Paper IV

Principles of MIR,
Multivariate Image Regression - II:

Cross validation - what you see *is what you* get

Thorbjørn Tønnesen Lied & Kim H. Esbensen

thorbjorn.t.lied@hit.no - kim.esbensen@hit.no
+47 35 57 51 53 - +47 35 57 51 50
fax: +47 35 57 52 50

Telemark University College
Dept. of technology
Kjølnes ring 56
N-3918 Porsgrunn
Norway

CONTENTS

ABSTRACT

This paper deals with generic problems regarding segmentation for cross validation in multivariate image regression. Multivariate images are characterized by a very large numbers of pixels which usually are highly redundant. When several thousand (ten thousand) pixels or more represent the same object, special considerations are required for proper cross validation segmentation.

A new approach for *guided segmentation* is introduced, in which the validation segments are specifically delineated by the informed user in score space. The practise of "blind", automated segmentation, which is dominating 2-way cross validation, is found to be useless in the 3-way MIA regimen. Problems concerning which order of components to use for the segmentation delineation are illustrated and the necessary precautions needed to ameliorate this approach are discussed. A general solution to the problem, called *higher-order components guided random sampling* , is described in detail, which may even also shed new light on current chemometric cross-validation practises in the conventional 2-way realm.

This new cross-validation approach is illustrated with multivariate image data sets which are known from the pertinent literature for easy comparison.

INTRODUCTION

This paper is the second in a series regarding Multivariate Image Regression, MIR, which has been developed to create regression models between multivariate images [1]. For a general introduction to this field, please see part 1 [2], in which the complete phenomenology of the three principal cases of multivariate image regression was detailed.

A multivariate image is a 3-D OOV matrix [3], i.e. two ways are objects (pixels in rows and columns), while the variable-way is comprised by different channels, e.g. colours. There are quite distinct differences between this 3-way domain and the complementary OVV domain, well-known from the three-way decomposition. These two domains do not in general make use of the same data modelling methods [2]. Here we treat OOV (MIA, MIR) exclusively.

In any multivariate model that will be used for prediction, it is important to know the predicting powers of the model. This is usually done by estimating the prediction errors as a measure between known and predicted values. A popular prediction measure is RMSEP (Root Mean Square Error of Prediction) which is defined as

Equation 1

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_{y,ref})^2}{n}}$$

where $\hat{y}_i$ refers to the predicted value, and $y_{i,ref}$ is the known value [4].

The procedure of testing prediction performance is known as *validation*. To perform this optimally, at least two sets of data are required, one for calibration and one for validation. When a model has been established, using the calibration set, the validation set is subsequently used for predicting the ŷ-values of the validation set for comparison, e.g. according to equation 1.

At least two variations for this type of validation exist, one is known as "*test set validation"*, the other as *"cross validation"*. In test set validation, a completely new, independently sampled dataset is acquired, in addition to the calibration set. This demands that an identical sampling procedure is used for both data sets.

If this is not feasible, a different, less optimal, approach will have to be resorted to. *Cross validation* extracts a pseudo-validation set from the calibration set before building the model on the remaining complement of data. The extracted data is now used for validation. This approach may take several different forms, but all are closely related, in that they per force must correspond to a number of so-called segments in the list: 2,3,4,5....N, where N stands for the total number of objects in the original calibration set. After prediction errors have been estimated for the one left-out segment, it is replaced back into the modelling base and a new model is created in which a different segment is being kept out of the modelling etc. This is continued until every segment, and object, has been used for validated, hence the term *cross* validation [5].

To get realistic validation estimates, it is important that the calibration and validation datasets represent two independent samplings from the target (parent) population. The degree of difference between them should reflect the variations that can be expected associated with the future measurement situation in which the regression model is to be

used for prediction purposes [4]. It is easy to see that test set validation is the only approach which honours all these requirements, ibid.

In 2-way chemometrics there are steadfast different opinions regarding how exactly to divide the data in calibration and cross validation sets or segments [6]. From so-called full cross validation (leave one out) on the one hand, to two-segment, so-called "test set switch" on the other; the latter represent a singularly unsatisfactory choice of terminology, as there is no "test set" present at all. It is always possible to use any intermediate number of segments from the list: 2,3,4 ....N. The relationships between test set validation and these systematics of cross validation remain an area of some confusion in conventional 2-way multivariate calibration [4]. In multivariate image analysis, however, distinct and special considerations are required to which this paper is dedicated.

There are two major characteristics in image data that are rarely found in 2-way data. Most striking is the number of “objects”. In a conventional video image (~500x700 pixels), there are more than 350.000 “objects”, i.e. pixels, in the range [0..255]. Removing any single object from this amount of data is not going to change the model adequately to perform any useful validation [4]. Also, calculating 350.000 sub-models, full cross validation, is not very tempting.

Secondly, and much more important to consider, is the large *redundancy* that exists in image data. Pixels lying close together in the image space are likely to represent the same object, and therefore often have closely similar values. Two-block data sets, for example in which every second pixel, say, is to be used for validation, would necessarily produce two almost identical images, clearly leading to inferior validation, *ibid.* This would correspond to some spatial (image space) segmentation scheme. With knowledge of object selection traditions in 2-way data analysis, the reader might well alternatively ask: “Why not simply use random sampling then?” This would correspond to a notion of a fair "blind", automated segmentation strategy. Again, consider the very large amount of data (pixels) present. Sampling 50% randomly out of 350.000 objects would most likely again simply produce two practically identical datasets. The last refuge from frustration of trying to generalise from the well-known 2-way regimen into the 3-way MIA/MIR realm will probably be to throw ones hands in the air: "Then use a larger number of segments, 10 or so!" - We shall show below that all such "blind" segmentation strategies are doomed to failure in the multivariate image regimen, irrespective of the actual number of segments chosen - if not specifically related to the covariance structure in the multivariate image.

In fact, multivariate image analysis requires a complete reconsideration of relevant strategies for selecting relevant data sets for calibration and validation. A new a strategy called "*guided random sampling*" is suggested below. In guided random sampling the *user* decides how the data is to be divided into the pertinent sets. This is neither done randomly, nor by a pre-specified "blind" number of segments, but with very specific respect for the *empirical data covariance structure* present (in the score feature space). A different angle from which to attack the data segmentation problem is required. Following the MIA experience this angle is to be found in the score-space.

