8. Principles of Proper Validation (PPV)

This chapter builds directly on the previous chapters on sampling (chapter 3), PCA (chapter 4) and multivariate regression (chapter 7). Validation is presented using the example of multivariate calibration/prediction, but the general principles apply to all data modelling for which a performance is to be assessed, e.g. model fit, classification, prediction, time-series forecasting etc.

8.1 Introduction

A set of generic Principles of Proper Validation (PPV) is presented, based on five distinctions:

i) The PPV are universal and apply to all situations in which validation assessment of performance is desired: modelling, prediction, classification, time series forecasting.

ii) The key element behind PPV is the Theory of Sampling (TOS), which provides insight into all variance-generating factors, especially the "Incorrect Sampling Errors" (ISE). If not properly eliminated, these are responsible for an *inconstant* sampling bias, for which no correction is possible, contrary to the widespread statistical tradition for "bias correction" (see also chapter 3). A sampling bias wreaks havoc with all types of validation—except test set validation. The sampling strategy should always address how known (and unknown) sources of variation can meaningfully be included in the first data-set which is to be modelled, the training set. Thus, sampling includes qualitative information about time, location, batch ID and other qualitative information related to the sampling strategy, and not only the physical procedure. Such a comprehensive overview is imperative for a reliable estimation of future prediction or classification performance.

iii) Validation cannot be understood solely by focusing on the *method(s)* of validation—it is not enough to be acquainted only with a particular validation *scheme, algorithm* or *implementation* as has been a longstanding tradition in chemometrics. Validation must be based on full knowledge of the underlying definitions, objectives, methods, effects and consequences of the full sampling–reference analysis–data analysis process.

iv) Analysis of the most general validation objectives leads to the conclusion that there is one valid paradigm only: test set validation. In this chapter, the most important alternative validation approaches are discussed, critiqued and rejected in this perspective.

v) Contrary to contemporary chemometric practices and validation *myths*, cross-validation is unjustified in the form of a *one-for-all* procedure for all data sets. Within its own methodological scope, cross-validation is shown to be but a suboptimal *simulation* of test set validation, as it is based on one data set only (the training data set). However, such a "first" singular data set *could*, occasionally, be representative of future variation, but one would never know whether this is the case within the one-sample-set context alone; only external experience or evidence would be decisive. Many re-sampling validation methods suffer from this principal deficiency. Thus, while there are *cases* in which cross-validation finds excellent use, for example, categorical segment cross-validation (in which categories can be batches, seasons, alternative models or pre-treatments), see below, such cases represent only special circumstances and no generalisation regarding validation principles can be made hereupon.

This chapter shows how a second data set (test set, external validation set) constitutes a minimum critical success factor for inclusion of the sampling errors incurred in all "future" situations in which the validated model is to perform. From this follows that all re-sampling validation approaches based on a singular data set only (for example a data set sampled on one day only, or with one instrument only, or addressing one batch of raw materials only) should logically be terminated, or used only with full scientific understanding

and disclosure of their detrimental limitations and consequences. This chapter builds on a comprehensive analysis of validation in Esbensen and Geladi [1]; see also Westad and Marini [2].

In the central case of PLSR, a call is here made for stringent commitment to test-set validation based on graphical inspection of pertinent *t–u* plots for optimal understanding of the underlying **X**–**Y** data structures and interrelationships for validation guidance. The *t–u* visualisation, or similar regarding other multivariate data modelling methods in need of validation, is critical in order to stop continuation of decades of "blind" use of a one-for-all procedure cross-validation, a state of affairs that has led to quite some confusion among generations of chemometricians.

8.2 The Principles of Validation: overview

Throughout the history of chemometrics discussions have often surfaced as to what constitutes *proper* validation. There are few other topics which have led to a more marked and heated set of different *opinions*. Discussions on this basis can be broad-ranging and informative, while at other times personal, emotional and counter-effective. The matter at hand is a thoroughly scientific one, however. In chemometrics, validation is most known in the context of prediction validation, of which there are (at least) four types: test-set validation, cross-validation, "correction validation" (leverage correction is the prime example) and re-sampling validation methods (bootstrap, jackknife, Monte Carlo simulation, permutation testing). The standard cross-validation can be viewed as a single instance of the jack-knife re-sampling procedure (Chapter 7, section 7.17)—and it can also be viewed as a simulated test set validation, albeit a fatally flawed simulation.

This chapter illustrates the central PPV by a phenomenological analysis of prediction validation and its objectives in the specific multivariate calibration context.

The PPV are concerned with the question of how to establish a general validation approach that does not depend upon *assumptions* of any specific data structure(s), nor associated with any specific variant of the many validation method alternatives that can be found in the literature.

A few salient definitions are needed at the outset—strictness and preciseness is a much-needed commodity in the validation debate:

proper *adj.*: adapted or appropriate to the purpose or circumstance
valid *adj.*: sound; just; well-founded; producing the desired result
validate *v.t.*: to make valid; substantiate; confirm

One reason for much of the often deeply felt differences-of-opinion regarding what constitutes proper validation relates to the fact that it involves both statistical as well as domain-specific issues (e.g. chemical, physical, data analytical or other error issues). A significant proportion of the historical debate simply reflects a much too limited point of departure from which is typically attempted to draw far too sweeping generalisations—for example, the belief that validation is exclusively a statistical issue. Within this view "sampling" is simply a matter of drawing from a population of independently and identically distributed (i.i.d.) measurements; this understanding is termed *statistical sampling* ($sampling_{STAT}$) in what follows.

In the present context, a broader, TOS-based holistic understanding of the interconnected sampling, analysis and validation issues is advocated, while taking care not to fall into the opposite, equally simplistic position, viz. that all data matrices result from sampling from heterogeneous material. However, most of the activities in the field of data analysis and data modelling, in fact, do occupy a realm in which one has to assume the presence of significant data errors. If these errors are neglected, they will cause grave prediction and validation problems. It is, therefore, also necessary to use the term $sampling_{TOS}$. The cases in which pure statistical sampling suffices, these can simply be treated in an identical fashion alongside the much more prevalent sampling-error cases, allowing for unity in all validation considerations. The issues regarding validation are neither about opinions (personal, institutional), nor about following one or other established schools-of-thought or traditions (thereby dodging a personal responsibility for understanding and method selection). All validation issues are fully tractable and lend themselves to rational discussion and sound, objective analysis that ultimately lead to impartial conclusions.

8.3 Data quality—data representativity

The PPV need a few initiating discussion points related to the concepts of *data quality, data representativity* and *sample representativity*.

"Data quality" is a broad, but often loosely defined term; any definition that does not include the specific aspect of representativity is lacking, however. The term "data" is often equated with "information", but it is obvious that this can only be in a *latent*, potential form. It takes *data analysis* with appropriate, problem-context *interpretation* to reveal the structural "information" residing in data matrices. In chemometrics, the prime interest is, of course, on data analysis, while issues pertaining to the prehistory of a data table usually receive but scant attention. In fact: "*Chemometricians analyse data…*" is an often-heard statement going a long way towards rejecting any chemometric responsibility for data quality, and hence also of sample representativity. Nothing could be more dangerous, however. One exception is Martens and Martens [3] which addresses "multivariate analysis of quality", where the focus is stated to relate to the "quality of information", which is defined as "...dependent on reliability and relevance". However, *reliability* and *relevance* are open-ended, general adjectives which, must be given a specific unambiguous meaning from the problem context at hand. It has, therefore, been argued that a far more relevant characteristic is *representativity*, partly because a clear definition is at hand, but mainly because the specific derivation of this definition in TOS allows for comprehensive account of the underlying phenomenon of *heterogeneity*.

It is mandatory to contemplate the specific *origin* of any data set and in this context, data analysis is always dependent upon at least one primary sampling stage in order to produce the sample, PAT or sensor signal acquisition stage no exception, often including mass-reduction and sample preparation in later sampling stages. An analytical stage *i.e.* (chemical, physical, measurement etc.) is also required before data analysis can commence. It is, therefore, an inescapable conclusion that "reliable analytical information" must be based on *representative samples*. In this chapter, a critical distinction is made between *statistical sampling* and the kind of *physical sampling* addressed by TOS. In chemometrics, it is necessary to be competent with respect to both these kinds of "sampling".

There will always be large, significant or alternatively only small sampling/signal acquisition errors involved—the point being that at the outset this quantitative issue is *unknown*, and therefore cannot be dismissed without grave danger. In chemometrics, the type of errors colloquially known as "measurement errors" are mostly considered to be related to the **X**-data only, typically conceptualised in the form "instrumental measurement errors", while they logically also must refer to analytical errors pertaining to "reference measurements" (**Y**-data in calibration). These effects are all incorporated into the concept of Global Estimation Error (GEE) within the realm of TOS, thus allowing a rational discussion of all sampling and analysis errors and their impacts on data quality. By dealing universally with these issues as if sampling issues *were always* significant, all cases can be treated identically in a rational and efficient manner, covering all combinations of large and/or small statistical errors as well as large/small TOS-sampling errors (including the rare, pure statistical case alluded to above).

