**Leuven, Belgium**

**April 19, 2021**

**Subject: In response to the revision of CHEMOLAB-D-21-00066**

**Cost-efficient unsupervised sample selection for multivariate calibration**

**Point-by-point responses to the two reviewers.**

The authors want to deeply thank the reviewers for their time to revise our manuscript highlighting important aspects of debate for the current topic. Below we explain and address the issues point by point and clarify the changes incorporated in the revised manuscript. In general, it was clarified that we are framing the PLSR model in the context of orthogonal scores (SIMPLS or NIPALS).

**Reviewer #1: I concentrated on the concepts, and less on the results section.**

First of all, the text is well written. It is easy to follow most of it. My main objections are on the conceptual side. The points below are all related but relate to various aspects of what I consider the main weaknesses of the paper.

Apart from being a possibly useful empirical study and comparison of alternative methods, I do not think the paper gives very much. I suggest focusing more on this aspect instead of giving the reader the impression that there is something really new. At least this latter aspect has to be described and justified better, before the paper can be published.

**Answer:** In regard to the novelty of our work, we were initially motivated to study the problem of sample selection to build calibration models as this problem is still being presented in the literature for particular applications which does not enable an easy extrapolation for a general guideline. There are already several methods as included in our work, but at the moment of deciding how to actually select the samples in a new application, there is still doubt on how to act about this problem. Therefore, when exploring the problem from a more theoretical point of view, we found that there is a possibility to find generalization in terms of the sample size and the evaluation on the representativeness of the selected samples. We believe this guideline can be highly useful for practitioners when developing new applications in multivariate calibration. We clarified this in **Section 1, paragraph 5.**

Comparing S matrices may sound intuitively appealing, but linking this to an orthogonality in the PLSR, seems very strange to me.

**Answer:** We fear that the reviewer got confused with respect tothe S matrices and the criterion to evaluate the representativeness of the selected samples.

We clarified this aspect in the manuscript more deeply in **Section 2.2.**

If we understand this comment well, the confusion came from the definition of S-orthogonality of the loadings in the PLSR (SIMPLS or NIPALS). This definition is nothing else than a name given to the fact that the loadings are orthogonal through the S matrix. The S matrix is the covariance matrix of X which is not to be confused with the covariance matrix between X and Y (s=cov(X,Y), S=cov(X)). While SIMPLS or NIPALS solve the PLSR model by maximizing s=cov(X,Y), this is not related to the criterion we propose which is that if we are able to find *n*  out of *N*  samples for which S*\_n = cov(Xn)*  and S*\_N = cov (XN)* are equivalent, then the *n* selected samples are representative for the total set of *N* samples to build a PLSR model.

I simply to not understand that. Maybe the S criterion can be argued for another way? Comparing S-matrices has a relation to the DI-PLS method that should be commented on.

**Answer:** We assume that this comment refers to the domain-invariant PLS algorithm presented in

#### **Domain-Invariant Partial-Least-Squares Regression**

Ramin Nikzad-Langerodi, Werner Zellinger, Edwin Lughofer, and Susanne Saminger-Platz

Analytical Chemistry 2018 *90* (11), 6693-6701

DOI: 10.1021/acs.analchem.8b00498

We understand why the reviewer commented on DI-PLS as the algorithm intends to align 2 matrices X. However, our aim is not to develop a new PLSR algorithm taking into account the S matrix, our aim is to explain how a set of representative samples can be selected to build a PLSR model, as explained in the previous point.

PLSR does not optimize the MSE in the calibration set. That is done by LS. It merely tries to find a solution which can also be used safely for future samples.

**Answer:**

We clarify the definition of the PLSR model and the SIMPLS or NIPALS algorithm in **Section 2.1 last paragraph and Section 2.2 first and second paragraph.**

To clarify the definition of PLSR under SIMPLS or NIPALS, these algorithms first construct a set of latent variables by optimizing the covariance between X and Y. Once the set of latent variables is calculated, a Least Squares step follows to fit a regression between the latent variables and Y. These two steps take place using the calibration set. The optimal number of latent variables, which in our notation is ***d,*** is to be chosen so that the model performs well in future samples, typically done with cross-validation.

The criterion (2) is therefore strange here since I suppose n is the number of calibrations samples. This is therefore not the criterion that is optimised by PLSR.

