Journal Name

ARTICLE

Common Mistakes in Cross-Validating Classification Models

Shuxia Guo,[a, b, †] Thomas Bocklitz,[a, b, †] Ute Neugebauer,[a, b, c] and Jürgen Popp [a, b,d]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

In this contribution we investigated the common mistakes of cross-validation (CV) for the development of chemometric models for Raman based biological applications. We focused on two common mistakes: the first mistake occurs when splitting the dataset into training and validation data sets improperly; and the second mistake is regarding the wrong position of a dimension reduction procedure with respect to the CV loop. For the first mistake, we split the dataset either randomly or each technical replicate was used as one fold of the CV and compared the results. To check the second mistake, we employed two dimension reduction methods including principal component analysis (PCA) and partial least squares regression (PLS). These dimension reduction models were constructed either once for the whole training data outside the CV loop or rebuilt inside the CV loop for each iteration. We based our study on a benchmark dataset of Raman spectra of three cell types (MCF-7, BT-20, and OCI-AML3), which included nine technical replicates respectively. Two binary classification models were constructed with a two-layer CV. For the external CV, each replicate was used once as the independent testing data set. The other replicates were used for the internal CV, where different methods of data splitting and different positions of the dimension reduction were studied.

The conclusions include two points. The first point is related to the reliability of the model evaluation by the internal CV, illustrated by the differences between the testing accuracies from the external CV and the validation accuracies from the internal CV. It was demonstrated that the dataset should be split at the highest hierarchical level, which means the biological/technical replicate in this manuscript. Meanwhile, the dimension reduction should be redone each iteration of the internal CV loop. The second aspect relates to the optimization performance of the internal CV, benchmarked by the prediction accuracy of the optimized model on the testing data set. Comparable results were observed for different methods of data splitting and positions of dimension reduction in the internal CV. That means if the internal CV is used for optimizing the model parameters, the two mistakes are less influential in contrast to the model evaluation.

[a.] *Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, 07745 Jena, Germany.*
[b.] *Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany.*
[c.] *Center for Sepsis Control and Care, Jena University Hospital, Germany.*
[d.] *InfectoGnostics Research Campus Jena, Center for Applied Research, Jena, Germany.*
† *These authors share main authorship due to equal contributions.*
Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

