Discriminant Analysis and Feature Selection in Mass Spectrometry Imaging using Constrained Repeated Random Sampling - Cross Validation (CORRS-CV)

David Perezguaita  
*Technological University Dublin*, david.perezguaita@tudublin.ie

Julia Kuligowsky  
*Health Research Institute, Hospital La Fe*

Guillermo Quintas  
*Leitat Technological Centre, Barcelona, Spain*

Follow this and additional works at: https://arrow.tudublin.ie/scschphyart

Part of the Atomic, Molecular and Optical Physics Commons

Recommended Citation
Pérez-Guaita, D., Quintás, G., & Kuligowski, J. (2020) Discriminant analysis and feature selection in mass spectrometry imaging using constrained repeated random sampling - cross validation (CORRS-CV), *Analytica Chimica Acta*, DOI:10.1016/j.aca.2019.10.039.

This Article is brought to you for free and open access by the School of Physics & Clinical & Optometric Science at ARROW@TU Dublin. It has been accepted for inclusion in Articles by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, gerard.connolly@tudublin.ie, vera.kilshaw@tudublin.ie.
Discriminant analysis and feature selection in mass spectrometry imaging using constrained repeated random sampling - cross validation (CORRS-CV)

David Pérez-Guaita, Guillermo Quintás, Julia Kuligowski

PII: S0003-2670(19)31266-8
DOI: https://doi.org/10.1016/j.aca.2019.10.039
Reference: ACA 237173

To appear in: Analytica Chimica Acta

Received Date: 2 April 2019
Revised Date: 16 October 2019
Accepted Date: 18 October 2019

Please cite this article as: D. Pérez-Guaita, G. Quintás, J. Kuligowski, Discriminant analysis and feature selection in mass spectrometry imaging using constrained repeated random sampling - cross validation (CORRS-CV), Analytica Chimica Acta, https://doi.org/10.1016/j.aca.2019.10.039.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Elsevier B.V. All rights reserved.
CORRS-CV: 99% pixels included in > 1 test set after 150 iterations
Discriminant analysis and feature selection in mass spectrometry imaging using constrained repeated random sampling - cross validation (CORRS-CV)

David Pérez-Guaita\textsuperscript{a}, Guillermo Quintás\textsuperscript{b,c,*}, Julia Kuligowski\textsuperscript{d}

\textsuperscript{a}FOCAS Research Institute, Dublin, Ireland
\textsuperscript{b}Health & Biomedicine, LEITAT Technological Center, Barcelona, Spain
\textsuperscript{c}Unidad Analítica, Health Research Institute Hospital La Fe, Valencia, Spain
\textsuperscript{d}Neonatal Research Unit, Health Research Institute Hospital La Fe, Valencia, Spain

*Corresponding author: gquintas@leitat.org

Abstract

The identification of biomarkers through Mass spectrometry imaging (MSI) is gaining popularity in the clinical field. However, considering the complexity of spectral and spatial variables faced, data mining of the hyperspectral images can be troublesome. The discovery of markers generally depends on the creation of classification models which should be validated to ensure the statistical significance of the discriminants m/z detected. Internal validation using resampling methods such as cross validation (CV) are widely used for model selection, the estimation of its generalization performance and biomarker discovery when sample sizes are limited and an independent test set is not available. Here, we introduce for first time the use of Constrained Repeated Random Subsampling CV (CORRS-CV) on multi-images for the validation of classification models on MSI. Although several aspects must be taken into account (e.g. image size, CORRS-CV \( \delta \) value, the similarity across spatially close pixels, the total computation time), CORRS-CV provides more accurate estimates of the model performance than \( k \)-fold CV using of biological replicates to define the data split when the number of biological replicates is scarce and holding images back for testing is a waste of valuable information. Besides, the combined use of CORRS-CV and rank products increases the robustness of the selection of discriminant features as candidate biomarkers which is an important issue due to the increased biological, environmental and technical variabilities when analysing multiple images, especially from human tissues collected in clinical studies.

