Calibration Transfer, Part III: The Mathematical Aspects

June 1, 2013
Jerome Workman Jr.

Article **Spectroscopy**

Spectroscopy
Spectroscopy-06-01-2013
Volume **28** Issue **6**

Columns | Column: Chemometrics in Spectroscopy

Calibration transfer is a series of techniques used to apply a single spectral database, and the calibration model developed using that database, to two or more instruments. Here, we review the mathematical approaches and issues related to the calibration transfer process.

This column is a continuation from our previous two columns on the subject of multivariate calibration transfer (or calibration transfer) for spectroscopy. As we noted in the previous columns, calibration transfer is a series of approaches or techniques used to attempt to apply a single spectral database, and the calibration model developed using that database, to two or more instruments. Those instruments may be of like or different technical design. In this installment, we review the mathematical approaches and issues related to the calibration transfer process.

As we have discussed in this series of column installments (1,2), calibration transfer involves several steps. Basic reviews on the subject briefly described here are found in the literature (3–5). The basic spectra are initially measured on at least one instrument (that is, the parent, primary, or master instrument) and combined with the corresponding reference chemical information (that is, actual or reference values) for the development of calibration models. These models are

using the child instruments with minimal intervention. We note that the issue of calibration transfer disappears if the instruments are precisely alike. If the instruments are the "same," then one sample placed on any of the instruments will predict precisely the "same" result using the same model. Because instruments are not alike, and in fact change over time, the use of calibration transfer mathematics is applied to produce the best attempt at model or data transfer. As mentioned in the first installment of this series (1), there are important issues of attempting to match calibrations using spectroscopy to the reference values. The main principle is that spectroscopy measures the volume fractions of the various components of a mixture. The reference values may be based on one of several physical or chemical properties that are only vaguely related to this measured volume fraction. These include the weight fraction of materials, the volume percent of composition with unequal densities, the physical or chemical residue after some processing or separation technique, the weight fraction of an element found in a larger molecule (such as total nitrogen vs. protein), and other measured or inferred properties. The nonlinearity caused by differences in the volume fraction measured by spectroscopy and the reported reference values must be compensated for by using specific mathematics for nonlinear fitting during calibration modeling. This compensation often involves additional factors when using partial least squares (PLS), or additional wavelengths when using multiple linear wavelength regression (MLR).

Multivariate calibration transfer, or simply calibration transfer, is a set of software algorithms and physical materials (or product standards) measured on multiple instruments, and is used to move calibrations from one instrument to another. All the techniques used to date involve measuring product samples on the parent instrument and child instrument and then applying a variety of algorithmic approaches to complete the transfer procedure. Traditionally this has been accomplished by measuring 10–40 or more product samples for each constituent on both the parent and child instruments, comparing the average near-infrared (NIR) predicted value from the parent instrument predictions to the average predicted value from the child instruments, and then biasing the child instrument to the same average value as the parent instrument. Note this procedure is carried out for each product and constituent combination! After this has been accomplished the user of the child instruments may also compare the average NIR predicted values to their corresponding laboratory reference values for each constituent, and then again adjust each constituent model with a new bias value, resulting in an extremely tedious and unsatisfying procedure. This entire wearisome process is exacerbated if different preprocessing and calibration algorithms are used for each constituent

calibration or when one is attempting to transfer calibrations from spectrometers of different optical design.

The Mathematical Approaches to Calibration Transfer

A basic instrument correction must be applied to align the wavelength axis and photometric response for each instrument to make their measurement spectra somewhat alike. This process will create spectra and measurement characteristics that are most similar and repeatable. The correction (or internal calibration) procedure requires photometric and wavelength measurement reference materials that are stable over time and can be relied on to have accurate and repeatable characteristics. It is of paramount importance that the standards do not change appreciably over time. All instruments will change with time because of lamp color temperature drift, mechanical wear, electronic component and detector aging, and variations associated with the instrument operating environment, such as temperature, vibration, dust, and humidity.

