⟨1039⟩ CHEMOMETRICS

1. INTRODUCTION

1.1 Scope and Purpose

This chapter provides guidance regarding scientifically sound practices for the chemometric analysis and interpretation of typical multivariate data for compendial and industrial applications. Established chemometric practices, including calibration and validation, for applications using different analytical technologies (e.g., spectroscopic, chromatographic, and others) and for different purposes (e.g., fingerprinting, identification, classification, properties prediction, and others) are discussed under a lifecycle approach. Both qualitative and quantitative applications are described.

The chapter discusses how method quality and performance are ensured through the proper lifecycle management of a chemometrics-based model, including the selection of appropriate algorithms, calibration, validation, verification, transfer, and ongoing maintenance steps.

This chapter may be viewed as a supplement to other guidance chapters such as *Analytical Data—Interpretation and Treatment* ⟨1010⟩, which are mainly concerned with the analysis and interpretation of univariate data.

[NOTE—It should not be inferred that the multivariate analysis tools mentioned in this chapter form an exhaustive list. Other equally valid models may be used at the discretion of the manufacturer and other users of this chapter.]

1.2 Content Summary of Document

The mind map below (*Figure 1*) provides a visual representation of the content of this chapter. This diagram is meant to assist the reader by showing how the various concepts and practices of chemometrics relate to each other.

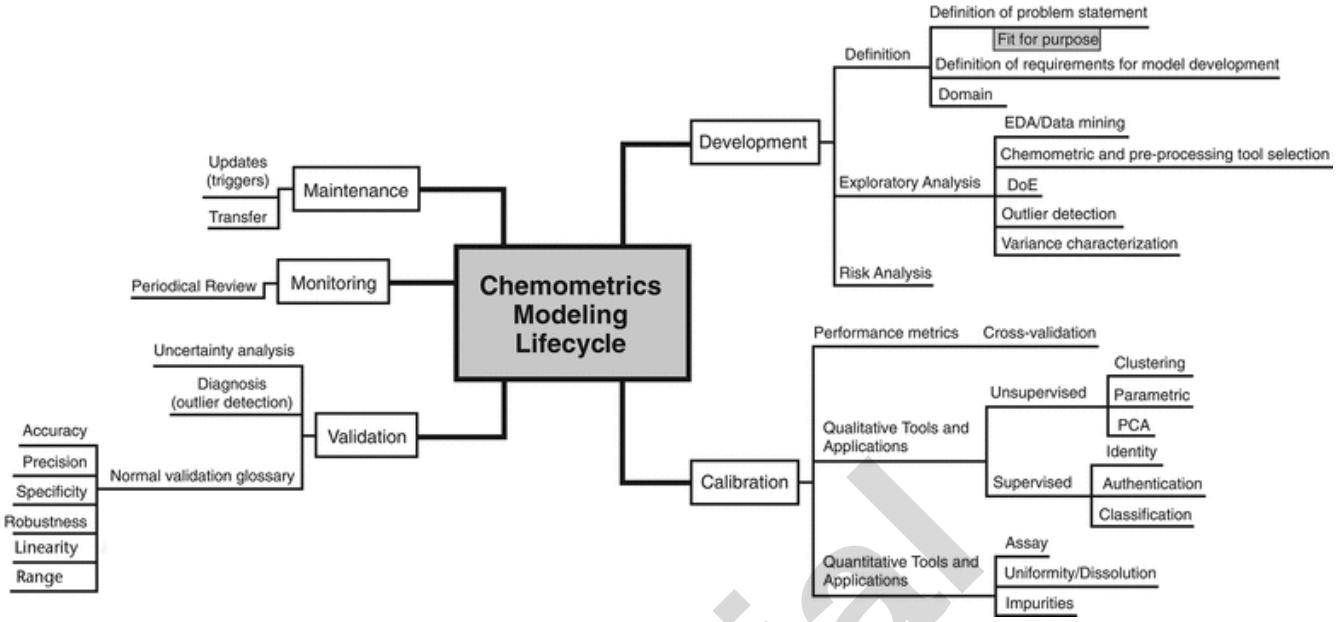

Figure 1. Content summary of document.

1.3 Audience

This chapter provides direction for both the chemometrician applying chemometric techniques to develop a model for a given application and the analyst that runs the model within an analytical procedure. The chemometrician will find support regarding algorithm selection for a given application and guidance for ensuring model performance throughout its lifecycle. The analyst will gain insight regarding the strengths and limitations of the chemometrics techniques as applied to support their applications.

2. WHAT IS CHEMOMETRICS?

Chemometrics was originally defined as the chemical discipline that uses mathematical, statistical, and other methods that employ formal logic to accomplish two objectives: 1) to design or select optimal measurement procedures and experiments, and 2) to provide the maximum amount of relevant chemical information by analyzing chemical data. (*1*,*2*) More specifically, "chemometrics" has come to mean the application of multivariate methods for the analysis of chemical or related data, although the algorithms in question may be used to extract information out of almost any measured data, regardless of origin—chemical, physical, biological, pharmaceutical, or others. This chapter does not focus on the development of optimal procedures or methods and the applied design of experiments (DoE), but rather on the analysis of multidimensional data collected from an analytical instrument, such as spectroscopic and chromatographic data. A multivariate data set (i.e., multivariate data obtained for a number of samples or objects) forms an $m \times n$ data table or matrix. The matrix is represented by **X**, with m as the number of samples or objects and n as the number of variables measured for each sample (*Figure 2*).

Consequently, the data analysis techniques considered in this chapter will be multivariate in nature. Depending on the purpose of the data treatment, different tools will be applied. Initially, the data handling techniques can be divided into two categories: unsupervised and supervised. The unsupervised tools use only the **X** matrix to extract information, but in supervised data analysis, in addition to the **X** matrix the samples are also described by a **y** vector. This is an $m \times 1$ table containing property information for each sample (e.g., concentration, enzyme inhibition activity). The supervised data analysis techniques are used to build a model between the **X** matrix and the **y** vector. In chemometric modeling, the equations provided are data driven to empirically describe the underlying variance in the data for either unsupervised or supervised purposes. The different techniques or tools applied for the different purposes are discussed in more detail in *4. Applications of Chemometrics*.

The most commonly used unsupervised technique (principal component analysis, or PCA) and supervised technique (partial least squares regression, or PLS) are by nature latent projection approaches (see *Figure 2*), which transform a large number of potentially correlated **X** variables, such as intensities at different retention times or wavelengths, into a possibly smaller number of uncorrelated variables (principal components, or PCs; latent variables). As can be seen in *Figure 2*, the original *n*-dimensional space was transformed to a two-PC space. When samples from the original data space are projected onto this lower-dimensionality space, the resulting sample coordinates are called scores (**T**). Visualization of the scores along two or more PCs forms a score plot that contains information on the relationships among different samples. The loadings (**P**) are a linear combination of the original variables and the coefficients or weights used in this linear combination. The loadings for specific latent variables also can be plotted in what is called a loading plot. The loading plot contains information on the relative importance among the original variables. Both scores and loading plots enable visualization and understanding of the underlying data structure (i.e., the presence of groups/clusters and/or outliers) within a reduced dimensional space. The variable matrix (**E**) produced by the model is defined as the residual error. The sample information along axes of common variance is

captured by the model's PCs. Variance unaccounted for by these PCs (residual error) is left for each sample at each variable, forming the residual matrix (**E**).

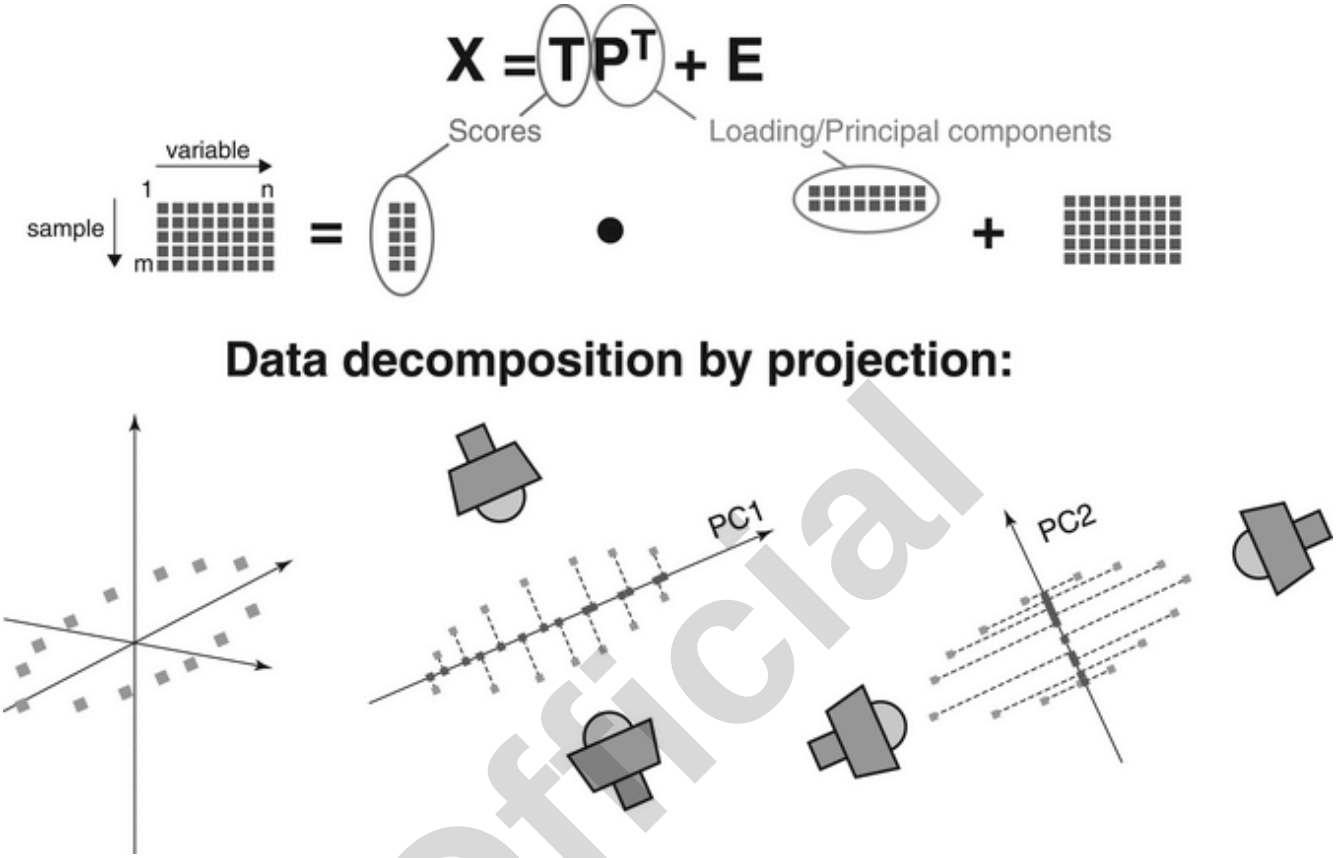

Figure 2. Schematic representation of latent projection techniques.