Keywords. mass spectrometry imaging (MSI); cross validation (CV); constrained repeated random sampling - cross validation (CORRS-CV); partial least squares-discriminant analysis (PLS-DA); feature selection
1. INTRODUCTION

Mass spectrometry imaging (MSI) is a powerful technique that acquires high dimensional mass spectra at known spatial points of a sample, providing quantitative information of the spatial (i.e. morphological) distribution of organic compounds such as proteins, peptides, lipids, drugs or metabolites in biological tissues. MSI is becoming an established tool in clinical and pharmaceutical research that might be soon be introduced into clinical practice thanks, at least partially, to technological advances in the development of novel ambient ionization techniques and MS instrumentation allows MSI with minimal sample preparation [1][2][3].

MSI generates large, complex datasets that require advanced high-throughput data processing and data mining techniques for a thorough biological analysis. Two of the main objectives of MSI data analysis are segmentation and classification. Segmentation aims at the unsupervised identification of pixels with characteristic spectral profiles. Classification aims at assigning pixels to pre-defined classes (e.g. control or disease) or to get further insight of underlying biochemical differences based on known information provided by a reference method such as pathological examination, thus facilitating the identification and interpretation of clinical biomarkers or the development of classifiers[4]. Segmentation is usually carried out using multivariate unsupervised methods that do not require a priori information or foreknowledge of the classes of the pixels and aim at revealing general data structure. Many methods have been exploited such as principal component analysis or clustering methods, including hierarchical, K-means, fuzzy k-means, and spatially aware data clustering [4][5]. Alternatively, classification requires a supervised analysis across comparable histological areas which is usually based on univariate tests (e.g. t-test, one-way ANOVA, Wilcoxon rank sum test), fold changes or on multivariate strategies such as linear discriminant analysis, partial least squares discriminant analysis (PLSDA), principal component discriminant analysis or support vector machines[2][4][6][7][8].

A critical step in the development of any classifier is the assessment of its generalization performance to estimate its prediction accuracy beyond the training data[9]. Besides, this step also supports the identification of discriminant MS features, prior to biomarker interpretation. Considering the large number of m/z on MS spectra, a significant challenge of MSI is the large data volume of the images generated. Its introduction into clinical practice would likely lead to an exponential growth in the computational requirements. Therefore, the development of algorithms for biomarker selection might contribute to reduce the data volume while improving the quality of the information retrieved, which will facilitate the development of hardware and software to meet the demand for automated image analysis.

Model development in an ideal situation involves the split of the MSI data set into three subsets: a training set to fit the models, a validation set to select the optimal model (i.e. selection of algorithm parameters that are not directly estimated from the data such as the number of latent variables (LVs) in PLS-DA) and a test set to estimate its predictive accuracy. Train, validation and test subsets should be independent, representative of the whole
population and include all sources of technical and biological variation[10]. However, if the total number of samples is small, this strategy may not be a feasible solution. In these situations, internal validation using resampling methods such as cross validation (CV) are commonly used for model selection and the estimation of its generalization performance.

K-fold CV is one of the most widely used methods for CV. During k-fold CV, the set of N objects (e.g. pixels or images) is split into k mutually exclusive subsets of size N/k. Then, k-1 subsets are used as training set to fit a model, which is used to predict the left-out validation subset. The process is repeated k times, each time excluding a different validation subset and then, an estimate of the model performance is calculated from the predicted values. Therefore, each pixel is included in a validation set once and k-1 times in the training sets. Lower k values typically lead to estimates of prediction error biased upward and higher k values minimize bias but increase variance[11].

As a general rule, CV data split should be done at the highest level of the sampling hierarchy and based e.g. on biological replicates, rather than technical replicates or individual pixels[12],[13]. However, this approach may be impractical if the number of images from independent biological sources is scarce. As a suboptimal alternative, the dataset can be randomly split into k equally sized folds. However, a selection of the k subsets completely at random results in a loss of spatial structure and pixel-neighbourhood information and spatially proximate pixels are distributed simultaneously into train and validation sets, leading to a situation resembling to the replicate trap and providing overly optimistic classification accuracy estimates.

