\,.$$

To interpret this, after measuring a spectrum x the probability $p(y_i|x)$ that the unknown reference measurement y corresponding to this sample has value y_i (the range of possible values of y has been discretized for computational convenience) is computed by multiplying $p(x|y_i)$, the

probability of observing this spectrum if the reference was equal to y_i, and $p(y_i)$, the so-called prior probability of y_i, and dividing the result by $p(x)$. The divisor is simply the scaling constant needed to make the $p(y_i|x)$ sum to 1. The end result of the computations will be a discrete probability distribution over the grid of values of y. To get an estimate of y and some measure of the uncertainty associated with this estimate, the mean and standard deviation of the distribution $p(y|x)$ may be calculated.

Taking a multivariate normal model with linear regressions of PC scores on reference values for $p(x|y_i)$ and a normal approximation to the distribution of reference values in the training set for $p(y_i)$ reproduces the PCR result. This fact can be demonstrated algebraically, and may also be checked numerically. To improve on the PCR, $p(y_i)$ was replaced by the empirical distribution of reference values in the training set, the bars in Figure 2, and $p(x|y_i)$ by multivariate kernel density estimates computed separately for each y_i in the grid. That is, for each y_i the multivariate distribution of the 30 PC scores was modeled as an average of spherical Gaussian distributions centered on each of the observed 30-dimensional scores for the samples in the training set with y-values

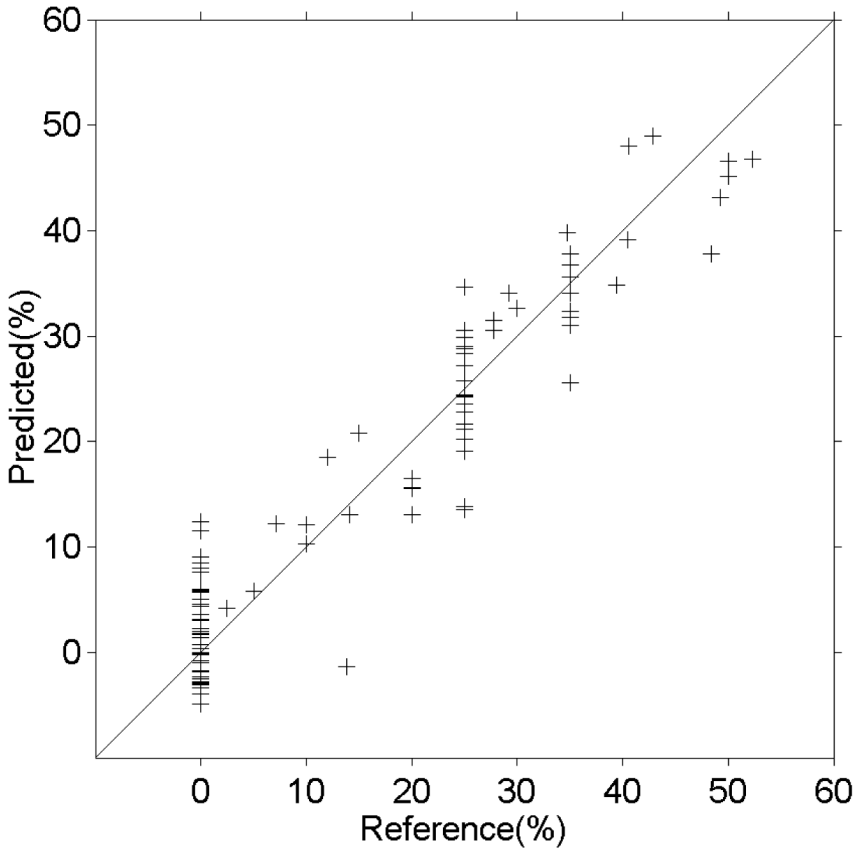

Figure 1. Predicted versus reference for % wheat in 100 validation samples using PCR.