In the process of transferring calibrations from a parent to a child instrument, one may take four different fundamental strategies for matching the predicted values across instruments. Each of these strategies varies in complexity and efficacy. One may adjust the calibration model (that is, the regression or b-vector), the instrument as it measures spectra (that is, the *x* and *y* axes), the spectra (using various spectral transformations, such as matching *x* and *y* axes and apparent lineshapes via smoothing), or the final predicted results (via bias or slope adjustments). All of these methods have been applied individually or in combination in an attempt to match the reported predicted results derived from parent and child instruments. Ideally, one would adjust all spectra to look alike across instruments, such that calibration equations all give the same results irrespective of the specific instrument used. This is the challenge for instrument designers and manufacturers, and a significant challenge it is.

What to Compare When Transferring Calibrations?

For spectroscopy-based measurement using multivariate calibrations, one may compare the standard or reference concentrations for a set of samples to the spectroscopy-based predicted values. One may also compare the response of the parent instrument to that of the child instrument. In making these comparisons, one may perform statistical tests for bias, correlation, and slope. A statistically significant difference in

bias should result in a change of the bias. A statistically significant result in correlation or slope should result in a basic multivariate recalibration, unless one can demonstrate that the differences between the compared values have some real slope variation between them because of fundamental scientific principles. In this column, we specifically address differences between parent and child instrument predictions and leave the reference laboratory values out of the discussion for now. We will address the issue of reference laboratory values vs. spectroscopy-based predicted values in the next installment of this series.

Virtual Instrument Standardization

The concept of virtual instrument standardization has been reported as successful and was demonstrated commercially (6–8). This technology was limited to a single instrument design and manufacturer and used a proprietary set of materials and algorithms. These methods and material definitions were never published in sufficient detail, and, thus, the exact elements remain a trade secret, except for information disclosed within an issued United States patent (9).

Bias or Slope Adjustments of Predicted Results Across Parent and Child Instruments

A significant bias between parent and child NIR predicted values results mainly from instrumental differences. Other sources of significant bias changes between reference values and spectroscopy-based predicted values are chemical or spectral interferences. These cause significant bias in the measured analyte concentration because of the effect of another component, property of the sample, or analytical measurement anomaly (10). We are addressing prediction bias as related to instrument differences in this discussion. There are more sophisticated tests for bias that will be addressed in future columns, but we introduce the topic here.

Bias (Mean) One-Sample t*-test Between Parent and Child Instruments*

As any statistician will tell you, the concept of mean differences (or bias) requires a test of significance to determine whether a bias adjustment should be made. An appropriate statistical test will tell us if the variation in the mean values (bias) between sample sets of predicted values is within the expected random fluctuation for a normally distributed population of measurements. The larger the sample set used to test calibration transfer bias, the more accurate the estimate of its value. Thus, using 20 test samples would give a more accurate

estimate of the bias than 10 samples. The smaller the standard error of the mean, the greater the confidence will be of the true bias value. The standard error of the mean (SEM) is given as equation 2 and so if we use 20 rather than 10 samples for bias checking, we will have a more powerful estimate of the true bias by a factor of $\sqrt{20}/\sqrt{10} = 1.41$. This is true only if there is no slope difference. Note: It is acceptable to make the statistically significant bias correction even if there is a slope difference.

The statistical test we will use to determine bias significance is a simple parametric one-sample *t*-test. For this test, we only need to have the average predicted mean value for a set of samples on the parent instrument and the set of predicted values for the same sample set on the child instrument. In the case of NIR data, we designate the mean NIR value for the set of reference samples measured on the parent (that is, calibration) instrument as $X(withbar)_{Parent}$. (For this test, we don't have all the measurements from the parent instrument, only the mean value is known.) We then compute the mean NIR value for the same set of reference samples measured on the child (that is, transfer) instrument as $X(withbar)_{Child}$. For this test we are setting our test hypotheses as:

We then compute the standard deviation for the child instrument (s_{Child}) NIR predicted data on the measured set of samples as follows. Where $\bar{y}_i$ is the mean predicted value for the sample set for the child instrument; and y_i values are the individual predicted values for the set of test samples for the child instrument.