In order to overcome this potential pitfall, a straightforward CV strategy, namely constrained repeated random subsampling – cross validation (CORRS-CV) has been recently proposed[11] and evaluated for the assessment of PLS-DA models of single infrared hyperspectral images where it avoided the optimistic effect of spectral oversampling. CORRS-CV is based on the random generation of training and test sub-sets using a constrained random sampling of training pixels without replacement. The constrained sampling ensures that the Euclidean distance among train pixels and between the test and train pixels is higher than a user defined threshold ($\partial$). After completing the training set selection, random repeated k-fold CV is used for model selection and optimization. In this inner k-fold CV, CORRS-CV also ensures that the distance among train and validation pixels is higher than $\partial$. Then, pixels included in the test set are used to estimate of the prediction accuracy of the trained model. The process of training set selection and model optimization is iterated to average the impact of the random selection of the train and test subsets. At the end of the process, a set of predicted values for each pixel is obtained which enables a better estimate of the class than a single prediction[14] and can be used to assess the performance of the classification models and to build class-predicted pseudo-image.

The objective of this study was to extend the application of CORSS-CV to MSI and to the discovery of biomarkers during the simultaneous analysis of multiple images obtained from different patients, a frequent scenario in biomarker discovery studies. This required the
adaptation of the CORSS-CV algorithm to multi-images, where boundaries of each image should be considered. Using the human renal cell carcinoma (RCC) data set[15] as benchmark, the use of biological replicates to define a k-fold CV split provided an accurate estimation of the generalization performance, but it was found inefficient when the number of biological replicates was scarce. We also evaluated different factors affecting the implementation of CORSSCV such as the images size, the \( \partial \) value, the similarity across spatially close pixels and the total computation time. In situations where holding images back for testing compromises the performance of the analysis, CORRS-CV provided accurate and more robust figures of merit. Besides, information extracted from the iterative development of discriminant models during CORRS-CV enabled a straightforward selection of discriminant biomarkers.

2. MATERIAL AND METHODS

**MSI data**

The RCC data set[15] is a relatively small MSI data set that consists of eight matched pairs of human kidney tissue (i.e. 16 images) including 6077 pixels (i.e. MS spectra) classified as cancer (2775 pixels) or normal (3302 pixels) by a pathologist. Each biological replicate consists of a normal tissue and a cancerous tissue samples with different absolute and relative pixel sizes (see Table 1).

Initial data pre-processing was carried out using Cardinal[13] and included total ion count standardization for pixel intensity normalization, and resampling to unit resolution.

**Constrained repeated random sampling - cross validation (CORRS-CV) for multi-images.**

CORRS-CV is based on the repeated split of the image into training and test subsets following a constrained random sampling of \( n \) training pixels without replacement[11]. Initially, a pixel is randomly sampled and included in the training set. Then, those pixels in the same image at a distance lower than a given threshold (\( \partial \)) from the selected pixel are excluded from the set of pixels suitable to be included in the training set. The process is repeated until \( n \) pixels have been included in the training set, and the test set is defined as the remaining set of pixels which were not included in the training set or excluded because of its closeness to a train pixel. The training set is used for the development of a discriminant model, which is applied for the prediction of the test set and calculation of AUROC value. Model selection is based on a repeated random \( k \)-fold CV of the training set, using the maximum mean classification accuracy as objective function. The whole process of selection of train and test sets, model development and test set prediction is repeated until a user defined value (In this study >99%) pixels have been included least once in a test set and a training set. The number of required iterations depends on the training set and image sizes and the selected \( \partial \). Finally, a mean estimate of the AUROC values obtained after each iteration is calculated.

In this work, PLS-DA was used as a classifier and the AUROC was selected as figure of merit for this binary classification. The number of LVs was selected considering the minimum mean classification accuracy of the inner \( k \)-fold cross validation, considering a maximum of 15 (see
The AUROC was selected because it does not depend on the prior class distribution and so, it facilitates comparing outcomes from data sets with different class distributions[16]. Five iterations of random 5-fold CV was used for CV of the inner PLS-DA models.

