Methods

Julius Krause*, Maurice Günder, Daniel Schulz and Robin Gruna

New active learning algorithms for near-infrared spectroscopy in agricultural applications

Neue aktive Lernalgorithmen für die Nahinfrarotspektroskopie in landwirtschaftlichen Anwendungen

https://doi.org/10.1515/auto-2020-0143
Received September 7, 2020; accepted February 8, 2021

Abstract: The selection of training data determines the quality of a chemometric calibration model. In order to cover the entire parameter space of known influencing parameters, an experimental design is usually created. Nevertheless, even with a carefully prepared Design of Experiment (DoE), redundant reference analyses are often performed during the analysis of agricultural products. Because the number of possible reference analyses is usually very limited, the presented active learning approaches are intended to provide a tool for better selection of training samples.

Keywords: near infrared spectroscopy, active learning, sample selection

Zusammenfassung: Mit Hilfe von chemometrischen Kalibrierungsmodellen können verschiedene Qualitäts- und Reifeparameter für Agrarprodukte aus Nahinfrarotspektren geschätzt werden. Die verwendeten Trainingsdaten bestimmen dabei die Güte des chemometrischen Kalibrierungsmodells. Für das Training wird deshalb ein Datensatz benötigt, welcher Proben im gesamten Parameterraum beinhaltet. In der Regel wird ein Versuchsplan zur Probennahme erstellt, jedoch können viele Parameter in der Herstellung von Agrarprodukten nicht eingestellt werden. Daher muss in der Regel eine große Menge an Proben gesammelt werden, wobei häufig zahlreiche Proben den Informationsgehalt des Datensatzes nicht erhöhen. Des Weiteren müssen die Qualitäts- und Reifeparameter der Proben im Trainingsdatensatz aufwändig durch chemische Referenzanalysen erstellt werden. Die vorgestellten aktiven Lernansätze dienen einer optimalen Probenauswahl anhand von Nahinfrarotspektren, wodurch sich die Zahl der benötigten Proben den damit verbundenen Referenzanalysen verringert.

Schlagwörter: Nahinfrarotspektroksopie, aktives Lernen, Versuchsplanung

***Corresponding author: Julius Krause,** Fraunhofer IOSB, Karlsruhe, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, 76131 Karlsruhe, Germany, e-mail: julius.krause@iosb.fraunhofer.de
Maurice Günder, Daniel Schulz, Fraunhofer IAIS Institute for Intelligent Analysis and Information Systems, Schloss Birlinghoven, 53757 Sankt Augustin, Germany, e-mails: maurice.guender@iais.fraunhofer.de, daniel.schulz@iais.fraunhofer.de
Robin Gruna, Fraunhofer IOSB, Karlsruhe, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, 76131 Karlsruhe, Germany, e-mail: robin.gruna@iosb.fraunhofer.de

1 Introduction

Near-infrared spectroscopy (NIRS) is used to estimate and quantify quality and other parameters in a wide range of applications [12]. A large field of NIRS applications is in agriculture. For example, the ripeness or freshness of fruits can be determined. Furthermore, parameters such as moisture, protein content or sugar content can be estimated [8]. So-called chemometric calibrations are required to estimate the above-mentioned parameters of quality or components. Using methods of statistics and machine learning, correlations are established between the optically measured data and reference measurements. However, the quality of the results of chemometric models and machine learning methods strongly depend on the data set used [3]. The presented methods should improve the creation of a data set.

1.1 Motivation

Depending on the desired parameter, the generation of reference data can be very costly and time-consuming. This is particularly the case if chemical analyses have to be done in laboratories. Therefore, the chemical reference analysis is a limiting factor in the amount of training data for chemometric calibrations. In contrast, the ac-

quisition of NIRS data is fast, non-destructive, and therefore cost-effective. Assuming that the desired parameters are already included in the NIRS data, the following active learning approaches could be applied when selecting samples for reference analysis. The methods make a selection based on unlabeled NIRS data. The goal of the active learning approach is to sample a subset of the unlabeled NIRS data that is as representative as possible for the whole data. An entropy-based method is used to estimate and optimize the information content of the training data set. In addition, another kernel-based method called *constrained vector quantization* (CVQ) will select representative data on the basis of spectral distances and thus also determine a training data set with optimal information content.

1.2 Related work

Design of experiments is a well-researched field of statistics and due to its great importance for near-infrared spectroscopy it is also the subject of current research in the field of NIR spectroscopy [5, 20]. Because, the performance of chemometric calibration models can be improved by optimized sampling [3, 19]. In addition, methods for continuous model adaptation can significantly reduce costs in process analysis technology [4]. Based on an existing model, new samples can be selected from unlabeled data [6]. Another approach is the Kennard-Stones algorithm [9], which is used to identify the samples with the largest distance. The Kennard-Stones algorithm can also be applied to unlabelled NIR spectra [11].

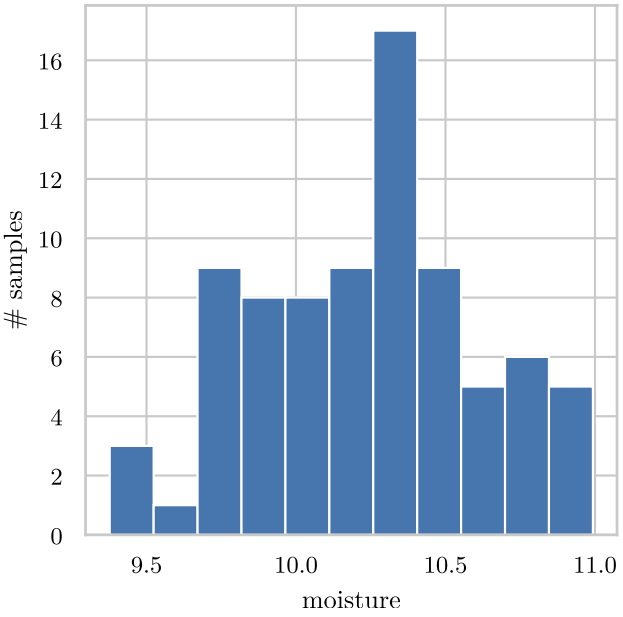


Figure 1: Distribution of the moisture content of the samples. There are significantly fewer samples with high or low moisture content.