We then compute the SEM for the child instrument NIR data as

Here, we are testing whether the predicted value mean is statistically the same for the parent and child instruments. There are many variations of testing mean differences, but this is a basic test for comparing means when the sample size of two groups is identical. This test determines whether the average predicted values are statistically the same for the test set used. We note that the reference values are not assumed to be known for the test set. For this *t*-test statistic, we are able to compute our *t*-test for mean bias significance as

If this resulting *t*-test is greater than the *t* critical value for $n - 1$ degrees of freedom, then we know the bias is significant and it should be changed. If the *t* value computed is less than the critical value of *t*, it is not significant and should not be changed. If the bias is significant, we can test the difference between the slope of the two lines or by comparing their correlation coefficients. Note that *n*, in this case, is the number of samples in the test set, for example, 20.

Bias (Mean) Two-Sample t*-test Between Parent and Child Instruments*

If we know the predicted values for the set of test samples from both the parent and child instruments, we could use a parametric two-sample *t*-test. Note that we are using an identical number of samples for each bias test, so the sample sizes for each test of the parent and child instruments are identical. We do not assume that we have the reference values for the set of test samples. For this test we compute the mean NIR value for the set of reference samples measured on the parent (that is, calibration) instrument as $X(withbar)_{\text{Parent}}$. We then compute the mean NIR value for the same set of reference samples measured on the child (that is, transfer) instrument as $X(withbar)_{\text{Child}}$. We set our test hypotheses the same way as in the one-sample *t*-test:

We then compute the standard deviations for both the parent (s_{Parent}) and child (s_{Child}) NIR predicted data on the measured set of samples as follows. Where $x(withbar)_i$ and $\bar{y}_i$ are the mean predicted values for the sample set for each instrument; and x_i and y_i are the individual predicted values for the set of test samples for the parent and child instruments, respectively.

Then we are able to compute our *t*-test for mean bias significance as

If this *t*-test value is greater than the *t* critical value for $n_c + n_p - 2$ degrees of freedom, we know the bias is significant (we accept H_A), and thus, the bias should be changed. If the *t* value computed is less than the critical value of *t* it should not be changed and is not significant (we accept H_o). If the bias is

significant, we know to test the difference between the slope of the two lines or to compare their correlation coefficients.

*Comparing the Correlation Coefficients Between Parent and Child Instruments Using the (*r*-to-*z *Transform) Significance Test*

We note that the slope should not be changed to adjust the predicted values following calibration transfer. However, the slope significance should be computed as an indication of a need to recalibrate the instrument using a new multivariate model. For this test, we compare the correlation coefficients of the sample set for the parent instrument to the correlation coefficient of the same sample set on the child instrument; this is Pearson's *r* statistic.

To compare the correlation coefficients between the parent and child instrument NIR values to test whether there is a significant difference between them, we regress the parent NIR values (as the *x*, reference, or independent values) against the child NIR predicted values (as the *y*, or dependent values). For the resulting correlation, expected to be near 1.0, we actually use a value of $r = 0.995$. Our hypothesis tests for a suitable correlation between the parent and child instrument predictions is given as

With this set of hypotheses tests we use the following to compute the test statistic

This Z_{Test} value is computed and compared for alpha = 0.05, and yields a Z critical value of 1.96. So if our computed Z_{Test} statistic is greater than 1.96, we accept the alternate hypothesis (H_A: $r \neq 1.0$). This indicates that the instrument predictions are not alike and the actual correlation is different between the child and parent instruments. If the $Z_{Test} < 1.96$ the correlations are alike and assumed to be near 1.0, so the null hypothesis is accepted (H_o: $r = 1.0$). If the null is accepted, we state that the correlation is essentially the same (or 1.0) for the child vs. parent predicted values.

So, how much variation is acceptable if we are using $n = 20$ samples as our test set to compare the predictions on the parent vs. child instruments? Table I is based on the expected match of the instruments. This is at alpha = 0.05, which is a typical test level.