Software

Initial data pre-processing was carried out using Cardinal (Release 3.8) [13] in R 3.5.1[17]. Further data analysis was carried out in MATLAB 2017b (Mathworks Inc., Natick, MA, USA) using in-house written scripts and the PLS Toolbox 8.7 (Eigenvector Research Inc., Wenatchee, USA). Calculations were carried out employing a standard laptop computer with Intel® Core i5 processor operating at 2.4 GHz. The MATLAB function used in this work for CORRSCV for multi-images (corrscv_msi.m) is available as supplementary material.

3. RESULTS AND DISCUSSION

Three strategies for CV were compared for the assessment of the discrimination of cancer and normal tissues and feature selection in the RCC data set. The first strategy involved a repeated (n=5) random split of the data set into \( k=5 \) folds (i.e. 5 iterations, 5-fold CV) for the purpose of illustrating how an inadequate CV strategy may easily lead to over optimistic results. The second one, \( k_{BIR} \)-fold CV, was based on the use of biological replicates (BIR) used to define the different \( k \)-fold CV subsets. Finally, the third strategy, CORRSCV, involved a repeated constrained random selection of training sets for \( k \)-fold CV.

Assessment of the predictive performance

False-colour images showing the distribution of a molecule in the sample can be built by plotting the intensity of a given m/z value at the pixels coordinates. Similarly, the outcome from the supervised or unsupervised analysis of the hyperspectral data cubes can be used to plot false-colour images by projecting information extracted (e.g. predicted class) at each pixel coordinates. For example, the input (Y) of the predefined classes on a PLS-DA is a dummy variable (generally +1 and -1) but the output is an unbounded continuous score that can take any values in the \([-\infty, +\infty]\) range. Figure 1A shows mean \( y \) predicted values by random 5-fold CV (5 iterations) of the data set (i.e. using 80% of the total sample size for training each time). Results show an almost perfect classification of cancer and normal tissues, in agreement with an AUROC-CV = 0.998 (LVs=3). Then, a permutation test was carried out for significance testing of the CV-figure or merit. Basically, this type of test measures the likelihood of obtaining an observed classification accuracy by chance[18]. During a permutation test, new labels are randomly assigned to the observations (e.g. pixels) and a new classifier is trained and cross validated. The process is repeated a number of times and the distribution of the accuracy estimates obtained using random class labels is compared to the accuracy estimate obtained using true class labels. Then, an empirical \( p \)-value is calculated as the fraction of values where the classifier provided better classification performance in the random than in the original data. If the \( p \)-value is lower than a user selected threshold (e.g. 0.05), the null hypothesis that there is no difference between the
means is rejected and it is assumed that the classifier is significant. The distribution of AUROC-CV values obtained using randomly permuted class label at the pixel level indicated a statistically significant classifier performance ($p$-value < 0.05, AUROC-CV range [0.48, 0.53]). However, it is expected that results presented were overly optimistic if there is a large redundancy across pixels lying together. On the one hand, a random 5-fold CV may produce train and test sets containing highly correlated spectra from spatially close pixels, leading to over optimistic prediction error estimates using true labels. Besides, the strategy selected to shuffle the class labels should reflect the structure of the data. A random permutation of the class of each pixel leads to situations in which spatially close pixels are differently labeled, leading to overly pessimistic performances for permuted models and so, to estimates of the statistical significance of the true model that were biased in an optimistic way. To avoid this potential pitfall, a second permutation test involving a random shuffle of the class label of each of the 16 tissues as a block was carried out. Although in this case, AUROC-CV values obtained using permuted class label varied in the [0.97, 0.99] range, the obtained $p$-value < 0.05 also indicated a statistically significant classifier performance. The distribution of AUROC-CV values from permuted models indicated that the random shuffle of the class label of each of the 16 tissues images provided a more reliable estimation of the statistical significance and was used for the assessment of $k_{BIR}$-fold CV and CORRS-CV results.