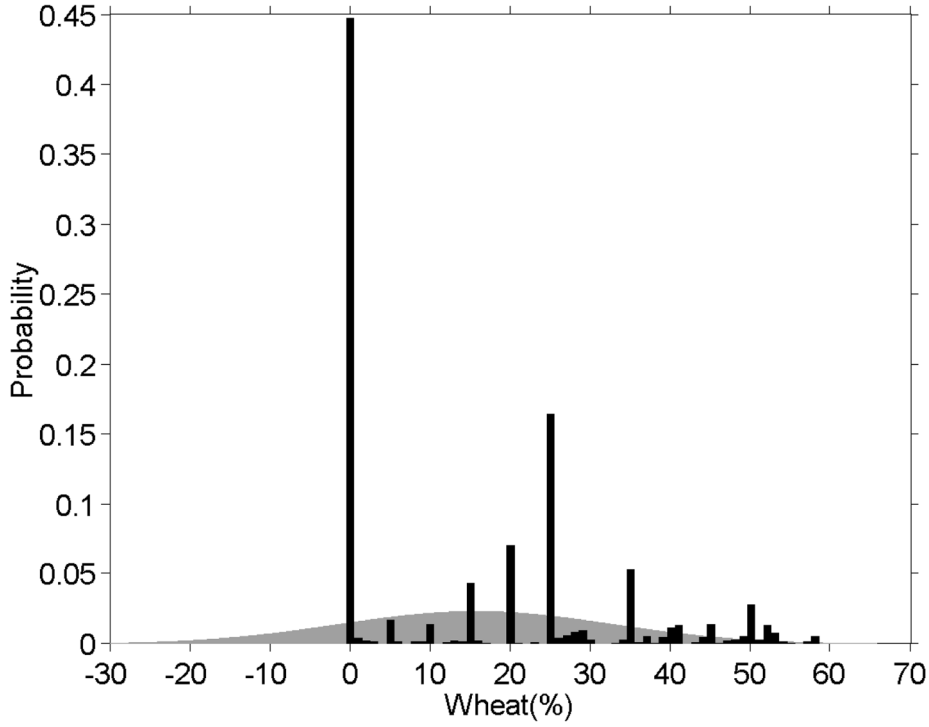

Figure 2. Empirical distribution (black bars) of % wheat content in the training samples, and the normal approximation to this distribution (grey area).

in the bin corresponding to the grid-value y_i. There is one parameter to tune, the spread of the Gaussian kernels, and this was done using a test set extracted from the training set (and not on the 100 validation samples).

Results and discussion

Using PCR with 30 factors the RMSEP on the validation set was 5.00% for wheat and 0.88% for sunflower. Figure 1 shows the predicted versus reference values for the validation set. Not only is the performance unacceptable, but there are many negative predictions.

Figure 2 compares the actual distribution of % wheat contents in the training set with the normal approximation implicitly used by PCR.

One of the reasons for the frequent negative predictions is apparent from this figure. The big spike of values at zero in the training set causes the normal prior distribution to put quite a lot of weight below zero. It is a very poor approximation to the actual distribution in this case. Using the empirical distribution for $p(y)$ in the Bayesian framework reduces the *RMSEP* for wheat to 4.23% and eliminates negative predictions. This is an improvement, but there is scope for a much bigger