Table I: ZTest of significance for r of 0.995 between parent and child predicted values

If one expects a more rigorous comparison test in which the minimum correlation between the parent and child instruments is tested to 0.999 then Table II applies.

Table II: ZTest of significance for r of 0.999 between parent and child predicted values

Use the URL http://www.vassarstats.net/tabs.html to compute your Z_{Child} values from your r value obtained by comparing the parent to child predicted values. From this URL, use the Fisher r-to-z transformation tab. Enter your n number for the test set (for example, 20) and the r value you are testing. Next, substitute this value into equation 7 for Z_{Child}. The Z_r values for each test are given in Tables I and II for your test level of minimum r. Note that n_c and n_p are the number of samples in your test set (for example, n = 20). From Table I we note that if we expect a correlation of 0.995 between instruments, then a correlation below 0.980 is significant and indicates a problem. If we expect a correlation of 0.999 between instruments, a correlation below 0.996 indicates a significant difference (Table II). This is a better test to use when comparing the parent to child instruments, instead of the simpler, but weaker, one-sample and two-sample t-tests described earlier. The one- and two-sample t-tests were included for illustrative purposes, but the regression test preferably should be used in all cases.

Slope Significance Limit Test Between Parent and Child Instruments

If we are expecting a child instrument to perform identically to the parent instrument for predictions, we should compute the confidence limits for comparison between the parent and child predicted values using criteria computed only from the parent instrument. So, for a slope significance test we need to look at the slope change between the parent instrument predicted values for a set of test samples over time and retain these results for this test. To accomplish this we would compute the predicted values for our set of 20 test samples on the parent instrument one week apart. Thus, we would have two sets of predicted values designated by different measurement times as x_i for time 0 (the reference values) and for the same set of samples measured one week later designated as (y_i). (Note that

these test samples must be chemically stable during the one-week period.)

For the computation of the acceptable slope confidence interval (C.I.) of the child instrument to be considered alike to the parent instrument, we need to compute three basic sets of numbers: the set of parent predicted values at time 0 (x_i), the set of parent predicted values at week one (y_i), and the set of predicted values from the simple linear regression between these x_i and y_i values, designated as ($\hat{y}_i$). Note that n is the number of x_i, y_i pairs (that is, 20 for this example). From these three sets of values we can compute the standard deviation of the residuals for the predicted values, from the regression ($\hat{y}_i = b + mx_i$) as

We then compute the standard deviation for the desired slope as

And now the confidence limit for the slope is given as

Where t is the critical value of the t distribution, two-tailed test, at alpha = 0.05 and with degrees of freedom as n — 2 = 18. This value is 2.10 and can be found in any table of the critical values of the t distribution. So, let's look at an example of data for a diverse ground wheat protein test set measured on the parent instrument at time 0 and at a one-week interval:

Time 0 (x_i): 12.4, 12.9, 14.0, 16.0, 13.2, 12.8, 14.5, 13.0, 13.6, 12.7, 14.2, 16.3, 17.8, 18.0, 14.5, 17.2, 14.4, 15.2, 16.6, 13.5

Time week one (y_i): 12.2, 12.5, 14.4, 16.3, 13.0, 12.4, 14.8, 13.3, 13.3, 12.4, 14.5, 16.6, 17.9, 18.3, 14.2, 17.0, 14.1, 15.4, 16.7, 13.2

The results of the regression analysis with the confidence intervals are shown in Table III.

Table III: Simple linear regression results for time 0 vs. one week (parent instrument)

The results indicate that the test criteria for testing a child instrument slope following calibration transfer would be within

the confidence limits of the test measured on the same parent instrument over this time interval. To complete this parent and child comparison one would designate the parent instrument results as x_i and the child measured results as y_i. Next, the slope would be computed for the regression and must fall within the computed confidence limits for the time 0 and week one parent tests. This is a test of equivalence in slope between parent and child predicted values. We note that the literature also includes such computations (11).

Developing Global or Robust Models Including Variation Between Instruments

Various methods have been proposed to produce the universal model or a calibration that is mostly robust against standard instrument differences or changes with time that are common to commercial instruments today. These are referred to as robust or global models. For computing a robust model, various experimental designs have been constructed to better represent the product, reference values, and instrument calibration space and to include typical changes and interferents that should be included within the model for it to be broadly applicable. Using this approach, one might design a factorial experiment for the composition of the calibration set to include multiple variations typically encountered during routine analysis. A list of some of these variations may consist of differences in sample pathlength, sample cup, sample temperature, sample moisture content, flow rate, particle size, interferent content, instrument type, constituent ratios, sampling parameters, and others (12). These approaches will work for a period of time until the instrument drifts or the product or constituent chemistry changes. These types of changes are expected and, thus, routine recalibration (that is, model updating) would be required as a standard procedure if the chemistry of the instrument changes are considered significant. A method for selecting specific robust wavelengths in MLR models that are more forgiving toward wavelength differences in interference filter–based instruments have been demonstrated as effective (13).

A method to reduce the effect of interference on NIR measurements has been demonstrated. This is a preprocessing method termed orthogonal signal correction (OSC). The goal of OSC is to remove variation from the spectral data, **X**, that are orthogonal to **Y** (14). This orthogonal variation is modeled by additional components for **X** and results in the decomposition, $X = t \cdot p' + t_o p_o' + e$, where t_o and p_o represent the scores and loadings for the orthogonal component and e represents the residual. By removing the **Y**-orthogonal variation from the data via $X - t_o p_o'$, OSC maximizes correlation and covariance

between the **X** and **Y** scores to achieve more accurate NIR predictions (15).

Augmenting Models Over Time

If instrument differences are significant, the predicted values between parent and child instruments will be unacceptably large. In these cases, one may begin to collect more samples on the child instrument and then recompute the multivariate model using the majority of high leverage sample data from the new instrument. This, in effect, uses the calibration transfer as a temporary solution or bridge to building an accurate model on the child instrument. Such a practice is often used when the instrument differences are too significant for direct and accurate calibration transfer.

Sample Selection to Improve Spectral data

Sample selection methods have been used and perfected since the beginning of chemometric methods and spectroscopy. There are many methods and a variety of nomenclatures for these techniques. The purpose of such methods are to remove the redundancy in spectral data so that the most repetitive samples do not have excessive influence on the regression model. This provides a basis so that the regression line is more appropriately fitted to extreme samples, including those with high and low analyte concentrations. Such methods of sample selection include random subset selection, manual subset selection, spectral subtraction methods for "uniqueness" tests, stratified sample selection, discriminant-based selection techniques using spectral distances, correlation matching techniques, and others. These methods are described in more detail in the literature (3). One of the first successful approaches for sample selection was a process that used spectral subtraction to remove the unusual spectra from all of the other spectra (16). Even early on, these methods significantly improved the standard error of prediction (SEP) for multiple constituents in forage analysis (17).

Spectral Data Transformation

This process consists of altering spectral data from the child instrument to be more like that measured on the parent instrument. Piecewise direct standardization (PDS) has been used most often for this procedure (18).

Special Standardization Mathematical Approaches

This topic was discussed and referenced in a previous column installment (2).

Local Methods

Locally weighted regression or local regression methods use spectral data and corresponding reference data to build a "local" calibration using only those samples near the unknown or test sample spectra. For example, the unknown spectrum is measured and the sample spectra most like the unknown are selected from a resident database. The multivariate calibration model is then computed using only the local samples. The samples can be down-weighted for use in the regression model based on distance from the unknown sample. This approach allows quite accurate prediction analysis when a variety of samples and instrument type data is incorporated into a spectral database. The first description of the use of this method for spectroscopy is based on original work from the statistics community (19,20).

Use of Indicator Variables

A method has been previously used that simultaneously optimizes the calibration for multiple instruments and provides *t*-tests for the differences between them. The method creates the calibration by running the samples on several instruments (the more the better). All the data are added into the calibration and indicator variables are used between the instruments. With this approach you obtain the best calibration that is optimized for all instruments, the coefficients of the indicator variables are the biases between the instruments, and the *t*-tests you report from this method are valid for the corresponding bias values. We will discuss this method in greater detail in the future.

Summary

There are many conventional and unconventional approaches to calibration transfer. However, the fact remains that significant differences in the instrument response between parent and child instruments causes the greatest variation in predicted results following calibration transfer. If instrument spectral profiles can be made statistically alike between instruments, then the transfer issue disappears. The additional challenges of relating specific reference laboratory results to results predicted using spectroscopy is another ongoing area of discovery and represents yet another problem still to be solved.

References

(1) H. Mark and J. Workman, *Spectroscopy* **28** (2), 24–37 (2013).

(2) H. Mark and J. Workman, *Spectroscopy* **28** (5), 12–25 (2013).

(3) J. Workman, P. Mobley, B. Kowalski, and R. Bro, *Appl. Spectrosc. Rev.* **31** (1&2), 73–124 (1996).

(4) P. Mobley, B. Kowalski, J. Workman, and R. Bro, *Appl. Spectrosc. Rev.* **31** (4), 347–368 (1996).

(5) R. Bro, J. Workman, P. Mobley, and B. Kowalski, *Appl. Spectrosc. Rev.* **32** (3), 237–261 (1997).

(6) J. Workman and L. McDermott, *JPAC* **2,** 444 (1996).

(7) J. Workman and J. Coates, *Spectroscopy* **8** (9), 36–42 (1993).

(8) J. Workman, in Proc. 2nd Oxford Conference on Spectroscopy, Franklin Pierce College, Rindge, New Hampshire, June, 1994.

(9) D. Tracy, R. Hoult, and A. Ganz, U.S. Patent No. 5,303,165, 1994.

(10) Interference Testing in Clinical Chemistry; Approved Guideline — Second edition, CLSI document EP7-A2 (Pennsylvania, 2005).

(11) H. Mark, *Principles and Practice of Spectroscopic Calibration* (John Wiley & Sons, Inc., Hoboken, New Jersey, 1991), pp. 152–156.

(12) D. Abookasis and J. Workman, *J. Biomed. Opt.* **16** (2), 027001-027001-9 (2011).

(13) H. Mark and J. Workman, *Spectroscopy* **3** (11), 28–36 (1988).

(14) S. Wold, H. Antti, F. Lindgren, and J. Ohman, *Chem. Intell. Lab. Syst.* **44,** 175 (1998).

(15) H.C. Goicoechea and A.C. Oliveri, *Intell. Lab. Sys.* **56,** 73 (2001).

(16) D.E. Honigs, G.M. Hieftje, H.L. Mark, and T.B. Hirschfeld,

Appl. Spectrosc. **57,** 2299 (1985).

(17) J. Workman, in Proc. 1986 Forage and Grasslands Conference, Athens, Georgia. American Forage and Grass. Council, Lexington, Kentucky, April, 1986.

(18) B.M. Wise, *Process Control Qual.* **5,** 73 (1993).

(19) T. Naes, T. Isaksson, and B. Kowalski, *Anal. Chem.* **62,** 664–673 (1990).

(20) W.S. Cleveland and S.J. Deviin, *J. Am. Stat. Assoc.* **83,** 596–610 (1988).

Jerome Workman, Jr. serves on the Editorial Advisory Board of *Spectroscopy* and is the Executive Vice President of Engineering at Unity Scientific, LLC, (Brookfield, Connecticut). He is also an adjunct professor at U.S. National University (La Jolla, California), and Liberty University (Lynchburg, Virginia). His e-mail address is JWorkman04@gsb.columbia.edu

Jerome Workman

Howard Mark serves on the Editorial Advisory Board of *Spectroscopy* and runs a consulting service, Mark Electronics (Suffern, New York). He can be reached via e-mail: hlmark@nearinfrared.com

Howard Mark