3. MODEL LIFECYCLE

The development of a chemometric model, as part of an analytical procedure, aims to fulfill a predefined, intended purpose. The intended purpose is typically a statement describing what to measure, the output format, and the level of performance needed with the result, and should be in accordance with the analytical target profile (ATP). The ATP may specify performance criteria for key characteristics of the analytical procedure output and/or decision risk probabilities expected in routine use of the analytical procedure.

Calibration of the model encompasses an iterative process that involves selection of the sample set to be used to develop the chemometric model, tuning of an appropriate chemometric algorithm with the necessary preprocessing algorithm, and evaluation of model performance according to predefined metrics for the ATP. During the validation stage, the method performance is demonstrated to fulfill the intended purpose and ATP. Both the knowledge gained during calibration and the assessment of specific metrics and corresponding limits are used to evaluate performance and define a method maintenance protocol that will be used for the monitoring stage, before deployment to routine use. Changes that may have an impact on model performance should trigger a defined set of activities to update it. Model update may also be triggered by the necessity of performing method transfer, in the course of its lifecycle, to different equipment or another site. The extent of the revalidation will be defined by the magnitude of the model update. *Figure 3* shows a schematic representation of the described workflow.

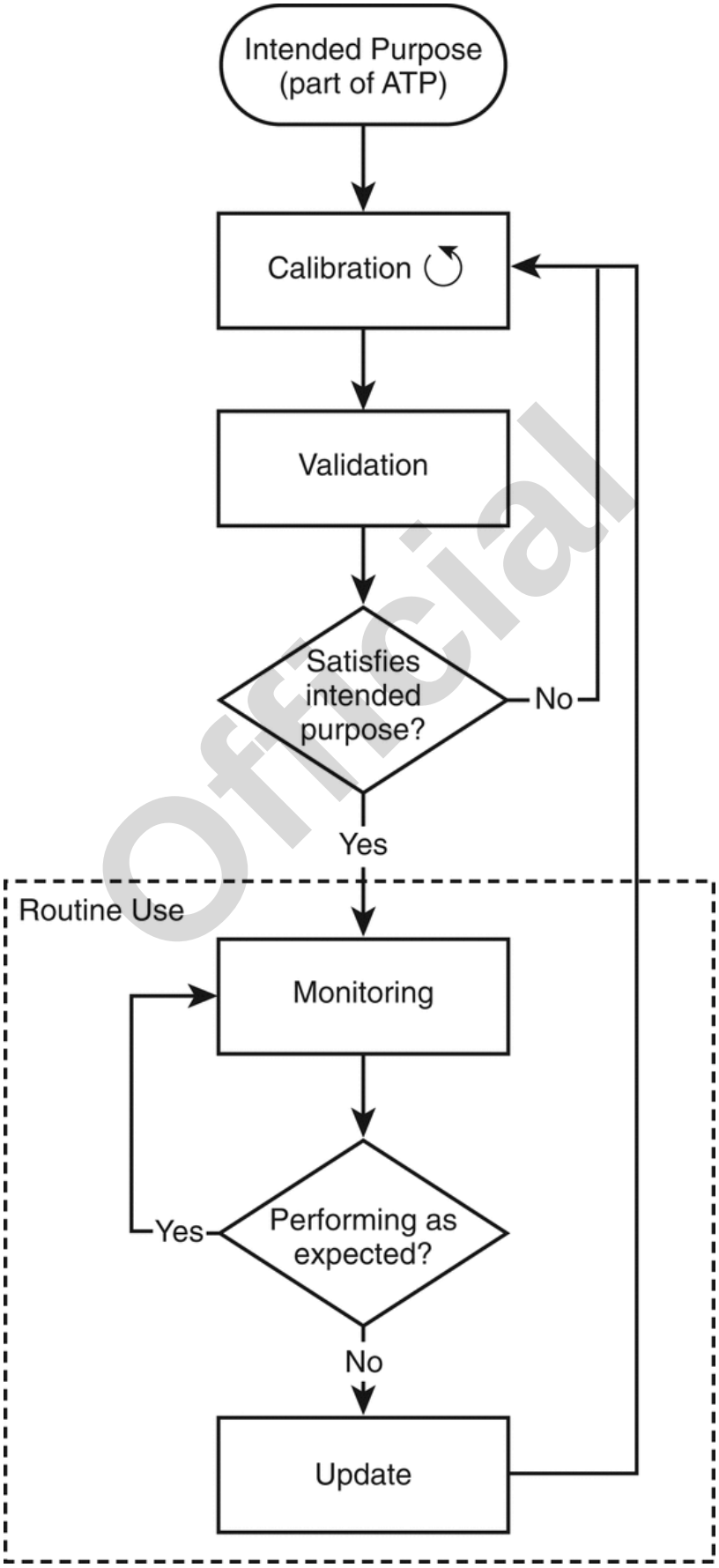

Figure 3. Schematic representation of the lifecycle workflow of a chemometric model.

3.1 Model Development: Calibration

The goal of multivariate calibration is to develop a model that links predictor variables (e.g., absorbance values for the range of spectral wavelengths) with response variables such as concentrations of certain chemical compounds. Response variables are typically more difficult or costly to obtain than predictor variables. Calibration is a critical step because details of the final model are influenced by many factors including the selection of samples, variables, an algorithm, and preprocessing options, as well as the quality of the reference method initially used (where applicable) to measure the response variable(s). Thus, the order of individual components of model calibration presented in this section does not always represent the order of operation, and individual steps might be repeated in an iterative fashion depending upon the outcome of subsequent steps. To perform well, the calibration model should be robust to intended variations.

SAMPLE SELECTION

Selection of samples for the calibration model is a critical step in model development. Scientifically sound rationales must be used to select representative calibration samples with respect to two criteria: the type/range of variability, and the number of samples. The range of response values in selected samples should cover the entire expected range of response variation for the intended application. The range and type of variation in predictors may include relevant variability in factors other than the property being predicted, such as particle size, different lots of ingredients used for sample preparation, analyst-to-analyst and day-to-day variation, and other sources of variation. In the manufacturing setting, batches of samples representing within and outside specification should be included. A risk-based approach is recommended for deciding which factors and variability will be included in the calibration samples. Performing a failure mode and effect analysis (FMEA) or a feasibility study to evaluate the effects of such factors is one of the commonly used approaches.

The required number of samples increases as the complexity of the calibration model (i.e., the type and the range of variability in both response and predictor variables) increases. In general, the larger the number of samples, the higher the probability that correct results can be achieved throughout the range of calibration. Meanwhile, the distribution of the calibration samples also deserves attention. A uniform distribution of the samples is preferred, although it is not required in all cases, depending on the specifics of the application. Scientifically sound rationales must be in place to justify the number and the distribution of calibration samples. When obtaining new calibration samples becomes costly, the DoE and/or historical database approaches are commonly used alternatives to build calibration models.

Data obtained from selected samples (i.e., a selected number of rows of the **X** matrix in *Figure 2*) should undergo further triage. Exploratory data analysis (see *Qualitative Application Examples*) may be valuable for understanding data structure, detecting potential clusters or groups of samples, identifying outliers, and selecting representative samples. Outliers may be detected using metrics that describe how well a given observation fits within model space (e.g., Hotelling's T^2 for latent variable methods) and distance to the model subspace (e.g., residual for latent variable methods).

PREPROCESSING

The goal of preprocessing is to deploy mathematical transformations on the column of the **X** matrix (*Figure 2*) to amplify variation of interest within the data set and attenuate extraneous variation for both variables and observations. Preprocessing is an important initial step for most data analyses, and can be applied iteratively with variable selection. The initial selection of preprocessing options should be guided by an understanding of the sample data and/or the underlying analytical technique. Guidance may also be obtained through exploratory multivariate data analyses (see *Qualitative Application Examples*). Initial selections can be modified or refined within the context of subsequent model performance optimization. This process is almost always cyclic in practice; comparison across different preprocessing strategies leads to a better understanding of the data, which can further refine the preprocessing to maximize the signal-to-noise ratio.

However, there are two points of caution. Preprocessing should be used judiciously because: 1) overprocessing can attenuate the signal and inflate undesirable variations; and 2) many chemometric methods can accommodate noisy and visually unappealing data. Furthermore, in some instances the preprocessing has been accomplished on the instrument; many algorithms used by instrument vendors are proprietary, and it may be impossible to know exactly what modifications have been made to the data. The chemometrician must be aware of, and cautious about, any preprocessing already applied in the reported raw data.

Preprocessing may consist of transformation, normalization, and/or other mathematical treatment of data. Preprocessing of samples (rows of data matrix) may include mean or median centering, scaling, and other procedures. Variables (columns of data matrix) can also be transformed, aligned (e.g., time warping), or scaled (e.g., autoscaling). Mean centering is probably the most common preprocessing method. For example, the use of mean-centered data in the regression model removes the intercept term, resulting in a potentially parsimonious model. It is also common to remove the linear baseline (e.g., bias/offset correction) or polynomial interferences or to apply digital filtering techniques such as derivatives, wavelets, Fourier filter, or a Savitzky-Golay filter to remove or attenuate linear offsets and noise contribution. Normalization is the scaling of observations, giving objects similar impact on model creation by: 1) accounting for factors other than the analyte of interest that impact measurement (e.g., sample path length variability caused by particle size, sample volume, or tablet thickness); 2) correcting for variation in concentration or mass of an analyte for qualitative models; and 3) attenuating measurement artifacts (e.g., instrument performance drift). When variables are highly correlated, such as in spectroscopic measurements, normalization is typically performed via standard normal variate (SNV) and multiplicative scatter correction (MSC). These normalizations are typically conducted on each sample (row).

In cases where variables have different units, sensitivities, or signal-to-noise ratios, variable scaling is typically applied to correct for that; autoscaling is typically used, but the analyst may consider other alternatives, such as using the pooled standard deviations of measurement replicates as scaling/weight parameters. These normalizations are typically conducted on each variable (column). No guidelines exist for deciding which preprocessing approach is the best for a given data set. It is often a

trial-and-error process as the analyst applies different preprocessing techniques to find the best one. Combinations of preprocessing techniques can be applied to a given data set as well.

ALGORITHM SELECTION

Many different algorithms have been tried with varying degrees of success in chemometrics applications. The methods include relatively simple ones such as multivariate linear regression and its modifications (robust or weighted regression); the widely used latent variables approaches such as principal components regression (PCR) and PLS; local methods such as k-nearest neighbors (kNN); and the more sophisticated methods such as support vector machines (SVM) and artificial neural networks. More details on commonly used algorithmic tools can be found in *4. Applications of Chemometrics*. In general, it is difficult to predict which algorithm will produce the best results for a particular data set. The multitude of choices for sample selection, variable selection, and data normalization and preprocessing—as well as the combination of tuning parameters for each algorithm—could significantly affect the performance of the calibration model. Sometimes, the choice of algorithm even iterates with the steps of sample selection, variable selection, and preprocessing. Thus, the choice of algorithm may depend on the task at hand, software availability, and the subject matter expert's familiarity with the method. Another consideration is that some algorithms provide useful tools for diagnostics and interpretation—such as PLS scores, loadings, and coefficients plots, whereas others are sometimes referred to as "black boxes" because their inner workings can be difficult to interpret (e.g., neural networks).