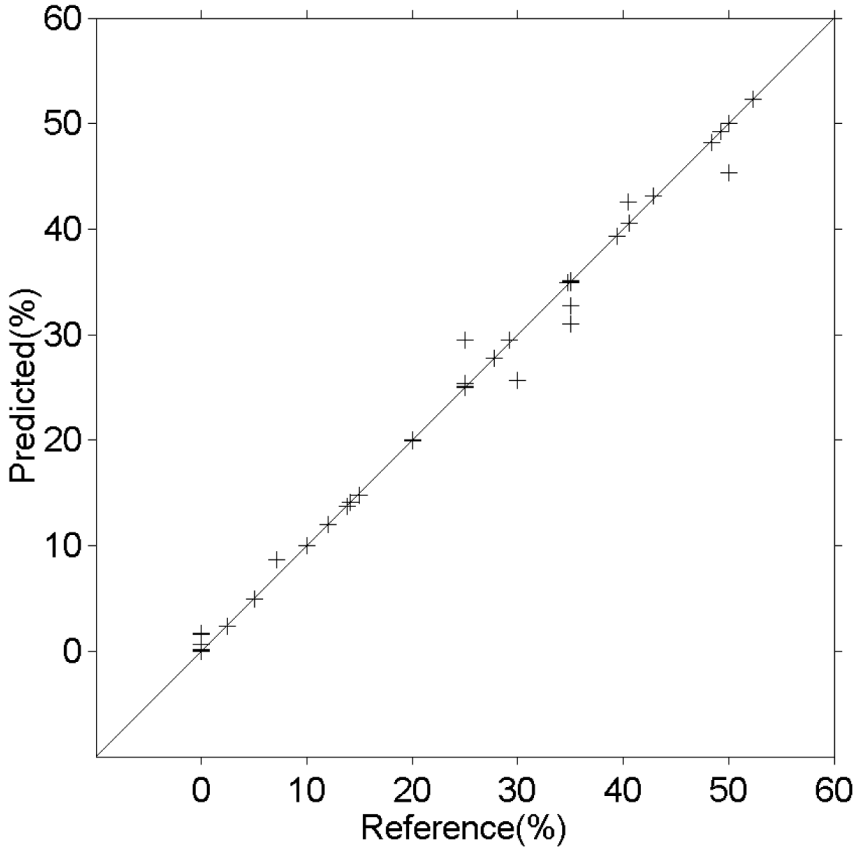

Figure 3. Predicted versus reference for % wheat in 100 validation samples using an empirical prior distribution $p(y)$ and a kernel density estimate for $p(x|y)$.

one. Replacing the regression model by the kernel density estimate further reduces the *RMSEP* to 0.98% for wheat and 0.34% for sunflower. The predicted versus reference values for wheat are shown in Figure 3.

The sunflower predictions, together with error bars derived from $p(y|x)$, are shown in Figure 4.

The error bars, which vary considerably in length, are a good measure of the uncertainty in the predictions.

This is just one example, characterized by a very non-normal distribution of reference values and a very large training set, and it remains to be seen how the methodology performs in general. It seems clear that the freedom to specify the prior distribution, rather than have it implicitly determined by the inverse regression, could be useful in many situations. To discover how well the kernel-density approach will work in general will require the study of other examples.

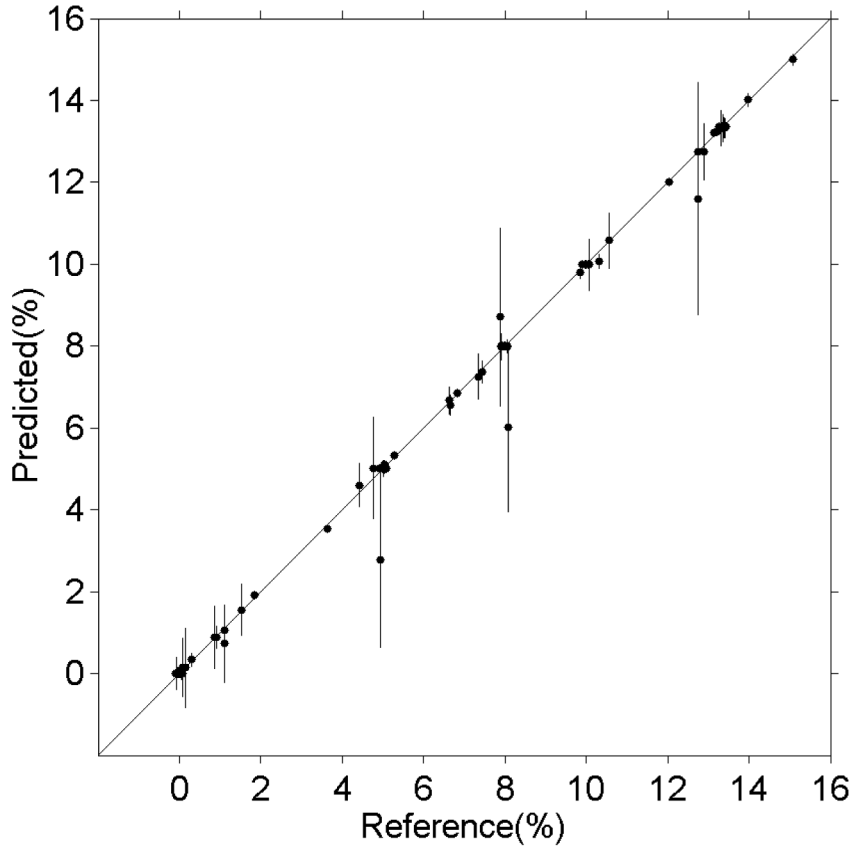

Figure 4. Predicted versus reference for % sunflower in 100 validation samples using an empirical prior distribution $p(y)$ and a kernel density estimate for $p(x|y)$. The error bars are ± 2 standard deviations, based on the distribution $p(y|x)$.

Reference

1. T. Fearn, D. Pérez-Marín, A. Garrido-Varo and J.E. Guerrero-Ginel, *J. Near Infrared Spectrosc.* **18,** 27 (2010).