Nomenclature

The following notation is used:

X	Matrix of predictor variables
Y	Matrix of dependent variables
y	Y-vector
T	Matrix of X-scores
U	Matrix of Y-scores

CASE STUDIES

For illustration purposes, several examples mostly based on already published multivariate image data sets will be used [7, 8]. The master dataset consists of a 512x512x8 image, the Montmorency Forest experimental data set [7, 8], where the channel with lowest wavelength is here chosen as the **Y**-image in the present context. This is not to be understood so that we suggest to predict this channel from the remaining seven others (although this actually might be an excellent solution for recovering a "corrupted" channel, which is often enough met with in remote sensing) - rather we make good Y-use of this particular channel in order to illustrate the special image regression case of Y-total, compare [2].

In figure 2 the pertinent T1-U1 score-plot from this application is shown. The cross validation challenge is here to divide this plot in, say, two sets (segments) that *both* are equally representative of the actually covariance structure present. A simple two-split in this plot may easily give rise to a significant difference between the subsets if the data structure does not comply well with a simple joint multivariate normal distribution

assumption. In multivariate image analysis we have yet to find such simple relationships! Some objects in one set will not be equally represented (if at all) in the other, and validation may easily tend to become unbalanced. A(ny) two-split - alone - would almost always be in danger of being unbalanced.

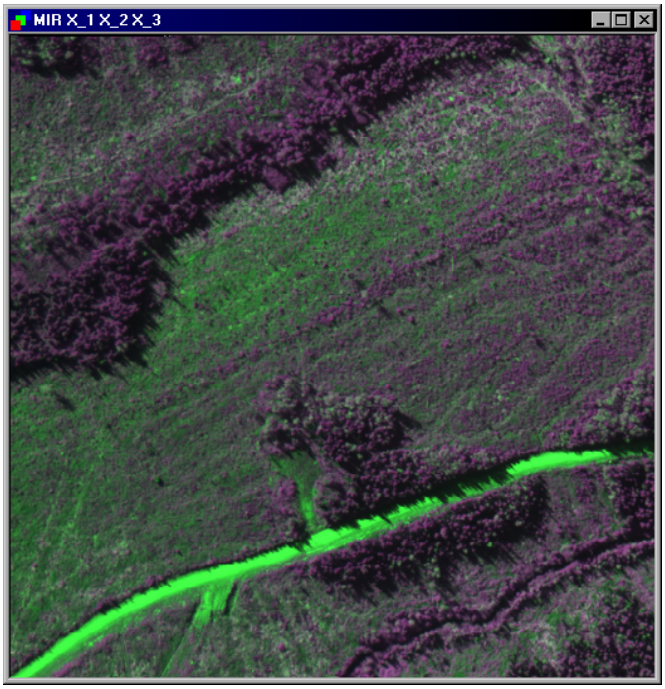


Figure 1. The scene space master image data which will be used for illustration, the Montmorency Forest data set [7,8]. The image consists of 512x512 pixels in eight bands (Channels, variables). Here it is represented by channels 1-2-3 as R-G-B.

To solve this problem, we suggest that the data set - generically - is divided in eight segments, sampling both *along* as well as *across* the dominant covariance data structures in the following way.

Initially the data is split in two halves along the main covariance direction. In figure 2 this would be a line passing through the two modes of the highest concentrations of pixels with similar score signatures, i.e. topographic “peaks”, compare [3,8] (figure 3), which are coloured red and orange in fig. 3. Each of these parts should now contain approximately 50% of the objects, and all main classes should be represented - at least the classes which go along to make up the dominating elongated covariance trend. Secondly, intersecting the first line, a new line should be drawn representing the second most important covariance direction, again as judged from the pertinent MIA

score cross-plot conventions, ibid. It is important that this second direction really corresponds to what the user perceives as the second most representative part of the overall covariance structure (more examples to be given below); thus there are no requirements for orthogonality of these two salient user-delineated covariance directions etc. This gives four segments, which each ideally should contain about 25% of the objects - barring whatever "surprises" may be in waiting in the higher-order components not captured in this first delineation. This, generally oblige, axis cross delineation is all the user has to supply in order for our new cross validation procedure to take over.

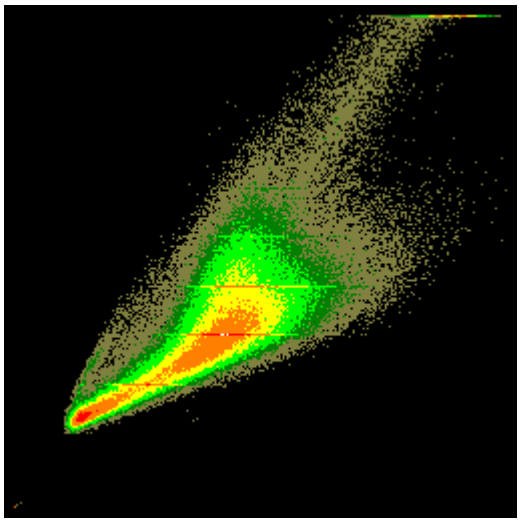


Figure 2. T1-U1 scoreplot from the MIR analysis of the image in figure 1.

After the user has drawn this second line, the software draws the four lines between the endpoints of these backbone intersecting lines. The software locates the intersection point, and finally calculates the midpoints between the corners of the outer frame. Lines are then drawn between the midpoints and the intersection point. An example of a resulting eight-segment mask is shown in figure 3. This configuration illustrates a generic eight-segment mask which it is the user's task to implement on top of a specific T-U, or T-T score plot.

With this type of mask, there are three functional combinations of subsets consisting of eight, four or two validation segments respectively. When selecting and combining sets, they should be opposite with regard to the centre point. Figure 4 shows the two compounded sets used in two-segment cross validation. In general each of these non-

overlapping two-fold division of the image covariance structure takes the form of a Maltese Cross, illustrated vividly in fig. 4.

Notice in Figs. 3 & 4 how some obvious outlying features have been excluded already in this first stage of cross validation segmentation (top right portions).