Figure 1B shows predicted values by $k_{BIR}$-fold CV. Results showed again a clear difference between the spectral profiles of cancer and normal tissues, providing an excellent classification rate (AUROC-CV = 0.985, LVs = 5). The lower AUROC-CV compared to that estimated by random $k$-fold CV, supported the hypothesis that random $k$-fold CV had an overly optimistic bias. Nonetheless, results from a permutation test also indicated that the AUROC-CV value estimated by $k_{BIR}$-fold CV was statistically significant ($p$-value < 0.05).

CORRS-CV was then used to assess the performance of a PLSDA model to discriminate cancer from normal tissues. The RCC data includes 16 images of human kidney tissues. However, each image is a relatively small and the maximum distances between pixels of the same class within each image are in the 17-41 AU (4.25-10.25 mm) range. One of the limitations of CORRS-CV is that the size of the training set limits the maximum $\partial$ values tested and vice versa[11]. Large $\partial$ values may lead to ill conditioned training sets formed by a very limited number of pixels not representative of the entire data set. Besides, the CORRS-CV algorithm iterates the training set selection and model development until a percentage of the pixels (99%) is included in, at least, one train and test sets and so, large $\partial$ values increase the number of iterations and the computing time.

An initial test was carried out to analyze the effect of $\partial$ and the train size on the CORRS-CV estimates by testing train sizes and $\partial$ values in the [100, 200] pixels and [1, 5] AU ranges, respectively. Figure 2A depicts the obtained AUROC values as a function of the train size and $\partial$ value. Results showed again an excellent discrimination between both types of tissues with AUROC values in the 0.970-0.988 range. However, for any train size considered, a gradual decrease in the AUROC value for test pixels at increasing $\partial$ was observed. Besides,
the classification accuracy was higher for pixels close to a train pixel for any train size considered (See Figure 2B). These results were in agreement with the hypothesis that the performance estimates observed by $k$-fold CV were overly optimistic.

Figure 1C shows the distribution of the $y$ predicted values by CORSS-CV for the test sets, using a train size of 200 pixels and $\partial = 4$. False color images were comparable to those obtained by $k$-fold CV and $k_{\text{BIR}}$-fold CV. It has to be noted that the dataset used represents as a ‘worst case scenario’ to show differences in the model performance between standard CV and CORRS-CV because the differences between control/cancer tissues were highly significant. Despite this fact, the mean AUROC = 0.982 obtained for the test sets was slightly higher than the AUROC-CV obtained by $k_{\text{BIR}}$-fold CV and lower than that provided by random $k$-fold CV. Moreover, as abovementioned, during CORRS-CV, the discriminant models are built and selected using training sets in which the minimum distance among pixels is also higher than a given threshold (i.e. $\partial$). The accuracy estimates obtained for the train and test sets were comparable (e.g. mean classification errors for inner-CV and testing sets were $0.052\pm0.01$ and $0.060\pm0.02$, respectively) indicating that the elimination of highly correlated spectra from spatially close pixels leads to generalizable inner models.

During CORRS-CV that the selection of training sets and model optimization is iterated, generating a set of predicted values for test pixels. In this work, we used the mean value of the predicted values, but other alternatives are available. For example, a majority vote using the classes predicted at each iteration for each testing pixel or the estimation of confidence intervals for the predicted values could also be applied.

**Effect of the sample size**

The use of $k_{\text{BIR}}$-fold CV is the most appropriate strategy for CV. However, it has several drawbacks: it can be inefficient when the number of biological replicates available is scarce, and also if the different sizes and class distributions of cancer and normal pixels in the biological replicates are highly unbalanced. To test the differences among random $k$-fold CV, $k_{\text{BIR}}$-fold CV and CORRS-CV, Figure 3 shows results obtained using the three CV strategies for every possible combination of biological replicates using between 3 and 8 images from the RCC data set. Results showed that repeated random $k$-fold CV provided the highest CV classification accuracy estimates for every data size with AUROC-CV ranging between 0.997 and 1.000. $k_{\text{BIR}}$-fold CV provided unstable accuracy estimates for limited data sizes as shown by AUROC-CV ranging between 0.634 and 0.997. In these cases, $k_{\text{BIR}}$-fold CV estimates of test error are easily biased because the models build during CV are not as representative as the model trained on the whole data set and, on the other hand, models typically show a poor consistency because prediction estimates are highly dependent on the selected training set. CORRS-CV provided more stable estimates as shown by lower variation observed in the AUROC values in the $[0.982, 0.999]$ range. CORRS-CV selects train pixels randomly across all the images and so, the training sets are more representative of the whole data set as they include signal (and noise) from different images and so, take into account additional sources of error. For this test, a $\partial = 2$ was used to allow the selection of training set of 200 pixels.
during CORRS-CV in small data sets (2-3 images) in which case higher $\partial$ would limit the train sizes.

