AdACLV FOR INTERPRETABLE VARIABLE CLUSTERING AND DIMENSIONALITY REDUCTION OF SPECTROSCOPIC DATA

Marion, R., Govaerts, B. and R. von Sachs

DISCUSSION PAPER | 2020 / 11

AdaCLV for Interpretable Variable Clustering and Dimensionality Reduction of Spectroscopic Data

Rebecca Marion^{a,*}, Bernadette Govaerts^a, Rainer von Sachs^a

^a*Université catholique de Louvain - ISBA, LIDAM
Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium*

Abstract

This paper introduces a new method, Adaptive Clustering around Latent Variables (AdaCLV), for simultaneous dimensionality reduction and variable clustering, the partitioning of variables into groups. This unsupervised method is particularly well suited for the exploration of spectroscopic datasets, such as Nuclear Magnetic Resonance (NMR) spectra, and can be used for the identification of potential biomarkers. AdaCLV is inspired by existing multivariate methods from the Clustering around Latent Variables (CLV) family, but it offers several key advantages with respect to these methods. First, AdaCLV allows variables to belong to multiple clusters with varying degrees. A cluster membership degree is estimated for each variable and cluster, and these memberships are used to define non-orthogonal latent variables that summarize the clusters. As a result, the clusters and latent variables identified by AdaCLV are more interpretable and representative of spectroscopic data, where peaks for different molecules (i.e. variable clusters) may overlap and variables within a cluster have different degrees of importance. Second, while the performance of existing methods depends greatly on hyperparameter selection, AdaCLV is less sensitive to its hyperparameters, adapting to the clustering structure present in the data. This paper compares AdaCLV with existing CLV methods and other competi-

*Corresponding author

Email address: `rebecca.marion@uclouvain.be` (Rebecca Marion)

tors in experiments involving real and semi-artificial NMR spectra. AdaCLV is shown to be more robust to hyperparameter choice and to have better precision than the other methods, for all cluster numbers, sample sizes and levels of signal tested, while achieving a comparable level of recall.

Keywords: Variable clustering, latent variables, dimensionality reduction, Nuclear Magnetic Resonance spectra

1. Introduction

In metabolomics and other -omics sciences, biomarker studies attempt to identify the biological entities (e.g. metabolites) whose fluctuation is predictive of a given health outcome [1]. One popular means of quantifying metabolite concentrations is through Nuclear Magnetic Resonance (NMR) spectroscopy. NMR spectra are highly reproducible, meaning that most of the variability between spectra represents biological differences between samples [1, 2]. Moreover, as NMR spectroscopy can quantify both known and unknown molecules, it is particularly useful for identifying and characterizing new compounds [3], making it an important platform for biomarker discovery. In an NMR spectrum, each molecule detected in the sample is represented by one or more peaks, and the height of these peaks is proportional to the concentration of the molecule [4, 3]. Identifying biomarkers boils down to identifying the sets of spectral variables or regions that most predict the health outcome of interest. This is illustrated with an example in Figure 1.

Figure 1: **Example spectrum for a biomarker study.** Mean intensities for each spectral variable in an NMR dataset representing the urine of rats. In this dataset, the only metabolites that differ in concentration between samples are citrate (variables highlighted in green) and hippurate (highlighted in red). Methods for biomarker discovery should be able to identify these highlighted variables as potential biomarkers, ideally separating them into two distinct groups.

However, identifying predictive spectral regions is complicated by the complex nature of NMR data. While the underlying dimension of the data is the number of measurable molecules in the mixture [5], the actual number of dimensions is much larger, which creates complex correlation structures in the data. Variables
 20 in the spectrum tend to be correlated, especially if they belong to the same peak or represent the same molecule. Moreover, in practice, these datasets often have many more variables than observations, making them very ill conditioned for classical multivariate analyses [6].

In this paper, we propose an unsupervised multivariate method, Adaptive Clus-
 25 tering around Latent Variables (AdaCLV), that addresses these challenges by simultaneously clustering variables and performing dimensionality reduction: AdaCLV automatically detects the spectral variables that most differentiate samples, assigns them to one or more variable clusters and estimates non-orthogonal latent variables, one for each cluster. Cluster membership is deter-
 30 mined by estimating membership degrees that indicate the relative importance

of variables within a cluster, making it possible to rank variables for subsequent biomarker validation. For a given cluster, these membership degrees are shrunk towards zero with a threshold that iteratively adapts to clustering structure in the data. As a result, only the variables that most discriminate samples
 35 (i.e. potential biomarkers) have non-zero cluster membership degrees. In addition to identifying potential biomarkers, AdaCLV generates a set of latent variables that summarize important information from the original dataset in a lower-dimensional form. These latent variables can be used in downstream analyses, such as regression or classification, as is done for methods like the
 40 Cluster Representative Lasso [7].