Journal Name

ARTICLE

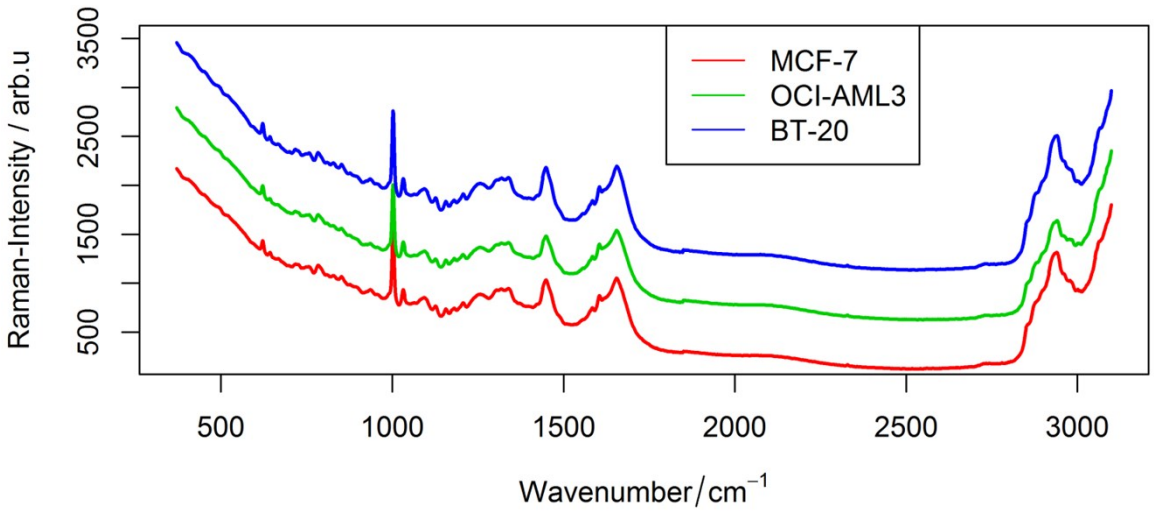

Fig. S1. Mean Raman spectra of the investigated three cell types.

```
////inside-CV
for i=1:N    //external CV
  ts_i=data without ith replicate
  vs_i=data of ith replicate
  fold=split(ts_i, replicate, nFold=N-1)  //split randomly or as replicate
  for nDim=3:50
    for j=1:(N-1) //internal CV
      ts_ij=ts_i[fold[-j],]   //jth training set for ith iteration of external CV
      vs_ij=ts_i[fold[j],]   //jth validation set for ith iteration of external CV
      pca=PCA(ts_ij, nDim)
      classifier=LDA(predict(pca, ts_ij))
      acc_nj=accuracy(predict(classifier, predict(pca, vs_ij)))
    End
  End
  optDim=nDim with the maximum in rowMeans(acc)
  pca=PCA(ts_i, optDim)
  acc_i=accuracy(predict(classifier, predict(pca, vs_i)))
End
```

```
////outside-CV
for i=1:N    //external CV
  ts_i=data without ith replicate
  vs_i=data of ith replicate
  fold=split(ts_i, replicate, nFold=N-1)  //split randomly or as replicate
  for nDim=3:50
    scores=predict(PCA(ts_i, nDim),ts_i)
    for j=1:(N-1) //internal validation
      tIndex_ij=fold[-j]   //jth training set for ith iteration of external CV
      vIndex_ij=fold[j]   //jth validation set for ith iteration of external CV
      classifier=LDA(scores[tIndex_ij,])
      acc_nj=accuracy(predict(classifier, scores[tIndex_ij,]))
    End
  End
  optDim=nDim with the maximum in rowMeans(acc)
  pca=PCA(ts_i, optDim)
  acc_i=accuracy(predict(classifier, predict(pca, vs_i)))
End
```

Fig. S2. . Pseudo code of the applied two-layer CV. Models with different component numbers *nPC* (*nLV*) were built and validated with an internal CV. Each replicate was taken out once and predicted within the external CV. For each iteration of the external CV, the model was built based on the overall training set with the *nPC* (*nLV*) featuring the highest averaged validation accuracy. (1) Within the Inside-CV, a dimension reduction method (PCA/PLS) was redone each iteration of the internal CV loop. Thus the PCA or PLS was executed after removing the validation set. The scores of the validation sets were predicted and then classified by the classification model (LDA or SVM). (2) For the Outside-CV, the dimension reduction method was carried out once for all data outside the internal CV loop. Therefore the validation set was involved in constructing the PCA/PLS model. Afterwards, the scores were split into training and validation sets for internal CV.

