Honigs regression, LOCAL and PLS: near infrared of dry and fresh forage

David Honigs[1]* and Peter Åberg[2]

[1]Perten Instruments Inc., 6444 South 6th Street Rd, Springfield, IL 62625, USA
[2]Perten Instruments AB, P.O. Box 5101, SE-141 05 Kungens Kurva, Sweden
*Corresponding author: dhonigs@perten.com

Introduction

The Honigs Regression (HR) has been mentioned previously but not fully described[1] as the owner, Perten Instruments, has chosen to keep the exact algorithm confidential. Still, it is possible to describe the general behaviour of this regression in terms of its inputs and outputs. The purpose of this study is to discuss this form of learning regression and review its results in the case of mixed dry and fresh forages from multiple sources. The Honigs Regression will be compared to PLS1 and the LOCAL regression approach (one calibration per sample).

Nomenclature

Learning regressions or machine learning is a relatively new field. It has several different nomenclatures depending on the particular problem being studied. In this paper, the following definitions are used, some of which are adapted from Reference 2:

Learning – Results are improved by experience
Neighbours – Samples with similar spectra
Supervised Learning – The case in which training pairs of spectra (x) and reference data (y) are known
Unsupervised Learning – The case in which only the spectra (x) are available
Eager Learning – The case in which the entire regression is changed with new training data
Lazy Learning – The case in which only neighbourhood changes are made with new training data

Materials and Methods

Spectra were collected using a Perten Instruments DA7200 instrument. Data were collected from 950nm to 1650 nm. The spectra were pre-treated by harmonisation (proprietary technique; Perten Instruments, Sweden)

Monogastric samples

The monogastric data contain NIR measurements and reference values from 10961 samples. The sample types include many varieties of poultry, swine, equine and fish feed from many locations including most of Europe, China and the United States.

Monogastric validation

In addition to the calibration and validation set a series of 112 samples (16 samples each from 7 specific site groups) were held out of the calibration set. These 112 samples were added one at a time to the library and the validation process was repeated to develop the error as a function of training samples. No samples from those 7 specific sites were included in the calibration set.

Forage samples

The forage data contain NIR measurements and reference values from 3499 samples. The sample types include alfalfa, alfalfa clover, corn silage, green forage, hay, haylage and meadow fescue. Samples originated from regions in Europe, China and the U.S.

Forage tuning, training and validation

The data set was divided into a calibration set with 2799 samples and a validation set with 700 samples. Not all calibration or validation samples have values for each analyte. The PLS and HR models were double cross-validated using the calibration data. The outcome of the inner cross-validation loop was used to tune the number of PLS factors, number of neighbours, and weight factor.

Reference paper as:
D. Honigs and P. Åberg (2012).Honigs regression, LOCAL and PLS: near infrared of dry and fresh forage, in: Proceedings of the15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 313-316.

Possible outliers in the calibration data were identified from the results of the outer loop using 98 percentiles of the prediction error, according to:

yErr = max(abs(yMeasured - yPredicted))
possibleOutlier = yErr > 98 percentile of yErr

Tuning of the number of PLS factors for each of the LOCAL sub-models was made using regular 8-folds cross-validation. The number of factors used was defined as the minimum RMSECV from one to twenty factors. The number of neighbours for the LOCAL models was not tuned, and a fixed number of 128 was used for all LOCAL analyses.

Results and Discussion

The Honigs Regression is a learning regression. Specifically, it is an example of a supervised-learning system using the lazy-learning approach. To develop a HR, one first develops a calibration from calibration samples in much the same manner as traditional PLS1. That calibration is applied to future samples by gathering neighbourhood training spectra (called library spectra), applying the calibration and making a neighbourhood-based learned modification. The calibration, specifically the beta coefficients of the regression, is kept constant in HR. A library of training samples is used to adapt the calibration to the specific circumstances. The learning behaviour of the HR is based in part on what some call the Manifold Assumption;[2] NIR data for a specific locality, region or crop year frequently lie on a lower-dimensional manifold of the entire data space. When this is true, one can see learning by simpler algorithms that avoid the curse of dimensionality. In applied mathematics, the curse of dimensionality refers to the fact that some problems become intractable as the number of the variables increases.[3] In this case it also means that it is easier to make and adapt limited calibrations rather than make a calibration that is robust to all possible future changes and effects.