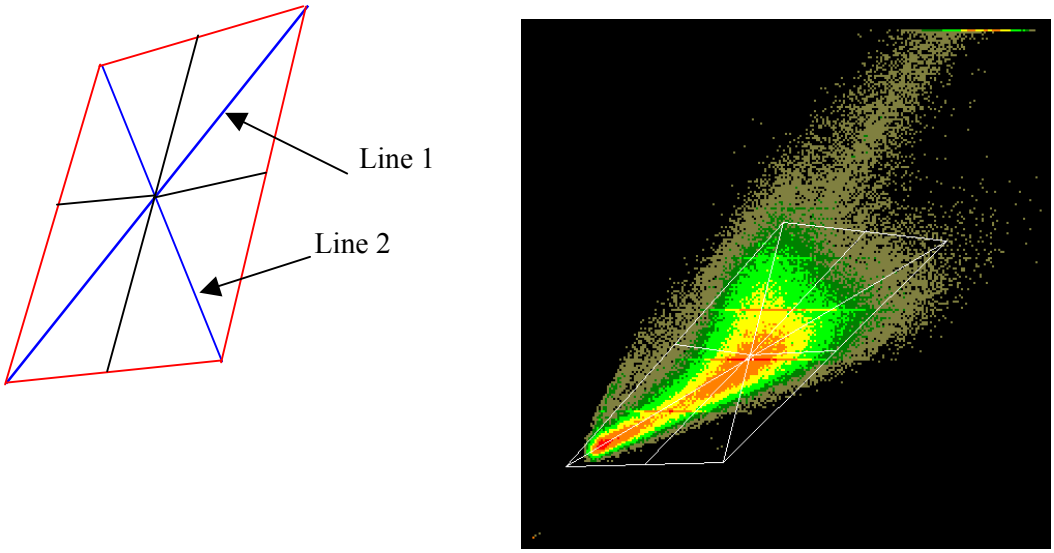


Figure 3. Left: Sketch of cross validation segment splitter initiated by two master lines drawn by the user. Right: Example of eight cross validation segments defined in a score-plot, T1-U1. Note that outlying pixels can be excluded already when delineating this mask.

Case 1: Full Y-image

The first case is a study of what was found in [2] to be a comparatively rare situation in image analysis; the full Y-image. In this situation, each object in X, each pixel, also has a corresponding representation in Y. This furnishes a particularly illustrating example of the new image analytical cross validation approach to be outlined. A more usual situation is studied in case 3.

While figure 4 shows the two validation data sets in the scoreplot, figure 5 displays the same data in image space. Pixels marked with white colour is used in the set.

Some outlying parts of the data was left out of the validation set entirely, because these pixels were identified as *outliers* already when delineating the problem-specified Maltese Cross region of interest. Alternatively this built-in outlier remover can be refined by making a local model [3,8] prior to the cross validation, allowing only the specific, problem-dependent objects of interest to be represented in the scoreplot.

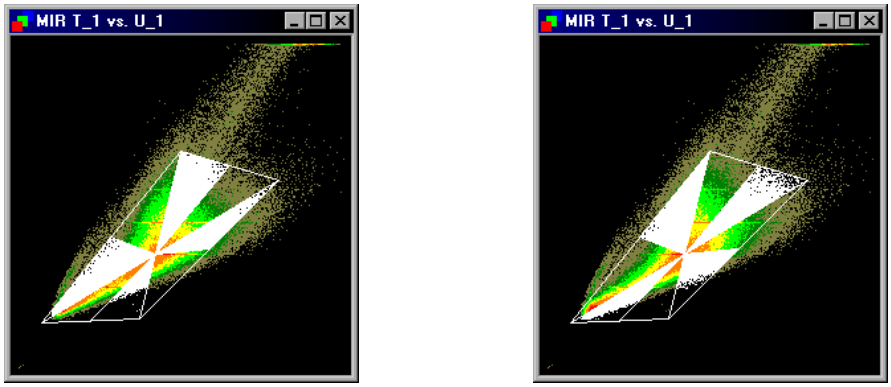


Figure 4. The two complementary "Maltese cross" validation data sets selected in the T1-U1 scoreplot shown in figure 2. Note how both achieve good data structure representation.

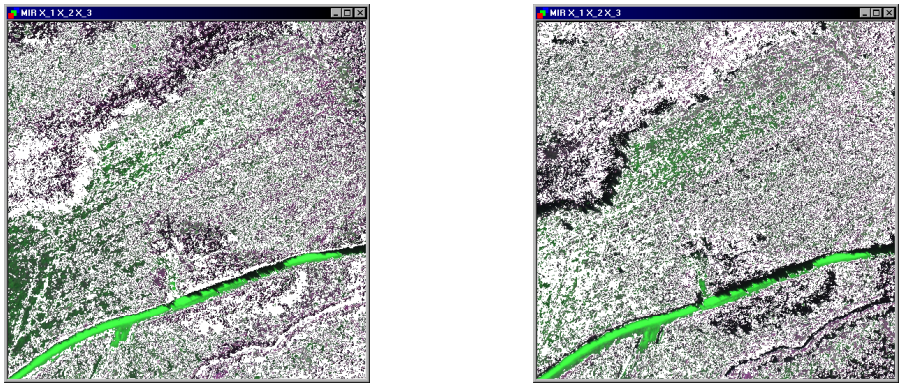


Figure 5. The two complementary validation segments selected figure 4 projected to image space. Note how both achieve satisfactory coverage and spatial representation.

Studying the images in figure 5, it should be fair to say that these two data sets represents approximately the same objects at the scale of the overall, full FOW image, with only a small difference at the most detailed levels. What you SEE in the score space rendition, fig. 4, is exactly what you GET, fig.5. The user has the full ability to iterate his or hers first tentative delineations of the Maltese Cross configuration by careful inspection of the RESULTING disposition of the two compound, non-overlapping scene space renditions, fig. 5, until a satisfactory results has been achieved.