It is important to keep in mind that many algorithms are empirical in nature. Almost always, some kind of model can be developed. Thus, the results should be evaluated critically to ensure that results generated by the model are relevant and correct, and that model performance meets the requirements of the ATP. This can be accomplished by cross-validation, and ultimately, by validation with the independent test data set that was not used to develop the model.

VARIABLE SELECTION

The intended purpose of variable selection is to identify a subset of predictor variables (i.e., the column of the **X** matrix in *Figure 2*) that can improve the performance of the model. The underlying rationale for variable selection in chemometrics is twofold. First, certain predictor variables could be irrelevant to the intended purpose, and ideally, these variables should be minimized or removed from the model. For example, only specific wavelengths of the whole range generated by the spectrometer may bear information relevant to the variation in levels of the response variable. Variable selection for this purpose should be based on first principles and experience. Inclusion of unrelated predictors, such as irrelevant spectral regions, could potentially degrade performance of the model. A smaller number of variables from the preprocessed data can be used to achieve superior performance, such as accuracy, precision, and robustness.

The second part of the rationale for variable selection is to avoid overfitting. Overfitting is the situation where the model is not only describing the intended variability in the data but also includes the modeling of the noise. The latter has a negative influence on the predictive properties of the model. Typically in chemometrics, the number of predictor variables (hundreds or thousands) is larger than the number of observations (dozens to hundreds). There are two general strategies for handling the issue of overfitting. One strategy is to manually or computationally select only a subset of predictor variables and use them for model development. Predictor variables can be selected manually (e.g., choosing certain spectral wavelengths characteristic of a given compound) or by using a variety of statistical selection methods such as stepwise regression and using simple univariate statistical measures such as the *F*-test, *t*-test, correlation coefficient, signal-to-noise ratios, intensity thresholds, variable importance in projection metric, or genetic algorithms. Multivariate approaches may use PC variable loadings, linear discriminants, or regression coefficients to define the key features.

In the second strategy for avoiding overfitting, variables are not selected at all. The latent variable methods—PLS and PCR—by their nature have the capability to selectively give more weight to the important predictor variables and de-weight the less important ones. These new latent variables, or components, are constructed as linear combinations of the observed variables. If combined with variable selection, the performance of latent variable models could potentially be improved, because some predictor variables contain only noise, thus perturbing the model-building process.

Selecting a subset of the predictor variables that are informative will reduce the size of the data set, provide computational efficiency, and obtain better models for postprocessing. One caveat for applications such as authentication is that eliminating predictor variables that are devoid of information may prevent the recognition of adulterants or novel objects that have features in these excluded predictor variables.

CROSS-VALIDATION

In practice, cross-validation is used to obtain an estimate of the model performance and to fine-tune an algorithm's parameters (e.g., the number of components for PLS). This is accomplished by repetitive splitting of the data (i.e., the number of rows of the **X** matrix in *Figure 2*) into a calibration set, which is used to develop the model, and a testing set (or internal validation set), which is used to make predictions and to compare actual and predicted values.

The approach of *n*-fold cross-validation is commonly used to split the data into *n* subsets to perform cross-validation. In each of the *n* iterations, *n*–1 subsets are used to develop the model, which is used to predict the remaining n^{th} data split. The procedure is repeated until all subsets are predicted. The error between reference values and predicted values during cross-validation is then recorded as root-mean-squared error of cross-validation (RMSECV). Multiple approaches exist for splitting the data into *n* segments. Regardless of how one decides to split the original calibration data, one aspect that deserves attention is that any batches used in calibration and cross-validation must not be considered or reused as an independent dataset for method validation. The most straightforward form is called leave-one-out cross-validation (LOOCV), where samples are removed one at a time. The LOOCV can involve intensive computations; its results can be heavily affected by outliers and are less consistent than the results of other forms of cross-validation. It is also possible to split the data according to prior information, such as one subset per batch given the available historical data set across multiple batches. Another option for

cross-validation is bootstrapping, where in each iteration a certain proportion of data is randomly sampled to create a calibration set, while the remainder of the data are used as a testing set. The procedure is repeated many times (typically 100 or more cycles) to obtain a stable estimate of prediction error.

The measure of error is the most common figure of merit used to characterize model performance during cross-validation. The intended purpose of the method determines the nature of the errors, such as the misclassification rate for qualitative methods and prediction error for quantitative methods. Irrespective of qualitative or quantitative applications, two metrics are commonly used to characterize error within cross-validation: root-mean-squared error of calibration (RMSEC) and RMSECV. The RMSEC is calculated for the samples when left in the calibration, which monotonically decreases with each additional factor (i.e., PC) added into the model. In comparison, the RMSECV that is calculated during cross-validation will decrease until the last meaningful (i.e., relevant signal-containing) factor is added. Then, as each additional factor is incorporated into the model, the RMSECV will increase, indicating that the calibration data are being overfit. The plot of the RMSEC and/or the RMSECV versus the number of factors in the model is referred to as a predicted residual error sum-of-squares plot. In general, the best practice is to avoid inclusion of factors beyond where the minimum of the RMSECV plot line occurs. In addition, correlation between observed and predicted values (e.g., R^2) is also commonly used to assess performance of quantitative methods. Finally, the cross-validating results are only meaningful when the calibration and testing sets are comparable (i.e., drawn from the same population). Otherwise, extrapolation may lead to incorrect predictions.

3.2 Method Validation

The objective of validation is to demonstrate that the performance of a method is suitable for its intended purpose. Validation of a model and validation of a method are two different activities. Model validation routinely involves the use of an internal validation set or cross-validation to assess the appropriate parameters of a model via identification or quantification error and uncertainty. Parameters often include the range of variables, the type of preprocessing, the model rank, the choice of the algorithm, and others. These activities have been addressed in detail in *3.1 Model Development: Calibration*. In comparison, method validation must be based upon a fully independent external validation set and must follow the validation requirement described in *Validation of Compendial Procedures* ⟨1225⟩, according to the method type category. The acceptance criteria should be justified for the intended purpose. During the lifecycle, method revalidation is necessary after a model transfer or model update. The method validation and revalidation strategy should be risk- and science-based and appropriate to its impact level.

The typical performance characteristics for method validation are specificity, accuracy, precision, linearity, range, and robustness for quantitative models, and specificity and robustness for qualitative models. The metrics and their descriptions are discussed below. In addition to those typical performance metrics, metrics such as limit of detection (LOD), limit of quantitation (LOQ), sensitivity, analytical sensitivity, and effective resolution may not be required for validation purposes but could be useful for understanding the boundary of the method performance for a specific analytical application and the analytical technique that a model is associated with.

PERFORMANCE CHARACTERISTICS FOR METHOD VALIDATION

The sections that follow provide method validation attribute descriptions and metrics.

Accuracy: Statistical comparison between predicted and reference values is recommended. For quantitative applications, root mean-squared error of prediction (RMSEP), squared error of prediction (SEP), and bias are the typical measures of method accuracy. For qualitative applications such as classification, sometimes misclassification rate or positive prediction rate could be used to characterize method accuracy. To ensure that a model is accurate enough when tested with an independent test set, RMSEP is often found to be comparable to RMSEC and RMSECV and to meet the requirements of ATP.

Precision: The routine metrics of RMSEP and SEP encompass both accuracy and precision. Assessment of precision could involve a determination of the uncertainty associated with the reportable result and a variance component analysis that determines an "error budget" that quantifies the sources of variation that contribute to the uncertainty. For an estimate of precision alone, the standard deviation across results for replicate and independent analyses on the same sample within method, across days, across analysts, across instruments, and across laboratories could be used.

Specificity: For both qualitative and quantitative methods, whenever possible the underlying chemical or physical meaning of the chemometric model should be demonstrated and validated. For example, the scientific meaning of the variables (such as spectral range) used for model construction, data preprocessing, regression vectors, and loading vectors [for PLS, PCR, PLS discriminant analysis (PLS-DA), and PCA] should be demonstrated. Specificity could also be validated by the tendency of sample components (matrix or other nonanalyte compounds present in the sample) to adversely affect the ability of the chemometric method to report results accurately and/or precisely. The level of specificity may be assessed by accuracy and precision in the presence of varying amounts of potentially interfering substances. Substances to be tested can be identified by considering the underlying physical/chemical methodology and modeling approach.

Specificity may be evaluated by including adulterated, substandard, or nonauthentic samples. Authentic samples, both within and out of the target criteria, may not be available for practical or economic reasons. For example, it may not be possible, due to cost, to obtain out-of-target samples for a controlled, large-scale process. Where possible, an exclusion panel of out-of-target samples that may be closely related to the target samples should be considered to increase the confidence in the specificity of the measurement. Simulated off-target samples (such as small-scale samples) may be used to validate the procedure's suitability for the intended purpose and range. In all cases, the suitability of validation samples must be justified with appropriate inclusion and exclusion criteria.

For qualitative methods, the method should demonstrate the capability to correctly identify or classify the samples. The receiver operating characteristic (ROC) curve and/or the probability of identification (POI) are commonly used metrics (detailed information on ROC curves can be found in *4.1 Qualitative* and *Figure 6*). The ROC approach is intended to illustrate true-positive rate (TPR) and false-positive rate (FPR) over a range of decision thresholds. A good identification method should generate a ROC curve with the area under the curve (AUC) close to 1. For most well-designed identification methods, the probability of a

positive identification is near zero when the analyte is not present, and the probability should approach 1 as the analyte concentration or mass increases.

Linearity: The algorithm used for chemometric method construction can be linear or nonlinear, as long as it is appropriate for modeling the relationship between the analytical signal and the analyte. The measures commonly used to assess either the model fit or its predictive properties are the correlation coefficient, slope, *y*-intercept, and residual sum of squares of the plot between the predicted versus observed results. Note that the plot between the residual versus observed results across the analytical range is expected to show no pattern.

Range: The range of a method should be appropriate for its intended use (e.g., specification).

Robustness: Typical factors to be considered include the normal variability of materials (e.g., lot-to-lot variability), operating environment variability, instrument variability, such as minor instrument maintenance, and method parameter variability, such as the number of spectra averaged in a spectrometer. In conjunction with the validation results, the method development strategy—such as the design of the calibration set or library and the choice of model parameters—can be taken into account to demonstrate the method robustness.

VALIDATION SAMPLES

Other general aspects to be considered for chemometric method validation include the validation samples. The validation samples should be independent of the calibration samples to demonstrate the ability of the model to predict. Being independent means that the validation samples were not used in the calibration set or used for model optimization. Internal validation or cross-validation is typically used during calibration for model parameter optimization, but is not considered sufficient for final method validation. The validation samples should be selected based upon the ATP and the desired performance characteristics of the model. Method robustness is based on the evaluation of authentic samples with typical sources of variance. For pharmaceutical or dietary supplement products, validation samples of nominal production scale, such as those routinely manufactured for in-process, release, and/or stability testing, may be included. For botanical articles, taxonomic identity, geographic origin, season of collection, and other variants may be included. For naturally sourced materials such as many food ingredients and excipients, variables such as the geographic or microbiological source of the material, processing conditions, impurity composition, and other relevant attributes may be included in the validation sample set.