The position most often met throughout the history of chemometrics is that of simply *assuming* all sampling errors are *insignificant*. This attitude represents a fatal illegitimate generalisation for which no proof has ever been presented. Chemometric data analysis without sufficient attention to the full context of relevant pre-data issues cannot be considered comprehensive, but is in fact incomplete, indeed unscientific. It is noteworthy how the complex validation scenario is all too often simply swept under the rug in the quest for a one-for-all method that automatically takes care of all the troublesome issues. This is called "Chemometrics without thinking".

8.4 Validation objectives

Validation, in the multivariate calibration context, means assessing that the prediction performance is *valid*, i.e. to confirm that a particular prediction model *is fit for purpose*. Usually the prediction performance is specified as a certain prediction uncertainty (error) maximum threshold, or similar. But it is known that the prediction accuracy is a characteristic that must relate

back all the way to the original lot/material (or population in certain cases), which translates to the prediction accuracy being a performance characteristic that refers to the features of the original lot, represented by the reference analytical values, which also require proper validation in order to be used to calibrate a model.

This objective does not refer only to the technical calibration/modelling and validation process, but first and foremost to the circumstances surrounding the future performance for new predictions when using the model on "similar data". This means that *both the training and validation data sets must be as similar as possible to those* ***new*** *data sets pertaining to the "future" working situation in which the model is to perform its task*. The training set and the test set should not be as identical to one-another, as is a current misconception.

Thus, already when designing and selecting a training data set for modelling it is imperative also to pay attention to how the model is to be validated. This means, that one must always try to be in a position also to be able to choose at least an additional, second independent data set, to be used only for validation. Such a data set is hereafter generically called the *test set*. This is the data set with which to represent the future working situation of the particular model. As shall be clear below, at times this may demand some work, sometimes a lot of work—there may not always be easy fixes here. Irrespective, however, all prediction models must be validated w.r.t. *realistic* future circumstances. It is simply not good enough to secure double as many objects (samples, measurements) as what appear sufficient for modelling and then slice off 50% (performing the so-called "test set split")*. It will become clear that there is much more involved in establishing a realistic, reliable validation foundation. This is not to say that the splitting of a pool of samples into calibration and validation sets is *always* bad, it just requires a relevant, problem specific approach to the design of the data sets involved; above all it requires a complete understanding of all the issues treated in this chapter.

In data analysis, statistics and chemometrics, ~20 years ago, there was a somewhat rude awakening to the fact that far too little prediction validation was on the agenda at the expense of mere modelling. In Höskuldsson's [4] reassessment of the entire realm of "Prediction methods in science and technology", it was described how modelling fit assessment reigned pretty much supreme as compared to the necessary, *complementary* prediction validation, for which he introduced the "H-principle" of balanced assessment of both modelling and prediction performance. Today, there is a much more widespread awareness that modelling fit optimisation is necessary, but there is still not a sufficiently well-known criterion for prediction performance. Høskuldsson pointed to the Heisenberg Uncertainty principle from quantum mechanics when naming his H-principle of balanced modelling and validation *complementarity*.†

8.4.1 Test set validation—a necessary and sufficient paradigm

The central theme of the present approach can be stated in quite unambiguous terms: All other validation methods are but *simulations* of test set validation, with various flaws.

simulate *v.t:* to *assume* or have the *appearance* of characteristics of

The objectives of test set validation are always structurally correct and complete. If a proper test set *was* always obtainable (and this is not *that* difficult, see further below), no other validation procedure need ever have been introduced; test set validation would then be the only validation method in existence.

* In chemometrics, there has been a key terminology confusion regarding the terms, sample, object, observation, measurement a.o. The duality regarding "sample" in the statistical vs the TOS contexts is covered comprehensively in this book. An "observation" can be understood as a *passive* measurement. A physical sample may result in several "replicate" measurements, which then would become individual objects when represented in a data matrix. Chapters 3, 8 and 9 have made a profound effort to clear up these confused terminology issues.

† Chemometrics owe Høskuldsson a very great depth of gratitude for his seminal 1996 treatise; the H-principle name may very well connote the last name of this chemometric author as well.

8.4.2 Validation in data analysis and chemometrics

Validation can be used for many different purposes; it is relevant to speak of *internal* as well as *external* validation scenarios. Below are discussed both legitimate and illegitimate approaches to validation focused on multivariate calibration for prediction purposes. Thus, cross-validation used on one data matrix (**X**) only, i.e. PCA (chapters 4 and 6) and SIMCA (chapter 10) is not covered in full, but neither is this necessary for the purposes of this explanation. It is straightforward to apply PPV for these cases, as most aspects of the analysis, discussion and conclusions from the prediction scenario can be carried over without loss of generality. It is the overarching *principles of validation* which are in focus, followed by examples and data analysis assignments which will allow the developing data analyst ample opportunity to become familiar with all the intricacies of practical validation.

8.5 Fallacies and abuse of the central limit theorem

Abraham de Moivre's *Central Limit Theorem* [5], also called *normal convergence theorem,* states: A collection of means of *reasonably large subsamples* taken from a *large* parent population form a population that is normally distributed around the mean of the parent population, no matter what the distribution of the parent population is.

It is critically important to note that *large* statistical subsamples are required and that the mean of the parent population is only found in the limit for many such statistical subsamples. On many occasions this point appears to be gravely misinterpreted or forgotten. Re-sampling on a single data set (e.g. the training data set), often of a significantly *small size*, or even on a fractional subset of a training set, can only lead to knowledge about this very set alone—and only very little, if any, useful knowledge about the parent population and much less as to future samples not yet sampled. Much of the popularity of cross-validation may be based on a too-swift dependence on the respectability accorded to the central limit theorem. See Wonacott and Wonacott [6] and Devore [7] or many current statistics textbooks for a full description of the central limit theorem.

The key issue here is that the singular training set is *supposed* to carry all possible and necessary information about the background population, including that it is also able to represent any other (new) data set(s) to be sampled in the future. In a very real sense the size of the training set is only but the first quality criterion for getting this statistical inference correctly started; much more important is the "coverage" of future data sets as compared with that of the training set. Small training data sets are often seen re-sampled/subdivided into even smaller segments which are supposed to perform the role of a "reasonably large subset" in the sense of de Moivre above. Clearly a very dangerous practice! It is questionable how often this fundamental limitation is known, far less respected, when doing a "routine cross-validation" on a typical training data set with but a relatively small number of objects, often less than, say, 50. Add to this the central message of chapter 3, i.e. significant presence of non-stochastic sampling/signal acquisition errors (TOS-errors). The most common argument encountered by typically inexperienced chemometricians is that a too small training set is available for test set validation, so… "Just perform cross-validation… and all will be well". Well, not so fast….

8.6 Systematics of cross-validation

It is advantageous to treat all cross-validation variants under a systematic heading, termed *segmented cross-validation*. This allows significant simplification in discussing the historically disparate variants: Leave-one-object-out (LOO), the plethora of differently segmented cross-validation variants, including the so-called "test set split" option (a particularly obfuscating terminology for an otherwise straight two-segmented cross-validation approach). Indeed, two of these names could not have been chosen in a worse fashion; "test set split" is a terrible misnomer as no test set can ever be created from this procedure—and "full cross-validation (LOO)" is actually the worst of all possible segmented cross-validations, to be explained below. The formal definition of "segmented cross-validation" is as follows.

Depending on the fraction of training set samples (totalling *N*) held out for cross-validation, an optional range of "*s*" potential validation *segments* will be available for the data analyst, the number of segments falling in the interval $s = [2, 3, 4, ..., (N-1), N]$. Various "schools-of-thought" of cross-validation have developed over history within chemometrics and elsewhere, some favouring "full cross-validation" (one object per segment, resulting in *N* segments in total), some defining 10 segments as the canonical number and others favouring similar *schemes* each with its own preference (e.g. 3, 4 or 5 segments), whereas a small, but steadily growing minority see more complexity in this issue than a more-or-less arbitrary selection from the full range of *s* options. Reflection reveals that there always exist (*N*−1) *potential* cross-validation variants for **any** given data set with *N* samples, but no set of principles for objective determination of the optimal number of segments has **ever** been offered in the data analysis, chemometric or statistical literature.

Below, some hard-won experiences with hundreds of (very) diverse data set types and data structures are presented that will allow an easy overview of the systematics of validation.

8.7 Data structure display via *t–u* plots

The canonical formulation of the PLSR-1 algorithm (**X** and **Y**: mean-centred and scaled as needed) stipulates:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 \quad (8.1)$$

$$\mathbf{u}_1 = \mathbf{y}q_1 \quad (8.2)$$

where $\mathbf{t}_1$ is the first PLS **X**-score, $\mathbf{w}_1$ is the first PLS **X** loading-weight, $\mathbf{u}_1$ is the first PLS **y**-score and $\mathbf{q}_1$ is the first PLS **y**-loading.