Figures 6 and 7 shows what happens if eight segments were to used *independently as in a conventional eight-segment cross validation*. Obviously there are very great differences between these eight datasets, in fact there is an absolute certainty that these sub-models will be totally incommensurable with each other. This is a dramatic illustration also of the general cross-validation "problem" when the relationships between the X and the Y-space is more complex. In the present image analysis example, it is evident what goes wrong, were one to use an eight-segment (12.5%) cross validation scheme.

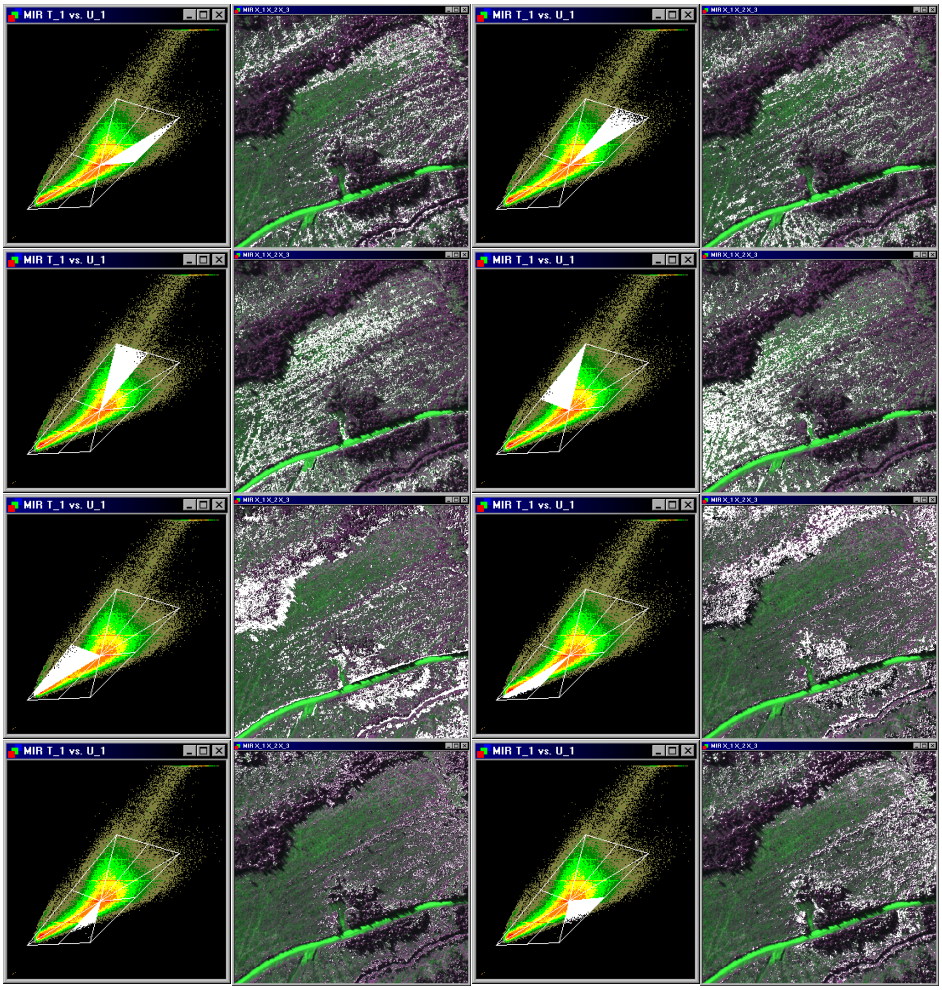


Figure 6. Eight individual validation segments in the T1-U1 scoreplot and the corresponding image space. Note how none of these achieve neither data structure nor spatial representativety.

Case 2: Problems

It is possible to run into problems with this approach if great care is not taken in the CV segmentation step however. If the segments are too small, they will very probably not be representative for the entire dataset. Another possibility, as will be shown in this case, is failure drawing the optimal guiding lines. Figure 7 and 8 shows what happens when the guiding lines split the data in a *off-centred* fashion. Clearly these two Maltese Cross configurations are NOT making up a good, balanced 50/50 cross validation bases. As can be seen even a small off-centred two-split has a dramatic effect on the two relative datasets because of the very high number of similar pixels making up the covariance backbone of the data structure. One dataset is provably very different than the other with very obvious poor, non-representative validation results to be expected. The current approach is thus very sensitive to the precision of - and the understanding behind - the user-interaction.

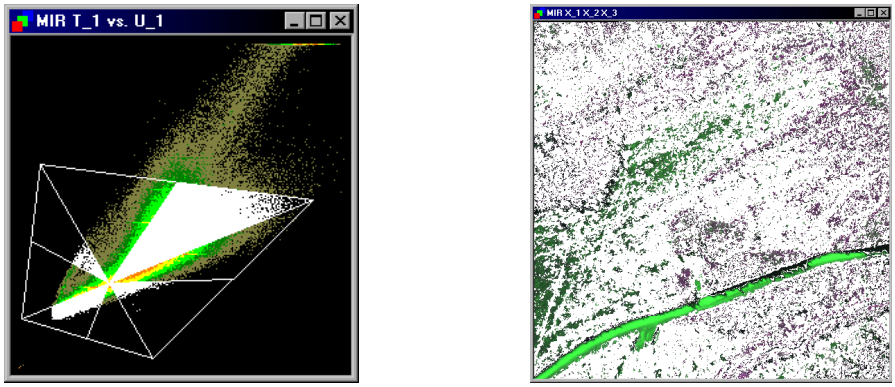
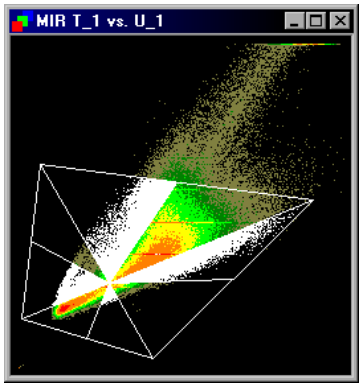


Figure 7. Corresponding scoreplot (T1-U1) and image for off-centred Maltese Cross. The complementary 50% segment is shown in figure 8.