**Feature selection**

A variety of algorithms have been proposed to interpret observed differences between classes in MSI and to select discriminant features such as bootstrapping or jack-knifing for estimating uncertainty in the model coefficients. Among them, Variable Importance in the Projection (VIP) scores are widely used to select the variables which contribute the most to the $y$ variance explanation in a PLS-DA model[19]. A VIP threshold value of one is generally used as a feature selection criterion. In this work, 3 and 5 LVs were selected based on repeated random 5-fold CV or $k_{BIR}$-fold CV results, leading to 244 and 243 MS features with VIP>1 in each model, respectively and 241 commonly selected in both models. Nonetheless, the type of CV used for model selection may lead to slightly different PLS-DA models and so, to different VIP score vectors.

CORRS-CV involves a repeated double cross validation in which, after selecting a training set, an inner PLS-DA model is selected and applied to a test set. The process is iterated providing a set of PLS-DA models resulting in different feature rankings according to their VIP values. Therefore, it seems reasonable to combine the information obtained from the set of train models to provide a more robust feature selection to aid in eliminating false positive biomarkers. Accordingly, a mean rank product (RP) was calculated as the geometric mean of the VIP ranks in the set of PLS-DA models. As described by Breitling et al. [20], “the rank products can be used to sort the features according to the likelihood of observing them so high on the list of differently expressed features by chance.” Low ranks (i.e. high VIP scores) consistently obtained in the PLS-DA models build for the different training sets lead to low mean VIP ranks, thus allowing the identification of the most discriminant features.

Although the rank of the VIP scores and their RPs are useful to sort the relative important of m/z features, they are not indicative of statistical significance and the selection of a threshold is essentially arbitrary. To determine the significance level of both, the VIP rank and the VIP RP obtained for each feature, values obtained in the model build using real class labels were compared to reference null distributions formed by the set of values found in the models build during the permutation tests [18]. Features that did not belong to the distribution of random values ($p$-value < 0.05) were classified as potential biomarkers. This strategy lead to three subsets of 27, 26 and 31 features based on $k$-fold CV, $k_{BIR}$-fold CV and CORRS-CV, respectively (see Figure 4), with 13 features commonly selected (m/z 215, 308, 353, 589, 612, 613, 716, 760, 782, 784, 808 and 809). Although CORRS-CV was slightly less conservative than $k$-fold CV, the subset included features previously reported to be associated with characteristic lipid profiles of RCC[15] such as increased intensities of PS(18:0/20:4) (m/z 810), PI(18:0/20:4) (m/z 886) and PI(22:4/18:0) (m/z 914) and decreased intensity of FA(12:0) (m/z 215), the latter with the lowest mean VIP rank product.
4. CONCLUSIONS

Assessment of discriminant models in hyperspectral MSI data sets is challenging and we always strive to validate at the most conservative level. The preferred strategy is independent testing, although in cases where the number of samples is limited, CV could be regarded as a sub-optimal alternative. K-fold CV is one of the most widely used methods for CV where, as a general rule, data split should be done at the highest level of the sampling hierarchy (e.g. biological replicate) rather than technical replicates or individual pixels. However, a low number of biological replicates represents a challenge for CV. Results show that the use of biological replicates to define k-fold CV split provides an accurate estimation of the generalization performance, but it is inefficient when the number of biological replicates is scarce. In these situations where holding images back for testing is a waste of valuable information, CORRS-CV can be used to reduce, or even eliminate, the overly optimistic bias due to the use of test pixels close to the training set, providing more accurate and robust figures of merit. Besides, the use of VIP rank products increases the robustness of the selection of discriminant features as candidate biomarkers which is an important issue due to the increased biological, environmental and technical variabilities when analysing multiple images, especially from human tissues collected in clinical studies. Nonetheless, several aspects must be taken into account. On the one hand, increasing∂ limits the number of training pixels and it might lead to an ill-conditioned trap where the underlying distribution is not represented in the train data. As the size of the training set and∂ determine each other, a trade-off must be ensured, taking into account the image size, the similarity across spatially close pixels and the total computation time. This study shows the performance of CORRS-CV involving PLS-DA; however, the method of subset- selection of training and test sets could potentially be beneficial also in combination with other algorithms for sample discrimination.