The PLSR algorithm also calculates higher order sets ($\boldsymbol{t}_a$, $\boldsymbol{w}_a$, $\boldsymbol{u}_a$ and $\boldsymbol{q}_a$) [$a = 2, 3, 4\ldots$], after suitable deflation allowing the following general appreciation. Plotting $\mathbf{u}_a$ against $\mathbf{t}_a$, [$a = 1, 2, 3, 4\ldots$] reveals the succession of the so-called "inner relationships", which are direct visual manifestations of the data structure present. It is precisely these plots that are used as a vehicle for visual assessment of outliers etc. in any regression context. The "*t–u* plots" also constitute a useful check of whether a possible next component seems meaningful or not, as evidenced by the strength of "inner" partial regressions as the dimensional index, *a,* is increased by one. The *t–u* issues are identical also for the PLS2 case, in which $\mathbf{u}_a$ is no longer a scaled (and sequentially deflated) version of **Y** alone, but a linear combination of all *Q* **Y**-variables.

Although the PLSR algorithm *can* be written without explicit projections in the **Y**-space, i.e. without equations 8.1 and 8.2, see e.g. Martens and Næs [8], Martens and Martens [3], there is a serious loss of potential information about the data set in doing so, as the purpose of getting visual insight into the empirical data structure is completely lost. It is immensely informative to assess the specific data structure by *t–u* plots. In order to be able to take proper action with respect to the actual data structure present in training, test or "future" data sets, a general *typology* of the principal types of data structures associated with multivariate calibration is presented in Figure 8.1.

There are three underlying features characterising the particular manifestations of any multivariate *t–u* data structure: i) the number of objects involved, *N*, ii) the degree of linear (or non-linear) correlation present between **X** and **Y** and iii) data clustering, grouping ("clumpiness") and/or significant outliers. The four first cases shown in Figure 8.1 constitute a systematic series of strong/weak correlation *vs* small/large *N*, outlining the full spectrum of typical data sets for which PLSR modelling is relevant and legitimate.

As an important contrast, three of the four latter cases in Figure 8.1 represent deviating covariance data structures for which PLSR-modelling should *never* even have been contemplated. It is obvious, that without *t–u* visualisation, such cases run a very high risk of going unnoticed. The validation literature is ripe with examples of over-generalised validation discussions and conclusions—which, when shown on the simple *t–u* plot simply to be related to such degenerate data structures, are in reality *nonsensical* and which should never have been published. Collegial considerations disallow specific references.

It cannot be overemphasised how much data structure information can be gained from diligent *inspection* of components cross plots. Many such plots are used throughout this book, complete with interpretations of the meaning of the data structures

revealed. This is perhaps one of the strongest attributes of PLSR. Contrary to the case for PCA, in which there only exist *t–t* plots, in PLSR (as in PCR), the *t–u* cross-plots displays highly relevant insight into the **X**–**Y** *inter-space* data structures. While *t–t* plots can always also be called up for PLSR model inspection, these are normally only consulted depending upon specific purposes behind the regression modelling in which the **X**-space data relationships are of interest in themselves, For the so-called "PLS-constrained **X**-space modelling" the objective is often specifically to observe the **X**-space projections, and here inspection and interpretation of *t–t* plots will play a central role, while for ordinary feed forward **X→Y** regression model building, *t–u* plots are the relevant information carriers.

A key issue of the highest value for data analysis beginners is the following: PCA-like *t–t* plots is not the proper, and very far from the most efficient, place to look for outliers when building a regression model (PLSR, PCR). Identification of outliers influencing PLSR modelling can most meaningfully, and shall here in fact only, be pursued on the series of relevant *t–u* plots. This is an insight that will spare novice data analysts a lot of grief and which will cut short an otherwise long learning phase.

With the help of *t–u* plots, it is easy to appreciate that an empirical close match between cross-validation and test set validation results (evidenced by "similar" A_{opt} and *RMSEP*) which is simply a manifestation of a particularly *strong correlation between* the **X**- and **Y**-spaces, e.g. cases a) and c) in Figure 8.1. But this holds exactly **only** for strongly correlated *t–u* data structures from which it follows that no generalisation is allowable to other data structures. This is an example of an *illegitimate* generalisation, because it is based on a particular data set structure only. The chemometrics literature is ripe with examples of such invalid transgressions within the realm of validation.

t–u plots must be inspected for every regression model to be validated as it will elucidate the underlying relationship between **X** and **Y** (sample groups, non-linearities) and also to indicate the correct model dimensionality.

There are several traditions within chemometrics that do not adhere to this flexible understanding, however, which instead prescribe "blind" adherence to one particular version of cross-validation with **no** graphic inspection, i.e. which rely on "blind" cross-validation with a fixed number of segments for **all** data sets (aptly called *straightjacket cross-validation*). But even a cursory overview of the principal correlation relationships delineated in Figure 8.1 leads to the insight that a fixed number of segments will work in markedly different ways depending on the *specific* data structure encountered. This goes a long way to explain why repeated cross-validation, identical but for alternative starting segment definitions, can lead to significantly different validation results. A fixed number of segments can never be said to pay the necessary tribute to the many *different* data structures met with in the realm of science, technology and industry. One, rigid scheme most emphatically **does not** fit all. There is ample justified reservation as to the plethora of *claims* in the literature, all hailing the so-called "robustness" of cross-validation. These claims are simply wrong.

By way of contrast, based on Figure 8.1, all types of varying validation results are completely *comprehensible*—a simple mental picture of selecting a fraction of the *N* objects displayed in a *t–u* plot (which is tantamount to selecting a segment in cross-validation) allows the data analyst to picture the resulting effect of sub-modelling of the remaining objects (perhaps first after a little experience, but that is exactly the reason for working through a chemometric textbook…).

Upon reflection, these relationships are but the reverse issue of the often claimed optimistic, but not fully thought through, cross-validation credo: validation is on safe ground as long as/if several variants of validation, including several different segmented cross-validations result in *similar* validation results (identical number of components, "similar" *RMSECV*). All is indeed well if-and-when this hopeful situation occurs—however, the only thing that has been demonstrated is a case of a strongly correlated **X**–**Y** relationship, as depicted in Figures 8.1a, c, f or h. In reality nothing has been revealed as to the *future* prediction potential, unless it has been independently proven beyond reasonable doubt, that this strong **X**–**Y** correlation remains the defining feature *also* of *all possible* "future" data sets on which the regression model is to perform. Such a relationship cannot ever be taken on blind faith, however, but should be substantiated based on theory and background knowledge about the actual application. This would correspond to believing that all training