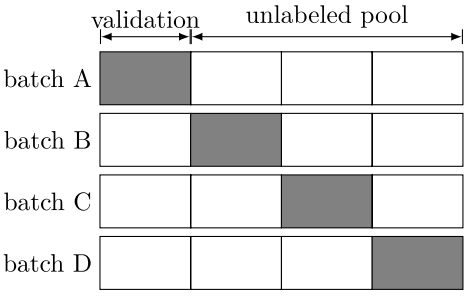


Figure 2: Partitioning of the data set. Partitioning of the data set. Four batches of 60 unlabeled samples in each batch are provided for active learning. Comparative validation of the methods is performed with 20 samples that are different in each batch.

2 Material and methods

The following methods are implemented in the *Python* programming language. Additionally, the goal is to keep the calculations as basic as possible by using commonly known libraries as *numpy* or *scipy*. For the later introduced machine learning part, the library *scikit-learn* is used.

2.1 Dataset

The methods introduced by this work are examined on a NIRS dataset of 80 corn samples in a wavelength range of 1100–2498 nm with a resolution of 700 bands. Originally, the data was measured at Cargill and three different NIR spectrometers (*m5*, *mp5*, and *mp6*) are used as well as four chemometrical features, i. e., moisture, oil, protein, and starch content. Here, the focus is on moisture content and NIRS data of the *mp5* spectrometer [7]. The distribution of moisture content is typical of agricultural sampling, which is usually uniformly or normally distributed (see Fig. 1).

In this work, the dataset is divided into four subsets, also called *batches* (see Fig. 2). Each batch consist of a pool of 60 unlabeled samples. The presented algorithms for active learning select the training data from this pool. To validate the performance of the methods, a set of 20 labeled samples is used. This allows comparison even with different selection of training data from the unlabeled pool. To make the results more independent of

the choice of validation data set, cross-validation was applied.

2.2 Regression

The focus of the presented work is the sample selection, therefore the methods were compared using a simple partial least square regression (PLSR). The PLSR is a commonly used regression technique in chemometrics. As the main idea, the predictors and responses are projected into a given number of principal components. These decompositions are made with the premise to maximize the covariance between the two projections. Before applying the PLSR on the data, the NIR spectra were smoothed by a Savitzky-Golay (7,2) filter [15]. Afterwards, the first derivative was formed and normalization was performed using the standard normal variable (SNV) transformation. This routine is a widely used preprocessing method in chemometrics and spectroscopy [14].

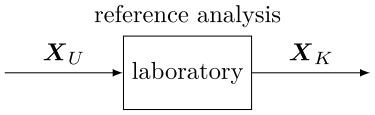
PLSR parameter estimation

The main hyper-parameter of the PLSR method is the number of principal components n_{comp} used in the transformations. In order to estimate the optimal choice of n_{comp}, an experiment is set up where a PLSR with different n_{comp} is trained on 75 % randomly assigned data and evaluated on the remaining 25 %. The *root-mean-squared error of validation* (RMSEV) is used as a metric to find the optimal value of n_{comp}. The number of components determined and used in the following is $n_{\text{comp}} = 7$.

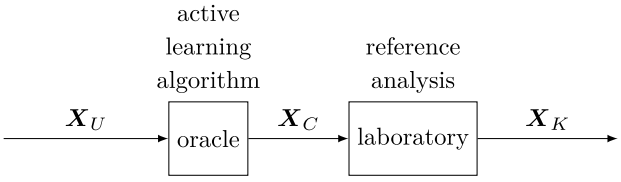
3 Active learning

In data science, active learning is a part of machine learning. An algorithm (oracle) is used to determine an information score from unlabeled data. The data selected by the oracle are then labelled to form the data set for training machine learning algorithms.

In the following, the method and idea of "active learning" is sketched. As a prerequisite, a dataset with ("cheap") NIR spectra is given. Additionally, one potentially is capable of doing the ("expensive") target value acquisition, i. e., chemical reference analysis, for all given spectra. Now, the goal is to do as few evaluations as possible to get a reasonably well-performing prediction model.



(a) In classical or batch wise learning, a complete set of unlabeled data $\boldsymbol{X}_U$ is analyzed. The selection of the samples is usually based on an experimental design.



(b) In active learning, an oracle selects a set of chosen samples $\boldsymbol{X}_C$ of the unknown samples $\boldsymbol{X}_U$. Only the selected samples are analyzed.

Figure 3: Principle of active learning. In comparison to classical methods, only a chosen subset $\boldsymbol{X}_C$ of the samples is analyzed in active learning methods. The oracle determines an informative score for a pool of spectral data of unknown samples $\boldsymbol{X}_U$. After the laboratory analysis the parameters of the samples are known, the result is a set of known samples $\boldsymbol{X}_K$.