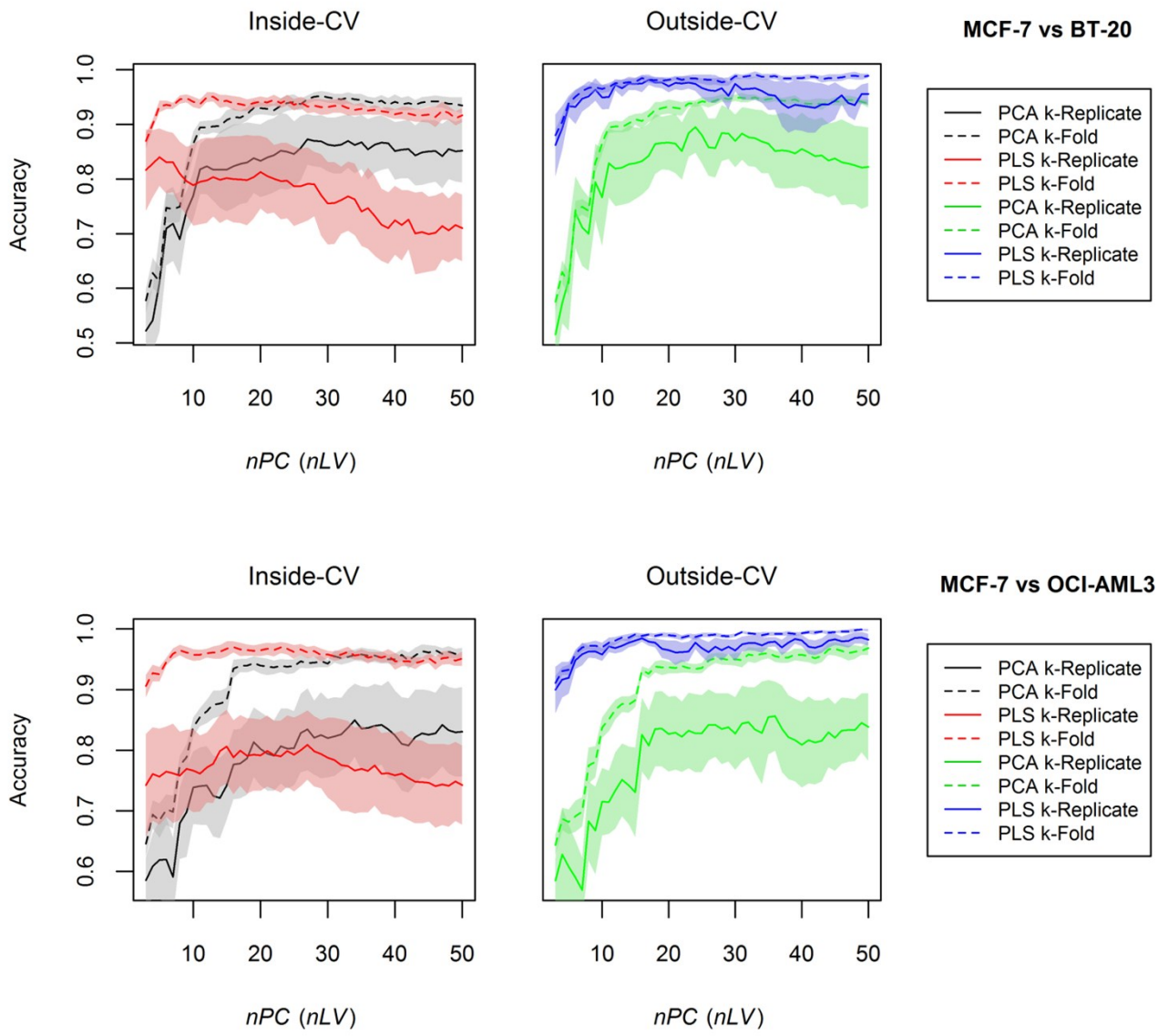

Fig. S3. Validation accuracies of the internal CV within the first iteration of the external CV, for two binary classification tasks (MCF-7 vs BT-20 (m vs b) and MCF-7 vs OCI-AML3 (m vs o)) based on SVM. In both tasks, the outside-CV yields higher accuracies. This is more obvious if supervised dimension reduction methods, such as PLS, are applied. The validation accuracies are always higher if a k-fold CV is used compared with a k-replicate CV. The over-estimation of the k-fold CV is due to the violated independence criteria between the training and validation sets. This effect of over-estimation is more enhanced for supervised dimension reduction.

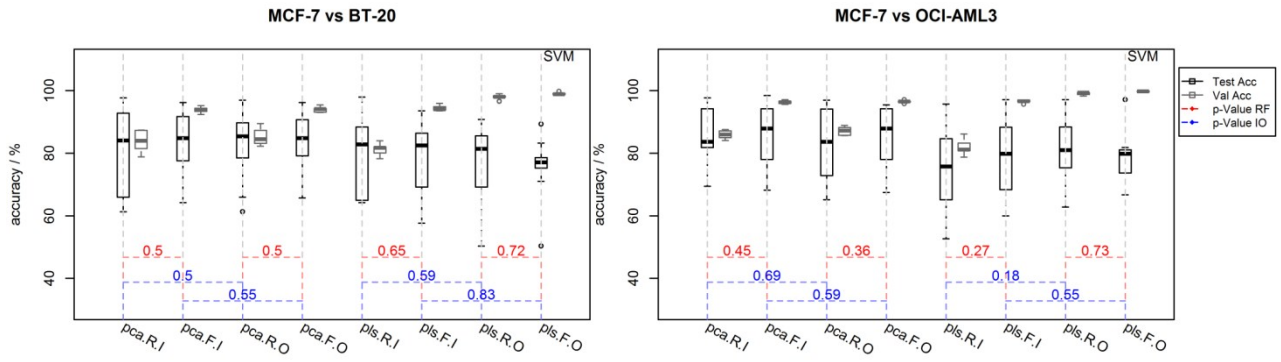

Fig. S4. Validation accuracies resulted from the internal CV and the independent testing accuracies from the external CV. Hereby the SVM was utilized for classification. The applied data splitting methods, dimension reduction methods and the position for the internal CV are referred to the x-axis labels, where 'R' and 'F' represent the k-replicate CV and k-fold CV, respectively; while 'I' and 'O' denote the inside-CV and outside-CV, respectively. The validation and testing accuracies were consistent for the k-replicate inside-CV, which means the model was evaluated reliably. On the contrary, the validation accuracies were significantly higher than the testing accuracies for k-fold CV and outside-CV. This demonstrated an over-estimation of the model. However, the over-estimation of k-fold-CV and outside-CV was ignorable if PCA was used for dimension reduction, demonstrated by the comparable validation and testing accuracies. In addition, in order to check the influence of the investigated two mistakes of CV with respective of model optimization, we compared the testing accuracies for k-fold CV against k-replicate CV (RF), and inside-CV against outside-CV (IO). The comparison was done by Wilcoxon-test. According to the p-values marked in the plot, no significant difference was observed. That means the investigated two mistakes were less influential if CV was used for model parameter optimization.