**Answer:** As commented above, our aim is not to develop a new PLSR algorithm and indeed the criterion (2) (i.e. S matrix comparison) is not the criterion optimized by PLSR. We are not claiming that the S=cov(X) is optimized in any way by PLSR. We intended to find a mathematical element inside PLSR which could be evaluated in an unsupervised way to define a criterion for sample selection, and this mathematical element is S=cov(X). We clarified this aspect in the manuscript more deeply in **Section 2.2.**

I also find it strange that a general function like (1) can be used a basis without assumptions. PLS can be formulated by a lower number comber of components by just adding some of them and this is in this framework a completely different solution than the full set of components. PCR can be written in the same way and many other methods as well. It seems that the S criterion comes out as special for PLSR, but I would suppose that this makes equally good sense also for other methods.

**Answer:** We believe that this comment also originated from the confusion on the S criterion. Therefore, we have clarified this criterion **in Section 2.2** which is a specific framework for PLSR.

A function like (1) is a general representation of the regression function as can be found in the literature by Vapnik. We particularly relied on Chapelle, O., Vapnik, V., & Bengio, Y. (2002). Model selection for small sample regression. *Machine Learning*, *48*(1–3), 9–23. https://doi.org/10.1023/A:1013943418833

The assumption of this representation is that the basis of variables is an orthonormal basis. In this regard, the definition of PLSR fits into this general representation. So does PCR and indeed other methods as long as they also rely on an orthonormal basis. The idea of carefully choosing an appropriate number of components in any of these models/algorithms/methods is to find a regression function that not only works on the training/calibration set but also on future samples and so the assumption is that such an optimal number of ***d*** components exists and fulfills this trade-off.

I simply do not believe that selecting a set of samples for a linear method in this context should vary very from method to method. A good set should be good for PLSR, PCR and for instance ridge regression.

**Answer:** We added a discussion about this very important point **in Section 6.4**

Indeed, when looking at the problem from a practical point of view, a representative set of samples should be good for these different methods. We aimed to look at the problem from a theoretical point of view to explain the usefulness of the current methods. From this theoretical point of view, the mechanics of PLSR, PCR, ridge regression and even variations like Lasso or Elastic Net, are all relying on S=cov(X). However, only for PCR an optimal design, like D-optimal, ensures to deliver an optimal set of samples as the PCR model is defined by the components of S=cov(X) and the sample selection problem is solved. This is not directly the case for the other methods. In PLSR, we have components derived from cov(X,y) and in ridge regression we have a regularization parameter accompanying S. Therefore, an unsupervised set of samples for which **S\_n** and **S\_N** are equivalent will indeed serve not only PLSR, but also ridge regression and the other methods.

Comparing S's only through eigenvalues is strange (if I understand this right). Two matrices can be principle have the same eigenvalue structure, but different eigenvectors.

**Answer:** We agree with this comment and defined the comparison through both, eigenvectors and eigenvalues. This can be found at the end of **Section 2.2,** in **Section 4.2.3**, and the results reported in **Figure 3.3**.

Whys not use a standard matrix comparison criterion, for instance the RV (or SMI) or something designed for symmetric matrices (if it exists).

**Answer:** RV or SMI are good metrics when we have the same samples measured in different variables, which does not correspond to the problem of sample selection. In the paper discussing RV,

Tomic, O., Forde, C., Delahunty, C., & Næs, T. (2013). Performance indices in descriptive sensory analysis - A complimentary screening tool for assessor and panel performance. *Food Quality and Preference*, *28*(1), 122–133. https://doi.org/10.1016/j.foodqual.2012.06.012

it is suggested to transpose the matrices when we have the same variables and different samples. However this use of RV is not widely discussed to our knowledge and some experiments in the context of calibration transfer did not provide us with a satisfactory insight of using RV in this way. SMI, on the other hand, can only be used for the same samples and different variables. Therefore, we did not find these metrics suitable for this problem.

It is important to be aware of the fact that not only the number of samples is important. Also the way they are selected is important. Space filling designs and clustering are different from optimal designs (for instance D-optimal) in a classical sense. robustness. A model is never perfectly linear and this should be taken into account when selecting samples. Evenly spaced designs are examples of more robust strategies for situations with non-linear tendencies. The first bullet point under Section 3 is not possible to achieve within this framework as far as I can see.