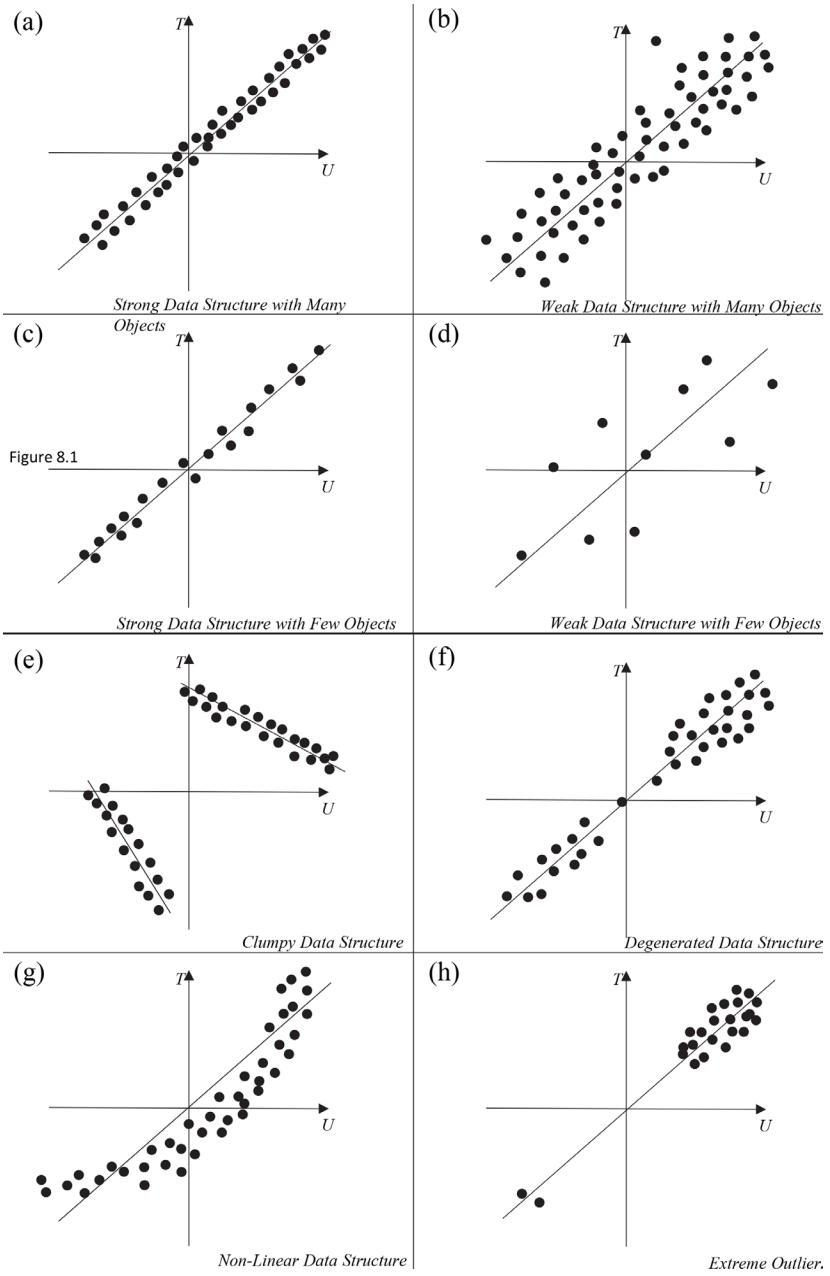

Figure 8.1

Figure 8.1: Eight correlation data structures as depicted in PLSR *t*–*u* plots. There are three features that determine the appearance of a *t*–*u* data structure: i) the number of objects, *N*, ii) the degree of linear (or non-linear) correlation present and iii) data clustering or grouping ("clumpiness"). An attempt has been made to cover as well as possible all principal data structure types met with in practical data analysis. Cases e) and h) should never have been subjected to regression modelling in the first place; case g) can not be modelled with a one-component model unless a satisfactory liearising transformation has been employed—case g) can be modelled straightforwardly, however, with an excess of components.

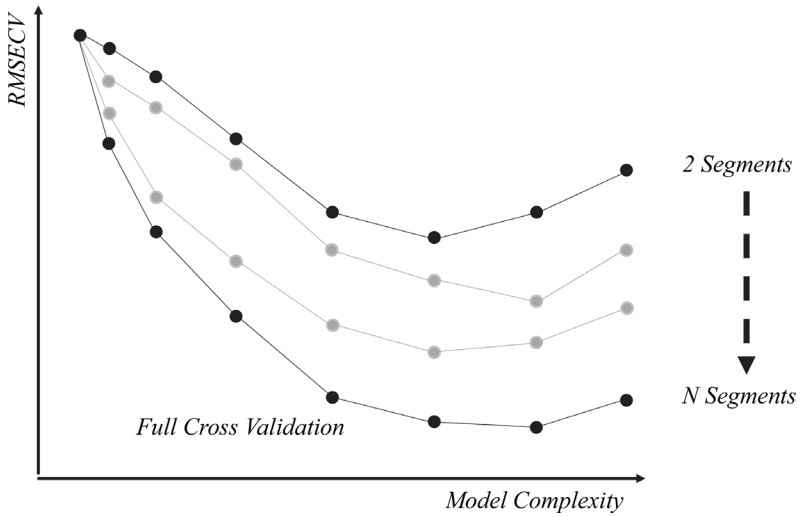

Figure 8.2: Schematic behaviour of *RMSECV*-estimation (prediction **Y**-error variance) based on the full range of cross-validation segments available to all particular data sets with *N* objects *s* = [2, 3, 4, ... *N*].

data sets are always and at all places… a 100% valid and reliable representation of the potential myriad of all future data sets $sampled_{stat}$ (from a population), or $sampled_{TOS}$ from a heterogeneous target. If this were indeed so, there would be no need for validation: the training set modelling fit would be all that was ever needed, as it would be *universal*—alas, reality checks in here with a very different lesson, see chapter 3; this is a hopelessly naïve belief.

Demonstration of whether such a situation only holds *locally*, i.e. for one *specific* data set, or not, is precisely the reason behind, the objective of, including the second data set (test set) in the validation, while all re-sampling approaches only deal with the singular $\mathbf{X}_{train}$ set exclusively.‡

Still more insight can be gained from careful inspection of Figure 8.1. For all data sets of the type like cases a–d) one observes a systematic regularity w.r.t. alternative segmented cross-validations with a varying number of segments [*s* = 2, 3, 4, … *N*]. Figure 8.2 depicts the systematics of "*RMSECV* vs # PLSR-components" plots, corresponding to the progression of all (*N* – 1) *potential* segmented cross-validations for a particular data set.

There will always be a *lowest RMSECV* when the number of segments is at its maximum, *N* (corresponding to leave one out cross-validation, LOOCV). Conversely, when *s* = 2, *RMSECV* will be at its *maximum*. These relationships hold for all reasonably *regular* correlation data structures when cross-validation is performed on one-and-the-same data set. Exceptions may occur, these relationships may be slightly less regular, but then only due to some influential data structure *irregularity* ("clumpiness" or presence of adverse outliers, especially "transverse" outliers), which will tend to blur the general pattern slightly. But the point is, again, there is never any generalisation potential beyond the particular *local* data structure.

For the same data set, a reduction of the number of segments will, in general, result in an increase in the *RMSECV* error estimate, and *vice versa*, but there is no reason for confusion: different validation setups will result in different validation outcomes; the number of components *may* change in more influential cases,

‡ One test set is the minimum requirement… but more can, of course, also be contemplated in specific situations, especially in cases where data set structures are varying more than what is comfortable. This issue is particularly relevant for PAT implementations and other process data modelling realms, see chapter 13.

and the numerical estimate of *RMSECV will* always change. Faulty conclusions may easily result if, for such a *particular* data set, the data analyst succumbs to the temptation to select the cross-validation setup that corresponds to the lowest *RMSECV*. This may at first *appear* as a legitimate cross-validation outcome, but it is only based on a subjective desire to select an "optimal model". Such a *voluntary* approach is untenable, indeed unscientific, biased and subjective, to say the least. In this book, the reader will be assigned the task personally to investigate and substantiate to which degree the general behaviour depicted in Figure 8.2 holds for the example data sets supplied. Figure 8.2 is based on jointly accumulated 50+ years of validation experience between the contributing authors (which constitute only a minute fraction of the very many practicing chemometricians who have had the same experiences).

Careful inspection of the pertinent *t–u* plot of any multivariate calibration model is thus the only way to fully understand and interpret results stemming from otherwise "blind" cross-validations. This is possibly a reason why some data analysis and chemometric traditions tend to avoid inspection of *t–u* plots; these may reveal an inconvenient truth in the form of a *complex* (as opposed to an *assumed simple*) **X**–**Y** data structure in regression modelling. Such potential information does not reach the data analyst if systematic inspection of *t–u* plots is not one of the first items on the model building agenda.

8.8 Multiple validation approaches

When using segmented cross-validation several times over with different seed sub-datasets, or when using a multitude of different validation approaches, there is often a tacit *assumption* that the majority of approaches will lead to practically the same optimal number of components also yielding very similar *RMSECV* results. When this happens, it is *claimed* that a successful validation has resulted, and that the model is "robust".§ Alas, from Figures 8.1 and 8.2, and the discussion above, it follows that this is a groundless claim and all that has been proven is that one particular data structure (the *local* data set) is characterised by a *strong* (**X**, **Y**) correlation. In this situation, nothing regarding the general future prediction performance nor the universal application of a variant of re-sampling approaches was in fact proven in any valid sense.

It must be noted, however, that particular scenarios *may* at times be so strictly bracketed that all future data sets will indeed behave more-or-less as typological clones—laboratory calibration of solutions abiding Beer's law could serve as a good example along with many others, however, generalisation to all systems is still **not** warranted. The occurrence of such cases *may*, or *may not* be met with over the entire career of any data analyst, however, it is always easy to find out the objective situation. Instead of being but a follower of *assumptions*—always inspect and interpret the relevant *t–u* plots and always perform test set validation where and whenever possible.

8.9 Verdict on training set splitting and many other myths

"Why is duplicated application of an identical sampling protocol in order to produce two distinct data sets, $\mathbf{X}_{train}$ and $\mathbf{X}_{test}$ *different* from splitting a twice-as-large $\mathbf{X}_{train}$ sampled in one operation?"

This is undoubtedly the most often heard remark in discussions on validation. Below it is argued forcefully why this is indeed so. What follows is a conceptual analysis—and a refutation of the many objections to test validation often raised as passive justification for continuing to apply cross-validation. The following comprehensive analysis has never before been given at this early stage of the education of new data analysts.

Taking care of the number of measurements (samples, objects), *N*, is the easiest obligation of the experimentalist/sampler/analyst/data analyst. But it is far more important to be in control of the variance-influencing factors when trying to secure a sufficiently representative ensemble of these *N* objects to serve as the all-important training data set. In the literature, and from chemometrics courses, there are usually few useful guidelines in this game—except the universal stipulation that the training set must *span* the range

§ Note that in reality the model is only "robust" with respect to alternative validation methods or strategies.

of **X** and **Y**-values in a "sufficient" fashion, which is a problem-dependent issue; often this is the only consideration given to the issue of training data set "representativity". This is a much too shallow understanding, however.