ACCEPTANCE AND DIAGNOSTIC CRITERIA

As with any other analytical procedure, the acceptance criteria for a chemometric method should be defined before execution of the validation protocol. If the chemometric model was developed using data from a secondary technique [e.g., near-infrared (NIR) or Raman] with reference values from a primary analytical procedure [e.g., gravimetric, nuclear magnetic resonance (NMR) spectroscopy, or high-performance liquid chromatography (HPLC)], the ATP for validation may not exceed that which is obtainable by the primary method. Some exceptions may occur, for example, it may be possible to achieve superior precision using a secondary procedure, although the accuracy will be limited to that of the reference technique.

In addition to those attributes addressed in *Performance Characteristics for Method Validation*, method validation must also take into consideration the setting of diagnostics limits for a multivariate model before deployment for routine use. Samples that are out of model space are considered outliers and not suitable for obtaining a reportable model prediction. The model diagnostic should have a demonstrated capability to flag any of the out-of-model-space samples. To set up the diagnostics limits, two cases must be considered. The first is to determine the statistical distribution of leverage and residual within-the-calibration data set. The second is to prepare intended in- and out-of-model-space validation samples to test the limit. For instance, for an NIR spectroscopy-based content uniformity method, the in- versus out-of-model-space samples could be samples at target label claim versus samples containing active pharmaceutical ingredient (API) concentrations outside of the intended range for the method.

3.3 Model Monitoring

Throughout the model lifecycle, changes that can affect model performance may occur. Procedures should be in place for continuous performance verification of the model and for model update and procedure revalidation if necessary. A specific order among model monitoring, model update, and model transfer does not exist. Scientifically sound rationales must be applied to determine an appropriate sequence for the intended application.

A control strategy for checking model performance over its lifecycle should be developed and documented as part of model development and procedure validation. The strategy should identify the necessary elements for ongoing monitoring and evaluation of model performance. In addition, a plan for monitoring, analyzing, and adjusting the model should be in place with a measurement frequency that allows identification of excursions related to critical aspects of the model. The level of maintenance should be considered part of the risk assessment activities and should be adequate for the model criticality. If applicable, analytical instrumentation used to generate the inputs for the model should be qualified and also subjected to a continuous verification plan (for relevant guidance, see general chapters related to the applicable analytical instrumentation).

Ongoing assurance of performance of the model throughout its lifecycle should include:

- Ongoing monitoring and review of the model
- Evaluation of risk assessment
- Evaluation of post-implementation changes and predefined model maintenance
- Model update and procedure revalidation as needed

The ongoing review of a model should occur at predefined intervals and also should be triggered by events that may have an impact on model performance. Examples of such events include changes in raw materials variability or manufacturer; changes in the upstream process that may alter the sample matrix (e.g., process equipment or operation settings); drifts in model prediction; and out-of-specification (OOS) or out-of-trend (OOT) results given by the model (dependent upon the root cause

of the OOS or OOT). In addition to triggers that are based on the prediction output of the model, triggers based on model diagnostics metrics and corresponding action rules should be included. Multivariate models may be more strongly affected by aberrant data signatures than are univariate models. Special care is needed to: 1) justify the multivariate model diagnostics statistically, 2) verify with data from the model development and procedure validation process, and 3) implement multivariate diagnostics for monitoring as part of the control strategy. Comparison of model predictions and reference or orthogonal procedures should take place on a periodic basis or as part of the investigation triggered by the review process.

The use of model diagnostics when applied to a new sample ensures that the model prediction is valid with regard to the calibration and validation sets used during model development and procedure validation, and also ensures that the result does not constitute an outlier. The observation of an outlier means that the result is invalid, but it is not a reliable indication of an OOS result; an OOS result is an observation produced by a model when the prediction falls outside the acceptance criteria and the model diagnostics are within the thresholds. In the case of qualitative models, nonconforming results should be treated as outliers and should trigger an investigation; the output of such an investigation will indicate whether the result is OOS.

The review process for a model should produce a decision regarding the need for an extension of model maintenance; such a decision may be the result of a risk-assessment exercise or the outcome of a predefined decision workflow. Model criticality and usage will define the extension of model maintenance, which can include restraining the model conditions; adjusting the calibration set (samples can be added, replaced, or removed); or even completely rebuilding the model. The decision, and corresponding rationale, must be scientifically sound and documented.

3.4 Model Update and Method Transfer

Model updating must be considered part of the analytical procedure verification, and both the justification and the activities must be documented as a part of the analytical procedure lifecycle. Before performing a model update, it is critical to understand the underlying causal factor that has prompted the update. The reasons for model updating can be roughly divided into two categories.

The first category is when the calibration set simply needs to be expanded. In this case, nothing has actually changed in terms of the response of the instrument to specific analytes. Instead, the original model is no longer valid because of the expanded range of original calibration, the addition of new analytes, or the occurrence of other, previously unseen variations (e.g., changes in particle size distribution, or the drift of a chemical process to a new steady state). Thus, the calibration space must be expanded with samples exhibiting this variation.

The second category is when the samples are the same but the measurement system response function has changed. This is often due to changes in the measurement components (new light source, clouding of optics, wavelength registration shift) or can be due to method transfer across different instruments. This, in essence, is an instrument standardization problem. Changes in instrument or measurement procedures over time can render a calibration model unusable, in which case a model update becomes necessary or a new model should be developed.

In practice, there are multiple model updating techniques that could be applied to each category of model update. Some updating techniques are relatively straightforward and simple to implement, whereas others are technically complex. Selection of an appropriate model updating approach must be based on a full understanding of the underlying causal factor. As a general rule, simple updating methods (e.g., slope and bias adjustment, as described below) should be considered first.

Before any model maintenance work is initiated, it is good practice to confirm that the fundamental construct of the original model—such as data preprocessing and variable selection—remains sound. For both qualitative and quantitative applications-based model updates, a straightforward preprocessing approach and a scientifically sound variable selection approach are recommended as the initial attempt to address challenges associated with updating the model. These approaches are equally applicable to differences caused by instruments, measurement conditions, or sample matrices.

SLOPE AND BIAS ADJUSTMENT

One of the simplest model-updating methods is to postprocess the predictions with a bias adjustment, a slope adjustment, or both. This approach is often used for quantitative applications. For some qualitative applications, this approach may also be useful, depending on the nature of the method. In a limited set of circumstances, bias/slope adjustments are expected to work. For example, if the matrix of the samples being predicted were systematically different from the calibration/validation sample set, the model predictions would be in error by a constant value. Bias/slope adjustments, however, would not correct for any new variation in the data, such as variation that would result from a different matrix (e.g., new analytes/interferents). Therefore, applying slope/bias adjustments without a full understanding of the underlying causal factor is not recommended. Guidance may be obtained via inspection of residuals corresponding to the new data obtained using the existing model or, alternately, from supplemental exploratory data analyses using techniques such as PCA (see *4. Applications of Chemometrics*).

CALIBRATION EXPANSION

When expanding the original calibration set, one should consider the number of new samples to be added, the impact (leverage) of the added samples on the overall composition of the new calibration set, and how to partition any new samples between the calibration and validation sets. It may be appropriate to simply augment the original calibration and validation sets with all the new samples, or it may be advisable to use a subset of the new and original samples. Multiple approaches, such as the Kennard-Stone algorithm and nearest-neighbor approach, are available to aid in the selection of new samples to add to an existing calibration set, but in general, all the approaches use methods to identify new samples on the basis of high leverage (i.e., model influence). Hotelling's T^2, multivariate score plots, and their corresponding limits are also effective approaches for accomplishing this goal.

CALIBRATION TRANSFER

Instrument standardization and calibration transfer methods are used to transform the response function of a measurement system so that it matches that of a reference measurement system. The reference measurement system could consist of a completely different analyzer, or it could be the same analyzer before it experienced a response-function shift. The vast majority of these methods generally requires the use of stable transfer samples that are measured on the original instrument and require the instrument to be standardized at the same time. In addition, the approaches commonly used for instrument standardization could also be applied effectively to address the challenges resulting from changes in sample matrix and measurement conditions.

3.5 Revalidation

Before the redeployment of a multivariate model, appropriate procedure revalidation should be established using criteria equivalent to those used in the original validation protocol. This revalidation is necessary to document the validity of the model as part of the analytical procedure verification. The nature and extent of the revalidation procedure, including aspects such as scientific justification and experimental approaches, must be based on the cause of the update and the nature of the corrective action required for establishing suitable performance. Revalidation should be documented as part of the analytical procedure lifecycle.

4. APPLICATIONS OF CHEMOMETRICS

As discussed in *2. What is Chemometrics?*, chemometric analyses may be performed in either a supervised or unsupervised manner depending on the availability of data and the specifics of a given application. This section provides an explanation of these different analysis scenarios, as well as the chemometric tools (i.e., algorithms) that are commonly used. Additionally, several specific applications will be described in detail.

4.1 Qualitative

GENERAL ASPECTS

Qualitative chemometric analyses may be performed by using supervised and/or unsupervised approaches. However, to be incorporated into analytical procedures that are alternatives to compendial methods, the performance of any chemometric model must be verified as described in *3. Model Lifecycle*. This may not be possible if very little is known about the samples being analyzed. Nevertheless, unsupervised analyses play an important supporting role during the development of chemometric alternatives to compendial procedures within the lifecycle framework, and are recommended for use before the development of subsequent supervised approaches (examples are provided in *Qualitative Application Examples*).

Qualitative compendial procedures are those that seek to supply the value of some categorical descriptor for an unknown sample. Examples of categorical properties include (among others):

- Chemical identity: microcrystalline cellulose versus API
- Morphology: polymorph A versus B; monoclinic versus triclinic
- Sample authenticity: authentic versus nonauthentic and/or adulterated
- Sample origin: facility A versus B; lot ABC-001 versus lot XYZ-002

Categorical properties can be modeled effectively using supervised chemometric algorithms that leverage either proximity or distance in multivariate space, resulting in a qualitative assignment of samples to one or more classes depending on the application and the technique used. According to ⟨1225⟩, procedures based on models of this type are suitable for incorporation into Category IV methods for identification and should be validated to demonstrate adequate specificity for the intended use. "Identification" is a term that is generally used to describe a range of analysis scenarios; several common ones are described in *Qualitative Tools*.

Supervised techniques used for classification purposes, besides the analytical sensor information, make use of discrete information related to the samples in the data set (e.g., class labels). It is intended to define the borders of the clusters (or classes), providing statistical criteria, in addition to the visualization techniques (e.g., scores plot of PCs from a PCA). The borders then can be used as acceptance criteria for inclusion or exclusion of new samples into a given class. Several useful tools are available to the analyst for building the model, such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), soft independent modeling of class analogy (SIMCA), kNN, and PLS-DA. More complex techniques also may be used with justification. Each type of model has its own strengths and weaknesses. If a particular algorithm does not give the desired level of performance, a different algorithm may be selected. These aspects are discussed further in *Qualitative Tools*.