Once there is an initial sub-sample to train the PLSR model, one can further improve it by including more spectra and their target values. Since the goal is to use as few as possible target value data, one needs a model that deliberately includes data that have the highest potential to improve the fit. This principle is referred to as *active learning* and there are several realization techniques. In the methods presented in this work, the focus is on a *pool-based* approach. The term *pool-based* indicates, that the process can choose the next spectrum from a bunch of unlabeled spectra in the pool represented by the matrix $\boldsymbol{X}_U := \{\boldsymbol{x}_i\}_{i=1..N}$ of N spectra. For the given data set, the pool contains all data that is not included in the model so far which means that potentially, every single spectrum $\boldsymbol{x}$ and its respective target value y can be evaluated. In comparison to classical methods, only a chosen subset $\boldsymbol{X}_C$ of the samples is analyzed (see Fig. 3). The reference analysis is used to build a set of known samples $\boldsymbol{X}_K$ with pairs of $(\boldsymbol{x}, y)$ for building regression or classification models.

3.1 PCA-entropy based sampling

The following approach of active learning aims to provide the most valuable data for a linear regression. The oracle iteratively selects a set of samples whose density distribution p is close to a desired density function q. The density function p of the sample distribution is obtained from a principal component analysis (PCA) of the spec-

tral data. The PCA gives the variance of the data in a low-dimensional representation. The desired density distribution q of the samples is derived from the optimal design. To minimize the regression error, the variance of the data must be maximized. The oracle's algorithm uses the Kullback-Leibler divergence (relative entropy) to measure the distance between the density functions p and q. Therefore it is named entropy based sampling.

3.1.1 Sample distribution using PCA

A NIRS measurement results in a spectrum, which is a high dimensional vector $\boldsymbol{x}_i$. Although, the dimension of the resulting vectors is high, the measurements are typically highly correlated since the general shape of spectra is quite similar. To determine the variance of the spectra in a simple way, a PCA can be applied on a collection $\boldsymbol{X}_U$ of unlabeled spectra. Usually, most of the variance of a NIRS measurement is contained in the first components of the PCA

$$\boldsymbol{T}_U = \boldsymbol{X}_U \boldsymbol{W}. \tag{1}$$

For further steps a probability density functions $p(\boldsymbol{t}_i) \in [0,1]$ of the distribution of the PCA scores $\boldsymbol{T}$ with $t_{i,j} = \boldsymbol{x}_i \boldsymbol{w}_j$ was generated by a normalized histogram and normalized scores $t_{i,j} \in [-1,1]$.

3.1.2 Optimal sample distribution

In the design of experiments there are optimal designs, which allow to minimize the effort, i. e., the number of samples. A common criterion is d-optimality by maximizing the determinant of the information matrix $\boldsymbol{X}^T\boldsymbol{X}$. The d-optimality is closely related to the least squares estimator. Under the Gauss-Markov assumptions, the covariance of the least squares estimator is proportional to $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. At this point, the similarity to the principal component analysis, which includes the eigenvectors $\boldsymbol{X}^T\boldsymbol{X}$, becomes clear.

In agricultural samples, however, it is only possible to control the parameters in experiments to a limited extent, and in some cases not all factors are known. Therefore, the information from the set of unlabeled spectral data $\boldsymbol{X}_U$ has to be optimized. Selecting samples with maximum PCA scores $\boldsymbol{t}_i$ is consistent with d-optimality and minimizes the covariance of the least squares estimator.

Therefore, the optimal sample distribution function

$$q(\boldsymbol{t}) = \frac{1 - e^{-\boldsymbol{t}^2}}{c} \quad \text{with} \quad c = \int_{-1}^{1}\int_{-1}^{1} 1 - e^{-\boldsymbol{t}^2} d\boldsymbol{t} \tag{2}$$

is chosen, which contains predominantly samples with high PCA scores.

3.1.3 Entropy based sampling

The entropy definition of statistical physics is closely related to information theory. Thus, entropy can be used as a measure for the information from a stochastic source. For example, an event with low probability has a high information content [18]. In the following the relative entropy, also known as Kullback-Leibler divergence, is used.

The idea of the presented approach is a sample selection, which is closest to the desired optimal $q(\boldsymbol{t})$. The oracle of the active learning algorithm determines the distance between the probability density function $p(\boldsymbol{t})$ of the measured spectra and the probability density function $q(\boldsymbol{t})$, which is considered optimal. To determine the similarity of two probability density functions, the Kullback-Leibler divergence

$$D(P||Q) := \int p(\boldsymbol{t}) \log \frac{p(\boldsymbol{t})}{q(\boldsymbol{t})} d\boldsymbol{t} \tag{3}$$

is calculated [10].

3.1.4 PCA-entropy algorithm

The algorithm is built in two parts. In the initialization two samples are selected and the Kullback-Leibler divergence is calculated for all combinations. After that, an iterative approach is used to choose the next optimal sample. This procedure causes less computation time for larger data sets. In both parts, the combination of samples with the minimum Kullback-Leibler divergence is chosen.

A qualitative result is shown in Fig. 4. The algorithm selects samples starting with the PCA scores in the boarder region of the first two dimensions. The sample selection is done symmetrically around the center. In local clusters of samples with similar PCA scores only a few samples are selected at the beginning.