The critical issue is, again, heterogeneity and herewith the obligation to be in full command regarding identification and elimination of the sampling errors, lest a sampling bias may dominate the measurement uncertainty budget. On this basis, any *t*–*u* plot must be seen as a fair reflection of the sum-total of **all** influencing factors on the measurement uncertainty. Sampling using a well-reflected, problem-dependent protocol should ensure that all circumstantial conditions are influencing the sampling process in a comparable manner regarding both $\mathbf{X}_{train}$ and $\mathbf{X}_{new}$, i.e. they are given the same opportunity to play out their role irrespective of who is doing the sampling, the sample preparation and analysis. This is the role of an objective sampling-and-analysis protocol. It is important that the protocol has both systematic requirements securing an effective *span*, as well as a modicum of random selection requirements, deliberately trying to represent possible unknown circumstantial effects and their impact on the correlation of the **X**–**Y** data structure.

Circumstantial conditions are capricious; however, they are *time-varying* and in general defy systematisation. But any second sampling from a bulk lot will better reflect the situation at the time, or at the place in the *future scenario*, pertaining to the application of the prediction model. This may, or may not, be characterised by the same set of conditions as governing the training set generation. The key issue is that the sampler has no control over which, and to which degree, these conditions may have changed *between* the time of sampling the training data set, $\mathbf{X}_{train}$ and the "future data set". The demand of a "second sampling at a future time/place" is prescribed so as to deliver a best possible glimpse of the future application situation. All the data analyst can rely on in this quest is to let the second data set capture the data structure pertaining to the test set as objectively as possible. And yes, when all of these intricacies are fully understood and acknowledged—two independent test sets *would* be (even) better. But there are limits to what one can do!

By focusing on test set validation, to the degree conditions have indeed changed, there is now a trustworthy representation hereof involved in the validation, as illustrated in Figures 8.3 and 8.4.

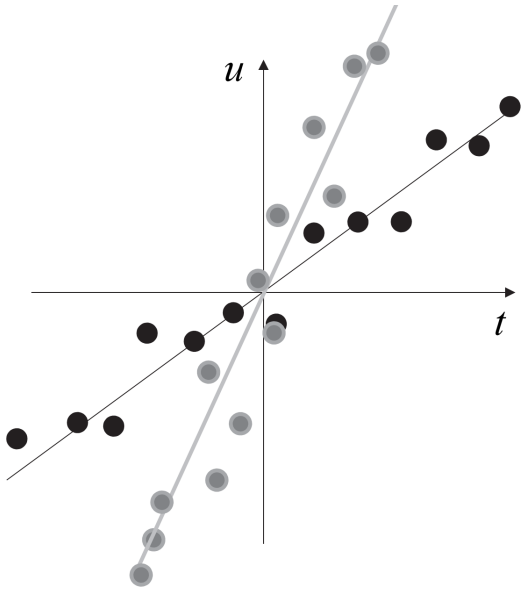

Figure 8.3: Synoptic display of $\mathbf{X}_{train}$ and $\mathbf{X}_{test}$ as a basis for evaluation of empirical data structure differences and their *t*–*u* expressions pertaining to two independent sampling events. The two data set models shown here display significantly different loading covariances (*t*,*u*) and the one data set (grey) displays a distinctly smaller variance in the **X**-space than the other. If the grey data set is the test data set, it is obvious that there is no similarity with the training data set (black), see also Figures 8.4 and 8.5.

In "some cases" it has been argued that the difference between the training and the test data set is not of sufficient magnitude to be of practical influence. The winning argument is that it is not possible to identify such cases *a priori*. Whatever the situation is at the time of decision of which validation to go for, the status of this issue is manifestly *unknown*, and in fact any of the scenarios shown in Figures 8.4 and 8.5 can potentially be on the agenda. To the degree that the two data sets depicted are bona fide training and test sets it is *vital* to include them both in the pertinent validation, which can only be performed using test set validation. The reasons such more and more marked disparities between training and test set can arise is, of course….

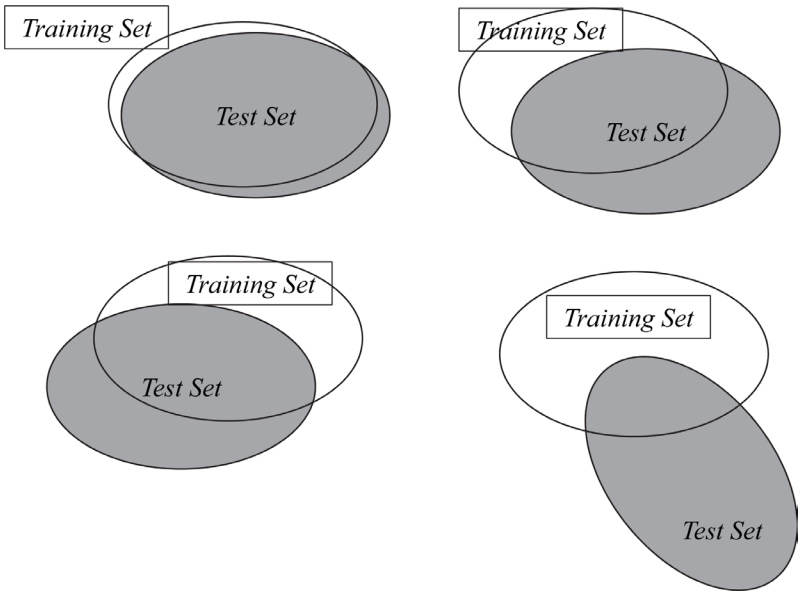

Figure 8.4: Illustration of a training set and test sets of progressively less-and-less overlap.

heterogeneity. How would one be able to ascertain if such is the case without going through the reasonable effort of (always) taking a second, independent test set?

This is a fundamentally unacceptable uncertainty, which is not resolvable within the one-data-set-only paradigm, again highlighting the dangers of an "auto-pilot" cross-validation approach depicted in the "one-click-model-development" options in less well-reflected offerings.

In fact, the only way this dilemma could ever be circumvented would be by carrying out both a test set validation as well as a particular cross-validation alternative. The cross-validation process in this case is used to assess the internal stability of the model (using stability plots) and the test set validation is used as an assessment of the future reliability of the model. If and when this approach is taken, the structurally inferior cross-validation would never be accepted as a performance indicator for a final model, one would **only** rely on the superior test set validation result. Still, for the reader's educational benefit, in this book there are plenty of data analysis assignments where it is required to carry out and compare *both* test set and the several principal variants of segmented cross-validation in an attempt to build up experience for both new and experienced chemometricians.

The *logical* conclusion to the everlasting "cross-validation"¶ dilemma is to declare a test set validation **imperative**. Nothing adverse will ever result from always applying test set validation, which delivers estimates of both A_{opt} as well as *RMSEP*, while everything is uncontrollably risky by basing a re-sampling cross-validation on the principally untestable *assumption* of representativity in the form of a timeless, constant data structure. Also, a random splitting of the objects in the first data set into a calibration and test set does not mean one is "home safe"—since the "second opinion" of the underlying data structure is missing. Calibration and validation is a systematic approach in which much understanding of the sampling background and the span of the **X**- and **Y**-space is required [5].

¶ In a recent meeting of the Australian Near Infrared Spectroscopy Group (ANISG), the respected chemometrician Professor Tom Fearn presented a paper entitled: "Crass-Validation"—a term that well aligns with the current argumentation in this textbook.