For model development, optimization, and validation, the classification and discrimination techniques follow the same process described in *3. Model Lifecycle*. These techniques develop a threshold or limits to make class assignments. The simplest form of classification is a two-class system (e.g., good/bad or A/B). When a new sample is classified as good or bad (A or B) by the model, this results from what is called a discrimination technique. Classification is the technique of assigning a new sample as A, B, either A or B, or neither A nor B. The error rate of the model is the total number of incorrect classifications divided by the total number of samples. Depending on the requirements of the application and the associated risk levels, the error rate(s) might be class-specific or pooled across classes. The specific details of the application will determine whether the assignment of a sample to more than one class would be an acceptable result. Where this must be avoided, classification thresholds should be set appropriately during model development to guarantee single-class results.

PCA: PCA is a commonly used exploratory analysis tool that was briefly introduced in *2. What is Chemometrics?*. PCA is a variable reduction technique and acts on the data by defining new variables, so-called principal components (PCs, see *Figure 2*). PCs are orthogonal and are defined in the direction of the largest (remaining) variance in the data. The results of PCA are discussed in terms of PC scores and loadings, which may be plotted graphically in two (and sometimes three) dimensions for visualizing clusters of samples or outlying samples (score plots, see *Figure 4*), whereas the loadings plots provide information on the original variables. The scores are the projections of the samples on the PCs, whereas the loadings provide the weights (coefficients) by which the values of the original variables are multiplied to generate the component score for a particular sample. Loadings are rich in information regarding which variables in the matrix are prominent and can be used to "understand" the information captured by the latent variables that provide the basis for the observed sample distribution in the scores plot (*Figure 4*). The PCs may also be used as the basis for quantitative models via PCR (see *4.2 Quantitative*).

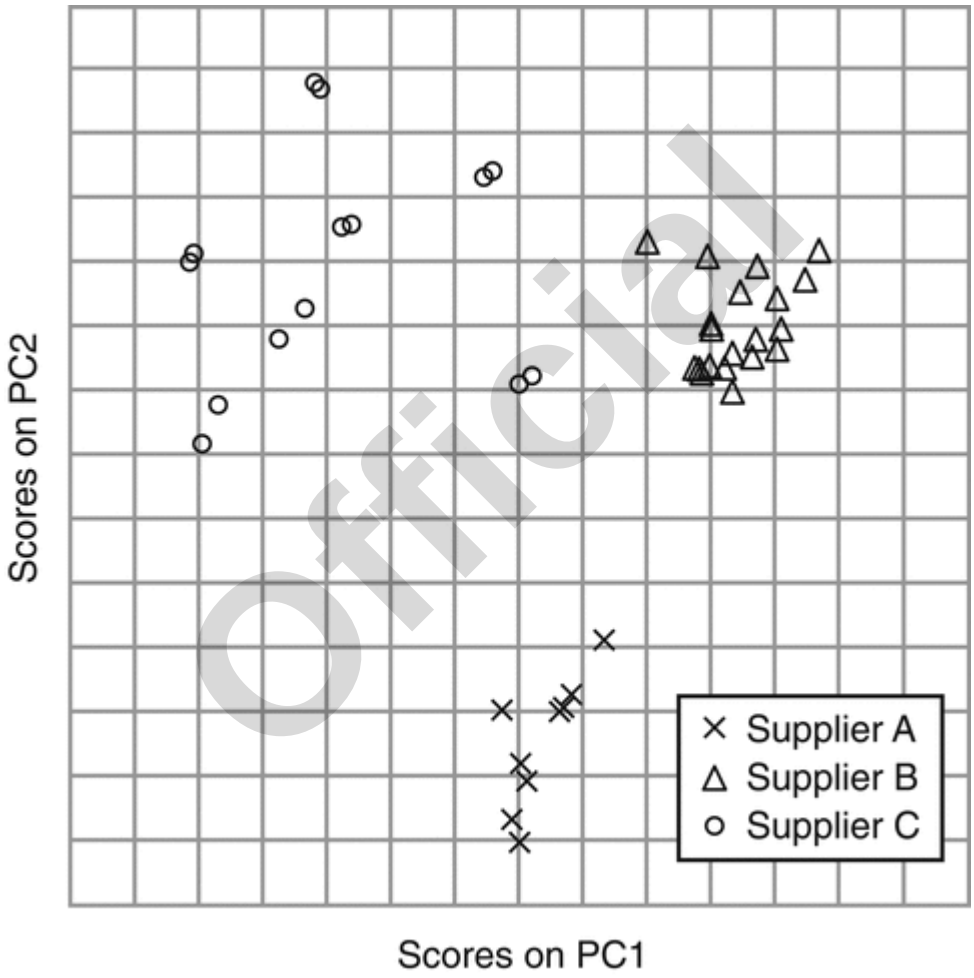

Figure 4. PCA scores plot: projection of PC1 versus PC2.

Clustering algorithms: The aim of clustering analyses is to partition data into different groups. Some methods, such as hierarchical cluster analysis (HCA), result in tree-like structures called dendrograms (*Figure 5*) that provide information on groups of samples or outlying objects occurring in the data set. Dendrograms can be built in many different ways and thus may show different aspects of the data set. A first possibility, called divisive or top-down methods, starts from the entire data set, which is split in consecutive steps until each object is in an individual cluster or all elements of a cluster fulfill a given similarity criterion. In the second scenario, called agglomerative or bottom-up methods, the opposite is done. Starting from individual objects/clusters, those most similar are merged until everything is in one cluster. Of the two options, bottom-up approaches tend to be computationally less intensive, are part of most computer packages, and are more frequently used.

The parameters used to express (dis)similarity between objects or clusters (see *y*-axis on *Figure 5*) can be either correlation-based (e.g., correlation coefficient, r for similarity or $1-|r|$ for dissimilarity), distance-based (e.g., Euclidian distance) measures, or a combination of both (e.g., Mahalanobis distance). Two clusters are linked in the dendrogram at the height related to this (dis)similarity measure. The parameter applied is expressed in such a way that clusters/objects linked low (i.e., close to the *x*-axis) are similar, whereas those linked high are dissimilar. Many different methods or criteria exist for use in deciding which objects/clusters are consecutively merged [e.g., single linkage, complete linkage, (weighted) average linkage, and centroid linkage]. Depending on the applied criterion, the dendrograms may look very different. Drawing arbitrary

horizontal lines (see *Figure 5*) splits the data set into different groups, which occasionally may be linked to sample properties. Outlying samples are linked very high in the dendrograms, most often as one-object clusters. The stability of clustering results can be affected by many factors, such as noise in the data, sample size, choice of algorithm, distance measure, and others. Divisive methods tend to produce more stable results than agglomerative methods, even though they are hardly ever used.

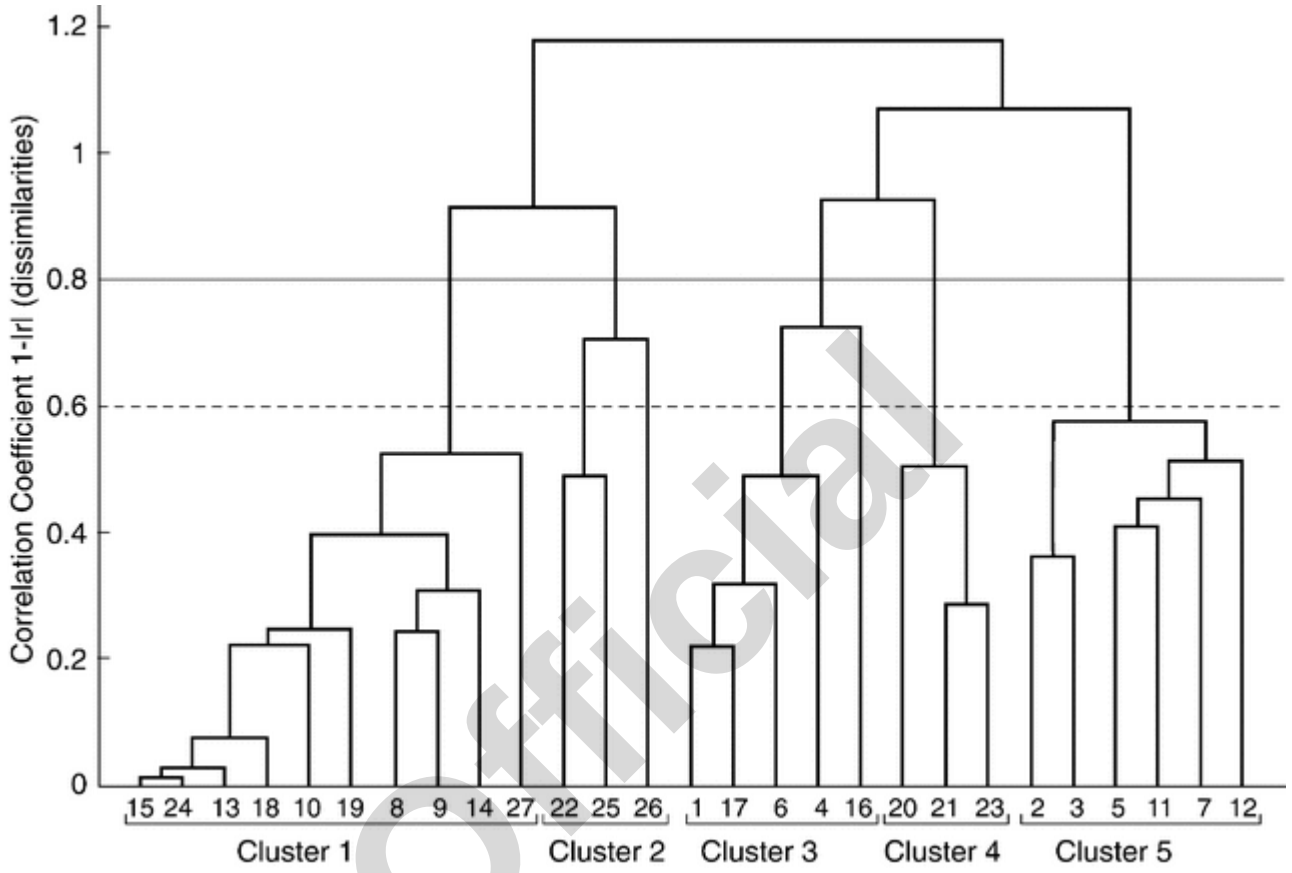

Figure 5. Hierarchical clustering based dendrogram for 27 objects. Abscissa: object numbers.

Parametric modeling: Authentication of a material or formulated product for quality assurance purposes may be achieved via multivariate parametric modeling of the chemical profile (e.g., spectrum or chromatogram). These methods are sometimes referred to as "one-class classifiers". Basically, a parametric model is developed statistically or empirically with a confidence boundary around the data from known authentic samples. Parametric models describe the behavior or structure of the majority of the data from a single class, assuming that a distribution (typically, a multivariate normal one) exists in those data. Predictions against a parametric model identify deviating points outside the significant boundaries of that distribution. Whereas classification or discrimination approaches provide identification (as previously described) of the unknown sample, parametric models will designate the unknown sample as either belonging to the authentic class or as an outlier, or exhibiting features other than those defined by the distribution. Some discrimination techniques will also optimize class separation, and may have greater selectivity, but will tend to fail when presented with an object from none of the classes. In contrast, parametric modeling methods will reject any sample that does not fall within the confidence boundary established for the model. Perhaps the most commonly used parametric modeling algorithm is SIMCA, where PCA is used to model the data and Hotelling's T^2 and Q residuals (DmodX) are used to define the limits. In some instances, parametric modeling approaches will result in "soft classifications" wherein a sample may be assigned to more than one class. To satisfy compendial usage requirements, appropriate action must be predefined for cases of ambiguous class assignment, and/or the α and β error rates must be established and justified according to risk.

