doi: 10.1255/nirn.1438

Myth: A partial least squares calibration model can never be more precise than the reference method...

Guest-buster: Peter Paash Mortensen
Ms.Chem, Ph.D. in PAT and process sampling

The Mythbuster column is always watching out for myths, old or new... In this column we get help from a friend and colleague, Peter Paasch Mortensen, who is an experienced industrial chemometrician. Peter has worked for 15 years at Novozymes (Denmark) and is now with Arla Foods (Denmark).

This myth usually crops up in the wide field of *practical chemometrics,* when we care to validate a given multivariate calibration model. Here we skip our usual grumpy comments re *proper* validation; we shall here picture a calibrator who has accepted the basic principles of validation, which will then be based on independent test sets (test set validation).

The Mythbuster column has previously displayed control charts tracking total measurement uncertainties for a secondary method, in this case near infrared (NIR) spectroscopy. In the applied chemometrics world, it is often implicitly assumed that differences between predicted NIR results and reference values are mainly associated with uncertainties in the NIR prediction. However, such differences are always differences between two values, each with many error sources, which collectively add up to the total observed deviations. This always leaves the user with the question: which contains the major uncertainty, the NIR prediction or the reference measurements? Only by basing our analysis on solid statistical descriptions of both series and their origin (especially with respect to all measurement errors involved, i.e. sampling, processing, calibration, prediction etc.) can we allow ourselves to make firm conclusions that, for example, may involve changing the state of an industrial production process. Who would want to do such a thing on anything less than well-documented evidence?

Primary analytical methods are commonly accepted as solid evidence but how often do we challenge this general assumption? The validity of a primary analytical method must mean, among other things, that it is under full statistical control, which again means that it is associated with quantified systematic as well as random uncertainty estimates. It is often *assumed*, with very weak theoretical foundation, however, that systematic errors are negligible for a primary analytical method. But this flies in the face of all established experience in the analytical laboratory. There is always a systematic bias, the question of course is... is the bias small enough to be acceptable?

In this column we show that even severe uncertainty does not harm a partial least squares (PLS) model as long as it is random in nature. But if the primary method contains systematic error effects (analytical bias), or worse, if these are not stable (sampling bias), any attempt to calibrate and validate will suffer a breakdown and become unreliable. Assuming that an analytical bias has been brought under control, there remains the prediction precision which most certainly can be enough of a problem in itself. Although always a potential killer, we shall here leave out the specific sampling issues (much covered elsewhere), except by stating categorically that they are always to be ignored entirely at your own peril.[1]

To establish proof that a primary analytical method is free from bias, it must be specifically tested against well-defined standard materials whereby it can ultimately be related to appropriate international or national standards through an unbroken chain of comparisons (metrological validation). Even in this case, the uncertainty associated with estimation of the reference method bias remains an essential component of the overall measurement uncertainty. We here bring a mainly visual illustration.

This problem is quite simple to *simulate*. In Figure 1, Excel was used to generate two series of analytical results, both with an average concentration = 42, with a coefficient of variation (CV%) = 1 between the labs, which can be considered quite good:

$$STD = 0.418 => CV\% = 0.418/42 \times 100\% \sim 1\%.$$

These numbers are meant to illustrate repeated analytical results generated over a

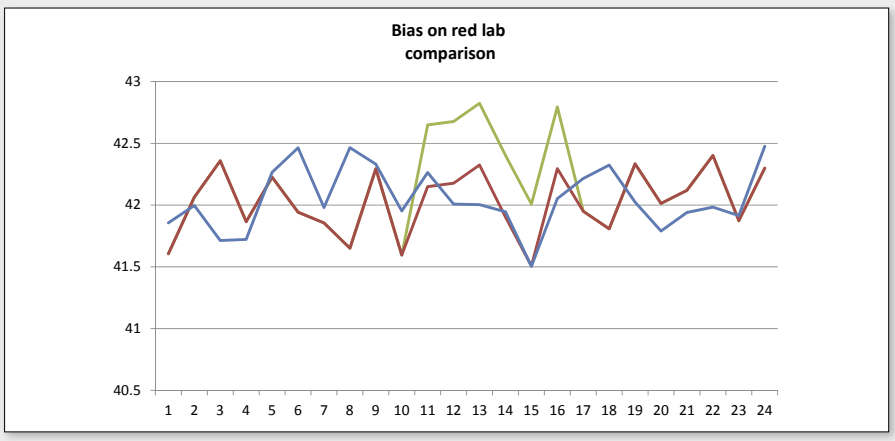

Figure 1. Blue lab and red lab analytical result series, based on duplicate samples extracted for control purposes. A bias of 0.5 has been added to the red lab results in the summertime (green period). Standard error of estimate (*SEE*) between labs (unbiased) is 0.32 but during the green subperiod it increases to 0.42. This is barely enough to detect the bias in most cases.

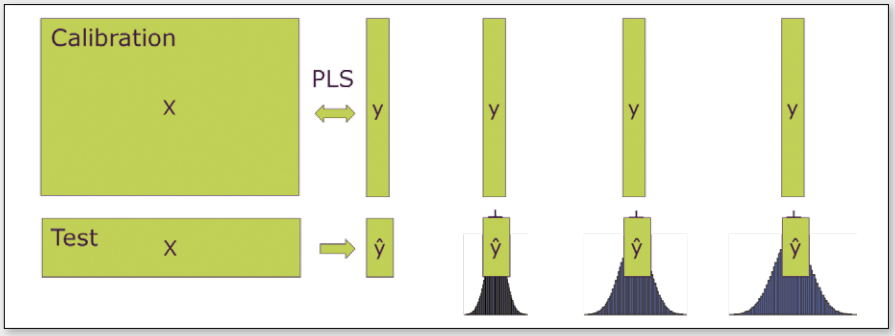

Figure 2. Adding increasing levels of symmetric noise to the same basic PLS model.

year stemming from two labs (red lab and blue lab). Duplicate samples are extracted every two weeks and both labs are then supposedly analysing the *same sample* (barring sampling issues). In the present simulations, a small bias of 0.5 to the results of the red lab has been added during the "green period". There is a very high possibility that this temporary bias would go unnoticed—it is here supposed to emulate a temporary aberrance for the blue lab during summer time.

Detecting a significant bias for a reference laboratory is not easy, but its impact on production, calibration and validation is easy to imagine. If we accept the possibility of a non-zero bias, this will lead to more careful control procedure interpretations when a mismatch between lab and NIR predictions is observed—i.e. it may in fact not only be NIR having the problem.

Another issue is repeatability, i.e. the ability to analyse/predict the same result repeatedly within an acceptable narrow range. It will often be the case that a NIR prediction method actually has better repeatability than the relevant primary laboratory reference methods, e.g. traditional wet chemistry methods. Can a NIR PLS model actually perform better in repeatability than the relevant primary method?

Let us build an instructive PLS model (X = NIR) with Y-data displaying zero bias but with various levels of noise added.

Experiment and simulations

The question at hand is how the intrinsic reference laboratory uncertainty will affect predictions and validation using PLS modeling. It was decided to test the Myth: "A calibration can never be more precise than the reference analysis" using realistic *simulation* by adding various, non-trivial levels of noise to the unbiased Y-data at levels of 2%, 5% and 10% *CV*, respectively. (A later Mythbuster column will illustrate adding identical noise levels as above but to distinctly *biased* Y-data.)

One hundred and thirty-six (136) data pairs (FT-IR spectra) of various pharmaceutical products was split into a calibration set of 114 samples; 22 samples were picked from a production run by stratified random selection to form the test set for the present simulations. The basic PLS model was based on full spectra [X] with an appropriate complexity (three PLS-components, no outliers present); the same basic model was used for all noise-added simulations below.

Figure 3 shows the calibration evaluations for the four models illustrated above in convenient predicted vs reference formats. For the specific comparison purpose needed here, one may use root mean square standard error of cross validation (*RMSECV*) [or one could use root mean standard error of prediction (*RMSEP*) based on the test—both options are illustrated here; Figures 3 and 4].