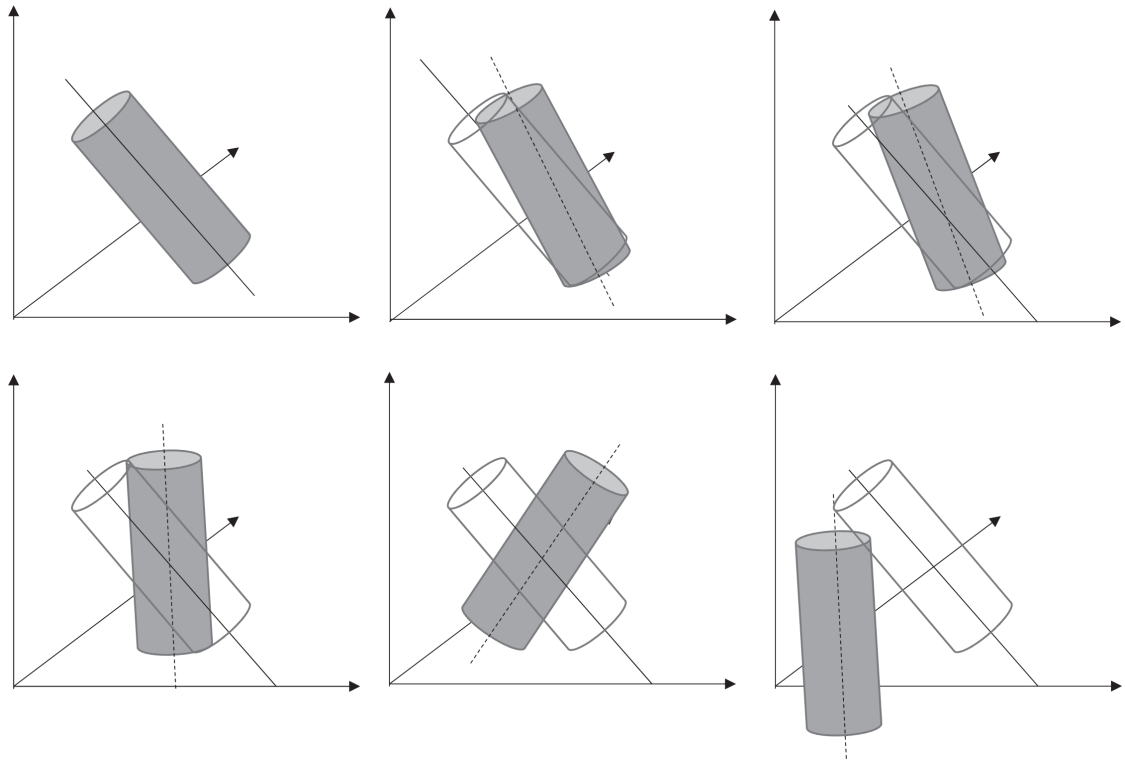

Figure 8.5: 3-D geometry renditions of progressively less-and-less similar covariance data structures. To the degree that these are bona fide training—and test sets—it is vital to allow them both to influence the outcome of validation, which can only be performed using test set validation.

As a pertinent example, the development of a spectroscopic analysis of pharmaceutical tablets is briefly presented below. When developing such a method, many batches of previously made tablets within their expiry dates are usually available for the sample pool. A protocol is defined that states how many random samples are to be taken from each lot, thus, a pool of manufactured lots is available covering the expected range of raw material and manufacturing conditions at the time of calibration development. This is the best estimate of the future conditions available and is supplemented with batches made at the time of method development.

Due to the tight specifications imparted on the production of pharmaceutical tablets, the variability of the **Y**-responses will typically be very low, thus the samples obtained should only represent the centre of the calibration line. In order to develop a linear model, the **Y**-response range must be expanded, typically through the manufacture of development samples that span 75–125% of the target **Y**-response, defined by the manufacturing set of samples.

To develop this extended range set, design of experiments (DoE) is employed to develop tablet formulations that vary the **Y**-response, but also change the other components in a designed way, so as not to lead to unrealistic "binary" mixtures of the constituent of interest and the rest of the tablet matrix. To test that the extended set and the manufacturing set of data are from similar populations, a set of replicated "manufacturing condition" tablets are also developed and the spectra are compared using methods such as PCA to assess if they are spectrally identical. If this is the case, the extended set can be combined with the manufacturing set to produce a representative sample pool for the development of a calibration model with

appropriately chosen validation set. To choose the validation set, the pooled objects are typically sorted in ascending order based on the **Y**-response. A systematic split of the data into a defined number of calibration and validation objects can now be made that best covers the **Y**-span. The selected objects are then tested for **X**-span and some row exchange is employed such that:

1) The calibration sample set spans the greatest variability of both **X**- and **Y**-space simultaneously.
2) The validation sample set completely lies within the calibration span, but covers the second greatest span. This defines the working range of the model to be developed.
3) Both the calibration and validation sets cover the widest variations in raw material composition, manufacturing dates and operating conditions as best as possible.

This validation set can be labelled the first *internal validation set*, as it is primarily used to assess the choice of preprocessing(s) applied and also to test the linearity of the calibration model. This internal validation set is the most representative set available at the time of model development.

Once internal validation has been performed and is complete, the model is assessed by its application to an *external validation set*. This set is typically made up of samples collected from **new** batches made **after** the development of the calibration model. It usually only assesses the centre of the model, as the production samples will only have a tight range of **Y**-response values, however, this is the overall objective of the model development, to assess the production of tablets for consistent manufacture at the target response value.

It is hopefully clear now that calibration development is **not** a simple random process in which samples to be selected for the validation sets are randomly selected from a manufacturing pool. This situation is highly applicable to the realms of industries that can design samples for making the calibration range larger, but what happens in the situation where this luxury does not exist, i.e. natural/agricultural systems?

In this case, the calibration development analyst is at the mercy of what nature delivers. This does not necessarily have to be a bad situation, however, most models developed on natural systems usually take many seasons of sample collection to become robust. There are ways of "smart" calibration development that can be employed for this situation. The process of natural system calibration development is actually quite similar to the pharmaceutical development described above, once a sample pool has been established.

The steps for successful natural system calibration are summarised as follows:

1) Collect representative composite samples from the expected strata (for example, geographical regions) that the calibration model is to be developed for.
2) Analyse all the collected samples using the **X** method (spectroscopic or other) and perform a PCA on the data to look for trends or groupings.
3) Perform a limited number of reference analyses (**Y**) on extreme objects found in the PCA (after removal of gross outliers, *should* such exist).
4) Develop an initial model. At this stage, due to the small number of calibration objects available, one might use cross-validation to establish a first "indicator" model complexity (only). If reasonable linearity can be established in a small number of components/factors, use this model on all new samples obtained to look for "holes" in the calibration line.
5) If a linear model cannot be obtained, then use the PCA model on all new samples to isolate objects that are different from what has been collected to date and submit those for reference analysis until a pool of samples is available to build extended calibration and test sets for "robust" model development.

It is stated categorically here that such protocols are the only way of developing reliable models. One of the current authors has used this protocol many times in industry and has developed models that are still in use today after being developed 10–15 years ago. Test set selection is systematic and requires great planning by the diligent analyst to develop robust calibrations. Cross-validation in these method developments is only used for the following reasons:

1) To establish *initial* models when not enough samples are available for test set validation. These models are **only** used to aid in the finding of more samples that can be used to build the sample pool.
2) To test the internal consistency of the calibration model. In particular:
   a) Random cross-validation is used to assess the stability of the model when random segments are taken out.

b) Systematic cross-validation is used to assess the quality of sample replicates.
c) Categorical cross-validation is used to assess the model stability under predefined conditions like growing seasons, manufacturing shifts, raw materials and other non-controllable factors.

In the "machine learning" community for example, random splitting has become a firm tradition, indeed repeated splits into calibration and test-sets and tuning model parameters to find the "best" model is often used here. It should be clear that this is a dangerous tradition based on the universal *belief* that any-and-all first data set is always representative of the future prediction situation. Such a fixed belief is totally unjustified by reality, however, as outlined in the calibration model development protocol described above.

It is often claimed that cross-validation is to be used to determine the "correct" number of factors, i.e. cross-validation is often accepted for *internal* validation purposes—but (interestingly) that for the most reliable estimation of *RMSEP* a "completely independent" set is also pointed to by the very same cross-validation proponents. The present treatment has no quarrels with the latter stipulation of course—but is in total disagreement regarding any use of cross-validation for determination of A_{opt}. Such internal use of cross-validation is the worst application imaginable, as it can only bring forth information as to the singular training set. It is never in anybody's interest to invoke a two-step, dual method validation approach. By using test set validation, one is presented the most reliable estimate of *RMSEP* (never structurally underestimated) based on the objectively correct number of components, all in one go.

This is the principal reason for not routinely splitting a training data set randomly, however large it may be. With random splitting, there is still no information pertaining to the future application situation, unless complete consideration to the design of the set can be given the attention it requires. A massive redundancy in the number of data available is mistaken for a realistic basis for future performance validation. Test set validation is the best possible way to remedy this predicament—by securing (at least) one new data set from as far in the "future" as is logistically possible, i.e. the *external validation set*.

By accepting that circumstantial conditions may well change (on occasion), but that information about this will usually be *unknown*, the test set validation approach is the best one can ever do. This also brings up the question to what extent an empirical model can be *extrapolated*, i.e. to a situation in which test-set samples as well as other future samples may be found to lie outside the full calibration space (if this was, somehow, under-represented). It is in order to guard against such undesirable situations that the demands for a *proper* training data set are so stringent.

From this discussion, it also transpires that a regimen of regular test set validation model checking is a wise approach within the arena of quality monitoring and quality control. The above discussion appears particularly easy to understand in the process technology, process monitoring and process control settings. Proper process sampling in this context is treated specifically by Esbensen and Paasch-Mortensen [9] and will be taken up in the chapter on PAT (chapter 13). In particular, the US FDA in its 2011 process validation guidance [10] has stated that every batch is now a validation batch in the realms of quality by design (QbD) and this requires continuous verification strategies. In other words, the use of process analytical technology and modern control/data management systems now allows manufacturers to test set validate **every** batch produced. This is where responsible chemometrics meets proper consumer protection.

8.10 Cross-validation does have a role—category and model comparisons

There is a role for cross-validation, however, several in fact—but they are all strictly compartmentalised and cannot be subjected to generalisation.

In the arena of model comparison (both regarding models of structurally identical nature, but of optionally alternative parameter settings: for example, different pre-processing alternatives, different **X**-variable selection alternatives… as well as more distinctly different models), cross-validation is in fact a particularly relevant approach. For this specific purpose, cross-validation furnishes precisely what is needed, a general identical performance framework within

which the effects from alternative models, parameter settings, preprocessing, categories (seasons, for example) can be objectively compared without having to deal with data set structure variations for each "segment". In this context, it is a necessity to use the same number of segments for all sub-validations, in which case it is strongly recommended to use a low number of segments, preferentially two, in order to impart the greatest possible semblance of realistic data set variability to influence the validation results—and never LOO cross-validation (full cross-validation). In this area of applied validation, there is very good reason to use cross-validation, although it is interesting to contemplate how one is to deal with the possibility of a different number of components A_{opt} for alternatively optimised models?

In the exemple of a prediction model encompassing distinctly different seasons, it is intuitively clear that in order for such a model to have truly predictive power, the only relevant category with which to segment the training data set, will use seasons as the cross-validation segments. A "robust" model, i.e. a prediction model able to predict all year round, should be stable with respect to seasonal perturbation. A standard "blind" cross-validation segmentation will invariably have representative objects from all seasons in the different segments, with the result that the model is not tested at all with respect to the individual seasons.

There exist several other categories that function *similarly* as seasons for many data sets, for which the exact same argument holds, as is laid out in full in Westad and Marini [2] in which it is shown how correctly applied cross-validation gives valuable information about these specific types of sources of variation. Thus, validating across relevant categorical object designations enables the analyst, for example, to evaluate the robustness across raw material suppliers, location, time, operators etc. These are meaningful segment definitions that qualify use of cross-validation.

Returning to the pharmaceutical tablet example given above, assume that the objective of a project is to develop a model for predicting the active ingredient in every tablet produced at multiple production locations. For this purpose, a spectrometer is used on-line to provide the necessary information in real time and it may be assumed that there exists an established reference analytical method. The experimenter then needs an estimate of the sources of variation for such a system to be implemented at the different production sites. Among the many considerations one needs to take, the final set of objects may be *stratified* into segments according to, for example:

a) replicated measurements on one side of one tablet
b) acquiring a spectrum on both sides of the tablet
c) changes over time for one production batch
d) changes between various batches of raw materials
e) changes due to equipment characteristics in the production line at one site
f) variation across production sites
g) variation due to the standard sampling and analytical procedures (covered in chapter 9)

By carefully setting up schemes for cross-validation according to this type of qualitative information about object groups, termed "conceptual cross-validation", the influence on the prediction results from these various sources of variation can be estimated and compared. If, for example a) above is the main source of variation, there is a fundamental problem with the measurement process. On the other hand, if cross-validation across instruments reveals large differences in the hardware components, the conclusion is that *individual* models for each instrument are needed, or some relevant method for instrument standardisation or model transfer is required.

Sometimes it is argued that since all the objects were acquired on a specific day with a specific instrument, on a specific batch of raw materials and by one person only, the above intricacies do not apply, and one can simply just get on with simple "blind" cross-validation. In such a case, however, estimates of the model performance will severely lack credibility because no *prospective* conclusions can be made, as per the many arguments above before section 8.10. Unfortunately, this situation is the typical basis for the validation presented in quite a number of technical reports, scientific publications and in oral presentations, which unavoidably must lead to overoptimistic validation assessments, findings that later cannot be reproduced. The history of chemometrics has very many such examples which has contributed to a certain measure of institutional confusion.

8.11 Cross-validation vs test set validation in practice

Usually there is more focus on strict adherence to one or another cross-validation procedure, complete with preferred number of segments (a fixed *scheme*), than openness with respect to what exactly are the assumptions and prerequisites behind cross-validation. This is troublesome, as no amount of discussion *pro et con* a specific number of segments will ever reveal the underlying structural problems associated with cross-validation using whatever number of segments "*s*". The general verdict, following from all of the above, is:

Cross-validation is, in general, not a validation which incorporates information as to the future use of the particular data model. Cross-validation is overwhelmingly an internal sub-setting stability assessment procedure; cross-validation here only speaks about the robustness of a particular (local) data model, as gauged by internal ***sub-setting*** *of the particular training set.*

Caveat: The latter feature can be turned to good use in specific cases, specifically, the case of "conceptual cross-validation", which often occurs in the process realm, as well as when comparing models. Cross-validation finds valid use for estimating the magnitude of resulting variabilities, provided all future samples lie inside the same conceptual modelling domain.**

The operative aspects of cross-validation versus test set validation are illustrated forcefully by the multivariate image analysis (MIA) examples in Esbensen and Lied [11]. Even though this publication is addressing MIA, the image analytical examples here throw unprecedented illumination on the general principles of cross-validation because of the extraordinary magnitude of the **X** and **Y** matrices involved. Because each pixel counts as an object, even a modest image illustrates truly huge data sets, i.e. 10,000 to 1,000,000 objects or more. Here the workings of cross-validation are visualised like nowhere else in chemometrics.

8.12 Visualisation of validation is everything

With the help of the relationships presented in Figure 8.2 above, further developed below as Figure 8.6, it is possible to delineate the universal deficiency displayed by segmented cross-validation, indeed also compared with leverage-corrected validation: Test set validation will always result in the *highest* estimate for *RMSEP* than any of the segmented cross-validation alternatives (and often very much higher than the leverage-corrected *RMSE* estimate) precisely because it incorporates all sampling, conceptual categories, model and analyses uncertainties. This will, therefore, always constitute the *most realistic* estimate. The point is not to search for the *lowest RMSE* outcome between a voluntary set of alternative validation methods/variants—the point is to estimate the *most realistic* future prediction error, and this is universally delivered by the test set estimate.

Figure 8.6 summarises experience with validation of many hundreds of projects and data sets. In the last two to three decades of chemometric experiences (teaching, professional, consulting) behind the present book, innumerable data analyses have dealt with all manner of types of data structure the general patterns of which are depicted in Figure 8.1, especially those of more regular appearance, types a–d). Occasionally partly deviating curves to the ones depicted may appear, but invariably related to a local, more irregular data structure. The "gap" between the test set validation and two-segment cross-validation curves represents the *missing* TSE-component, which can only be quantified by *comparing* test set and the cross-validation results. This represents the missing TOS-error components that can only be incorporated by sampling a second data set, the essential feature of which is that the sampling protocol is *identical* for both the training and the test data sets. From these universal relationships emerges one very

** However, there is a paradox when one set of objects is collected on one day, with one raw material etc., there is still no information about the samples as basis for systematic validation in the sense of "conceptual cross-validation". In this case only random cross-validation or setting aside one part of the objects as a test-set are real alternatives. A random split into a calibration and a test set will not reveal if the model is stable towards *future* sources of variation. Again, only a well thought out training data set will allow validation to address all the relevant issues; the specific cross-validation method alone will be insufficient.

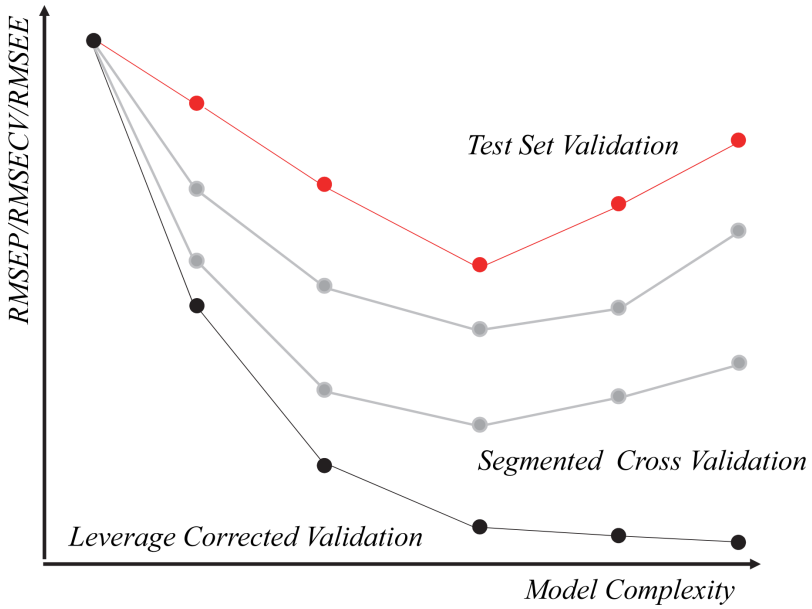

Figure 8.6: Relationships between the three principal *RMSE*-estimate procedures as a function of model complexity. Leverage-corrected estimates are universally lower than those pertaining to cross-validation, which are always structurally lower than those stemming from test set validation proper. For one-and-the-same training data set [**X**, **Y**], the systematic relationships between the different segment variants of cross-validation are indicated in principle; these were laid out in detail in Figure.8.2. Stronger (**X**, **Y**) correlation will result in more close-lying curves, but the principal relationships shown remain the same.

powerful conclusion: only test set validation can stand up to the logical and scientific demands of all the characteristics of *proper validation*. One should henceforth observe a test set validation mandate if and whenever possible based on the arguments provided in this chapter. Cross-validation used as a model validation technique when it is possible to perform a test set validation is **unacceptable** in every way and is a main cause of why chemometrics has been given a bad name in some situations. In one case, one of the authors witnessed a situation where a so-called "data analyst" performed cross-validation on a full set of 6000 samples. The mind boggles at such incompetence!

There has been a persistent chemometric tradition of validating all data sets, large or small—regular or chaotic w.r.t. data structure, which are often unknown *unless* visualised by *t–u* plots. This is especially dangerous when dealing with small data sets, see e.g. Martens and Dardenne [12]. In such situations where the number of objects is limited, setting aside a certain proportion of the objects as a test-set comes with the very likely cost of removing significant parts of the overall latent structure of the data.

Thus, there exist situations (so-called "small $sample_{STAT}$ cases") in which absolutely **all** objects are needed for optimal modelling and interpretation of the data structure, such as the relationships between the variables etc. In such cases, it is definitely better to allow for the possibility of **not** validating data sets when the conditions for proper validation are lacking. "To validate, or not to validate" is thus an evergreen valid question for the consummate chemometrician.

8.13 Final remark on several test sets

Arguments can easily be raised for invoking a postulated need for several test sets: of course, more than one test set will always allow for more valid assessment, since more test set realisations correspond to

more examples of the future in-work prediction scenario for the prediction model. One properly materialised test set will have a decent chance of incorporating the *principal* information from the future situation. This in contrast to the vociferous objections and postulated budget or effort constraints that are often claimed, not even allowing for a single test set. In a rational context, it is evident that a decision regarding the real need for several test sets will be based much more on problem-dependent specifics, always related to the complete problem-dependent background regarding the likely consequences, and the price to pay, for sloppy validation. Thus, it is here advocated **always** to plan for and materialise one test set, acknowledging the occasional need for more, but this decision is left well and safely in the hand of the informed data analysts who are closer to the relevant data and their background in all cases.

8.14 Conclusions

Re-sampling and cross-validation approaches work on one data set only, $\mathbf{X}_{train}$. The tradition of cross-validation is particularly strong in the realm of less experienced chemometricians. The current use of cross-validation and its huge popularity is based on tacit, unsubstantiated *assumptions* of the training set *always* being *fully* representative of the future scenario and future measurements on new samples. However, this belief finds itself in strong disregard of the extremely varying origin and the very diverse data structures in the real world. This widespread assumption was shown to be untenable in the light of the significant bias-generating sampling errors described in the Theory of Sampling (TOS).

On the other hand, in *some* cases, one cannot "wait forever" until all sources of variation for a given application are represented in the first training set of objects before the starting to establish a particular data model. Thus, it is of critical importance exactly what is represented in this singular data set and if it has been acquired within the framework of a "suitable sampling strategy" that forces it to reflect future variation.

Instead of the endless series of partial examples (based on *local* data set structures only) presented in the chemometrics and other literature, and from which no valid generalisation can be made, this chapter presented *first principles*, the principles of proper validation (PPV), which are universal and apply to **all** situations in which assessment of performance is desired—be this prediction, classification, time series forecasting or modelling validation. The underlying element in PPV is the Theory of Sampling (TOS), chapter 3, which is needed in order to identify and eliminate all bias-generating sampling errors, which are otherwise responsible for unnecessary, significantly inflated measurement errors, for which no statistical corrections are possible. Invoking the complete body of theoretical and practical experiences from ~60 years of application of TOS, it was shown to be untenable to continue with bland, unjustified *assumptions* regarding universal training data set representativity.

On the basis of chapter 3 and the present chapter, it was concluded that re-sampling and cross-validation approaches miss out with respect to the crucial $sampling_{TOS}$ variance. This variance can only be accommodated by a test set (a second independent sampling—more than one if deemed necessary by local, problem-dependent reasons), without which simple re-sampling validation on one-and-the-same data set will always structurally underestimate the *realistic* prediction error. No theoretical procedure exists to derive an approach that can estimate the magnitude of this missing part. For this reason, re-sampling and cross-validation should logically be terminated, except for the cases of exception described in section 8.10. Standard use of "blind cross-validation" only performs assessment of *internal sub-setting model stability*. Use of cross-validation must always be accompanied by full disclosure of the procedures used and the inherent method deficiencies described in this chapter.

The main purpose of establishing a model may not necessarily be for predicting or for classifying new objects, but simply to *understand* the inherent structure in the system under observation. The previous chapters describe methods that provide insights into the underlying structure of any process or system under observation, through *scores* and *loadings* relationships a.o. All model interpretation is highly dependent on the number of latent variables retained, and therefore it is vital to be able to determine the correct dimensionality (rank) of the model. It is important to distinguish between numerical rank, statistical rank and

the application-specific rank, which may not always be identical.

Regarding PLSR, a major chemometric regression method, a call was made for stringent commitment to test set validation based on graphical inspection of *t–u* plots for optimal understanding of the operative **X–Y** interrelationships. Simple visual inspection will also allow a reliable premonition of the outcome of any particular validation approach. There is no justification to reject the work effort involved in securing a test set for validation purposes, acknowledging that this is the only approach which eliminates the deficiencies outlined. The comparatively rare occasions when a test set is not an available option (historical data a.o.), have no generalisation power. The comprehensive understanding outlined in this chapter will stand the data analyst in good stead when, feeling forced to make use of some form of re-sampling. Complete understanding and full disclosure of the structural *RMSE* underestimation deficiency is mandatory in all such cases.

Many reasons are given in numerous traditional arguments for continued use of cross-validation and re-sampling for validation. The following arguments and reasons are **not** valid:

- Complacency: one cross-validation approach/method for all data sets is an easy buy, but one that completely disregards the gamut of vastly different data structures and correlations.
- Focus is on algorithms, implementation and software, without critical thinking.
- Unwillingness to investigate consequences of traditional statistical assumptions (*myths*).
- Resistance to the Theory of Sampling (TOS) for complementary understanding regarding heterogeneity and sampling process issues.
- Misunderstanding, or misplaced universal trust in the central limit theorem.
- No interest for how "data" and "data quality" originate.
- Blind adherence to traditions or schools-of-thought: "This is the way chemometrics has been doing validation for more than 40 years…"

This chapter mostly discussed how a system can be validated using the best available information about the origin of the data (objects), Esbensen and Geladi [1]. However, validation *may* have various meanings in different scientific communities. Questions like "do I use the expected chemical information in my instrumental variables to predict product quality", "do various methods give the same interpretation" or "do I find the same subset of variables with various variable selection approaches?" are examples where cross-validation in specific, bracketed situations may be useful in broader and more advanced contexts, see Westad and Marini [2] for more on these issues.

8.15 References

[1] Esbensen, K.H. and Geladi, P. (2010). "Principles of proper validation: use and abuse of re-sampling for validation", *J. Chemometr.* **24,** 168–187. https://doi.org/10.1002/cem.1310

[2] Westad, F. and Marini, F. (2015). "Validation of chemometric models—a tutorial", *Anal. Chem. Acta*. **893,** 14–24. https://doi.org/10.1016/j.aca.2015.06.056

[3] Martens, M. and Martens, H. (2001). *Multivariate Analysis of Quality. An Introduction*. Wiley, Chichester, p. 445.

[4] Høskuldsson, A. (1996). *Prediction Methods in the Sciences*. Thor Publishing. Copenhagen.

[5] Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (2007). *Numerical Recipes, The Art of Scientific Computing*, 3rd Edn. Cambridge Press, NY, p. 777.

[6] Wonacott, T., and Wonacott, R. (1990). *Introductory Statistics*, 5th Edn. Wiley, New York.

[7] Devore, J. (1995). *Probability and Statistics for Engineering and the Sciences*, 4th Edn. Duxbury Press, Belmont, CA.

[8] Martens, H. and Naes, T. (1989). *Multivariate Calibration*. Wiley, Chichester.

[9] Esbensen, K.H. and Paasch-Mortensen, P. (2010). "Process sampling (Theory of Sampling)—the missing link in process analytical technologies (PAT)", *Process Analytical Technologies*, 2nd Edn, Ed by Bakeev, K. Wiley-Blackwell, Oxford. https://doi.org/10.1002/9780470689592.ch3

[10] US FDA, *Guidance for Industry—Process Validation: General Principles and Practices*. http://www.fda.gov/downloads/Drugs/Guidances/UCM070336.pdf [Accessed 6 January 2017].

[11] Esbensen, K.H. and Lied, T.T. (2007). "Principles of image cross-validation (ICV): representative segmentation of image data structures", in *Techniques and Applications of Hyperspectral Image Analysis*, Ed by Grahn, H.F. and Geladi, P. Wiley, Chichester, Ch. 7, pp. 155–180. https://doi.org/10.1002/9780470010884.ch7

[12] Martens, H. and Dardenne, P. (1998). "Validation and verification of regression in small data sets", *Chemometr. Intell. Lab. Syst.* **44,** 99–121. https://doi.org/10.1016/S0169-7439(98)00167-1