Another potential problem is when the modes (the "peaks") in the scoreplot does not lie on a straight line. If there are more than two peaks of interest, drawing a representative two-split line through them is practically next to impossible. This problem is illustrated well by a scoreplot from a different representative data set, also from [2], illustrated in figure 9. This example illustrates with all clarity why multivariate image analytical endeavours usually are of an order-of-magnitude more complex than in the ordinary two-way regimen.



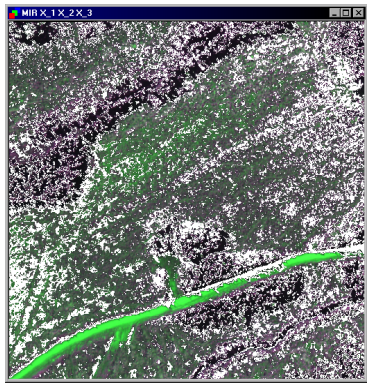


Figure 8. Corresponding scoreplot and image for off-centred Maltese cross. The contrary segment is shown in figure 7.

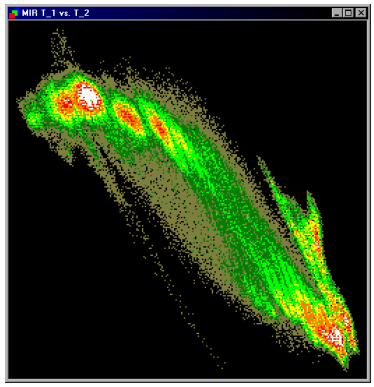


Figure 9. T1-T2 Scoreplot from a complex dataset showing a 7-8 mode ("peaked") curved data structure. Observe how it is apparently impossible to apply a Maltese Cross segmentation on a data structure as complex as this.

Case 3: Y-grid

More commonly than the full Y-image, is when X and Y are constructed as grids from several smaller images. This is a useful approach when making a reference dataset as a basis for a regression model. A typical grid image is shown in figure 11. This image consists of 6 smaller images of different sausages. The corresponding Y-image contains the overall fat-content for each sub-image. The fat content is represented as a grey-level as shown in figure 10. This data set-up was discussed extensively in [2] where used as a vehicle for explaining the concepts of MIR, Multivariate Image

Regression. In this particular case the objective is the be able to predict the average fat content in the six heterogeneous sausages (left in fig. 10).

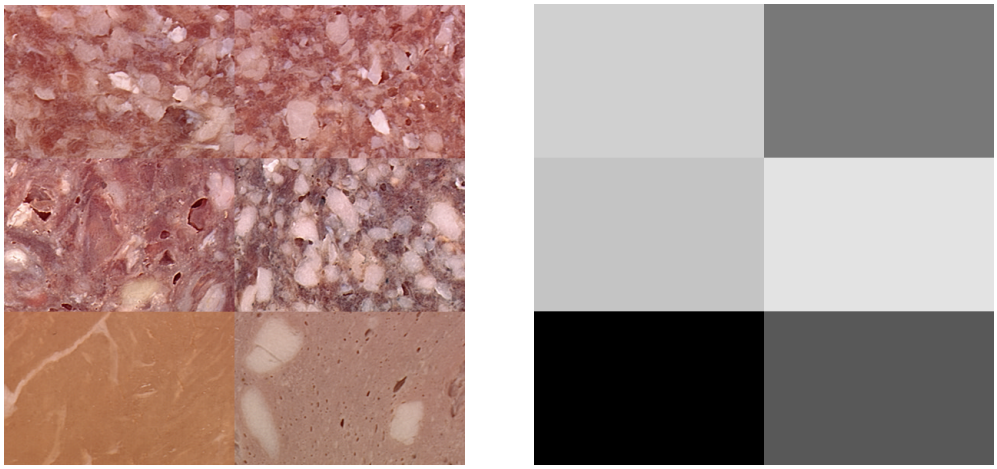


Figure 10. Illustration of the Y-grid MIR case. Six sausages, left: (X-image, variables 1, 2 and 3) and corresponding fat-content, right (Y-image).

As can be seen from the Y-image in figure 10, there is no unique Y-value for each pixel in X. This phenomenon occurs when an overall value is to be predicted from an image, and it has a somewhat negative effect on the T-U scoreplot. This effect is shown in figure 11 the pertinent T1-U1 scoreplot from the sausage data.

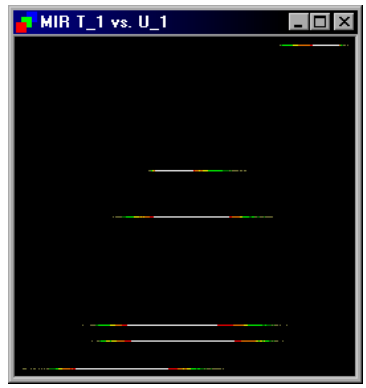


Figure 11. T1-U1 scoreplot from the sausage fat prediction case. Each line represents a specific Y-value, or sub-image, compare figure 10 (right).

In figure 12 it will be demonstrated that applying the eight segment Maltese Cross scheme in a T-U plot, as the one in figure 11 is not at all straight forward. The nature

of the T-U plot in grid cases will force an uneven distribution in the image-space, almost no matter how the eight-fold segmentation mask is delineated. Also observed how the score space delineations are very difficult to evaluate because of the extremely discrete nature of the Y-levels present in a Y-grid case; for full details, see [2].