Adding noise to the Y data clearly diminishes the prediction performance of the

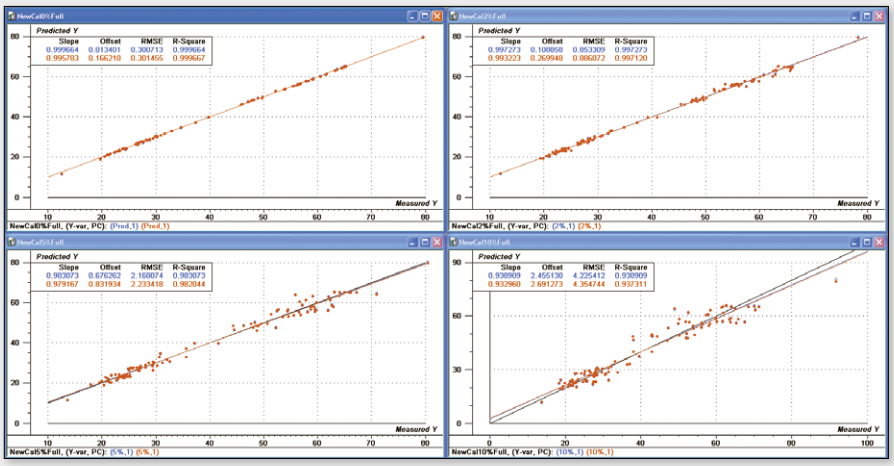

Figure 3. PLS predictions showing predicted vs reference values for the four models in Figure 2.

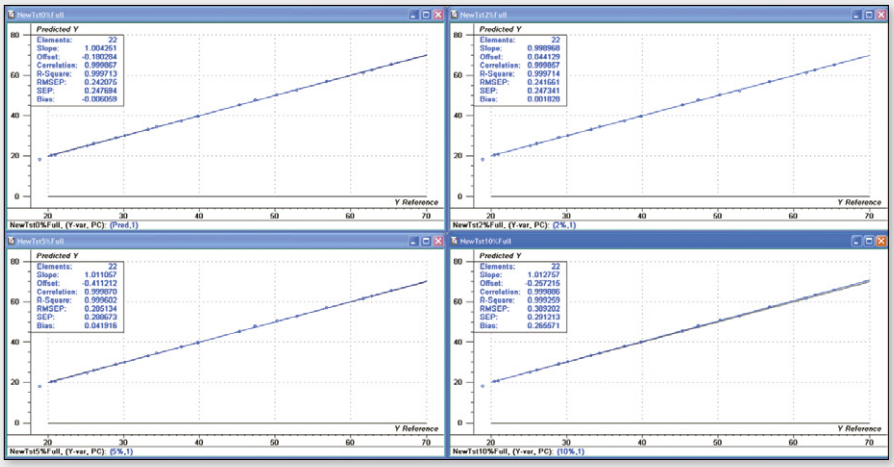

Figure 4. Prediction of the independent test set (22 samples) using the same four models depicted in Figure 2. Test set validation shows that *RMSEP* only increases from 0.24 to 0.38—expressed in the same units as the *RMSECV*.

model. *RMSECV* ranges from 0.30 (basic model, no noise added yet) to 4.36 for the highest noise addition (10%). Significant noise additions are corrupting the possibilities for detailed prediction precision.

The same four models were subsequently also tested on the independent test set (22 samples), with results shown in Figure 4.

The remarkable effect is that all model performances are now almost identical. The reason is obvious: the calibration consists of 114 data pairs (X-spectrum and corresponding Y-reference value) which tends to *average out* the added random analytical errors.

Conclusion

As long as the PLS model is built on data with an acceptably small analytical bias, prediction precision is actually smaller than that of a noise-prone Y-reference method. It may often provide robust and acceptable results even when the reference method is fraught with huge random fluctuations.

The take-home message from the guest buster is that PLS calibration is not sensitive to high laboratory uncertainties (random measurement noise). PLS models with reasonably strong X–Y correlation, to which is added *symmetric noise* (i.e. unbiased noise), reveal that this type of noise stabilises model building and results in less prediction uncertainty than that which characterises the Y-reference uncertainties themselves.

Myth busted!

P.S. A later column shall address the identical issue when played out against the background of both an analytical bias (*constant*) and one reflecting an *inconstant* bias (sampling bias). Readers can find a sneak preview of problems surrounding measurement error effects with and *without* proper attention to specific sampling issues in Reference 2.

References

1. DS 3077 (2013) DS 3077 Representative sampling—Horizontal Standard. Danish Standards. http://webshop.ds.dk/Files/Files/Products/M278012_attachPV.pdf

2. K.H. Esbensen and C. Wagner, "Theory of sampling (TOS) versus measurement uncertainty (MU)—A call for integration", *TrAC Trends Anal. Chem.* **57,** 93–106 (2014). doi: http://dx.doi.org/10.1016/j.trac.2014.02.007