Acknowledgements

The project received funding from the Spanish Agencia Estatal de Investigación (AEI) and the Fondo Europeo de Desarrollo Regional (FEDER) (CTQ2016-79561-P). DPG acknowledges support from the European Research Council MSCA grant (Spectro-metrics, 020-MSCA-IF-2017 Project ID:796287). JK acknowledges her personal Miguel Servet grant (CP16/00034, Instituto de Salud Carlos III, Ministry of Economy and Competitiveness, Spain).
Figure 1. PLSDA $y$ predicted values by $k$-fold CV using repeated random 5-fold CV (A), biological replicates to define the $k$-fold CV subsets (B) and CORRS-CV ($\vartheta$: 2, train size: 200 pixels) (C).
Figure 2. AUROC values calculated by CORRS-CV for testing sets (A) and for pixels close to the training sets (B) excluded from the test sets, as a function of the maximum distance to a train pixel and train size in the RCC data set.

Figure 3. Effect of the number of CRC images on the predicted performance of PLS-DA models for the discrimination between cancer and normal tissues in the RCC data set. For each number of images, results using all possible combination of RCC images were calculated using $k$-fold CV (green circles), $k_{BIR}$-fold CV (red circles) and CORRS-CV (yellow circles). NOTE: Points are slightly shifted on the x-axis for a better visualization.
Figure 4. Number of features selected as discriminant based on results obtained by $k$-fold CV, $k_{BIR}$-fold CV and CORRS-CV.
Table 1. Number of pixels classified as cancer or normal tissue in each matched pair of human kidney tissues included in the RCC data set.

| Image (biological replicate) | Cancer | Normal | %Cancer pixels within image | %Cancer pixels across images |
|------------------------------|--------|--------|----------------------------|----------------------------|
| 1                            | 297    | 514    | 37                         | 9.0                        |
| 2                            | 190    | 204    | 48                         | 5.7                        |
| 3                            | 202    | 161    | 56                         | 6.1                        |
| 4                            | 409    | 392    | 51                         | 12.4                       |
| 5                            | 527    | 229    | 70                         | 16.0                       |
| 6                            | 533    | 81     | 87                         | 16.1                       |
| 7                            | 351    | 586    | 34                         | 10.6                       |
| 8                            | 793    | 608    | 57                         | 24.0                       |
References

[1] B. Rocha, C. Ruiz-Romero, F.J. Blanco, Mass spectrometry imaging: a novel technology in rheumatology, Nat. Rev. Rheumatol. 13 (2017) 52–63. doi:10.1038/nrrheum.2016.184.

[2] A. Rae Buchberger, K. DeLaney, J. Johnson, L. Li, Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights, Anal. Chem. 90 (2018) 240–265. doi:10.1021/acs.analchem.7b04733.

[3] M. Dilillo, B. Heijs, L.A. McDonnell, Mass spectrometry imaging: How will it affect clinical research in the future?, Expert Rev. Proteomics. 15 (2018) 709–716. doi:10.1080/14789450.2018.1521278.

[4] E.A. Jones, S.-O. Deininger, P.C.W. Hogendoorn, A.M. Deelder, L.A. McDonnell, Imaging mass spectrometry statistical analysis, J. Proteomics. 75 (2012) 4962–4989. doi:10.1016/j.jprot.2012.06.014.