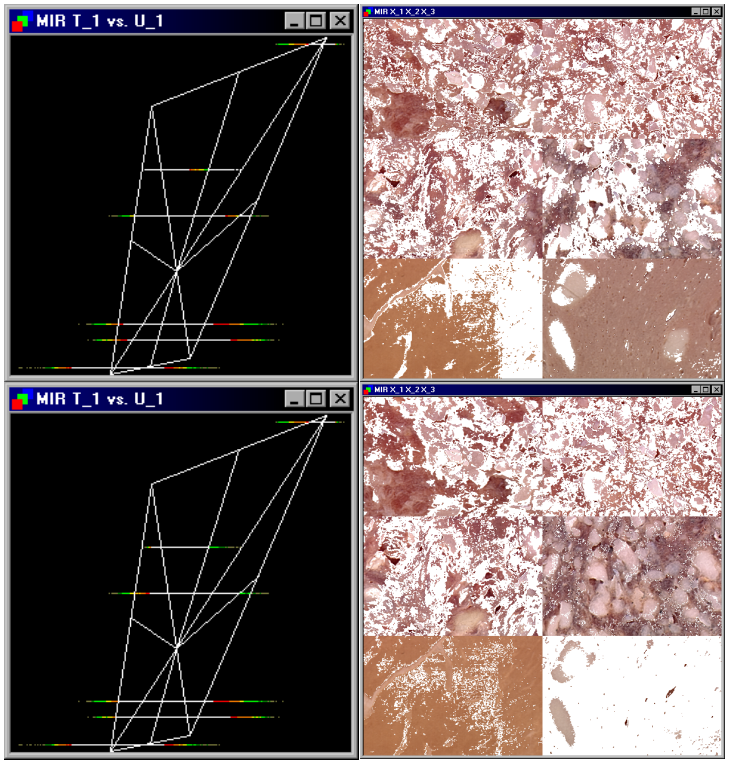


Figure 12. Y-grid case, two-block segments from selection in T1-U1 plot. Note extensive unbalance in the image space (right).

In figure 12 it is evident that especially the two lower and the middle right X sub-images are very poorly represented in the complementary validation segments. This is even more so if eight individual segments were to be used, as was shown in the first example in figure 6. To save space, this is not repeated for the current example.

Thus what seemed initially to be a good idea, i.e. the "Maltese Cross" eightfold cross validation segmentation in the TU-score space, on further inspection has proved to be at best a very sensitive approach - in fact in would be wrong to say that it has proved its reason for existence convincingly.

It can be shown, however, that this is merely a question of application. The critical point is not so much *how* the lines are drawn in the plot, it is *what plot* the lines are drawn in. So far, the procedure has been applied to plots where there are strong

correlation in the data, and physical objects have specific locations too, i.e. the familiar low-order score plot(s), e.g. T_1-U_1 etc. which all play a very dominating role in conventional 2-way multivariate calibration [4]. Chemometricians will be familiar with the fact that in the score space, the first dimensions contain the most structured parts of the data, while for the higher-order components there is bound to be less and less variance etc.

With this in mind, the next, perhaps surprising step in the present image exploration will focus explicitly on this higher-order score space.

Figure 13 shows T-scores 4 vs. 5 from the master Montmorency Forest example. What is interesting in this plot, is that most of the *structural* information is now *orthogonal* to the data delineated in this figure. This indicates that the current plot is *well* suited as a starting point for the cross validation data segmentation.

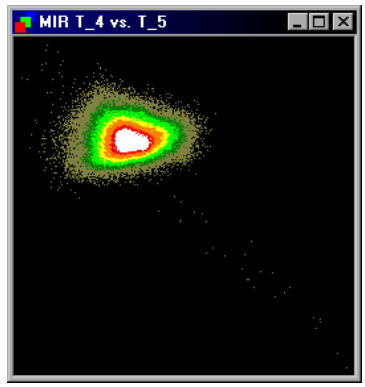


Figure 13. Alternative higher-order components scoreplot (T4 vs. T5) from the Montmorency Forest (figure 1).

Below, a Maltese Cross cross validation segmentation has been applied to the scoreplot in figure 13. Figure 14 shows the resulting two non-overlapping segments both in score space and image space. As can bee seen from the figure, there is now a very satisfactory even distribution in the two segments (and only with very close investigation, some minor differences can be found between the image-space representations though, which have to do with **shadows** mainly).

In figure 15 this is further illustrated by examining the eight segments separately. The conclusions from figs 13 - 15 are very clear: when delineating the new image analytical eight-fold cross validation segmentation in some appropriate higher-order score space rendition, in which most of the substantial data structure is orthogonal, the documented sensitivity has been controlled completely.

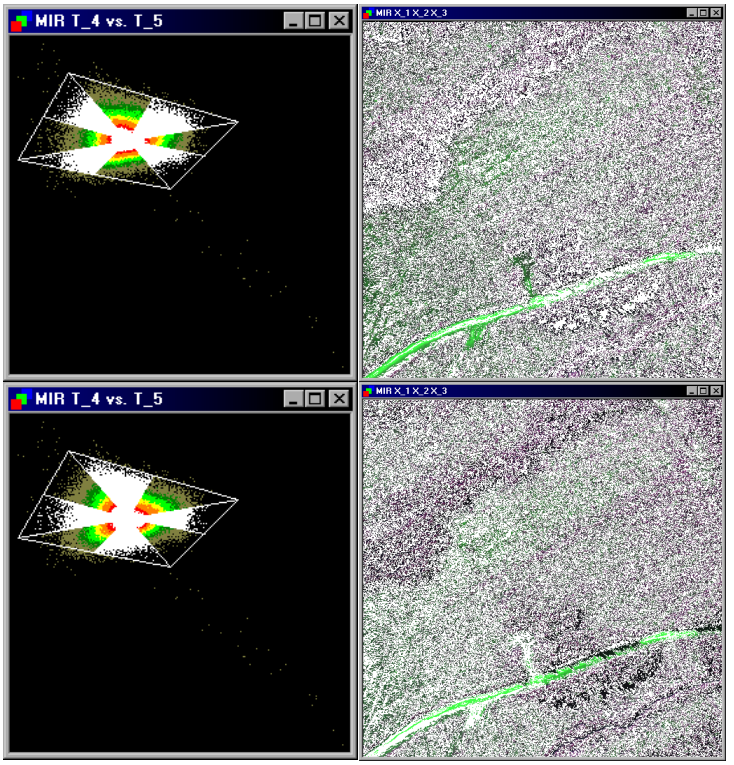


Figure 14. . Maltese Cross validation segments selected in higher-order T4-T5 scoreplot in figure 13. Note excellent data structure as well as image (spatial) coverage and representativity.

It is still evident that an "even" rotation of the segments in score-space, leads to extremely opposing unbalanced pixel divisions in the corresponding image space. From this it is necessary to conclude that many such segments must always be *combined* to form larger fractions of the entire field-of-view, e.g. two 50% segments as in figure 14.