3.2 Constrained vector quantization based sampling

In the following, a method of determining representative sub-samples of NIRS data by using *constrained vector quantization* (CVQ) is presented. The idea exploits the kernel trick—which is a widely used concept especially

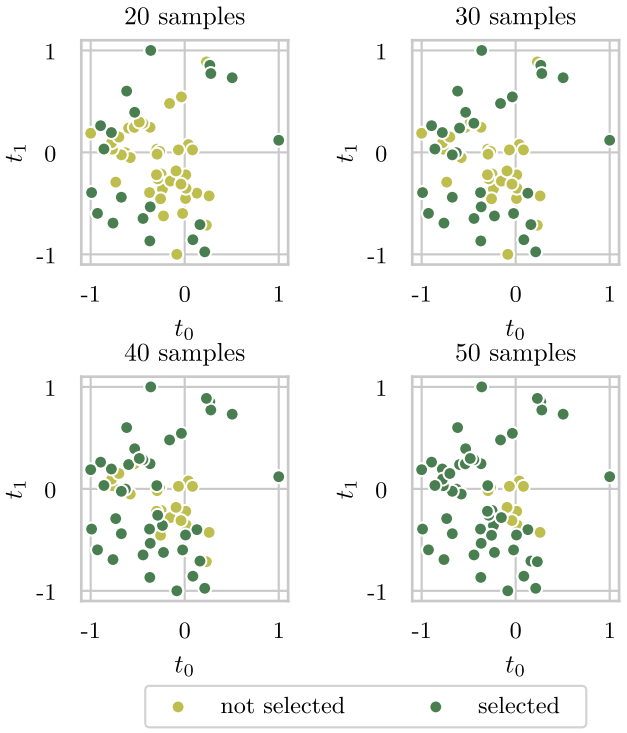


Figure 4: Principle of entropy based sampling. In this example, the normalized scores of the first two principal components are shown. It can be seen that the procedure selects the samples symmetrically, starting at the edges. It should also be mentioned that initially only individual samples from a local cluster are selected.

Algorithm 1: Entropy based sampling.

Data : Matrix $\boldsymbol{X_U}$ of spectra from unknown samples;
number N of samples to chose;
optimal sample distribution q

Result: List of chosen samples c

$T_U \leftarrow \mathtt{PCA}(X_U, n_{components} = 2)$
combinations $\leftarrow \mathtt{AllCombinations}(2, N)$
foreach comb *in* combinations **do**
  select T_{comb} from T_U
  p $\leftarrow \mathtt{hist}(T_{\mathsf{comb}})$
  $d_{\mathsf{comb}} \leftarrow \mathtt{KLDivergence}$ (p, q)
end
c $\leftarrow \min_{\mathsf{comb}}(d_{\mathsf{comb}})$
remove c from T_U
add c to T_C
for $n \leftarrow 3$ **to** N **do**
  foreach t *in* T_U **do**
    $T_{\mathsf{comb}} \leftarrow T_C$ and t
    p $\leftarrow \mathtt{hist}(T_{\mathsf{comb}})$
    $d_t \leftarrow \mathtt{KLDivergence}$ (p, q)
  end
  c $\leftarrow \min_t(d_t)$
  remove c from T_U
  add c to T_C
end

for non-linear machine learning concepts—to find a sub-sample of the data which represents its distribution preferably well. The term *constrained* is actually abbreviated from *subset-constrained* and shall denote, that the representative sub-sample $\mathcal{W}$ is a real sub-sample of the m-dimensional dataset $\mathcal{X}$, namely

$$\mathcal{W} \subset \mathcal{X} \subset \mathbb{R}^m . \tag{4}$$

The sub-sample $\mathcal{W}$ is referred to as the *codebook* of the whole dataset. In order to find the codebook that represents the distribution of our data best, an objective function is needed to be minimized. Here, the objective function of *mean discrepancy* is used.

3.2.1 Mean discrepancy

As mentioned beforehand, the choice of a optimal codebook relies on the minimization of *(squared) mean discrepancy* MD^2 given by

$$MD^2(\mathcal{X}, \mathcal{W}) = \|\bar{\boldsymbol{\varphi}}_{\mathcal{X}} - \bar{\boldsymbol{\varphi}}_{\mathcal{W}}\|^2 , \tag{5}$$

where $\mathcal{X}$ is the dataset and $\mathcal{W}$ a codebook sub-sample. $\bar{\varphi}_{\mathcal{X}}$ and $\bar{\varphi}_{\mathcal{W}}$ are the mean feature space vectors of the respective set defined by

$$\bar{\boldsymbol{\varphi}}_{\mathcal{X}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_j) , \qquad \bar{\boldsymbol{\varphi}}_{\mathcal{W}} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{\varphi}(\boldsymbol{w}_i) , \tag{6}$$

where $\boldsymbol{\varphi}(\boldsymbol{x})$ is the feature space vector of $\boldsymbol{x}$. This relies on the idea to express the probability densities as inner products in a feature space (also known as Hilbert space) [1, 2]. The inner products come into the picture by expanding equation (5) which is

$$\|\bar{\boldsymbol{\varphi}}_{\mathcal{X}} - \bar{\boldsymbol{\varphi}}_{\mathcal{W}}\|^2 = \langle\bar{\boldsymbol{\varphi}}_{\mathcal{X}}|\bar{\boldsymbol{\varphi}}_{\mathcal{X}}\rangle + \langle\bar{\boldsymbol{\varphi}}_{\mathcal{W}}|\bar{\boldsymbol{\varphi}}_{\mathcal{W}}\rangle - 2\langle\bar{\boldsymbol{\varphi}}_{\mathcal{X}}|\bar{\boldsymbol{\varphi}}_{\mathcal{W}}\rangle . \tag{7}$$

3.2.2 Kernel trick

The step of extending the problem in the feature space follows the principle of the "kernel trick" which is widely

used in non-linear machine learning methods. For a special class of kernels, the so called *Mercer kernels*, the feature map $\boldsymbol{\varphi}$ as introduced before exists. Additionally, we can express the inner products of feature space vectors as a kernel [16]

$$\phi(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\varphi}(\boldsymbol{x}) | \boldsymbol{\varphi}(\boldsymbol{x}') \rangle \,. \tag{8}$$