QUALITATIVE APPLICATION EXAMPLES

Exploratory analysis: Unsupervised algorithms are routinely used for initial exploratory data analyses, which precede formal calibration and validation exercises. Unsupervised data exploration can rapidly indicate, in the absence of any prior knowledge, whether distinct classes or outliers exist within the pool of available data. For instance, in *Figure 4* a graphical example is shown in which PCA applied to spectroscopic data was used to differentiate between samples from different suppliers. Exploratory models yield latent variables that may be inspected to identify the original variables with the greatest amounts of relevant analyte signal to carry forward into subsequent models or to guide and/or verify the results of variable selection approaches. This might occur during the analysis phase of a DoE study to assess which factors to include in subsequent designs and/or calibrations. Further, exploratory analyses may be used to test and empirically optimize various signal preprocessing options to

maximize the relative contribution of desired analyte signal versus signals from sources of interference (e.g., scattering effects, optical path length differences, sample physical properties, and instrumental drift).

Change to read:

Material identity testing: Identification testing is routinely performed for all pharmaceutical active ingredients, excipients, drug products, and packaging materials to authenticate them and verify that the correct materials are used in manufacturing, release testing, and packaging operations. ▲*Spectroscopic Identification Tests* ⟨197⟩ ▲ (CN 1-May-2020) illustrates one of the most widely used analytical techniques for identification, Fourier transform IR spectroscopy. Typically, identification testing involves comparing the sample spectrum to the spectrum of a reference sample and assessing the level of agreement between the two. Different chemometric algorithms can be applied for identification, and the general model development and maintenance should follow the guidance in *3. Model Lifecycle*. All methods for identity testing are supervised classification methods because either a single reference spectrum or a set of reference spectra from material(s) with known identity is utilized by these algorithms. Algorithms can be applied in either original variable space or transformed variable space (such as latent variable space) after data preprocessing.

According to ⟨1225⟩, models of this type are suitable for incorporation into Category IV methods for identification and should, at a minimum, be validated to demonstrate adequate specificity for the intended use. An example approach for performance verification is the documentation of TPR and FPR. TPR is the percentage of samples that are correctly identified, whereas FPR is the percentage of samples that are incorrectly identified. TPR is also referred to as sensitivity, and FPR is referred to as specificity. Regardless of which approach is taken, there is always a trade-off between FPR and TPR, and this trade-off behavior and overall procedure performance are best visualized in a ROC curve plot (see *Figure 6*).

A classical ROC curve is generated by plotting the TPR on the ordinate and the FPR on the abscissa. Each point on a ROC curve represents a specific TPR and FPR pair obtained at a specific threshold. The threshold is typically a number generated by the algorithm, and when it is exceeded, this corresponds to a positive identification. This threshold value is determined during the procedure development process. The threshold could be the *P* value from hypothesis testing, hit quality index (HQI) threshold, Mahalanobis distance threshold, PLS discriminant score threshold, or some other threshold.

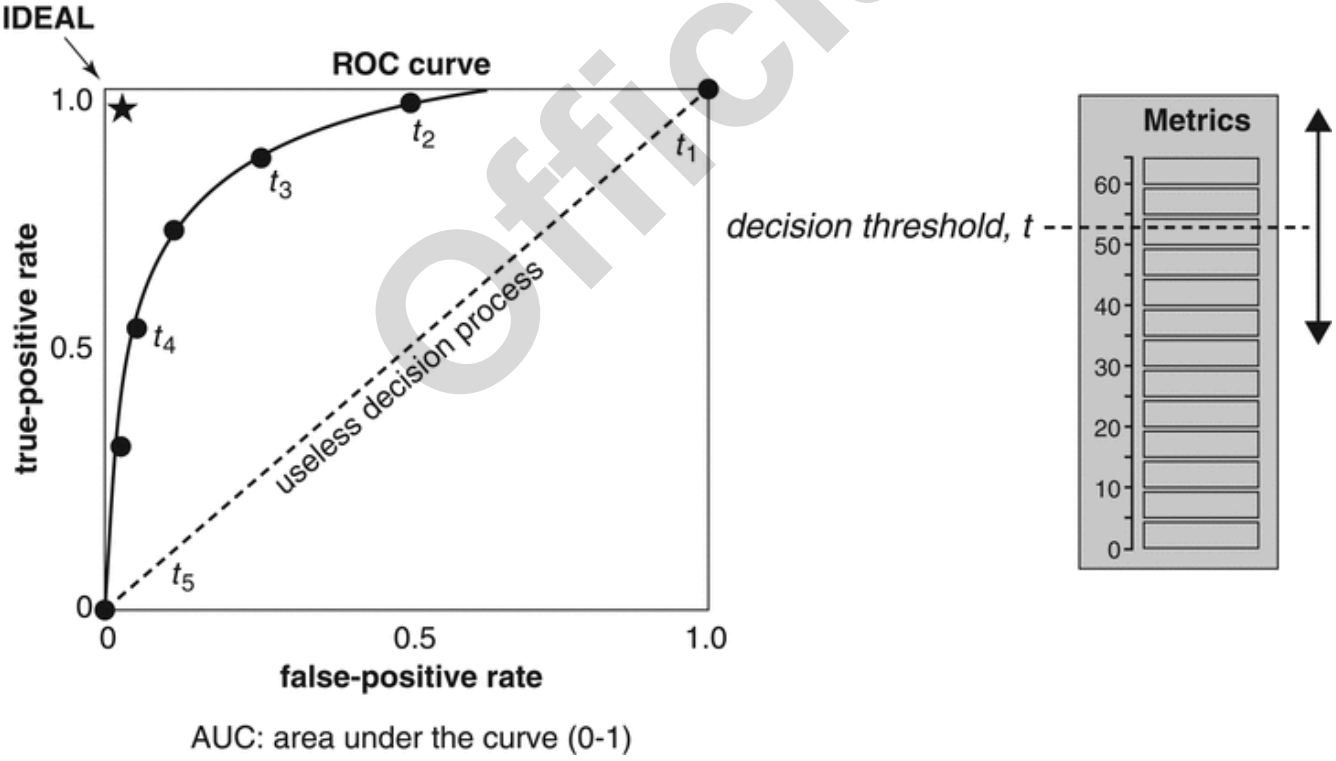

Figure 6. ROC curve plot showing the AUC (*3*).

During procedure development, a number of known positive and negative samples are examined, and the algorithm output for each sample is recorded. The performance of the method when operating at a particular threshold *t* is characterized by a single TPR/FPR pair, and the full operating characteristic curve can be produced by plotting TPR/FPR pairs over a range of decision thresholds. A false-positive identification corresponds to a decision threshold that is too low. False positives can be minimized by raising the decision threshold. However, excessively high decision thresholds can produce a false-negative identification.

A good identification method should generate a ROC curve that produces one or more TPR/FPR pairs in the top-left corner of the plot, as shown in *Figure 6*. This characteristic is best captured by the AUC. A random decision method is represented by the diagonal line with an AUC of 0.5, whereas a good identification method should have an AUC closer to 1. The shape of ROC

curve depends upon a combination of spectral properties of the sample and the chemometric algorithm used for identification. Characterization of the ROC curve is a fundamental task associated with method development and validation.

For a two-class model, false positives and false negatives are the incorrectly assigned samples. Together with the true positives and true negatives, they provide the basis for the sensitivity and specificity of the model. The analyst may wish to care about one type of error (α or β) more than another, and therefore could change the model to be more conservative when choosing a threshold. Thresholds can be used as action criteria or evaluation criteria.

Complex material authentication testing: Nutraceuticals and samples of herbal origin may be complex mixtures of numerous components. Authentication relies upon assessment of the presence of multiple analytes and their relative concentrations as well as potentially demonstrating the absence of adulterants and/or related materials. Thus, authentication in this scenario is a multivariate decision process that is appropriate for chemometric approaches. Typically, a characteristic "fingerprint" pattern incorporating information from the various analytes (and adulterants, potentially) that are present in the sample is obtained by applying a combination of separation and characterization techniques or spectroscopic measurements. The chemical composition information in the sample fingerprint provides a basis for multivariate pattern recognition. In particular, techniques with exceptional analytical selectivity, such as mass spectrometry and NMR spectroscopy, have emerged as powerful fingerprinting tools, especially when operated as detectors for chromatographic separations of multicomponent samples. Chromatographic fingerprinting is recognized as a viable identification procedure for herbal medicines by both the World Health Organization and the U.S. Food and Drug Administration. (*4*,*5*)

Depending on the origin and purity of the sample, it may be appropriate to develop a multivariate classification model using either the entire fingerprint or a selected subset of peaks. Variable selection approaches may be useful in determining the optimal approach, depending on the nature of the samples and/or the ATP. It is crucial that the samples have an associated class label or quantitative property (activity) obtained by an objective reference technique (separate from that used to generate the fingerprint) to enable supervised modeling and subsequent model performance verification. Care should be taken to ensure that appropriate sources and levels of variability are included in model development and performance verification exercises based on the intended use of the test.

Several useful tools are available to the analyst for building the model, such as LDA and QDA, SIMCA, kNN, and PLS-DA. More complex techniques may also be used with justification (e.g., SVM).

Chemometric methods for multicomponent sample authentication are consistent with *Validation of Compendial Procedures* ⟨1225⟩*, Validation, Data Elements Required for Validation, Category IV*. Various numerical methods may be used to verify predictive performance. As discussed in *Qualitative Tools* for pure material identification, predictive performance may be verified by comparison of the TPR and FPR once a classification threshold appropriate to the application has been established. It is crucial to finalize this threshold before formal validation of the overall analytical method.

Classification: This is a supervised learning technique. The goal of calibration is to construct a classification rule based on a calibration set where both **X** and **Y** (class labels) are known. Once obtained, the classification rule can then be used for the prediction of new objects whose **X** data are available.

Classification models may be divided into two main groups: hard and soft. Hard classification directly targets the classification decision boundary without producing a probability estimation. Examples of hard classification techniques are kNN and expert systems. Soft classification estimates the class conditional likelihood explicitly and then makes the class assignment based on the largest estimated likelihood. SIMCA is a commonly used soft classification technique. In either scenario, the thresholds to determine the class boundaries are important and should be established during method development based on the resulting TPR and FPR, the requirements for which are determined by the ATP. These rates should also be verified as part of the full procedure validation.