Stepping back to the difficult Y-grid example (sausage fat-prediction), it is now interesting to see how this higher-order components approach will behave. Using T-scores 5 vs. 6 and drawing the two lines that split this data set in as equally representative fashion as possible produces the segments shown in figure 16.

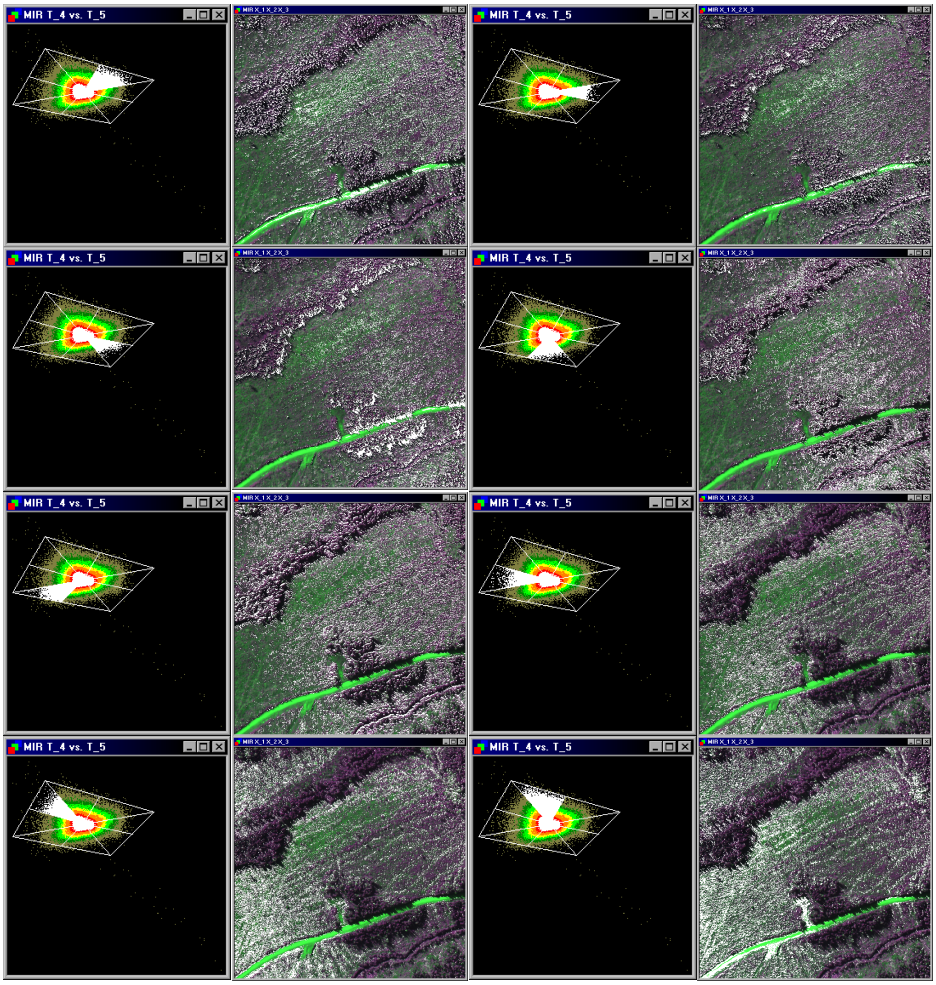


Figure 15. Eight segments from the T4-T5 scoreplot cross validation splitter shown in both score- and image space. Note that an acceptable representation has now been achieved in both score- and image space; compare figure 6.

Compared with figure 12, figure 16 now shows a strikingly more uniform distribution of one validation segment in the image with respect to the complementary calibration set - and there are only a few, minor differences. Overall, this partition should lead to a realistic validation of the prediction model performance even for this very complex difficult data structure.

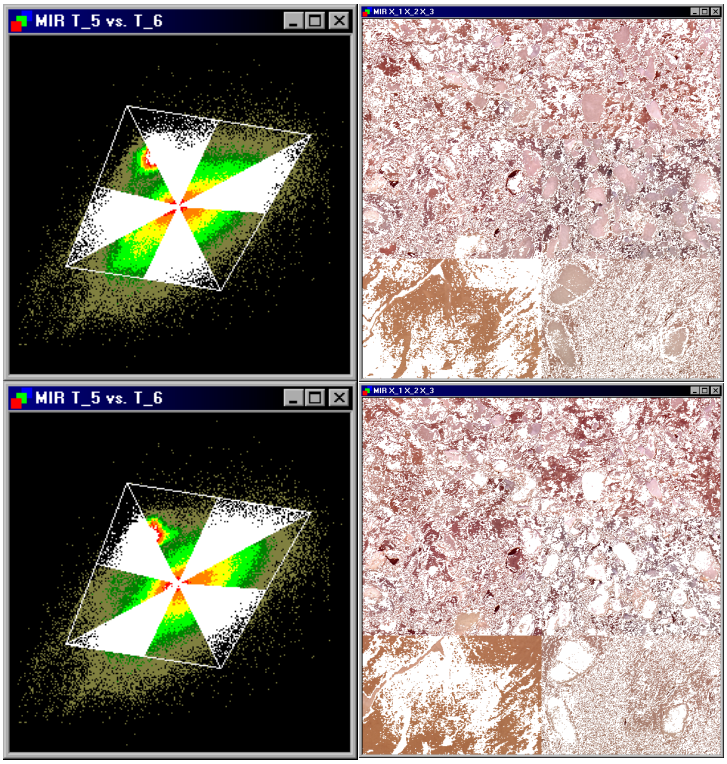


Figure 16. Cross validation segments in [T5T6] score- and[1,2,3] image space for the sausage fat prediction example. Note that an acceptable representation has now been achieved in both score- and image space. Compare figure 12.

Case 4: Cutting to the bone

One of the key features in image analysis, mentioned in the introduction, is the huge redundancy in this type of data. Having 350.000 objects describing, say typically, 10-20 classes is obviously an overkill. In MIR-cases where reducing this redundancy is essential, it is possible to reduce the number of objects dramatically by a simple procedure, compare also [8] in which this case was described for MIA. The suggestion is shown in figure 17 in the form of the curved (hand-drawn) line, where the number of objects have been reduced to a small fraction of the original, but deliberately covering all the important classes of interest in the image. This is so because it has been drawn specifically to "cover" the most dominating global covariance trend of the image feature space. Since this mask is positioned directly along the "topographic" highs, compare [8] for full details, it will - per force - be maximally representative for the essential data structure present while at the same time allowing for the exclusion of all similar pixels lying outside its width (typically 1-3 pixels wide) without any risk of

loosing out on the most representative pixels. Observe how we have made use of this feature in the so-called "pred-meas" plot (predicted vs. measured), well-known from conventional 2-way multivariate regression validation. Thus for fig. 17 below:

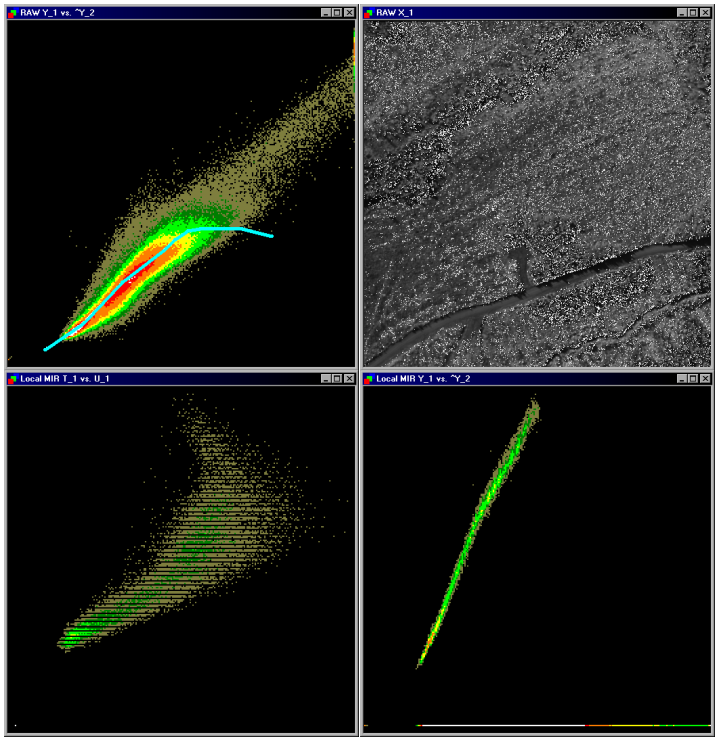


Figure 17. Freehand line covering the essential covariance structure in a Pred.-Meas. plot. After a local MIR model has been created on this basis, the corresponding T1U1 (lower left) and the Pred.-Meas. (lower right) plots are shown, validating this type of representative sampling of MIA/MIR data.

Starting out in the strongly correlated pred-meas plot, a one-pixel-wide line is drawn covering the main data of interest. This line emulates the global covariance trend as best as at all possible. All objects (pixels) covered by this line only, are then used as objects in a new, *local model* [3,8]. This model will contain far less objects, and the redundancy in the data will be strongly reduced. In figure 17, a T1-U1 plot is shown at the lower left. This can now be used as a starting point for the cross validation segmentation. The corresponding *local model* pread.-meas plot is shown at the right in the figure.

Some comments are required for the last figure. The points and line that can be observed in the lower part of the plots, represent the objects that have been **left out of the model**. In the calibration procedure, they have been removed from the data modelling, but for image

displaying purposes, it is necessary to include these pixels. To avoid them from interfering with the image, they are set to zero-value, and are displayed black in the image. The lower-left point is hence the (0,0) coordinate, as all score-values are scaled in the range [0..255] to optimise their display.

DISCUSSION AND CONCLUSIONS

We have shown that the new approach in which segmentation is done based on the *orthogonal* data representation in *higher-order* score components, is of a powerful and general nature, which in most cases will enable a realistic two-split cross validation (approximately 50/50). Segmentation following this approach takes the form of two non-overlapping "mirror" Maltese Cross configurations, each made up of four "arms". The Maltese Cross is designed specifically to allow equal (but non-overlapping) neighbouring segments in parallel along both the user-defined axes of the mask (figure 3). This enables a near-optimal representative split of the training data set across all covariance structure directions, precisely because of this *compound* nature.

We have also shown that considerable care is needed when employing this feature on the alternative lower-order component plots available (e.g. T1-U1), in which a rather large "off-centre" sensitive was demonstrated.

In general it is not recommended to use cross validation in multivariate image analysis with a number of segments higher than two, and then *only* in the form of the Maltese Cross (sic) - due to the much higher complexity of the covariance structures for this type of data relative to the experiences from the conventional 2-way realm.

In multivariate image analysis, there is usually a high degree of redundancy in the data. In such cases with relatively few physical objects (classes), data reduction with local modelling should be considered prior to validation. We have delineated a simple approach for this – the one-pixel-wide swath across the backbone of the dominating data covariance structure(s).

REFERENCES

1 Esbensen K., Geladi P. & Grahn H. 1992: Strategies for multivariate image regression (MIR). Chemometrics and Intelligent Laboratory Systems vol 14, pp. 67-86

2 Lied T.T. & Esbensen K. 200?: Principles of MIR, Multivariate Image Regression -I: Regression typology and representative application studies. In preparation.

3 Geladi P. & Grahn H. 1996: Multivariate image analysis. (Chichester: John Wiley & Sons) ISBN 0-471-93001-6

4 Esbensen K. 2000: Multivariate Analysis in Practice, 4th edition. CAMO ASA. ISBN 82-993330-2-4

5 Lindgren, F. 1994: Third Generation PLS. PhD thesis, Umeå University. ISBN 91-7174-911-X

6 Martens H. and Næs, T. 1989: Multivariate Calibration. Wiley & Sons . ISBN 0-471-93047-4.

7 Esbensen K., Edwards, G. & Eldridge N.R. 1993: Multivariate Image Analysis in forestry applications involving high resolution airborne imagery. 8th Scandinavian Conference on Image Analysis- SCIA'93, pp. 953-963

8 Esbensen K., Lied T.T., Lowell K. and Edwards G. 200?. Principles of Multivariate Image Analysis (MIA) in remote sensing, technology and industry. In preparation.