The radial basis function (RBF) kernel

$$\phi(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{h}||\boldsymbol{x} - \boldsymbol{x}'||^2\right) \tag{9}$$

indeed is an instance of a Mercer kernel. Its respective kernel matrix is $\boldsymbol{\Phi} \in \mathbb{R}^{n \times n}$ with the elements

$$(\boldsymbol{\Phi})_{ij} = \phi(\boldsymbol{x}_i, \boldsymbol{x}_j) \,. \tag{10}$$

The bandwidth parameter h has to be estimated wisely since it has a high impact on performance of the CVQ method. It turns out that it can be estimated by using the dataset's covariance matrix $\boldsymbol{\Sigma}$ and by Scott's rule of thumb [17]

$$\alpha_{\text{Scott}} = n^{-\frac{1}{d+4}} \,, \tag{11}$$

with the number of spectra n and the number of dimensions d. Therefore, we estimate the bandwidth by

$$h = \det(\alpha^2_{\text{Scott}} \boldsymbol{\Sigma}) \,. \tag{12}$$

3.2.3 Finding an optimal codebook

Moreover, a binary vector $\boldsymbol{z} \in \{0,1\}^n$ is introduced to show which of the n spectra are part of the codebook (= 1) or not (= 0). Introducing the kernel matrix in equation (10) finally allows to express the mean discrepancy by matrix vector notation

$$\begin{aligned} MD^2(\mathcal{X}, \mathcal{W}) = & \frac{1}{n^2} \sum_{j=1}^{n} \sum_{l=1}^{n} \phi(\boldsymbol{x}_j, \boldsymbol{x}_l) \\ & + \frac{1}{k^2} \sum_{j=1}^{n} \sum_{l=1}^{n} z_j \phi(\boldsymbol{x}_j, \boldsymbol{x}_l) z_l \\ & - \frac{2}{nk} \sum_{j=1}^{n} \sum_{l=1}^{n} \phi(\boldsymbol{x}_j, \boldsymbol{x}_l) z_l \\ = & \frac{1}{n^2} \mathbf{1}^T \boldsymbol{\Phi} \mathbf{1} + \frac{1}{k^2} \boldsymbol{z}^T \boldsymbol{\Phi} \boldsymbol{z} - \frac{2}{nk} \mathbf{1}^T \boldsymbol{\Phi} \boldsymbol{z} \end{aligned} \tag{13}$$

with codebook size k and an all-one vector $\mathbf{1} \in \mathbb{R}^n$. Since this derivation here is quite reduced, please find the complete elaboration in [2]. The first term in equation (13) is constant, calculation of the last two terms is sufficient for a valid objective function. Note, that both sums in the last two terms of equation (13) are running from 1 to n. However, only k terms contribute to the sums since the binary components of $\boldsymbol{z}$ are zero for non-codebook samples. Thus, the index running up to n is still valid and should intend that all sums can be performed over the complete data which is crucial to formulate the problem via matrix vector notation. The codebook with minimal mean discrepancy in this work is calculated by a greedy algorithm elaborated in [1] as well as the full theory that is sketched above. In the algorithm, codebook candidates are successively explored by switching the corresponding elements of $\boldsymbol{z}$ to 1 and calculating the objective function value. The candidate, whose inclusion results in the minimal objective function value is then considered as the next codebook sample. Similarly to the entropy based method, this procedure is applied on the PCA-transformed NIRS data $\boldsymbol{T}_U$ by only using the first two PCA components. Additionally, the coordinates are normalized to be in the range $[0,1]$. A pseudocode of the complete method is shown in Algorithm 2. Moreover, the sampling procedure with the given corn dataset is shown in Figure 5. In general, greedy algorithms are relatively time-efficient but do not necessarily find the optimal solution. Nevertheless, the greedy approach is sufficient in this method since this method should present a more scalable and less time-consuming alternative to the entropy based sampling method (cf. Section 3.1).

4 Results

The performance of the methods is evaluated using the *root mean squared error* (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \,. \tag{14}$$

The RMSE for each training set of 10 to 60 samples was determined from unlearned data (batch A–D) using k-fold cross validation. The RMSE determined for the data set and accessible in the literature [13] is about 0.15, with small deviations already caused by the selection of the validation data set.

4.1 PCA entropy

Already 30–40 samples selected with the PCA entropy method are sufficient for the training data set used to

Algorithm 2: Subset-constrained vector quantization sampling ("greedy").

Data : Matrix $\boldsymbol{X_U}$ of N spectra from unknown samples;
number k of samples to chose
Result: List of chosen samples $\boldsymbol{c}$

$\boldsymbol{T_U} \leftarrow \mathtt{PCA}(\boldsymbol{X_U}, n_{components} = 2)$
$\boldsymbol{\Sigma} \leftarrow \mathtt{CovarianceMatrix}(\boldsymbol{T_U})$
$\alpha_{\mathrm{Scott}} \leftarrow N^{-\frac{1}{6}}$ (cf. eq. (11) with $d = 2$)
$h \leftarrow \det(\alpha^2_{\mathrm{Scott}}\boldsymbol{\Sigma})$ (cf. eq. (12))

$\boldsymbol{\Phi} \leftarrow \mathtt{KernelMatrix}(\boldsymbol{T_U})$ (cf. eqs. (9), (10))
$\boldsymbol{\phi} \leftarrow \boldsymbol{\Phi 1}$
initialize $\boldsymbol{c} \leftarrow [\ \ ]$
initialize $\boldsymbol{z} \leftarrow \boldsymbol{0}$

for $i \leftarrow 1$ **to** k **do**
  initialize $\boldsymbol{H} = \infty$
  for $j \leftarrow 1$ **to** N **do**
    if $z_j == 0$ **then**
      $z_j \leftarrow 1$
      $H_j \leftarrow \frac{1}{k^2}\boldsymbol{z}^T\boldsymbol{\Phi}\boldsymbol{z} - \frac{2}{Nk}\boldsymbol{\phi}^T\boldsymbol{z}$
      $z_j \leftarrow 0$
    end
  end
  $j_{\min} \leftarrow \mathtt{argmin}(\boldsymbol{H})$
  $z_{j_{\min}} \leftarrow 1$
  $c \leftarrow \mathtt{append}(j_{min})$
end

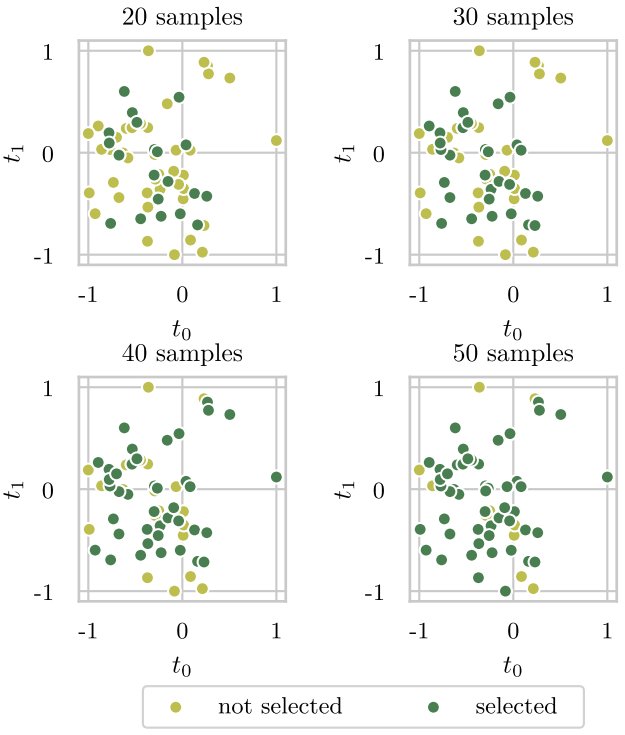


Figure 5: Principle of CVQ sampling. In this example, the normalized scores of the first two principal components are shown. The procedure samples a codebook trying to represent the complete dataset with a small sub-sample. Denser regions are preferred to outliers.

create a PLSR model. Adding more samples to the training data set does not result in a significant improvement, which is the case in all four experiments (see Fig. 6).

Another important aspect of the PCA entropy method is the possibility of an termination criterion based on unlabeled data. It is shown that the Kullback-Leibler divergence determined for the training data set also has a minimum between 30 and 40 samples for all four experiments (see Fig. 7). This can be interpreted to indicate that from this point on there is no more information gain if further samples are added from the available pool of unmarked data. The PCA shows (see Fig. 4) that from this point on the algorithm selects more data near the center, i. e., with low variance.

4.2 Constrained vector quantization

Similar the entropy sampling method, the CVQ sampling method shows a convergence at about 30–40 samples. This emphasizes, that there is no significant improvement for the PLSR model beyond 40 included training samples. The k-fold cross-validation shows that there is a slightly lower variance between the different batches compared to the entropy method (see Fig. 8).

4.3 Summary

A direct comparison of the two methods (see Fig. 9) shows that the RMSE for the CVQ sampling method decreases faster with the number of samples that the entropy method. Additionally, the RMSE variance between the four batches is significantly lower. Beyond the common convergence region between 30 and 40 samples, the entropy method leads to a lower RMSE value than the CVQ method.

Compared to random sampling, the reliability of the results of is significantly increased by active learning. For example, the results with 35 randomly selected samples can be significantly worse than with a set of 33 samples. There is also no reliable criterion for the number of samples to be analysed in random sampling.

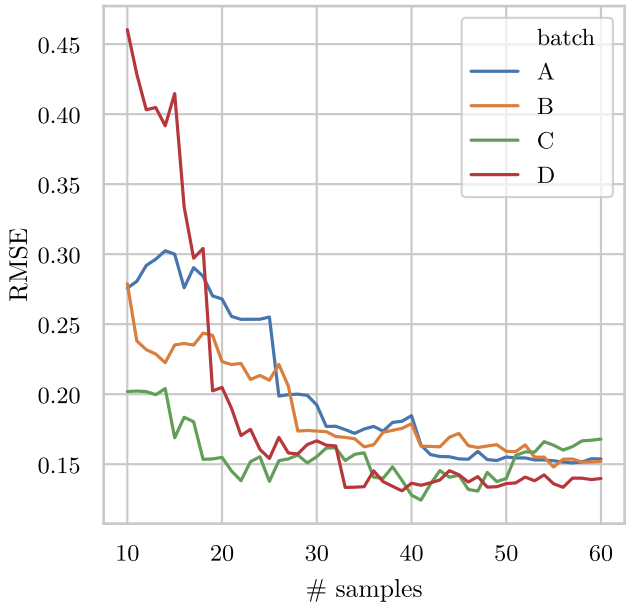


Figure 6: RMSE of PCA-Entropy based sampling. In all four experiments, the RMSE initially decreases with increasing sample number of the training data set. From a number of 30–40 samples in the training data set, there is no significant improvement with a larger number of samples in the training data set.

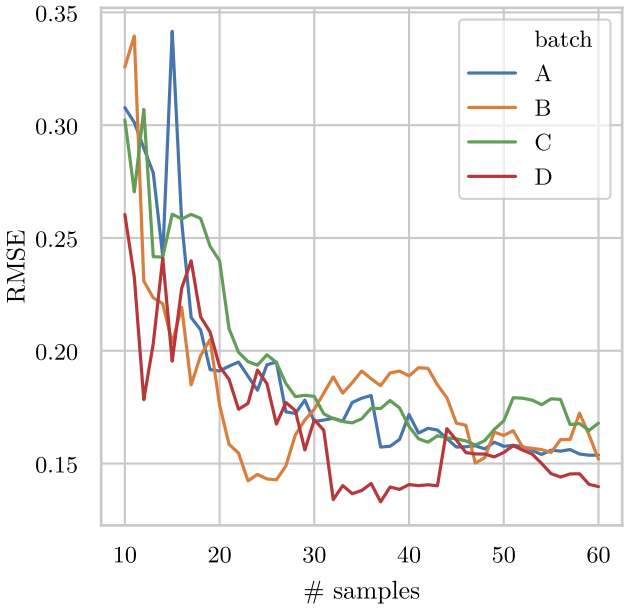


Figure 8: RMSE of CVQ based sampling. All four experiments show a similar trend to the entropy method (see Fig. 6). The RMSE variance between the experiments is lower compared to the entropy method.

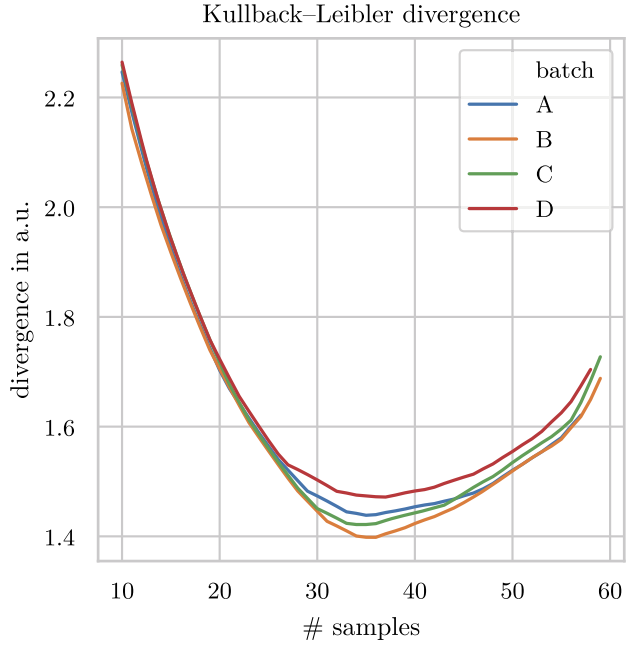


Figure 7: Kullback-Leibler divergence with increasing number of samples. In all four experiments the Kullback-Leibler divergence decreases with increasing number of samples until a minimum in the range of 30–40 samples is observed. The minimum can be used as a termination criterion for sample selection.

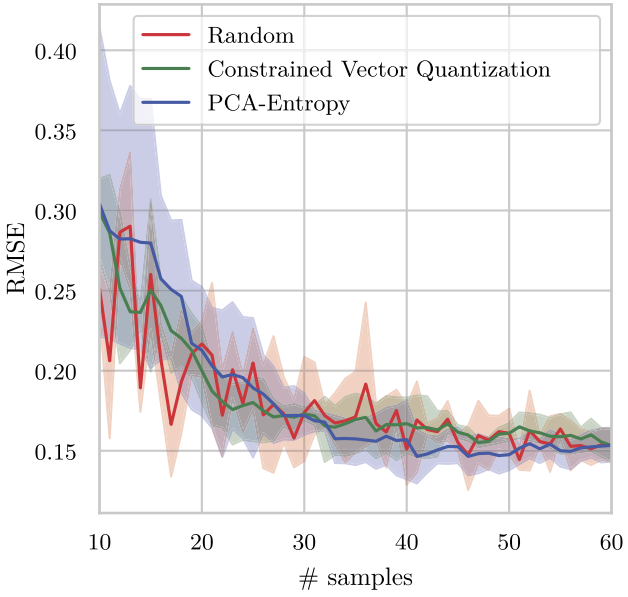


Figure 9: RMSE of both sampling methods. Both methods show a similar trend, from around 35 samples there is no significant improvement in RMSE by adding more samples to the training data set.

5 Conclusion

With both methods presented, a training data set with about 35 selected samples could already be generated from a pool of 60 unmarked spectral data. This is nearly a reduction by half of the required data. Active learning methods therefore offer the possibility to reduce the effort and thus also the costs of creating chemometric models.

In addition, a termination criterion could be determined by evaluating the Kullback-Leibler divergence.

Funding: The authors of this work were supported by the Fraunhofer Center for Machine Learning within the Fraunhofer Cluster for Cognitive Internet Technologies (CCIT) and the PhenoRob project which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 – 390732324, and by the German Federal Ministry of Education and Research (BMBF) within the context of the Software Campus project SmartSpectrometer under grant No. 01|S17047.