[5] T. Alexandrov, M. Becker, S.-O. Deininger, G. Ernst, L. Wehder, M. Grasmair, F. von Eggeling, H. Thiele, P. Maass, Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering, J. Proteome Res. 9 (2010) 6535–6546. doi:10.1021/pr100734z.

[6] J.G. Swales, G. Hamm, M.R. Clench, R.J.A. Goodwin, Mass spectrometry imaging and its application in pharmaceutical research and development: A concise review, Int. J. Mass Spectrom. 437 (2019) 99–112. doi:10.1016/j.ijms.2018.02.007.

[7] F. Falcetta, L. Morosi, P. Ubezio, S. Giordano, A. Decio, R. Giavazzi, R. Frapolli, M. Prasad, P. Franceschi, M. D’Incalci, E. Davoli, Past-in-the-Future. Peak detection improves targeted mass spectrometry imaging, Anal. Chim. Acta. 1042 (2018) 1–10. doi:10.1016/j.aca.2018.06.067.

[8] J. He, L. Huang, R. Tian, T. Li, C. Sun, X. Song, Y. Lv, Z. Luo, X. Li, Z. Abliz, MassImager: A software for interactive and in-depth analysis of mass spectrometry imaging data, Anal. Chim. Acta. 1015 (2018) 50–57. doi:10.1016/j.aca.2018.02.030.

[9] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning. Data mining, inference, and prediction, 2nd ed., Springer Verlag, 2017.

[10] R.G. Brereton, Chemometrics for pattern recognition, Wiley, Chichester, U.K, 2009.

[11] D. Pérez-Guaita, J. Kuliwowski, B. Lendl, B.R. Wood, G. Quintás, Assessment of discriminant models in infrared imaging using constrained repeated random sampling - Cross validation, Anal. Chim. Acta. 1033 (2018) 156–164. doi:10.1016/j.aca.2018.05.019.

[12] S. Hugelier, O. Devos, C. Ruckebusch, On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis, J. Chemom. 29 (2015) 557–561. doi:10.1002/cem.2742.

[13] K.D. Bemis, A. Harry, L.S. Eberlin, C. Ferreira, S.M. van de Ven, P. Mallick, M. Stolowitz, O. Vitek, Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments, Bioinforma. Oxf. Engl. 31 (2015) 2418–2420. doi:10.1093/bioinformatics/btv146.

[14] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics. 4 (2008) 81–89. doi:10.1007/s11306-007-0099-6.

[15] A.L. Dill, L.S. Eberlin, C. Zheng, A.B. Costa, D.R. Ifa, L. Cheng, T.A. Masterson, M.O. Koch, O. Vitek, R.G. Cooks, Multivariate statistical differentiation of renal cell carcinomas based on lipidomic analysis by ambient ionization imaging mass spectrometry, Anal. Bioanal.
[16] I. Tsamardinos, A. Rakhshani, V. Lagani, Performance-Estimation Properties of Cross-Validation-Based Protocols with Simultaneous Hyper-Parameter Optimization, in: A. Likas, K. Blekas, D. Kalles (Eds.), Artif. Intell. Methods Appl., Springer International Publishing, 2014: pp. 1–14.

[17] R: A language and environment for statistical computing. R Foundation for Statistical Computing, R Development Core Team, n.d. http://www.r-project.org/.

[18] C.M. Rubingh, S. Bijlsma, E.P.P.A. Derks, I. Bobeldijk, E.R. Verheij, S. Kochhar, A.K. Smilde, Assessing the performance of statistical validation tools for megavariate metabolomics data, Metabolomics. 2 (2006) 53–61. doi:10.1007/s11306-006-0022-6.

[19] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. 29 (2015) 528–536. doi:10.1002/cem.2736.

[20] R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, FEBS Lett. 573 (2004) 83–92. doi:10.1016/j.febslet.2004.07.055.
HIGHLIGHTS

The validation of MSI biomarkers is challenging due to correlation among pixels.
CV based on pixels provide over-optimistic results.
CV based on biological replicates improves reliability, but it is no viable at small sample sizes.
CORSSCV provides a suitable alternative, providing trustworthiness to the biomarkers found.
Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: