Cost-effective model building in multivariate calibration

Valeria Fonseca Diaz,[*,†] Bart De Ketelaere,[*,†] Ben Aernouts,[*,†,‡] and Wouter Saeys[*,†]

†*KU Leuven, Kasteelpark Arenberg 30, Leuven, Belgium*

‡*Biosystems TC, KU Leuven, Kleinhoefstraat 4, Geel, Belgium*

E-mail: valeria.fonsecadiaz@kuleuven.be; bart.deketelaere@kuleuven.be; ben.aernouts@gmail.com; wouter.saeys@kuleuven.be

Abstract

Multivariate calibration models conceived as virtual sensors that are used to measure chemical compositions in products are built based on spectral data of samples and their corresponding reference chemical values. The cost of these reference analyses is a major ~~of~~ feature of interest to be minimized in industrial applications for the sake of more efficient analytical processes. The present work aims at characterizing the problem of sample selection based on spectral measurements to build calibration models. We mainly focused on evaluating optimal sample sizes, evaluation of different selection methods and we give recommendations on how to assess the suitability of a set of samples to build bilinear calibration models.

Introduction

Multivariate calibration models have been the primary analytical tool to indirectly measure the chemical composition of products making processes such as those of quality control more cost-efficient. Many ~~are still the~~ challenges ~~that~~ are present when building and maintaining these models. The present work presents an exhaustive evaluation of the ~~problem~~ of unsupervised sample selection to build successful calibration models. Within the range of applications of multivariate calibration, the stated problem is particularly relevant for indirect inspection of collected primary resources using near-infrared (NIR) spectroscopy, as it is the case of the agrofood industry (AFI).[1–4]

The ~~problem~~ of unsupervised sample selection ~~consists in~~ determining what the best methodology or strategy is to select the samples that would be worthy of reference analysis (chemical composition values) only based on spectral measurements to build calibration models. As it is widely known, spectral measurements are easy and cheap to collect, therefore a vast quantity of units can be submitted to these measurements with low effort. On the contrary, collecting the reference analyses is a task that requires a ~~lot of~~ effort, high costs and possibly high waste. This has been the motivation to pay attention to the optimization of model building costs while gaining model building effectiveness. Historically, the interpretation for this gain has been rephrased as the optimal spanning of variability with a minimum number of samples.[5–7]

Because of the importance of this problem, ~~up to now there is still~~ research about features such as optimal sample sizes and suitable strategies to select these samples~~.~~[1,8] Through the last decades, many methods have been proposed for sample selection within a finite set of units, some of them becoming widely accepted for their intuitively accurate approach and good performance.[9–11] The two most popular methods in the context of NIR spectroscopy and multivariate calibration are the Kennard-Stone[7] and Puchwein[12] algorithms. As a strategy, clustering techniques are widely accepted in order to account for multiple sources of variability,[5] particularly, when groups of samples exist in batches of collected products.[4] These popular algorithms have the common feature of relying in distance between the samples. Less popular in NIR applications but highly effective in product design are the optimal design of experiments, which translate the concept of variability from distance between samples to the variance of the coefficients of an assumed model.[13] Yet, no clear understanding exists about which of all these methods are more suitable for NIR applications with multivariate calibration and no exhaustive competition of them has been disseminated to the authors' knowledge. The most recent study found ~~attempting to study~~ this problem ~~made use of~~ only the two most popular methods mentioned before.[1]

Multivariate calibration is a concept that in principle can involve any type of statistical model. Moreover, calibrations are made for classification as well as regression tasks. While the present work focuses on regression models, the scheme here presented can serve as guidance for classification models. Regarding the model architecture, as of now, there is little doubt on the effectiveness of the type of bilinear models that are widely used for NIR applications. Unless the spectral values have a strong nonlinear relationship with chemical reference values, bilinear models such as partial least squares regression (PLSR) or principal component regression (PCR) remain the workhorse model architectures. Most of the work that can be found addressing the current problem of interest provides answers about the minimum required sample size using PLSR models.[1,5,9,14] The diversity on the answers is still large leaving researchers ~~still~~ with a general answer out of the scope.

The present work aims to provide a more general answer on how to approach the selection of calibration samples by understanding the properties of PLSR models. The general framework of statistical learning theory by Vapnik provides specific answers on the sample size needed depending on the model architecture to be used.[15,16] To deepen into this question for NIR applications, it is necessary to understand PLSR within the framework explained by Vapnik. In a similar way, to understand the required features to take into account for a successful PLSR model, it is essential to seek for the elements of the PLSR architecture that can be controlled with unsupervised measurements. These aspects are to be described and explained for the state-of-the-art methods to select calibration samples.

A general and a specific framework of PLSR is presented in order to set a context of analysis to solve the problem of the best strategy for unsupervised sample selection. The work is organized as follows. First, a description of the general and specific frameworks is presented prior to the research questions of the current work. Afterwards, the experimental work and results are presented along with the discussion and conclusions about the aspects that lead to a more successful unsupervised sample selection to build calibration models.

Frameworks to understand PLSR

The general framework

A general framework to understand PLSR models takes place in the context of statistical learning theory. In the general regression task, the response variable is assumed to be a linear combination of basis functions plus an error variable. With that model architecture and using a square loss, the objective to estimate the model is to minimize the square error.[15] More concretely, let $y \in \mathbb{R}$ be the random variable representing the chemical constituent of interest and $\mathbf{x} \in \mathbb{R}^p$ the predictor vector of spectral measurements. The regression model is defined as $y = f(\mathbf{x}, \boldsymbol{\beta}) + \epsilon$. The general regression problem with square error can be established as:[16]

$$\min E\left[(y - f(\mathbf{x}, \boldsymbol{\beta}))^2\right]; \quad f(\mathbf{x}, \boldsymbol{\beta}) = \sum_{k=1}^{\infty} \beta_k \phi_k(\mathbf{x}) \tag{1}$$

where $\{\phi_i(\mathbf{x})\}$ constitutes a basis of L_2 for which its elements can be ordered by some criterion. In the context of statistical learning theory by Vapnik, $E\left[(y - f(\mathbf{x}, \boldsymbol{\beta}))^2\right]$ is called the *expected risk*, which in practice is replaced by the so-called *empirical risk* in the presence of a set of n observations to estimate function f.[16] This empirical risk is what has been known for centuries as the sum of square errors. The important feature of this optimization task is that to ensure a small *expected risk*, the *empirical risk* is to be minimized only over a limited number of basis functions $\{\phi_k(\mathbf{x})\}_{k=1}^{d}$. Therefore, the regression problem becomes:

$$\min \frac{1}{n}\sum_{i=1}^{n}(y_i - f_d(\mathbf{x}_i, \boldsymbol{\beta}))^2; \quad f_d(\mathbf{x}, \boldsymbol{\beta}) = \sum_{k=1}^{d} \beta_k \phi_k(\mathbf{x}) \tag{2}$$

It is in this regard that the problem as stated in eq. (2) corresponds to the definition of the PLSR model,[17] where the basis $\{\phi_k(\mathbf{x})\}$ is constructed by means of maximizing the covariance between y and $\mathbf{x}$. The order in this basis is established by the covariance deflation at each step of the PLSR algorithm. The key element in this framework is the fact that the number of chosen basis functions d corresponds to the so-called VC dimension in Vapnik's theory, which measures the capacity control of a learning machine.[15] The importance of the capacity control parameter d is that it serves as the reference for determining a suitable sample size when aiming to build a regression machine, or, as known in the context of chemometrics, a multivariate calibration model. In the work by Vapnik, it has been stated that the ratio between the sample size and the VC dimension determines whether the sample size is large or small. Although there is no absolute threshold, it is stated that the sample size is *large* when $n/d > 20$.[16]

The specific framework

In the jargon of multivariate calibration, the basis of L_2 is what is known as the set of latent variables. Based on a set of n observations stored in the matrices $\mathbf{X}_{n\times p}$ and $Y_{n\times 1}$, the underlying idea of the PLS algorithm is to calculate latent variables $\{\phi_k(\mathbf{x})\}$ such that $\phi_k(\mathbf{x}) = \mathbf{X}\mathbf{v}_k$, where $\mathbf{v}_k$ results from maximizing the covariance between $\mathbf{X}$ and Y at the k-th deflation step.[18]

For a given value of d, the set of resulting latent variables constitute a set of orthonormal variables $\{\phi_k(\mathbf{x})\}_{i=1}^{d}$. However, the set of loading vectors $\{\mathbf{v}_k\}$ is regarded as $\mathbf{S}$-orthogonal, i.e. $\mathbf{v}_k'\mathbf{S}\mathbf{v}_j = 0 \quad (j < k)$, where $\mathbf{S}$ is the covariance matrix of $\mathbf{X}$. The $\mathbf{S}$-orthogonality property of the PLSR algorithm clarifies that the estimation of this type of regression model depends highly on the covariance matrix $\mathbf{S}$. This property suggests that if $n < N$ samples are found such that $\mathbf{S}_n$ and $\mathbf{S}_N$ are equivalent, there is no pure unsupervised information discarded that serves the PLSR model in the $N - n$ remaining samples.

There are several methods and indexes to study the equivalence between two matrices.[19] For the sake of bilinear regression models such as PLSR, it becomes manifest to evaluate this equivalence via the spectral value decomposition (SVD) of the matrices due to the rank deficiency of $\mathbf{S}$ in NIR applications. The reason is two-fold. On the one hand, matrix congruence constitutes an equivalence relationship and it holds true when two matrices have the same eigenvectors.[20] On the other hand, as the SVD of $\mathbf{S}$ constitutes the theory of principal component analysis (PCA), the eigenvalues of $\mathbf{S}$ account for the variability that the different dimensions contain and their relative predictive power for a regression model.[21] The latter reference is indeed the depiction of PCR models. Notwithstanding, the present work focuses on the role of matrix $\mathbf{S}$ for PLSR models.

Research questions

The present work aims at presenting a scheme to approach the problem of unsupervised sample selection for PLSR models models. This scheme is defined in terms of three factors, namely, selection methods, sample sizes and the dimensionality of the matrix $\mathbf{S}$ to obtain satisfactory performance of calibration models. The objective is to provide the following answers:

- Can particular thresholds be found regarding the optimal sample size for satisfactory PLSR models in chemometrics based on the ratio n/d?
- What is the equivalence achieved between $\mathbf{S}_N$ and $\mathbf{S}_n$ by selection methods, input dimensionality and sample size?
- What are the most optimal conditions of the three factors for satisfactory PLSR models?

Experimental

Case studies

Two AFI real case studies were taken to demonstrate the aspects previously discussed about unsupervised sample selection. The first case corresponds to the inspection of milk composition where data of two periods of time were available both for spectral signals and reference analysis.[2] During the first period, 316 samples were collected followed by 79 new samples collected in the second period. The time frame between the periods was two weeks. The spectral measurements correspond to transmittance mode gathered in the range 900 nm - 1700 nm with a resolution of 3 nm. In this case the chemical variable of interest was lactose. The second case corresponds to pig manure samples for the inspection of the manure composition.[3] One set of 420 samples was measured for calibration and a separate set of 164 samples was measured for validation. The chemical constituent chosen for inspection through calibration was Dry Matter (DM). The spectral measurements were taken in reflectance mode in the range 426 nm - 1686 nm with a resolution of 9 nm. For the purpose of analyzing the sample selection problem, the first sets in both cases are referred to as selection set. The samples chosen by the different methods constitute the calibration set and the samples on the second sets are taken as the test set. The descriptive statistics of the data sets are shown in Table 1.

Preprocessing: Mean centering preprocessing was used in all cases. Initial experiments were carried out to decide whether to preprocess the data with other specific filters, but unsatisfactory results on the selected calibration samples were obtained. It was seen that assuming certain preprocessing filters for the spectral measurements prior to any knowledge of the y values may be harmful to build calibration models.

Table 1: Descriptive statistics

Case study	set	size	mean(y)	std(y)
Milk (y: lactose (%))	selection	316	4.7371	0.1547
	test	79	4.7044	0.1554
Manure (y: DM (gl^{-1}))	selection	420	66.0224	34.7173
	test	164	64.2887	38.5147

Methodology

Exhaustive evaluation for unsupervised sample selection

An exhaustive evaluation was set up ~~consisting in~~ combining three main factors involved into the problem of interest: Method, input dimensionality and sample size. This allowed to evaluate the impact of each of these factors on model performance by selecting subsets of samples for each possible combination of these factors. The definition of each factor is explained below and their possible values are listed in Table 2.

Selection methods

There are multiple methods in the state of the art for sample selection based on the available matrix $\mathbf{X}$. After revising the literature on chemometrics and calibration models, the methods selected for the present work were: Kennard Stone (KS),[7] Duplex (DUP),[22] Puchwein (PUCH),[12] complete linkage hierarchical clustering (CL)[5] and D-optimal designs based on the Federov algorithm (D-OPT).[13] In addition, random selection (RAND) was included.

Input dimensionality

The available spectral measurements stored in matrix $\mathbf{X}_{N\times p}$ ~~constitute an~~ input dimensionality ~~equals~~ p. However, due to the rank deficiency of $\mathbf{X}$, which is equivalent to the rank deficiency of $\mathbf{S}$, the input dimensionality was taken as the second factor involved into sample selection. For an input dimensionality a, the samples were selected using the PCA scores $\mathbf{T}_{N\times a} = \mathbf{X}_{N\times p}\mathbf{R}_{p\times a}$ where $\mathbf{R}_{p\times a}$ contains the first a eigenvectors of $\mathbf{S}$. Based on the evaluation of the current cases and results seen in the literature of chemometrics, the range of a was set from 1 to 25 in addition to $a = p$. The value of a should not be confused with

Table 2: Sample selection settings

Selection method	Input dimensionality	Sample size
KS	1 PC	30
DUP	.	40
PUCH		.
CL	.	.
D-OPT	25 PC's	.
RS	p	N

the VC dimension for the PLSR model d.

Sample size

Based on the maximum number of principal components ($a = 25$), the minimum sample size was set to 30 samples in order to leave a small margin. The sample size range was considered in steps of 10 from 30 to the maximum number of samples available in the selection set for each case study.

Constraints

For the distance-based selection methods (KS,DUP,PUCH,CL), a mahalanobis distance was used for $a = 1, ..., 25$ and a euclidean distance measure was used for $a = p$. This was motivated due to instability of the inverse of the covariance matrix $\mathbf{S}$ for $a = p$ that is used in the mahalanobis distance. For the same reason, the selection based on D-OPT could be applied only for $a = 1, ..., 25$.

Equivalence analysis of covariance matrix S

The equivalence or congruence between the covariance matrices $\mathbf{S}_N$ and $\mathbf{S}_n$ was evaluated through the correspondence of their eigenvectors and the eigenvalues. Given a value of $a << p$, the spectral value decomposition of $\mathbf{S}_N = \mathbf{V}_N \mathbf{\Delta}_N \mathbf{V}'_N$ and $\mathbf{S}_n = \mathbf{V}_n \mathbf{\Delta}_n \mathbf{V}'_n$ was calculated. The a eigenvalues were compared by calculating the ratio $\mathbf{\Delta}_n / \mathbf{\Delta}_N$ and the a eigenvectors were compared by computing the absolute value of the determinant for the matrix $\mathbf{V}'_n \mathbf{V}_N$. As the set of eigenvectors constitutes also an orthonormal basis, this determinant takes an absolute value between 0 and 1. In addition, Pearson correlations between y and the PC's were calculated to show the impact of the value of a.

Multivariate calibration models

The PLSR models were trained using the SIMPLS algorithm.[18] For each selection setting as explained in Table 2, a PLSR model was calibrated with the selected samples and applied on the test set for each case study. The performance of the model corresponded to the root mean squared error on 10-fold crossvalidation (RMSECV) and the test set (RMSEP) in a range for d from 1 to 30 latent variables.

Computational tools

The complete analysis was programmed in Python 3.7. The selection methods, excluding D-OPT, were programmed in house as well as the SIMPLS algorithm. The D-OPT method was used in a connection from R to Python, using the function `optfederov` from the R-package *AlgDesign*.[23] The exhaustive sample selection and subsequent PLSR calibrations was embedded into a loop which was highly optimized using `numba` for Python.[24] It was possible to fit more than 5000 PLSR models including crossvalidation results in 10 minutes ~~in~~ a~~n~~ 64-bit Inter Core i7 vPro 8th generation with 16 GB of RAM.

Results

General framework

Can particular thresholds be found regarding the optimal sample size for satisfactory PLSR models in chemometrics based on the ratio n/d?

The results on model performance from all the calibration models in the exhaustive evaluation were put together in order to find thresholds for the optimal sample size regardless of the selection method or the input dimensionality. In total, for the Milk case there were 5360 models and for the Manure case there were 7210 models. With the purpose of showing the effect of the ratio n/d on model performance, an *optimal* complexity d for the corresponding chemical constituent needed to be established. Such an optimal complexity was chosen based on the performance in crossvalidation (RMSECV) and prediction error (RMSEP) using the total number of samples available in the selection set. Figure 1(a) shows the RMSE curves for crossvalidation and test where it can be seen that for lactose the optimal complexity happens to be between 16 and 18.[2] Therefore, to analyze the ratio n/d using the lactose prediction, d was fixed as 16,17 and 18 latent variables. In the case of Manure, the same choices were made based on the RMSE curves shown in Figure 2(a). Taking the separation between RMSECV and RMSEP and the increasing point for each curve, the optimal complexity for DM was set as $d = 11, 12, 13$ and 14.[3]

Figures 1(b) and 2(b) show the RMSEP values as a function of the ratio n/d for the chosen optimal complexities in each case study. To answer the current question, it was observed that there is a clear stabilization of the prediction error as a function of the ratio of interest, regardless or the other factors for sample selection. Before a ratio of 10, there is clear uncertainty in the prediction performance of the models, where therefore the sample size was small compared to the optimal complexity of the calibration model. However, in the prediction of lactose, a stabilization threshold could be obtained as $n/d > 12$ obtaining similar performance as reported in previous works.[2,25] Because of the available number of samples (N=316), for the chosen optimal complexity, the maximum ratio could not surpass 20. Interestingly, a similar result was observed in the prediction of DM. A jump was observed after a ratio of 12 and then a final stabilization point was detected as $n/d > 16$ reaching comparable performance as obtained in the reference work, provided that the data was only mean-centered in the current work.[3] In both of the cases, the highest uncertainty for the performance of the calibration model was obtained in $n/d < 10$. For these specific cases, this insight suggested that if the model complexity for lactose would not be less than 16 latent variables, a strict minimum sample size should be $16 * 10 = 160$. For the case of DM, a strict minimum sample size would be $11 * 10 = 110$.

Specific framework

What is the equivalence achieved between $\mathbf{S}_N$ and $\mathbf{S}_n$ by selection methods, input dimensionality and sample size?

Due to the vast quantity of results, the degree of equivalence between $\mathbf{S}_N$ and $\mathbf{S}_n$ is presented here for the different selection methods, sample sizes and a few values of the input dimensionality a. The impact of the entire range of a for model performance is presented in the next section. Figure 3 shows the comparison of the eigenvectors of $\mathbf{S}_N$ and $\mathbf{S}_n$ for both of the case studies when the sample selection was made with $a =$ 15, 20 and 25 with each method.

Across the different methods, the behavior of the determinant at a fixed rank a was noticeably related to the sample size, yet differently for the different selection methods. All the methods in both case studies showed similar results for $a = 15$ and 20. Determinants above 0.8 were achieved with sample sizes below 30% of N. When setting $a = 25$ large differences were observed by the selected calibration sets across the different methods and sample sizes. In both case studies, CL and D-OPT selection showed a jump in the determinant around the same sample sizes. In the Manure case, KS related to CL and D-OPT in this behavior while it was PUCH in the Milk case. As expected, due to the nature of selection with DUP, the determinant presents a declination after crossing 50% of the samples for selection. However, this decline was obtained for $a = 20, 25$ and not for $a = 15$ making evident the effect of the input dimensionality. RAND selection did not achieve a clear stabilization of the determinant compared to that of the other selection methods.

For KS, PUCH, CL and D-OPT, the stabilization of the determinant for $a = 25$ at a value of n was in line with the threshold found with the ratio $n/d > 12$. The comparison of the individual eigenvalues is therefore presented for this input dimensionality for both case studies in Figures 4 and 5. Each curve represents the eigenvalues ratio for a given value of n. In this comparison, the 50% effect of DUP selection was detected as the ratio of the eigenvalues went under 1 after such a sample size n. In both case studies, KS, D-

OPT and PUCH selection showed that there is generally no underestimation of variability as the eigenvalues ratio stayed above 1 with very few exceptions. No systematic behavior in the variability by eigenvalues was revealed by RAND selection. For all the other methods, the selection with small sample sizes clearly showed no uniform spanning of the variability occurred given that some dimensions resulted in eigenvalues ratios twice as large as others. Most interestingly, high peaks and drops happened at the same dimensions indistinctly of the selection method.

As CL is a more robust method compared to the other methods in terms of control over possible outliers, the eigenvalues ratio is overall more stable and closer to 1 in comparison with the other methods. In the Milk case, for very small sample sizes, some dimensions resulted in ratios between 0.5 and 1. D-OPT showed more linearity for the eigenvalues ratio as a function of the sample size compared to the other methods in both case studies.

The possible risk of the non-uniform spanning of variability across the different dimensions was supported by the correlations between PC dimensions and the chemical constituent. Table 3 shows the Pearson correlations for both case studies. In the case of the Milk dataset, PC's 18 and 21 showed higher correlations with lactose than even PC's 1 and 2. In the Manure case, DM had a higher correlation with PC 5 than PC 1, but the first 10 PC's accounted for the highest correlations. The impact of diminishing or underestimating the variability at certain dimensions is to be presented with respect to the PLSR model performance. In regard to the equivalence between $\mathbf{S}_N$ and $\mathbf{S}_n$, the degree of equivalence depended on the rank value of a and a more uniform span of variability according to the eigenvalues could be obtained in both case studies with sample sizes according to $n/d > 12$.

Model performance

What are the most optimal conditions of the three factors for satisfactory PLSR models?

The simultaneous effect of the three selection factors on the performance of PLSR models was analyzed by comparing the RMSEP curves in each case study. The grid in Figures 6 and

Table 3: Pearson correlations between PC's and chemical component based on the selection set

Milk (lactose)		Manure (DM)	
pc	correlation	pc	correlation
1	-0.1337	1	-0.3132
2	-0.1839	2	-0.6231
3	-0.0239	3	-0.1587
4	0.2380	4	0.2220
5	0.0496	5	0.4036
16	0.0060	6	0.0299
17	0.0266	7	-0.3102
18	0.4208	8	0.0521
19	0.1202	9	0.1543
20	0.0759	10	0.0259
21	0.3819		

7 shows the PLSR model performance for the Milk and the Manure case study, respectively. The selection methods are accommodated row-wise, punctual sample sizes are positioned column-wise and the color of the curves stands for the input dimensionality. The main insight when comparing both case studies in terms of the impact of the sample selection factors is that the stabilization or convergence of model performance occurred at the thresholds found for n at the optimal complexity d.

In both case studies, RAND selection resulted in higher performance uncertainty as the sample size decreased. For higher sample sizes surpassing the thresholds found in the general framework, RAND selection rendered calibration sets that were equally optimal as the sets delivered by any other selection method. In the prediction of lactose, the model performance for the selected sets of size $n = 60$ was highly uncertain and no clear relation was found between a satisfactory performance and the selection method or the input dimensionality. Nonetheless, for such a small n, the selected set by D-OPT with a large number of a dimensions produced consistently models with better performance. The same result was detected for D-OPT for other small sample sizes ($n = 90$ and $n = 120$). This differentiation of the effect of the input dimensionality was not equally clear for small sample sizes with the other methods. Moreover, when selecting samples based on the original $\mathbf{X}$ matrix (i.e.

$a = p$), there was no satisfactory result for the small sample sizes compared to values of a between 15 and 25. When surpassing the threshold and selecting more than 160 samples, the performance of the models with the different methods and input dimensionality stabilized and no important or systematic difference was observed.

The model performance for the Manure case suggested similar conclusions on the effect of the factors as in the Milk case study. In particular, for a small sample size as $n = 60$, D-OPT selection showed a more stable performance around $d = 11$ for large values of a than the performance stability by the other methods. This stabilization by D-OPT was observed more clearly for $n = 90$ from 8 to 16 latent variables. Once the threshold $n/d > 12$ was surpassed, which happened between 130 and 140 samples, all the selection methods except DUP with a between 20 and 25 rendered calibration sets with satisfactory performance. RAND selection also proved here to render similar results as other selection methods as the sample size increased.

Discussion

When positioning the PLSR algorithm in the general framework of statistical learning theory, it was found that indeed, the sample size to be considered for multivariate calibration models cannot be thought outside the VC dimension (i.e. model complexity). This means that a sample size n will be differently optimal for an easy-to-predict chemical constituent than for another one that is harder to predict. This was revealed by the analysis of the ratio n/d and the optimal sample sizes obtained in each of the case studies. The level of easiness to predict is what the VC dimension or model complexity represents. It was confirmed that a ratio $n/d > 20$ accounts for a sample size that is rather large for a satisfactory calibration model. Based on the results of the present study, for which a hard-to-predict component (lactose) and an easier component (DM) were analyzed, evidence was obtained that a satisfactory calibration model can be built with $n/d > 12$. Although in practice the problem of unsupervised sample selection is encountered when there is no data available for the target variable y, it is possible to make an estimation based on the literature and expertise of the problem at hand.

In Au (2020),[1] thresholds for the optimal sample size were found in absolute numbers. However, as the sample size steps taken were large, the detected thresholds corresponded to sample sizes for which the ratio $n/d > 20$, as the optimal model complexity shown there was about 7 to 10 latent variables. Those results still confirm the theory based on the n/d given by Vapnik.[16]

The leading feature to address the current problem is that of spanning as much as possible the variability contained in the available selection set so that a representative subset of samples is obtained. In the context of chemometrics, this feature has not been concretely translated into a mathematical criterion, which at the same time is to be defined based on the model architecture that best describes the relationship between $\mathbf{X}$ and y. The specific framework analysis shows the role that the matrix $\mathbf{S}$ plays in the PLSR model, making it the mathematical element that can be controlled in an unsupervised setting. Therefore, finding a subset of n samples that renders $\mathbf{S}_n$ equivalent to $\mathbf{S}_N$ is a proposed definition about the representativity of the selected samples.

The results about the comparison of eigenvectors and eigenvalues shows that, provided the input dimensionality of $\mathbf{S}$, there is convergence in the equivalence of these matrices. In practice, if the number of samples can be decided based on ratio n/d, the rank of $\mathbf{S}$ and the samples that render an equivalent $\mathbf{S}_n$ can be detected by evaluating the determinant and eigenvalues ratio criteria. Based on the model performance results when selecting samples with a low input dimensionality, it was detected that keeping dimensions of low variability out of the sample selection strategy may greatly compromise the performance of the model after gathering reference analyses. This insight was supported by the model performance achieved with high values of a at least for KS, CL, PUCH and D-OPT and the high correlations between PC dimensions of low variability and y in both case studies. In the works that are found in the literature of chemometrics applications, the unsupervised sample selection has been applied with different methods and a reduced input dimensionality, but no analysis is generally found on the effect of the latter factor.[1,5,10,26] Complementary, it is not conclusive that selecting the samples with the original dimensionality of $\mathbf{X}$ was particularly advantageous than reducing it to $a << p$. This is particularly relevant because D-OPT does not support the original matrix $\mathbf{X}$ when its rank is deficient.

The effect of the selection method is largely diminished once the sample size and the dimensionality a are controlled. Nonetheless, each of the selection methods operates under their own criterion to span the variability. DUP is the least suitable method for unsupervised sample selection as it is concretely ~~design~~ to separate the set in exactly 50% parts. It functions by finding two replicate submatrices in the matrix $\mathbf{X}$, which might not be the ideal separation according to the threshold for the optimal sample size as confirmed by the presented results. D-OPT selection, on the other hand, proved to deliver an optimal selection consistently for high values of a. This method, however, is characterized by overestimating the variability, representing a possible challenge in presence of bad outliers or an advantage for good leverage points. CL demonstrates to be the most robust strategy for sample selection. Together with KS and PUCH selection, representative calibration sets can be delivered by these methods, but the results on model performance showed that this aspect is highly dependent on punctual values of the dimensionality a, rather than ranges of it. A suitable criterion would then be one that, provided a rank a, delivers a set for which $\mathbf{S}_n$ and $\mathbf{S}_N$ are equivalent in terms of their eigenvectors and eigenvalues, subject to a ratio of the eigenvalues above 1, which can be evaluated with any strategy for sample selection.

An important feature that is of question when using D-OPT selection is the so-called *effects* to include, concretely, the polynomial degree of the input variables.[13] When including only the main effects, D-OPT unavoidably selects the outer layer of the data dispersion in the space. For particular purposes of the theory and applications of Experimental Design, interactions and higher order effects are commonly considered for sample selection.[26] However, in multivariate calibration and unsupervised sample selection, the absence of prior information does not justify the inclusion of other types of effects based on the model architecture of the PLSR model. Including high order effects is a practical strategy to select more central points, but there is no theoretical support for it in the context of the present work.

A final additional aspect comes about preprocessing. Several authors still consider specific preprocessing techniques to filter out certain spectral attributes from the signals. For sample selection, previous work was found commenting on the possible advantages of preprocessing the spectral data for unsupervised sample selection.[8] However, initial experiments in the current case studies suggested that assuming advanced preprocessing such as scattering correction resulted in sets rendering calibration models with poor performance. Therefore, strong assumptions on preprocessing prior to obtaining reference analysis is generally not advisable.

Conclusions

The exhaustive comparison of the factors involved into unsupervised sample selection allowed to define a scheme to approach this problem in NIR applications and multivariate calibration. Optimal sample sizes can be calculated based on the threshold $n/d > 12$. A maximum value for the PLSR model complexity d can be established based on literature. Once n is calculated, a value for the input dimensionality a can be found by evaluating at what value of a there is convergence in equivalence between $\mathbf{S}_N$ and $\mathbf{S}_n$. This ensures that the dimensions included in the selection account for PC's of small variability that can be beneficial for the PLSR model if a hard-to-predict constituent is involved. KS, PUCH, CL and D-OPT all proved to deliver calibration sets for satisfactory models. D-OPT proved to have better consistency when the sample size is strictly restricted to the minimum value according to the thresholds.

This scheme was developed in the context of PLSR models. However, most model architectures, including nonlinear models such as neural networks or support vector machines relate to a VC dimension.[15] This is also not restricted only to regression models, the same ideas apply for classification models. This means that optimal sample sizes can be calculated based on the ratio n/d if an estimation on the model complexity d can be found for the problem at hand.

The criterion based on the equivalence between $\mathbf{S}_N$ and $\mathbf{S}_n$ is specific for the PLSR model architecture or similar models such as PLS discriminant analysis, linear regression, among others. For every model architecture, it is advisable to identify the mathematical components that could be controlled in unsupervised sample selection.

Data availability

All the methodology that was used in the present work is available for public share.

Acknowledgement

Valeria Fonseca Diaz is funded as aspirant to doctoral fellow of the Research Foundation-Flanders (FWO Brussels, Belgium). The authors thank Dr. Raffaele Vitale and professor Peter Goos (KU Leuven, Belgium) for their suggestions on the conducted analysis given their expertise in Chemometrics and Experimental Design.

References

(1) Au, J.; Youngentob, K. N.; Foley, W. J.; Moore, B. D.; Fearn, T. Sample selection, calibration and validation of models developed from a large dataset of near infrared spectra of tree leaves. *Journal of Near Infrared Spectroscopy* **2020**,

(2) Diaz-Olivares, J. A.; Adriaens, I.; Stevens, E.; Saeys, W.; Aernouts, B. *Computers and Electronics in Agriculture*; 2020; Vol. 178.

(3) Saeys, W.; Mouazen, A. M.; Ramon, H. Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. *Biosystems Engineering* **2005**, *91*, 393–402.

(4) Bobelyn, E.; Serban, A. S.; Nicu, M.; Lammertyn, J.; Nicolai, B. M.; Saeys, W. Postharvest quality of apple predicted by NIR-spectroscopy: Study of the effect of biological variability on spectra and model performance. *Postharvest Biology and Technology* **2010**, *55*, 133–143.

(5) Næs, T.; Isaksson, T. Selection of samples for calibration in near-infrared spectroscopy. Part II: Selection based on Spectral Measurements. *Applied Spectroscopy* **1990**, *44*, 1152–1158.

(6) Saeys, W.; Nguyen Do Trong, N.; Van Beers, R.; Nicolaï, B. M. Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: A review. 2019.

(7) Kennard, R.; Stone, L. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 13.

(8) Liu, Y.; Liu, Y.; Chen, Y.; Zhang, Y.; Shi, T.; Wang, J.; Hong, Y.; Fei, T.; Zhang, Y. The influence of spectral pretreatment on the selection of representative calibration samples for soil organic matter estimation using vis-NIR reflectance spectroscopy. *Remote Sensing* **2019**, *11*.

(9) Shetty, N.; Rinnan, Å.; Gislum, R. Selection of representative calibration sample sets for near-infrared reflectance spectroscopy to predict nitrogen concentration in grasses. *Chemometrics and Intelligent Laboratory Systems* **2012**, *111*, 59–65.

(10) Nawar, S.; Mouazen, A. M. Optimal sample selection for measurement of soil organic carbon using on-line vis-NIR spectroscopy. *Computers and Electronics in Agriculture* **2018**, *151*, 469–477.

(11) He, Z.; Li, M.; Ma, Z. Design of a reference value-based sample-selection method and evaluation of its prediction capability. *Chemometrics and Intelligent Laboratory Systems* **2015**, *148*, 72–76.

(12) Puchwein, G. Selection of Calibration Samples for Near-Infrared Spectrometry by Factor Analysis of Spectra. *Analytical Chemistry* **1988**, *60*, 569–573.

(13) Goos, P.; Jones, B. *Optimal design of experiments. A case study approach*; John Wiley & Sons, 2011; p 305.

(14) Rodionova, O. Y.; Pomerantsev, A. L. Subset selection strategy. *Journal of Chemometrics* **2008**, *22*, 674–685.

(15) Vapnik, V. N. Complete Statistical Theory of Learning. *Automation and Remote Control* **2019**, *80*, 1949–1975.

(16) Vapnik, V. N. *The nature of statistical learning theory*; Springer, 2000.

(17) Stone, M.; Brooks, R. Continuum Regression : Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares , Partial Least Squares and Principal Components Regression Author (s): M . Stone and R . J . Brooks Published by : Blackwell Publishing for the. *Journal of the Royal Statistical Society* **1990**, *52*, 237–269.

(18) de Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **1993**, *18*, 251–263.

(19) Tomic, O.; Forde, C.; Delahunty, C.; Næs, T. Performance indices in descriptive sensory analysis - A complimentary screening tool for assessor and panel performance. *Food Quality and Preference* **2013**, *28*, 122–133.

(20) Horn, R. A.; Johnson, C. R. *Matrix Analysis*; 1985.

(21) Artemiou, A.; Li, B. Predictive power of principal components for single-index model and sufficient dimension reduction. *Journal of Multivariate Analysis* **2013**, *119*, 176–184.

(22) Snee, R. D. Validation of Regression Models: Methods and Examples. *Technometrics* **1977**, *19*, 415–428.

(23) Wheeler, R. optFederov.AlgDesign. The R project for statistical computing. 2019; `https://cran.r-project.org/web/packages/AlgDesign/AlgDesign.pdf`.

(24) Lam, S. K.; Pitrou, A.; Seibert, S. Numba: a LLVM-based Python JIT compiler. *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15* **2015**, 1–6.

(25) Aernouts, B.; Polshin, E.; Lammertyn, J.; Saeys, W. Visible and near-infrared spectroscopic analysis of raw milk for cow health monitoring: Reflectance or transmittance? *Journal of Dairy Science* **2011**, *94*, 5315–5329.

(26) Brandmaier, S.; Sahlin, U.; Tetko, I. V.; Öberg, T. PLS-optimal: A stepwise D-Optimal design based on latent variables. *Journal of Chemical Information and Modeling* **2012**, *52*, 975–983.

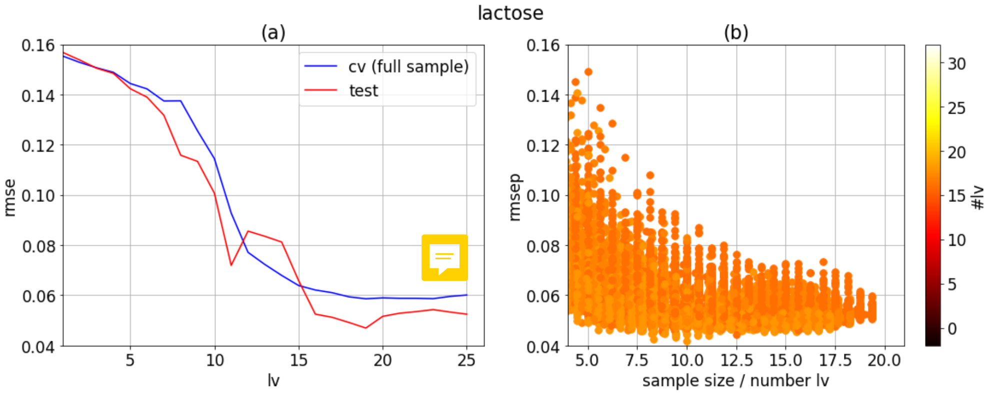


Figure 1: (a) RMSECV based on the complete selection set and RMSEP for Milk data set. (b) RMSEP values as function of the ratio n/d for 16, 17 and 18 latent variables as indicated in colorbar.

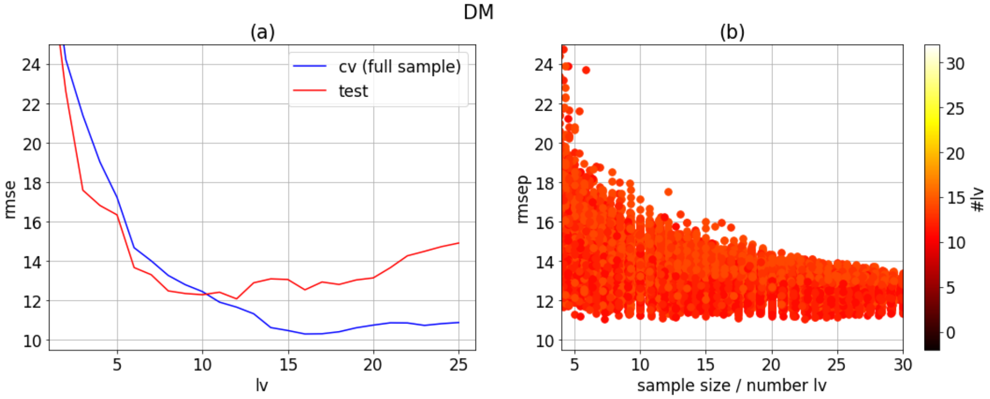
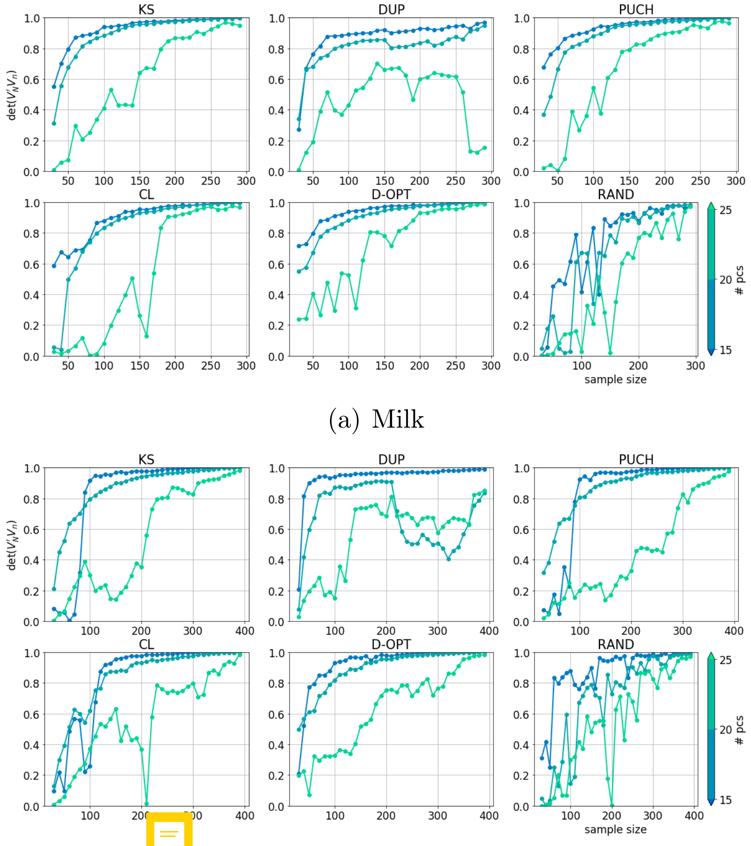


Figure 2: (a) RMSECV based on the complete selection set and RMSEP for Manure data set. (b) RMSEP values as function of the ratio n/d for 11 to 14 latent variables as indicated in colorbar



(a) Milk

(b) Manure

Figure 3: Comparison of eigenvectors when selecting samples with every method and input dimensionality $a = 15, 20, 25$.

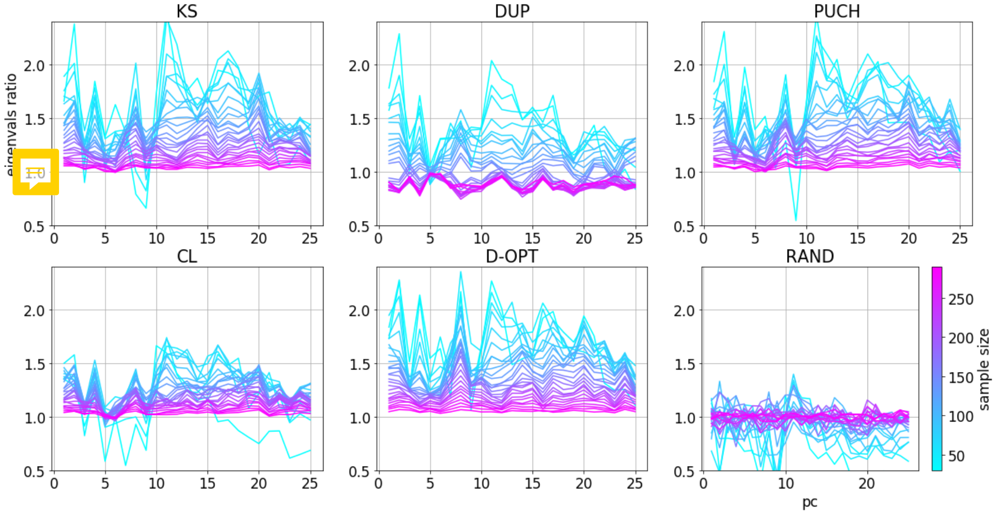


Figure 4: Comparison of eigenvalues for Milk when selecting samples with every method and input dimensionality $a = 25$.

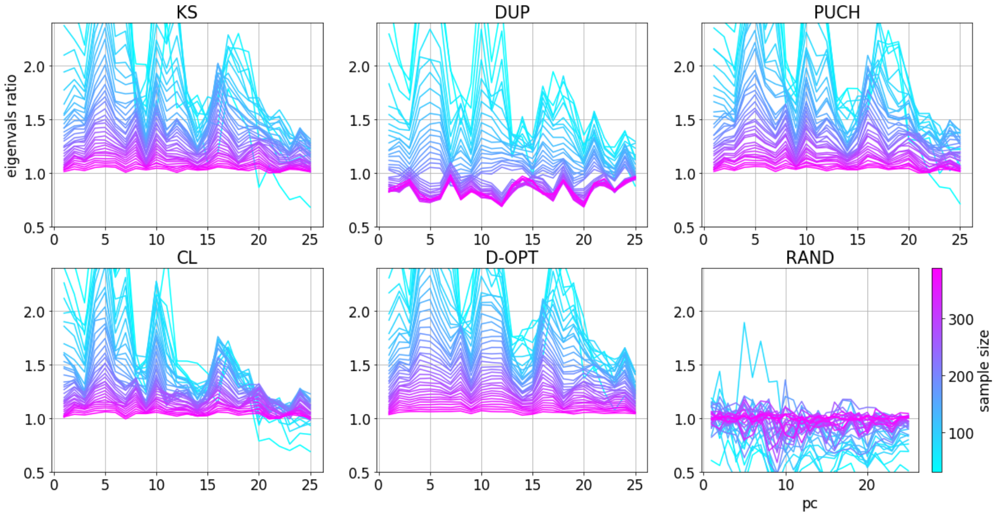


Figure 5: Comparison of eigenvalues for Manure when selecting samples with every method and input dimensionality $a = 25$.

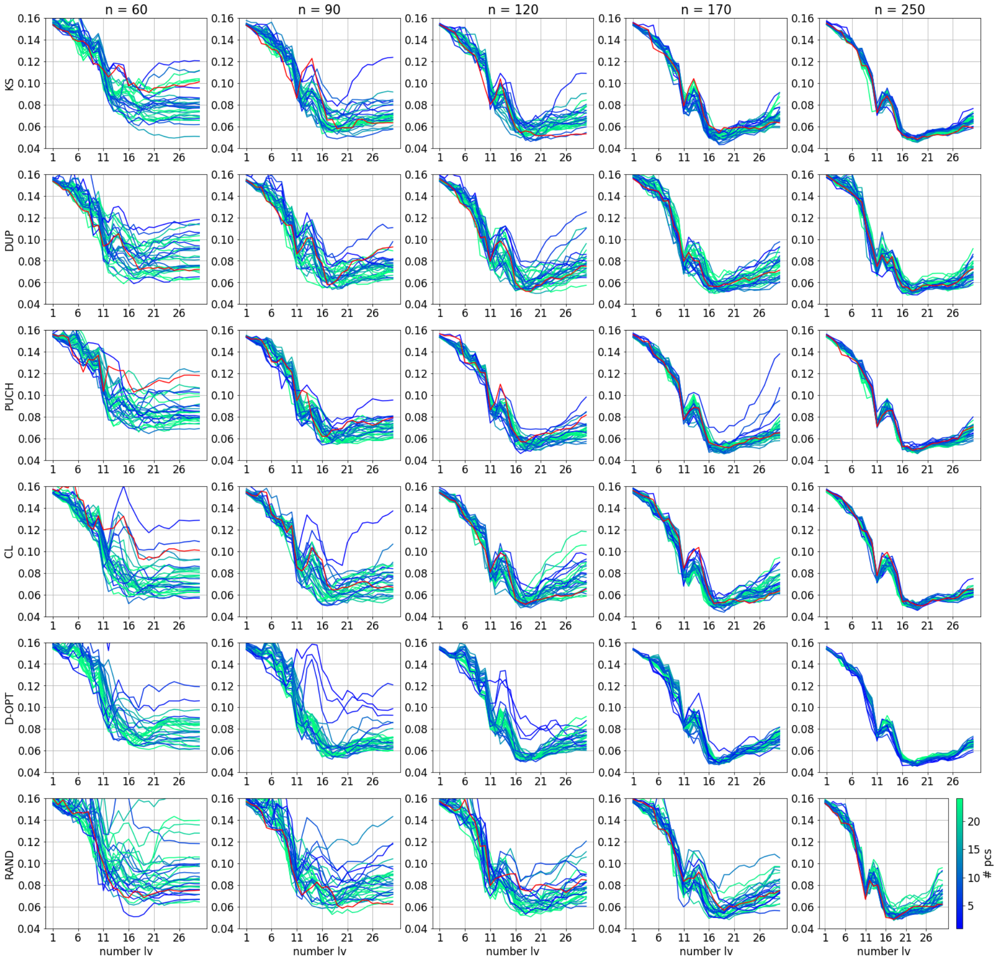


Figure 6: Model performance to predict lactose according to selection method, input dimensionality and sample size. The red line represents input dimensionality $a = p$.

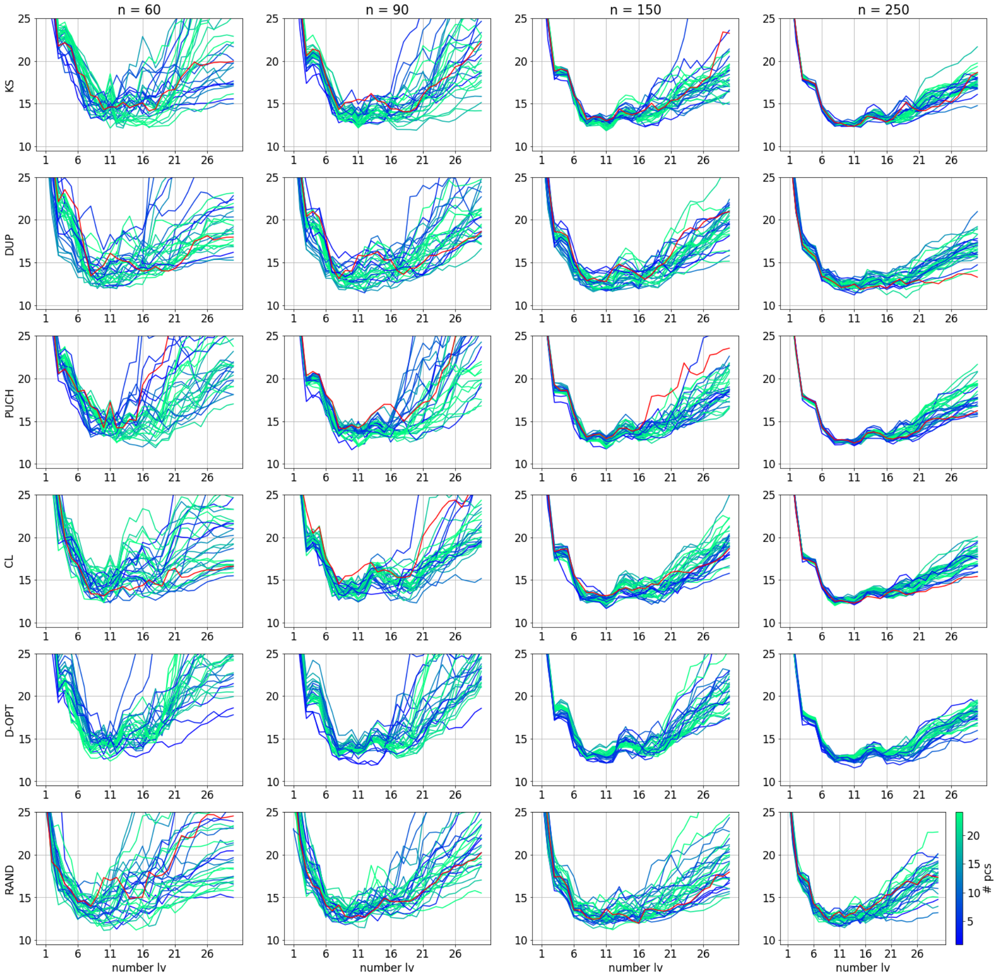


Figure 7: Model performance to predict DM according to selection method, input dimensionality and sample size. The red line represents input dimensionality $a = p$.