References

1. Christian Bauckhage. Numpy / scipy recipes for data science: Mean discrepancy minimization for vector quantization. February 2020.
2. Christian Bauckhage. Numpy / scipy recipes for data science: Subset-constrained vector quantization via mean discrepancy minimization. June 2020.
3. Robert W. Bondi, Benoît Igne, James K. Drennen and Carl A. Anderson. Effect of experimental design on the prediction performance of calibration models based on near-infrared spectroscopy for pharmaceutical applications. *Applied Spectroscopy*, 66(12):1442–1453, 2012.
4. Carlos Cernuda, Edwin Lughofer, Georg Mayr, Thomas Röder, Peter Hintenaus, Wolfgang Märzinger and Jürgen Kasberger. Incremental and decremental active learning for optimized self-adaptive calibration in viscose production. *Chemometrics and Intelligent Laboratory Systems*, 138:14–29, 2014.
5. Lorenzo De Benedictis and Christian Huck. New approach to optimize near-infrared spectra with design of experiments and determination of milk compounds as influence factors for changing milk over time. *Food Chemistry*, 212:552–560, 12 2016.
6. Fouzi Douak, Farid Melgani, Edoardo Pasolli and Nabil Benoudjit. SVR active learning for product quality control. In *2012 11th International Conference on Information Science, Signal Processing and their Applications, ISSPA 2012*, pages 1113–1117, 2012.
7. Eigenvector Research Inc. Data sets. https://eigenvector.com/resources/data-sets/#corn-sec. Accessed: 2020/18/08.
8. Simon Goisser, Julius Krause, Michael Fernandes and Heike Mempel. Determination of tomato quality attributes using portable NIR-sensors. In *OCM 2019 – Optical Characterization of Materials: Conference Proceedings*. Ed.: J. Beyerer, F. Puente León, T. Längle, page 1, 2019.
9. R. W. Kennard and L. A. Stone. Computer aided design of experiments. *Technometrics*, 11(1):137–148, 1969.
10. Doris Krahe and Juergen Beyerer. Parametric method to quantify the balance of groove sets of honed cylinder bores. *Architectures, Networks, and Intelligent Systems for Manufacturing Integration*, 3203(December 1997):192–201, 1997.
11. Anna Palou, Aira Miró, Marcelo Blanco, Rafael Larraz, José Francisco Gómez, Teresa Martínez, Josep Maria González and Manel Alcalà. Calibration sets selection strategy for the construction of robust PLS models for prediction of biodiesel/diesel blends physico-chemical properties using NIR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 180:119–126, June 2017.
12. Celio Pasquini. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. *Analytica Chimica Acta*, 1026:8–36, October 2018.
13. Dominic V. Poerio and Steven D. Brown. Dual-domain calibration transfer using orthogonal projection. *Applied Spectroscopy*, 72(3):378–391, 2018.
14. Åsmund Rinnan, Frans van den Berg and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201–1222, 11 2009.
15. A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, January 1964.
16. B. Schölkopf and AJ. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, December 2002.
17. David W. Scott. *Multivariate Density Estimation*. Wiley, August 1992.
18. C E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.
19. Han Tian, Linna Zhang, Ming Li, Yue Wang, Dinggao Sheng, Jun Liu and Chengmin Wang. Weighted SPXY method for calibration set selection for composition analysis based on near-infrared spectroscopy. *Infrared Physics & Technology*, 95:88–92, December 2018.
20. V. Wiedemair, M. De Biasio, R. Leitner, D. Balthasar and C. W. Huck. Application of design of experiment for detection of meat fraud with a portable near-infrared spectrometer. *Current Analytical Chemistry*, 14(1), January 2018.

Bionotes

Julius Krause
Fraunhofer IOSB, Karlsruhe, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, 76131 Karlsruhe, Germany
julius.krause@iosb.fraunhofer.de

Julius Krause recived his M.Sc. degree in physics in 2016 from the Karlsruhe Institute of Technology (KIT). Since 2016, his research has taken place at the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB in cooperation with the Vision and Fusion Laboratory at the Karlsruhe Institute of Technology. His research interests are hyperspectral signal processing and imaging for optical inspection and quality control, and machine learning.

Maurice Günder
Fraunhofer IAIS Institute for Intelligent Analysis and Information Systems, Schloss Birlinghoven, 53757 Sankt Augustin, Germany
maurice.guender@iais.fraunhofer.de

Maurice Günder recieved his M.Sc. degree in Experimental Particle Physics at RWTH Aachen University in Aachen, Germany. Since 2020, he has been a Data Scientist at the Fraunhofer Insitute for Intelligent Analysis and Information System IAIS in Sankt Augustin, Germany, while pursuing the PhD degree at the University of Bonn, Germany. His research interests comprise time series analysis, knowledge extraction from sensorical data, and inclusion of expert knowledge in machine learning processes.

Daniel Schulz
Fraunhofer IAIS Institute for Intelligent Analysis and Information Systems, Schloss Birlinghoven, 53757 Sankt Augustin, Germany
daniel.schulz@iais.fraunhofer.de

Daniel Schulz studied geography, geology and soil science at the universities of Cologne, Bonn and Gothenburg. After his graduation he began his work at the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Sankt Ausgutin. There he worked as a project manager in various large-scale projects with industry and public clients. His research focuses on Machine Learning (Informmed Machine Learning) and Artificial Intelligence. Currently, he heads the office of the Fraunhofer Research Center for Machine Learning.

Robin Gruna
Fraunhofer IOSB, Karlsruhe, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, 76131 Karlsruhe, Germany
robin.gruna@iosb.fraunhofer.de

Robin Gruna obtained his PhD from the Karlsruhe Institute of Technology (KIT) in the field of Machine Vision and Computational Imaging. Currently, he is the research group manager at the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB in Karlsruhe. His research interests include machine learning, hyperspectral imaging and spectral sensing.