The effect of HR learning via new library samples can be seen in Figures 1 and 2. These figures show the root-mean-square error improvement for a protein in monogastric feed calibration as additional training samples are added to the library. The left-most data point on each graph is the measured error using only the calibration sample set. This is plotted at 0 samples (technically 0 additional samples beyond the calibration set). Then, example spectra and lab values are added to the library one by one. The resulting RMSE can be seen to decrease relatively rapidly for even small numbers of added library samples. In the case of Figure 1, initial calibration results were poor but improved by more than 50% with only 16 added library samples. In the case of Figure 2, the error in protein prediction started off much better but still improved considerably. This learning effect is immediate (as soon as a new example spectrum (x) and constituent concentration (y) are added to the library). Nevertheless, HR uses the same calibration for prediction, making sure that the previous validation or certification of that calibration still holds. Both Figure 1 and Figure 2 use the same calibration.

The rapid learning nature of HR can be difficult to comprehend as it is contrary to past experience. If one adds a single new sample to a calibration set of over 10,000 samples, the conventional wisdom is that there will be little or no change in the calibration. However, adding as little as one sample, 100 parts per million of the data set, has a noticeable effect and improvement in the HR results. In this case, 16 added samples represents about a 1.5 parts per thousand additions to the library. The effect of such small relative additions of information is an example of the power of the lazy-learning approach.

Forages provide an example in which the data do not necessarily lie on a lower-dimensional manifold. The water present in wet forages is chemically different from the water in the dry forages. Water is still the same molecule, but the nature of hydrogen bonding and water activity are much different in those two different sample matrices. This problem is a challenging test case for HR.

Results of the forage test and the corresponding RMSEPs are given in Figure 3. This figure presents a summary of the improvements seen when HR and LOCAL are compared to PLS (0% baseline). It was found that the HR and/or LOCAL approaches improve the accuracy over PLS for at least 10 of the 12 parameters, up to a maximum improvement of 31% for the LOCAL regression for moisture when compared to PLS. In each case, LOCAL performed better than HR. For forages the data suggest that the importance of having multiple calibrations is more important than having a single calibration. However, this data is only for the initial calibration. More work is required to see if and how HR might begin to approach LOCAL as future learning samples are added. These results are case-specific. One would expect that a problem which behaves in a more linear manner might give different results.

Reference paper as:
D. Honigs and P. Åberg (2012).Honigs regression, LOCAL and PLS: near infrared of dry and fresh forage, in:
Proceedings of the15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley,
C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 313-316.

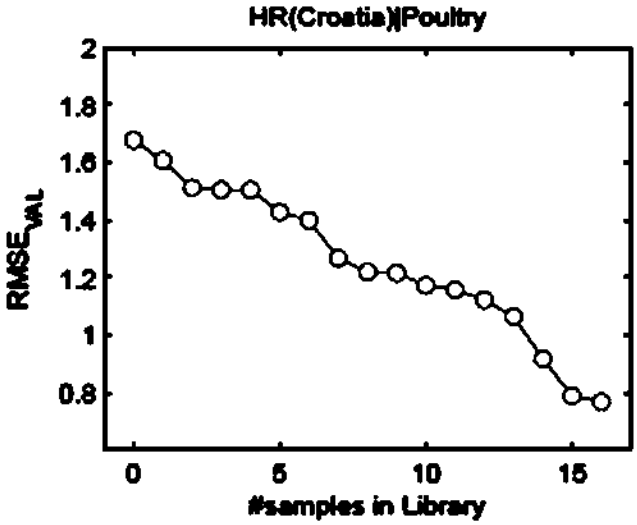

Figure 1. Protein error in Croatian poultry feed as a function of added library samples.