**Answer:** Indeed, the way the samples are selected is crucial and this is what we define by the S comparison criterion. Because there are already several well defined sample selection methods or designs, in order to bring a structured approach and give a guidance on which of all these possible sets to choose, we found that the evaluation of the quality of a selected set through the S comparison brings a final decision whether the sample set satisfies the unsupervised information or not. The S comparison works under the assumption that the relationship between X and Y can be captured by bilinear models, be it rather linear or with nonlinearities that can still be modelled by these model architectures. Therefore, we believe that the first bullet point under Section 3 can be answered under this model assumption.

We commented in **Section 8** on the conclusions on the need for further research for well known and strong nonlinear relationships to address this problem. Nonetheless, if a very strong nonlinear relationship indeed exists, it can still be revealed by a first solid unsupervised sample selection as proposed by our analysis.

Related to previous: I am not sure I buy the idea that it is generally possible to find general statements about number of samples other than proposing useful strategies. Everything will depend on degree of linearity, noise level, the representativity of calibration samples, dimensionality of X etc.

**Answer:** We aimed to provide a general strategy for sample selection and a generalized value for the sample size, under the assumption that the relationship between X and Y can be very well captured by these bilinear models. We added this discussion in **Section 6.4.**

The value of the model complexity ***d*** is the one in charge of controlling the degree of linearity, noise level and dimensionality of X. The representativity of the calibration samples is indeed what we aim to tackle with the S comparison criterion.

More explanations or a more solid foundation must be provided before the paper can be published. Another option is to turn it into an empirical comparison without any pretentions of providing something really new.

**Answer:** Our purpose was to find a theoretical framework that could define how to address the problem of sample selection for this type of bilinear models. We clarified this in **Section 1.**

The most important aspect was to find a common ground for this approach that has been described in many empirical studies. We believe this common framework can help researchers and industry users react to this problem in a more concrete way in new applications.

**Reviewer #2: CHEMOLAB-D-21-00066**

This is a really interesting paper. It is publishable as is, but below are a few comments the authors might want to consider.

p3, discussion of the role of d. I think it is a mistake (of Vapnik's not yours) to try to find completely general rules for d. In the context of this application, I would expect the dependence of performance on n/d to be different for PCR and PLSR for example, because PLSR uses y to construct the phi.

**Answer:** Depending on the degree of linearity between X and Y and their corresponding variance, PLSR typically needs less components (smaller d) than PCR. If this is the case, PCR would need more samples to achieve the same performance as PLSR.

In terms of looking at the problem from the point of view of the ratio ***n/d*** , we studied the perspective by Vapnik and found it valid to think of the sample size relative to the complexity of the model. While in his theory it is stated that n/d>20 accounts for a large sample size, we found motivation to put this in the perspective of multivariate calibration for NIR applications

To take an extreme case, if I use MLR with a generalised inverse on the full X to construct a phi, I can have d=1 but still have too small a sample for any n.

**Answer:** We would like to ask for more clarification on this statement. We believe that if a generalized inverse of the full X is used, the value of ***d*** would be the full rank of X. On the other hand, by the proposed generalized inverse to construct phi, we remain in doubt whether the basis is still an orthonormal basis.

p3, section 2.2. There are (alas) lots of PLSR algorithms and some have orthogonal loadings rather than orthogonal scores. You say later that you are using SIMPLS, but that isn't clear when you refer to "the" PLSR algorithm.

**Answer:** This was clarified starting from **Section 2.1.**

We are indeed within the context of orthogonal scores to be able to look at the problem from the generalized framework.

Can you say a bit more about what went wrong when you used pre-processing? My instinct would be that if I know in advance I am probably going to use second derivative spectra then I should select my samples after that pre-treatment. It will certainly change all the distances between spectra and thus result in a different selection. Do you have any insight into why my instinct is wrong? Maybe the key point is that it is important to select spectra that adequately represent all aspects of the variability in the samples, but it isn't clear to me why!

**Answer:** A more detailed discussion was added i**n Section 6.3.**

As we were exploring this problem in a general view, when trying derivatives or MSC in different datasets previous to the sample selection, in several cases we obtained deteriorated performance on the test set. Schoot et al. (2020) commented also on the effect of preprocessing for these models. In general, if a calibration model is to be built with new and not well known instruments, it may be a risk to assume that we need to filter out certain features of the signals for the prediction of the Y variables prior to this information. We find it suitable in general to recommend selecting the samples with raw spectra, and further maintenance of the calibration model could be done with the suitable preprocessing. Nonetheless, if the instrument and case are well known and there is sufficient information to motivate preprocessing the data, this is an option for the user, but we don’t have evidence of this to report in the present study.