Abstract

A new formulation for estimating the variance of model prediction

E. Fernandez-Ahumada, J.M. Roger* and B. Palagos

UMR ITAP, Cemagref, BP5095, 34196 Montpellier France
*Corresponding author: jean-michel.roger@cemagref.fr

Introduction

There are two basic ways of estimating prediction uncertainty, namely, error propagation or resampling strategies. Error propagation leads to closed-form expressions where some hypotheses are made but which provide a platform for evaluating the different sources of uncertainty. Resampling is essentially a "black box" approach which, however, is often more accurate because fewer assumptions and approximations are made. Some analytical expressions can be found in the literature for error propagation methods but all consider local linearisation and other important assumptions. In particular, errors in the predictors are assumed to be independent and to have constant variance. This latter assumption is never fulfilled in spectroscopy. So, this paper proposes a new expression for prediction uncertainty estimation based on the error propagation strategy, using as few assumptions as possible.

Let $\hat{y} = \mathbf{x}^T\mathbf{b}$ be a model. One of the most complete existing and published expressions of prediction uncertainty is:

A: $Var(\hat{y}) = \left(1 + \frac{1}{N}\right)\|\mathbf{b}\|^2\sigma_x^2 + \mathbf{z}^T\mathbf{Var(b)z} + \frac{\sigma_y^2}{N}$

With very few hypotheses, we calculate a new expression:

B: $Var(\hat{y}) = \left(1 + \frac{1}{N}\right)\mathbf{b}^T\mathbf{Var(X)b} + \mathbf{z}^T\mathbf{Var(b)z} + \mathbf{Var(z}^T\mathbf{b)} + \frac{\sigma_y^2}{N}$

Where: **z** is the spectrum centred against calibration set and **Var** represents the variance covariance matrix. The main differences between the two expressions are: (i) the first term is more general in expression B, as it takes into account the complete variance / covariance of **X**; (ii) the third term of expression B is new; it represents a kind of covariance between the variations of **z** and those of **b**.

Material and Methods

The terms of the two expressions were calculated on a dataset of N=385 x 10 repetitions of NIR spectra of feed, regressed against protein content.

Results and Discussion

Estimations of the prediction variance with expressions A and B were performed considering different pretreatments and no pretreatment of **X** data. Results show that expression A overestimates prediction variance, especially when no preprocessing is applied to data:

Expression A: $Var(\hat{y}) = 15.78$ (no pret.); $Var(\hat{y}) = 1.11$ (2der); $Var(\hat{y}) = 4.57$ (SNV);
$Var(\hat{y}) = 0.76$ (2der+SNV); $Var(\hat{y}) = 1.07$ (detrend);

Expression B: $Var(\hat{y}) = 0.57$ (no pret); $Var(\hat{y}) = 0.72$ (2der); $Var(\hat{y}) = 0.99$ (SNV);
$Var(\hat{y}) = 0.60$ (2der+SNV); $Var(\hat{y}) = 0.59$ (detrend)).

The overestimation is mostly due to the first term. Since expression A considers σ_x^2 instead of the complete variance/covariance of **X**, it does not take into account systematic variance, due for example to the baselines. It is also noticeable that the third term introduced by expression B is not at all negligible. Depending on the preprocessing, it represents between 11.6% and 19.2% of the total variance.

Conclusion

A new formulation for the estimation of the prediction variance, that is more suited to the mathematical specificities of the spectral data, is proposed.

Reference paper as:
Fernandez-Ahumada, E., Roger, J.M. and Palagos, B. (2012). A new formulation for estimating the variance of model prediction (abstract), in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy,
Edited by M. Manley,C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, p. 80.