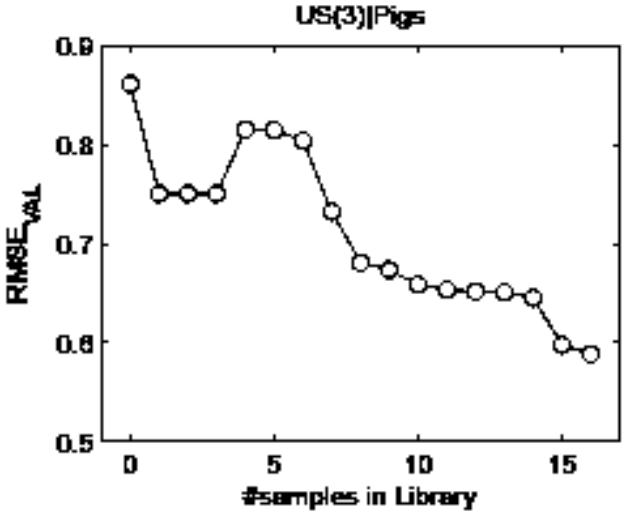

Figure 2. Protein error in US pig feed as a function of added library samples.

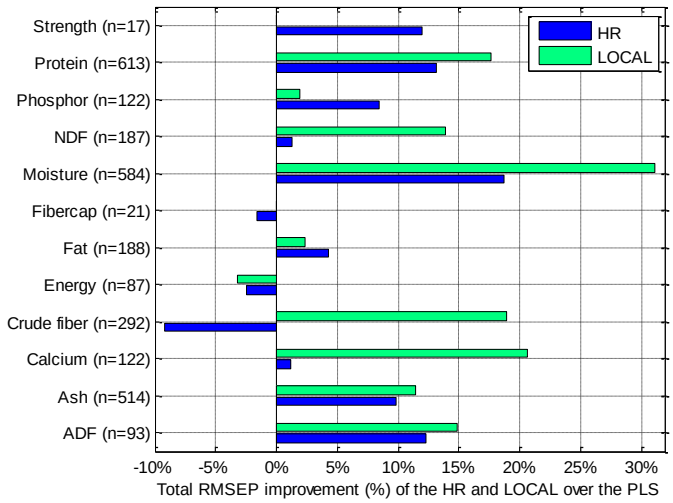

Figure 3. Total improvement (%) in RMSEP (Validation) for HR and LOCAL over PLS1.

Reference paper as:
D. Honigs and P. Åberg (2012).Honigs regression, LOCAL and PLS: near infrared of dry and fresh forage, in: Proceedings of the15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 313-316.

Conclusion

Learning regressions in general and the HR in particular hold promise for the automation or regularisation of calibration updates. This feature promises to be most powerful for users who maintain multiple instruments or instrument companies which maintain large groups of instrument calibrations. HR shows itself to be a quick study as it rapidly learns and adapts to new situations. HR is not a panacea. One still needs to pay attention to the chemistry and spectroscopy behind the calibrations themselves. Unlike LOCAL, HR still uses the expertise of an analyst to review samples, spectra and regression vectors. Unlike PLS, HR then leverages that information forward by learning and adapting to local circumstances without outside intervention. No single calibration technique is best for every situation. However, by broadening the view of the NIR calibration problem to recognise that it is a learning problem, one may be able to develop mechanisms and routines that go beyond fixed PLS1 equations which are so commonplace today.

References

1. B. Igne, P. Dardenne, D. Honigs, J.T. Kuenstner, K. Norris, Z. Shif and M. Westerhaus, *NIR news* **21**, 14-16 (2010).
2. O. Chapelle, B. Schölkoph and A. Zien (eds), "Introduction to semi-supervised learning" in S*emi-supervised learning*, The MIT Press, Cambridge, Massachusetts, p.1 (2010).
3. R.E. Bellman, *Adaptive Control Processes: a Guided Tour*. Princeton University Press (1961).

Reference paper as:
D. Honigs and P. Åberg (2012).Honigs regression, LOCAL and PLS: near infrared of dry and fresh forage, in: Proceedings of the15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 313-316.