In many applications, it is unacceptable to have a sample classified into two classes. Further, if a sample is not classified into any class, the sample is deemed an outlier to the calibration set. For example, a classification model built from NIR spectra of four materials (A, B, C, D) has four classes. Any material that has a source of variance (e.g., different water content) not previously seen and included in the model calibration may not be classified, and therefore should be identified as an outlier and investigated. Similarly, when a spectrum of another material (E) is applied to the model, the model should not be able to classify it into a given class. If similar samples are expected in the future, then the model should be revised to include the outlier samples in the calibration set, either to supplement the calibration data for existing classes or to create an entirely new class, depending on the nature of the outlier(s) and the specific requirements of the application.

4.2 Quantitative

GENERAL ASPECTS

When the intended use of an analytical procedure is the determination of one or more sample components or properties that have values that can vary continuously rather than categorically, then a quantitative model is required. Similar to the process for qualitative modeling, quantitative chemometric models are produced in a supervised manner using both the independent **X**-block and dependent **y** variables to ensure optimal predictive performance. The difference arises in the nature of the **y** values, which are continuous/numerical (rather than categorical) for quantitative applications. Per ⟨1225⟩, quantitative chemometric models are suitable for incorporation into Category I, II, or III methods for determination of the content of drug substance or finished drug product, impurity levels, or performance characteristics, respectively. In-process tests applied in an offline manner or an in/on/at-line manner that produce the value for a continuous variable would also represent appropriate uses of quantitative chemometric models. The specific requirements of the analysis per the ATP (see *3. Model Lifecycle*) should be used to guide the validation metrics used to demonstrate performance.

QUANTITATIVE TOOLS

The algorithmic tools needed for quantitative applications are wide ranging; for more complete descriptions see *Appendix, Additional Sources of Information*. Tools that are commonly used are briefly described here. As is the case with any chemometric

model, the tool must be matched to the ATP, and vice versa. Certain advantages and disadvantages are characteristic of each tool, and the chemometrician must be mindful of these when justifying their use.

Multiple Linear Regression (MLR): This tool establishes a correlation between a dependent variable (or **y**, response variable) and multiple independent variables (or **x**, explanatory variables) by fitting a linear equation to observed data. MLR has two broad applications, as follows.

The first broad application of MLR is to establish a model that explains the importance or magnitude of effect on the response variable, and this is typically used in the DoE screening or response analysis. Typically the input variables are mutually independent and by design represent large magnitudes. Often, the higher order of terms such as square terms and interaction terms are included, and a step-wise procedure is used to exclude those that make a smaller contribution to the model. Caution is advised to avoid having too many variables in the model, because including too many terms will overfit the model. The adjusted R^2 statistic is typically used to estimate whether there are too many factors in the model, although other metrics may also be appropriate. Adjusted R^2 is a modification of the standard coefficient of determination that takes into account the number of model terms and the number of samples, resulting in a plot of R^2 (adjusted) that has a maximum value beyond which additional terms result in overfitting.

The second broad application of MLR is to establish a prediction model for analytical purpose. MLR can be used to fit a predictive model to an observed data set of **y** and **x** values. The established model can be used to predict the value **y** from the **x** of new samples. The model for prediction purpose created from correlated **x** inputs is suitable to explore the relationship between **y** and individual **x**'s.

The prediction MLR model can be used in a procedure or method for a pharmaceutical compendial test. For example, a filter NIR, which is designed for a specific use such as water testing in the product, produces a "spectrum" of only a few data points.

The development and validation can follow the same procedures as are used with latent variable models such as a PLS. During the development and validation of an MLR model: 1) appropriate data preprocessing needs to be used; 2) cross-validation is used to estimate the model performance; and 3) an independent data set is used to assess the model.

A limitation of MLR is that model predictive performance may be negatively impacted when the modeled variables are collinear. The issue of collinearity is not unique to MLR. It is mitigated in PLS modeling via generation of a new basis set of orthogonal (noncollinear) latent variables that are linear combinations of the original variables. However, this is not relevant for MLR models. In MLR models, variable selection or transformation (e.g., mean centering) may be needed to avoid or reduce collinearity in the calibration set. Even though MLR models have some diagnostic metrics to identify outliers and highly leveraged samples in the calibration, they do not have diagnostic metrics (such as Hotelling's T^2 and **X** residual test) that PLS models offer during the prediction of new samples.

PCR: This is a two-step modeling approach in which a PCA is first performed on the **X**-block data to obtain a basis set of PCs. The PCs are then used in combination with the **y** data to develop a regression model to predict the **y** values.

PLS: This is one of the most commonly used chemometric algorithms for both quantitative and qualitative (PLS-DA, see *Qualitative Tools*) modeling scenarios. The computational advantage that PLS offers over MLR and PCR is that the PCs are derived in the PLS algorithm using the **X** and **y** data simultaneously. This often results in predictive models that require fewer latent variables compared to PCR, thereby improving robustness. PLS is capable of handling mildly nonlinear relationships, whereas PCR and MLR are linear modeling techniques. Further, more than one **y** variable can be modeled using the same **X**-block data via different algorithmic approaches. PLS-1 results in a separate model for each **y** variable. PLS-2 will produce a single model capable of simultaneously predicting the values of two or more **y** variables. The relative advantages of each approach may vary from one application to another. For instance, if different mathematical preprocessing is necessary to resolve each component to be predicted, separate PLS-1 models may provide improved performance over a single PLS-2 model.

QUANTITATIVE APPLICATION EXAMPLES

Pharmaceutical dosage form assay and/or uniformity of content: Assay of the API content in the final dosage form is commonly performed via measurement of a sample drawn from a homogenous composite of a number of individual dosage units (e.g., tablets or capsules). The quantitative result is then reported as an "average content" (expressed as percentage of target API amount) for the batch tested. Uniformity of content may be tested using a method similar to the assay, but measurements are performed on individual samples. The reported result is primarily based on the variability in assay results across the individual dosage units analyzed. Assay and uniformity of content methods are intended to characterize the "major" component in the sample(s) in question and are considered Category I tests per the definitions established in ⟨1225⟩. In the majority of cases, the target analyte will be the active ingredient in the sample, although quantitative assays for other components exist throughout the compendia as well.

For pharmaceutical solid dosage forms, the majority of assay procedures are based upon HPLC. HPLC analysis is time consuming and often involves the use of large volumes of solvent for the mobile phase. However, HPLC is highly linear and is typically calibrated using univariate mathematics (i.e., one measured variable is directly and uniquely proportional to sample concentration).

In contrast, alternative methods for assay that are based on spectroscopy (e.g., NIR and transmission Raman spectroscopies) offer the advantage of increased analysis speed and are nondestructive. However, in many cases the analyte signal may not display the same degree of linearity or signal-to-noise as with the corresponding HPLC method. Moreover, no single variable is directly and uniquely related to the concentration of interest. For this reason, multivariate models, often based on PLS regression, are commonly used in spectroscopic procedures for assay and uniformity of content.

For Category I tests, it is crucial to demonstrate the accuracy, precision, specificity, linearity, and range of results during the validation of the overall analytical procedure. Each of these performance aspects involves considerations during multivariate model development or validation that are distinct from those related to univariate techniques.

Most spectroscopic methods are not quantitative in an absolute sense. For example, models developed using NIR or Raman data must be calibrated relative to a primary reference technique (often HPLC or NMR spectroscopy). Given the nondestructive nature of NIR and Raman, it is straightforward to analyze a given set of samples using one of these techniques and then subsequently analyze the same set of samples via HPLC or NMR spectroscopy to obtain the reference content values. This

approach is typical for assays of solid dosage forms, however, it is worth noting that reference values for assays of some dosage forms, APIs, polymorphs, or excipients might be derived from other techniques such as gravimetric data recorded during the calibration sample preparation. The chemometric models are developed via a multivariate calibration using the NIR or Raman spectral data and the content values from the reference technique. Accuracy is determined by using the absolute difference between the model predictions and the reference values. Precision is determined by using the standard deviation of procedure results alone, such as from replicate analyses of the same sample. Thus, precision is a measure of performance for the analytical procedure as a whole. The model itself will be absolutely precise. That is, given the same spectral input, the same numeric output will always be generated. That said, precision may be influenced by the specifics of the chemometric modeling approach (e.g., preprocessing, modeling algorithm, number of latent variables) and therefore precision should be evaluated during model development as well as during validation. It is important to note that the accuracy of chemometric models will not be able to exceed the accuracy of the reference analytical technique used to calibrate the model. However, it may in certain cases be possible to exceed the precision of the reference technique.

The range of the method will be determined by the calibration samples that are used to develop the multivariate model. According to ⟨1225⟩, it is recommended that assay calibration samples should have analyte contents ranging from 80%–120% of the target amount. For evaluating uniformity of content, the recommended range is 70%–130% of the target amount. In some cases, both the assay results and uniformity results may be obtained from the same set of sample measurements. Averaging of a requisite number of individual dosage-unit assay results will provide a value equivalent to the "average content" parameter. Calculation of the "acceptance value" (see *Uniformity of Dosage Units* ⟨905⟩) using this mean result combined with the standard deviation for the same set of individual dosage unit results will provide a uniformity of content result. In this scenario, a single calibration set that spans the wider of the two recommended ranges (70%–130%) should be used.

If it is not practical to obtain a sufficiently wide range of content values in calibration samples produced at commercial manufacturing scale, calibration samples may also be produced in a smaller scale or in the laboratory. In this scenario, all attempts should be made to replicate the physical properties representative of the commercial scale. Additionally, care should be taken to verify that minimal bias exists between samples from different scales. Commercial-scale samples may need to be incorporated into the model (and/or procedure) validation process to verify the accuracy and precision of the model (and/or procedure) results against any differences and/or variability in physical properties. Mathematical processing algorithms (e.g., normalization, second derivatives) should be optimized as much as possible during initial exploratory data analysis (see *Qualitative Application Examples*) to mitigate prediction bias resulting from these factors. For instance, it is known that thickness, particle size, and density differences can lead to spectral slope changes and baseline offsets in NIR spectra. Depending on which specific effect is present, improperly selected preprocessing may not fully correct for spectral differences, leading to errors in prediction.

Fit and predictive performance of chemometric models should be demonstrated via the linearity of a plot of the model outputs versus the reference or nominal values of the target analyte(s), ideally using results from a set of independent test samples. The raw analytical signal may demonstrate a nonlinear relationship with the reference analyte concentrations. Likewise, multiple latent variables may be utilized for the chemometric model building. However, the key aspect of the model that must be demonstrated to vary linearly with analyte concentration is the model outputs, not the inputs. A plot of the prediction residuals versus concentration may assist in revealing any systematic lack of fit. Any observations of patterns in this residuals plot may indicate a need for revision of either 1) the number of latent variables included in the model, 2) the preprocessing mathematics, or 3) the algorithm type used for modeling.

Demonstration of the specificity of the procedure for assay may be based on likely sources of interference or material substitution based on an understanding of the material properties of the sample components and the manufacturing process and/or supply chain. One approach might be to verify that low assay results that are OOS are consistently obtained for a placebo version of the dosage unit. Another approach might be to verify that dosage units of a different product (especially one manufactured in the same facility and/or tested in the same laboratory) result in assay values that are consistently OOS.

Impurity limit tests: Routine testing is required for determination of impurities and/or degradants in intermediates, bulk drug substances, and finished pharmaceutical products. Testing for this purpose may take the form of a limit test (Category II, ⟨1225⟩).

Typically, these impurities tests are based upon HPLC or other methods (e.g., Karl Fischer titration for water). These analytical procedures are highly linear and are typically calibrated using univariate mathematics. However, these methods are often time consuming and involve manual sampling, which is disadvantageous, especially for products having high potency or for those with toxic intermediates produced during API synthesis.

In contrast, alternative methods for assay based on spectroscopy (e.g., NIR and Raman spectroscopy) offer the advantage of increased analysis speed, and more importantly, total elimination of human sampling due to the noninvasive nature of the measurement. However, these advantages are typically offset by decreased sensitivity (LOD) and potentially nonlinear sensor response across the required analyte concentration range as the corresponding reference method. For this reason, multivariate approaches combining spectral preprocessing and latent variable models are often employed in spectroscopic methods for these limit tests. (*6*) However, due to the potential challenges involved in achieving adequate sensitivity for a spectroscopy-based limit test, careful development and feasibility studies must be employed.

For a limit test in Category II procedures, a calibration set containing varied concentrations of the analyte of interest is often used to characterize the performance of such an analytical procedure. Although it is not required, these performance metrics could include accuracy, range, linearity, and others. The use of the performance metrics depends on the nature of the limit test.

In contrast, it is critical to illustrate specificity and LOD during validation of a limit test according to ⟨1225⟩. For specificity, a comparison between loading/regression vector and a spectrum representative of the pure component of interest is often used. Additionally, measuring matrix effects on the determination of the analyte within the specified range is another useful approach to demonstrate that the analytical procedure is specific to the analyte of interest without impacts from other variables introduced in the calibration set. Moreover, evidence of method specificity can be demonstrated by calculation of a selectivity ratio from the calibration data set, defined as the variance of model-reconstructed spectra divided by that of residual spectra. The selectivity ratio at the analyte absorption band is expected to be higher relative to other spectral regions.

Regarding LOD, there is no generally accepted estimator for PLS models. A common practice to determine LOD involves calculating a spectral signal-to-noise ratio. The model results for replicate analyses of a blank sample(s) may be used to represent

the noise. Alternatively, the noise may be approximated using the standard error of a multivariate model. Other more elaborate approaches were suggested to address the multivariate nature of PLS models. (*7*) With any approach, the estimated LOD should be verified experimentally using samples with analyte levels at or near that concentration. Because the LOD obtainable via a spectroscopic method will likely be higher than that obtainable via a chromatographic approach, it becomes crucial to demonstrate that the LOD based on the chemometric model meets the requirements of the ATP.

Dissolution testing: In vitro dissolution is a crucial performance test for solid dosage forms that is used to predict in vivo drug-release profiles during drug development and to assess batch-to-batch consistency for quality control purposes. Chemometric models can be used to predict the drug product dissolution from relevant measured product properties, and the method that uses such a model may be established as an alternative procedure. Alternatively, the dissolution profile can be modeled via several responses from different time points on the dissolution curve using individual models.

Dissolution data may be collected as a profile (i.e., a series of values obtained at various time points), even though the acceptance criteria may rely on a single time point. When building a chemometric model, it is essential for the model to be able to predict the entire profile. The number, and spacing in time, of points in the dissolution profile will vary depending upon the product type (e.g., immediate versus modified/extended release). A common approach is to transform the profile into a single value, which then serves as a dependent variable (or **y**, response variable). This variable, combined with another measured variable, can be used to restore the dissolution profile. The transformation can be accomplished by fitting the dissolution profile data into a mechanistic model (such as a first-order rate equation or its variants) or an empirical model (such as a Weibull function). After transformation, a few variables (typically, two or three variables) can be used to represent the dissolution profile. Generally, the two variables are a dissolution rate factor and a plateau factor; a third variable such as lag time may be necessary. The dissolution rate factor, which represents the rate of dissolution, will be the dependent variable for the modeling. The plateau factor represents the final amount of the drug in the solution, which should be equal to the drug content for most products. These two factors are used to restore the profile. Therefore, the model for dissolution prediction is used to predict the dissolution rate.

The key to designing a successful model is defining the input variables. The plateau factor may be obtained from drug content data, whereas the dissolution rate factor may be modeled from a data set of chemical and physical attributes of the product or intermediate product. The necessary knowledge of physicochemical properties and engineering concepts should be used to identify and justify the relevant inputs that have potential impact on the dissolution of the product. This typically involves risk assessment and an additional data collection step such as a small-scale DoE study. The inputs typically include the measured material attributes at various processing stages (such as particle size data for granules or blend; NIR measurements for blend or tablets; and physical tablet properties).

There are two approaches for defining samples in the model calibration: batch samples and individual unit dose samples. For the batch-sample approach, one product batch is treated as a sample, and the input variables are the attribute means of the batch. The individual unit dose approach is to measure the attributes of individual unit doses (such as NIR measurement on the individual tablets). The input variables from both approaches may be supplemented with other raw material properties as justified by the critical quality attributes.

The validation of the model-based dissolution prediction method could be different from the typical HPLC dissolution method validation, depending on the specific approach used. The HPLC dissolution method puts more emphasis on precision, as the dissolution is treated as a Category III method in ⟨1225⟩.

In the batch-sample approach, the model-based method treats dissolution as a batch property instead of a property of individual tablets, and accuracy is the focus of method validation or verification. For example, each batch forms a sample, and it is not realistic to have a large validation sample set. This approach generates a single profile for a batch, and the result at a certain time point can be used to assess the product batch quality. Evaluation of variability must be aligned with the criteria in *Dissolution* ⟨711⟩. The variability of the final dissolution result can be evaluated by analyzing the variations of inputs. Possible ways to evaluate the variation of the sample (batch) include the analysis of variation of inputs, simulations, and dividing a batch into multiple sub-batches.

In the individual unit dose approach, the model-based method evaluates individual dosage units. If many tablets are tested, this becomes a large *n* situation in which the acceptance criteria in ⟨711⟩ should be used with caution. The chemometrician may propose acceptance criteria as long as these criteria are demonstrated (by simulation or other means) to have equivalent or tighter passing criteria than those in ⟨711⟩. In the method validation or verification, it is desirable to have at least one sample that shows low dissolution (near or below the specification), and the chemometric model-based method should demonstrate the capability to distinguish this low-dissolution sample from normal samples.

GLOSSARY

Many common chemometric terms have been defined within the text of this chapter. Some additional terms are defined here and serve as a reference to their usage in the text. For a more complete description, consult the texts listed in *Appendix, Additional Sources of Information*.

Calibration model: A mathematical expression used to relate the response from an analytical instrument to the properties of samples, or to capture the underlying structure of a calibration data set.

Calibration set: Collection of data used to develop a chemometric classification or model.

Derivatives: The change in intensity with respect to the measurement variable (i.e., abscissa). Derivatives are useful for removing baseline offsets (constant or linear) due to sample properties, or for highlighting small changes in a signal, helping to enhance selectivity and sensitivity (e.g., one spectroscopy and chromatographic peak with a shoulder peak adjacent).

Internal validation: The application of resampling statistics such as cross-validation. Subsets of the calibration data set are subjected to a variety of statistical processes to identify which calibration model best fits the available data. Each model is characterized by a statistical parameter. For cross-validation, the entire data set of samples is split into individual samples or groups of samples, which are removed individually from the rest of the samples and tested as unknowns against a calibration model constructed using the rest of the samples. The characteristic statistic is the standard error of cross validation (SECV).

Matrix: A two-dimensional data structure comprised of columns and rows used to organize inputs or outputs during chemometric analyses. Common practice is for each row to correspond to an individual sample and each column to correspond to an individual variable.

Validation set: The data set that challenges the performance attributes of the model. The validation data set is independent of the training data set, although testing on the training data set gives an optimistic view of performance and thus allows for iteration of preprocessing and model tweaking until this optimistic view meets the expectations. Validation or cross-validation is then necessary to adequately gauge the performance of the model. The model performance with the validation data set will always be equal to or worse than its performance with the training data set. If the performance attributes are not met, then the chemometrician must assess whether the model underfits or overfits the data and whether iterations of the model are needed to meet the acceptable error rate. If these efforts fail, it is recommended to back up one step, alter the preprocessing conditions, and perform the same task again.

APPENDIX

Additional Sources of Information

Many books have been written on the subject of chemometrics and multivariate analysis. Various terms not found in the glossary may be found in this short list of additional sources of information.

- Massart DL, Vandeginste BGM, Buydens LMC, De Jong P, Lewi PJ, Smeyers-Verbeke J, eds. *Handbook of Chemometrics and Qualimetrics: Part A.* 1st ed. Amsterdam: Elsevier Science B.V.; 1997.
- Vandeginste BGM, Rutan SC, eds. *Handbook of Chemometrics and Qualimetrics: Part B.* 1st ed. Amsterdam: Elsevier Science B.V.; 1998.
- Mocak J. Chemometrics in medicine and pharmacy. *Nova Biotechnologica et Chimica.* 2012;11(1):11–25.
- Singh I, Juneja P, Kaur B, Kumar P. Pharmaceutical applications of chemometric techniques. *ISRN Anal Chem.* 2013; article ID 795178. http://www.hindawi.com/journals/isrn/2013/795178/. Accessed 25 May 2016.
- Brown S, Tauler R, Walczak B, eds. *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis.* 1st ed. Amsterdam: Elsevier Science B.V.; 2009.
- Varmuza K, Filzmoser P. *Introduction to Multivariate Statistical Analysis in Chemometrics.* Boca Raton, FL: CRC Press (Taylor & Francis Group); 2009.

REFERENCES

1. Massart DL, Vandeginste BGM, Buydens LMC, De Jong P, Lewi PJ, Smeyers-Verbeke J, eds. *Handbook of Chemometrics and Qualimetrics: Part A.* 1st ed. Amsterdam: Elsevier Science B.V.; 1997.
2. Massart DL, Vandeginste BGM, Deming SM, Michotte Y, Kaufman L. *Chemometrics: A Textbook*. 1st ed. Amsterdam: Elsevier Science B.V.; 1988.
3. Brown CD, Davis HT. Receiver operating characteristics curves and related decision measures: a tutorial. *Chemometr Intell Lab Syst.* 2006;80:24–38.
4. World Health Organization. General guidelines for methodologies on research and evaluation of traditional medicine. Geneva, Switzerland: World Health Organization; April 2000. http://apps.who.int/iris/bitstream/10665/66783/1/WHO_EDM_TRM_2000.1.pdf. Accessed 25 May 2016.
5. Food and Drug Administration. Guidance for industry: botanical drug products. Rockville, MD: Food and Drug Administration; June 2004. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm070491.pdf. Accessed 25 May 2016.
6. Lambertus G, Shi Z, Forbes R, Kramer TT, Doherty S, Hermiller J, et al. On-line application of near-infrared spectroscopy for monitoring water levels in parts per million in a manufacturing-scale distillation process. *Appl Spectrosc.* 2014;68(4):445–457.
7. Allegrini F, Olivieri AC. IUPAC-consistent approach to the limit of detection in partial least-squares calibration. *Anal Chem.* 2014;86:7858–7866.