AdaCLV is based on the Clustering around Latent Variables (CLV) approach, which was previously developed in [8] and [9]. Like AdaCLV, Clustering around Latent Variables (CLV) [8] clusters variables into K groups using a covariance-based criterion and estimates K non-orthogonal latent variables. However, CLV
 45 clusters are non-overlapping and cluster memberships are binary: a variable either belongs to a cluster or it does not. Thus, these memberships are less informative than AdaCLV memberships, which indicate the relative importance of variables within a cluster and reflect possible cluster overlap. Moreover, when estimating latent variables, CLV does not account for differing degrees of
 50 membership within a given cluster, and it is relatively sensitive to the choice of the hyperparameter K , the number of clusters to detect.

One further limitation of CLV is that cluster assignment and latent variable estimation is adversely affected by the presence of noise variables (i.e. low-intensity variables in noisy regions of the spectra). Two methods were recently
 55 proposed in [9] that attempt to address this issue. CLV $K + 1$ identifies noise variables using a correlation-based threshold at each iteration of its algorithm and removes these variables from the analysis. As a result, noise variables can be assigned to a “garbage” cluster (cluster $K + 1$) and are not used to estimate latent variables. The second method proposed in [9], CLV Sparse LV, also uses
 60 a correlation-based threshold to exclude noise variables during the estimation

of latent variables. However, CLV Sparse LV remains sensitive to noise for its clustering step, as it assigns each variable, even noise variables, to one of the K clusters. Moreover, for both methods, the filtering of noise variables depends on an extra hyperparameter that must be tuned by the user. As with CLV, these
65 methods depend on binary memberships, cannot find overlapping clusters and are sensitive to the choice of K .

Other existing unsupervised methods also have limitations with respect to AdaCLV. Principal Component Analysis (PCA) is by far the most popular method for dimensionality reduction, generating orthogonal latent variables that represent
70 major sources of variance in the data. While it is possible to identify important variables (i.e. potential biomarkers) for each of the latent variables, the requirement of orthogonality between latent variables may be too rigid for many natural phenomena, which are not necessarily orthogonal [6]. This is particularly true for spectral datasets, where the latent phenomena of interest –
75 fluctuations in molecule concentrations – are inherently correlated. As an alternative to PCA, methods built for observation clustering, such as K-means or hierarchical clustering, can be applied to cluster variables. However, like existing CLV approaches, these methods define cluster membership as binary, and they are not adapted for datasets with overlapping clusters.

80 This paper makes several contributions. First, it introduces AdaCLV, an unsupervised method for performing dimensionality reduction and identifying potentially overlapping clusters of variables. In order to better appreciate and understand our novel contribution, we complete the paper by giving a fairly detailed review of related methods. Finally, a rigorous comparison of AdaCLV
85 with existing CLV methods and baseline approaches (K-means and Ward’s hierarchical clustering) is performed using simulations where the ground-truth is known and performance can be evaluated using numerical criteria. This numerical study demonstrates that AdaCLV estimates target cluster memberships and latent variables better than its competitors. It also shows that AdaCLV,
90 unlike the other methods, is relatively insensitive to the choice of the number

K of clusters to detect.

The notation used throughout this paper is explained in Section 2, AdaCLV is presented in Section 3 and related works are summarized in Section 4. Numerical experiments and results are described in Section 5 and the conclusion and suggestions for future work can be found in Section 6.

2. Preliminaries

Let $\mathbf{X}$ ($n \times p$) be a data matrix where n is the number of observations and p is the number of variables. In what follows, we will assume that the columns of $\mathbf{X}$ are mean-centered. Observations are indexed with the letter $i \in \{1, \dots, n\}$, variables with the letter $j \in \{1, \dots, p\}$ and clusters with the letter $k \in \{1, \dots, K\}$, where K is the total number of clusters. $C_k := \{j \in \text{cluster } k\}$ is the set of p_k variables j belonging to a given cluster k , and $\mathbf{X}_k$ is an $n \times p_k$ matrix containing the variables from $\mathbf{X}$ that are assigned to cluster k . Scalar elements from matrices are denoted using lower-case letters with two subscripts (e.g. x_{ij}). Row and column vectors from matrices are indicated using bold, lower-case letters.

The indicator function $\mathbb{I}[z]$ evaluates to 1 if the argument z is true and to 0 otherwise, and the non-negative truncation function $(z)_+$ takes the value z if z is positive and 0 otherwise. The matrix $\text{diag}(\mathbf{z})$ is a diagonal matrix whose diagonal elements are contained in $\mathbf{z}$, and the ℓ_2 -norm of a vector $\mathbf{z}$ is defined as $\|\mathbf{z}\|_2 = \sqrt{\sum_{k=1}^K z_k^2}$. The functions $\text{mean}()$, $\text{sd}()$, $\text{var}()$, $\text{cov}()$ and $\text{cor}()$ return the sample mean, standard deviation, variance, covariance and correlation, respectively. For centered vectors $\mathbf{z}_j$ and $\mathbf{z}_\ell$, $\text{sd}(\mathbf{z}_j) = \hat{\sigma}_j = \sqrt{\mathbf{z}_j^\top \mathbf{z}_j / (n-1)}$, $\text{var}(\mathbf{z}_j) = \hat{\sigma}_j^2$, $\text{cov}(\mathbf{z}_j, \mathbf{z}_\ell) = \mathbf{z}_j^\top \mathbf{z}_\ell / (n-1)$ and $\text{cor}(\mathbf{z}_j, \mathbf{z}_\ell) = \text{cov}(\mathbf{z}_j, \mathbf{z}_\ell) / [\hat{\sigma}_j \hat{\sigma}_\ell]$.

3. Proposed Method: AdaCLV

AdaCLV is a method for dimensionality reduction and variable clustering inspired by the Clustering around Latent Variables (CLV) approaches introduced in [8, 9] (see Section 4). In contrast to existing CLV methods, AdaCLV accounts

for differing degrees of cluster membership within a cluster and estimates more easily interpretable cluster memberships. It also identifies the variable clusters that most differentiate samples, making it a candidate method for unsupervised identification of potential biomarkers.

AdaCLV simultaneously estimates two matrices: $\mathbf{A}$, an $n \times K$ latent representation of the data, and $\mathbf{S}$, a $p \times K$ cluster membership matrix. Each column vector $\mathbf{a}_k$ in $\mathbf{A}$ is a “latent variable” representing the essential information in cluster k , and each column vector $\mathbf{s}_k$ in $\mathbf{S}$ is a cluster membership profile for cluster k . For a given variable j and cluster k , s_{jk} represents a non-negative membership degree: $s_{jk} \in [0, 1]$, where $s_{jk} = 0$ indicates lack of membership in cluster k . Within a given cluster, the vector $\mathbf{s}_k$ is constrained to have unit ℓ_2 -norm, so the membership values s_{jk} indicate the relative importance of variable j within cluster k . These memberships are “soft” in the sense that they are non-binary, but they are slightly different than the typical soft memberships estimated for soft or “fuzzy” clustering problems [10]. Indeed, for fuzzy clustering problems, the scale of the rows in $\mathbf{S}$ is fixed ($\sum_{k=1}^K s_{jk} = 1, \forall j$), whereas for AdaCLV the scale of the columns is fixed ($\sum_{j=1}^p s_{jk}^2 = \mathbf{s}_k^\top \mathbf{s}_k = 1, \forall k$). Thus, fuzzy clustering memberships indicate the degree to which an object belongs to each of the K clusters, but they do not indicate the relative importance of objects within a cluster.

For a given true cluster k , variables in spectral data generally fall into one of three categories: 1) variables in noisy regions (“noise variables”) that do not contain any discriminating signal, 2) variables in signal regions that belong to a cluster other than cluster k and 3) variables in signal regions that belong to cluster k . Noise variables (category 1) should not be assigned to any of the clusters or used to estimate latent variables. The covariance of these variables with the true latent variables is generally of small magnitude and distributed around 0. Variables in category 2 tend to have negative or small positive covariance with the latent variable for cluster k . This is due to the fact that, in spectral datasets representing mixtures, peaks representing different molecules can be somewhat

negatively correlated, even in the absence of a biological or chemical relationship between the molecules: the compositional nature of mixtures implies that increases in the concentration of one molecule are accompanied by decreases in the concentration of the others. Finally, the third category of variable, in contrast to the first two categories, has relatively high positive covariance with the latent variable for cluster k . Taken together, these insights suggest that the covariance of variables with a given latent variable k can be used to distinguish relevant variables (category 3: $j \in C_k$) from irrelevant variables (categories 1 and 2: $j \notin C_k$).

AdaCLV takes advantage of this phenomenon when assigning variables to clusters: variable j is assigned to cluster k (i.e. $s_{jk} > 0$) if and only if $\text{cov}(\mathbf{a}_k, \mathbf{x}_j) \geq \tau_k$, where $\tau_k \geq 0$ is a threshold parameter that is updated adaptively at each iteration of the algorithm. Note that this approach makes it possible to detect overlapping clusters. If a variable has sufficiently high covariance with multiple latent variables, it can be assigned to multiple clusters. Overlapping peaks are very common in NMR spectra [1], so this property is particularly useful.

For fixed $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)^\top$, the criterion to maximize with respect to $\mathbf{A}$ and $\mathbf{S}$ is

$$J(\mathbf{A}, \mathbf{S}) = \sum_{k=1}^K \sum_{j=1}^p s_{jk} [\text{cov}(\mathbf{a}_k, \mathbf{x}_j) - \tau_k] \quad (1)$$

subject to the following constraints:

1. $s_{jk} \geq 0, \forall j, k$ (nonnegativity)
2. $\mathbf{s}_k^\top \mathbf{s}_k = 1, \forall k$ (membership profiles have unit length)
3. $\text{var}(\mathbf{a}_k) = 1, \forall k$ (latent variables have unit variance).

Constraint 1 ensures that the cluster memberships are non-negative, and it also has the convenient property of encouraging sparseness in the cluster membership vectors $\mathbf{s}_k$. Indeed, s_{jk} will be set to zero if $\text{cov}(\mathbf{a}_k, \mathbf{x}_j) < \tau_k$. Thus, negatively

correlated variables will not be clustered together. Constraints 2 and 3 are imposed to avoid scale ambiguity.

For fixed τ_k and $\mathbf{a}_k$, a vector of memberships $\mathbf{s}_k$ can be estimated, where each element is calculated as follows:

$$s_{jk} = \frac{(\text{cov}(\mathbf{a}_k, \mathbf{x}_j) - \tau_k)_+}{\sqrt{\sum_{\ell=1}^p (\text{cov}(\mathbf{a}_k, \mathbf{x}_\ell) - \tau_k)_+^2}}. \quad (2)$$

For fixed $\mathbf{s}_k$, the optimal latent variable $\mathbf{a}_k$ for the constrained problem above is

$$\mathbf{a}_k = \sqrt{n-1} \mathbf{X} \mathbf{s}_k / \|\mathbf{X} \mathbf{s}_k\|_2. \quad (3)$$

175 For fixed $\mathbf{s}_k$ and $\mathbf{a}_k$, the adaptive threshold τ_k is determined using a two-step procedure. In the first step, noise variables are identified by taking advantage of the covariance properties discussed above. Given an initial vector $\mathbf{r}_k = (\text{cov}(\mathbf{a}_k, \mathbf{x}_1) \dots \text{cov}(\mathbf{a}_k, \mathbf{x}_p))^\top$ and a user-defined hyperparameter $\eta > 0$, a noise identification algorithm iteratively removes extreme elements from $\mathbf{r}_k$
 180 until only non-extreme (i.e. noise) variables remain. The algorithm, which is inspired by [11], proceeds as follows:

1. Calculate the sample mean and standard deviation of $\mathbf{r}_k$: $\bar{r}_k = \text{mean}(\mathbf{r}_k)$ and $\hat{\sigma}_k = \text{sd}(\mathbf{r}_k)$
2. Remove all elements j from $\mathbf{r}_k$ for which $|r_{jk}| > \bar{r}_k + \eta \hat{\sigma}_k$
- 185 3. Repeat steps 1 and 2 until no values $|r_{jk}| > \bar{r}_k + \eta \hat{\sigma}_k$

In the second step of the threshold estimation procedure, the output $\mathbf{r}_k$ from the first step is used to calculate τ_k , where

$$\tau_k = (\max_j(r_{jk}))_+. \quad (4)$$

Note that the noise identification algorithm assumes that the covariance of noise variables with a given latent variable k is symmetrically distributed. The maximum of this distribution is also assumed to be smaller than $\text{cov}(\mathbf{a}_k, \mathbf{x}_j)$ for most variables $j \in C_k$ and greater than $\text{cov}(\mathbf{a}_k, \mathbf{x}_j)$ for signal variables $j \notin C_k$.

190 The entire AdaCLV algorithm is summarized below in Algorithm 1. At the beginning of the algorithm, $\mathbf{S}$ is initialized based on an initial partition $C_1, \dots, C_K$ of K variable clusters, where $s_{jk} = \mathbb{I}[j \in C_k] / \sqrt{p_k}$. Then, at each iteration of the algorithm, the latent variable matrix $\mathbf{A}$, the threshold vector $\boldsymbol{\tau}$ and the membership matrix $\mathbf{S}$ are updated successively.

```

Data: A centered matrix  $\mathbf{X}$  ( $n \times p$ ), initial partition  $C_1, \dots, C_K$ ,
        hyperparameter  $\eta > 0$ 
1 // Initialization
2 for each cluster  $k$  in  $1, \dots, K$  do
3     for each variable  $j$  in  $1, \dots, p$  do
4          $s_{jk} = \mathbb{I}[j \in C_k] / \sqrt{p_k}$ ;
5     end
6 end
7 // Main Algorithm
8 while not converged do
9     for each cluster  $k$  in  $1, \dots, K$  do
10         // Update latent variable  $k$ 
195          $\mathbf{a}_k = \sqrt{n-1} \mathbf{X} \mathbf{s}_k / \|\mathbf{X} \mathbf{s}_k\|_2$ ;
11         // Update adaptive threshold  $k$ 
12         Given the hyperparameter  $\eta$ , update  $\tau_k$  using the noise
            identification algorithm described above
13         // Update membership vector  $k$ 
14         for each variable  $j$  in  $1, \dots, p$  do
15              $s_{jk} = \frac{(\text{cov}(\mathbf{a}_k, \mathbf{x}_j) - \tau_k)_+}{\sqrt{\sum_{\ell=1}^p (\text{cov}(\mathbf{a}_k, \mathbf{x}_\ell) - \tau_k)_+^2}}$ ;
16         end
17     end
18 end
19 return Latent representation  $\mathbf{A}$  and sparse cluster membership matrix  $\mathbf{S}$ 
20

```

The maximum value for η can be identified when the vectors $\mathbf{r}_k$ are calculated

for the first time (i.e. at the first iteration of the AdaCLV algorithm). The noise identification algorithm will label all variables as noise for cluster k if $\max_j(|r_{jk}|) \leq \bar{r}_k + \eta\hat{\sigma}_k$. If this occurs, the vector of memberships $\mathbf{s}_k$ will be set to zero. The entire matrix $\mathbf{S}$ will be set to zero if $\max_j(|r_{jk}|) \leq \bar{r}_k + \eta\hat{\sigma}_k, \forall k$.
 Thus, the maximum value for η can be defined as follows:

$$\eta_{max} = \max_k \left[\left(\max_j(|r_{jk}^{(1)}|) - \bar{r}_k^{(1)} \right) / \hat{\sigma}_k^{(1)} \right], \quad (5)$$

where $\mathbf{r}_k^{(1)}$ is the input to the noise identification algorithm at the first iteration of AdaCLV, and $\bar{r}_k^{(1)} = \text{mean}(\mathbf{r}_k^{(1)})$ and $\hat{\sigma}_k^{(1)} = \text{sd}(\mathbf{r}_k^{(1)})$.

4. Related Work

AdaCLV is closely related to the Clustering around Latent Variables (CLV) family of methods. These methods also estimate a latent variable matrix $\mathbf{A}$ and cluster membership matrix $\mathbf{S}$. However, elements of this latter matrix are binary: $s_{jk} = 1$ if $j \in C_k$ and 0 otherwise. These memberships are therefore less informative than AdaCLV cluster memberships.

4.1. Clustering around Latent Variables (CLV) [8]

Clustering around Latent Variables (CLV) [8] was introduced for the purpose of data exploration through simultaneous dimensionality reduction and variable clustering. Vigneau et al. [8] proposed two different modes for CLV: local and directional mode. The objective function for the local mode is based on covariances, whereas it is based on squared covariances for the directional mode. This means that when the criterion for the directional mode is maximized, only the magnitude of covariances is considered, not their signs.

For NMR spectra representing mixtures, the directional mode may not always be appropriate. This is because variables are generally positively correlated when they represent the same molecule (i.e. belong to the same cluster) but can be somewhat negatively correlated when they represent different molecules

(i.e. belong to different clusters). Thus, the sign of pairwise variable covariances provides useful information about the cluster membership of variables. While the local mode can capitalize on this information, the directional mode cannot.

For the local mode, the matrices $\mathbf{A}$ and $\mathbf{S}$ are found by maximizing the function

$$J(\mathbf{A}, \mathbf{S}) = \sum_{k=1}^K \sum_{j=1}^p s_{jk} \text{cov}(\mathbf{a}_k, \mathbf{x}_j) = \frac{1}{n-1} \sum_{k=1}^K \mathbf{a}_k^\top \mathbf{X} \mathbf{s}_k \quad (6)$$

with respect to $\mathbf{A}$ and $\mathbf{S}$, subject to the constraints

1. $s_{jk} \in \{0, 1\}, \forall j, k$ (binary memberships)
2. $\sum_{k=1}^K s_{jk} = 1, \forall j$ (each variable assigned to one cluster)
3. $\text{var}(\mathbf{a}_k) = 1, \forall k$ (latent variables have unit variance)

Variable clusters are a function of the membership values and vice versa: $C_k := \{j \mid s_{jk} = 1\}$ and $s_{jk} = \mathbb{I}[j \in C_k]$. Given a binary membership vector $\mathbf{s}_k$, the optimal solution $\mathbf{a}_k$ to this constrained problem is the same as for AdaCLV (see Eq. (3)). For a given latent component $\mathbf{a}_k$, the function in Eq. (6) is maximized when $s_{jk} = \mathbb{I}\left[k = \arg \max_{\ell} \text{cov}(\mathbf{a}_{\ell}, \mathbf{x}_j)\right]$.

The matrices $\mathbf{A}$ and $\mathbf{S}$ are estimated using alternating optimization. Starting from an initial set of variable clusters $C_1^{(0)}, \dots, C_K^{(0)}$ and membership values $s_{jk}^{(0)} = \mathbb{I}[j \in C_k^{(0)}]$, the following steps are repeated for $t = 1$ until convergence:

1. Calculate $\mathbf{a}_k^{(t)} = \sqrt{n-1} \frac{\mathbf{X} \mathbf{s}_k^{(t-1)}}{\|\mathbf{X} \mathbf{s}_k^{(t-1)}\|_2}$ for all k
2. Update cluster assignments for all k s.t.
 $C_k^{(t)} := \{j \mid k = \arg \max_{\ell} \text{cov}(\mathbf{a}_{\ell}^{(t)}, \mathbf{x}_j)\}$
3. Update memberships for all j and k s.t. $s_{jk}^{(t)} = \mathbb{I}[j \in C_k^{(t)}]$

At iteration t , the only variables that contribute to the estimation of latent variable $\mathbf{a}_k^{(t)}$ are the variables j for which $k = \arg \max_{\ell} \text{cov}(\mathbf{a}_{\ell}^{(t-1)}, \mathbf{x}_j)$ (i.e.

$s_{jk}^{(t-1)} = 1$). Thus, each variable contributes to the estimation of only one of the K latent variables (see Table 1 for a summary).

The directional mode of CLV is summarized in Appendix A.

4.2. CLV $K + 1$ [9]

Vigneau et al. [9] proposed CLV $K + 1$ to iteratively filter variables so that variables representing noise (“garbage” variables) do not contribute to the estimation of $\mathbf{A}$ and $\mathbf{S}$. Let $\gamma \in [0, 1]$ be a user-defined hyperparameter for filtering variables. For the local mode, CLV $K + 1$ maximizes the modified CLV objective function

$$J(\mathbf{A}, \mathbf{S}) = \sum_{j=1}^p \left\{ \sum_{k=1}^K s_{jk} \text{cov}(\mathbf{a}_k, \mathbf{x}_j) + \gamma \left(1 - \sum_{k=1}^K s_{jk} \right) \sqrt{\text{var}(\mathbf{x}_j)} \right\} \quad (7)$$

with respect to $\mathbf{A}$ and $\mathbf{S}$, subject to the same constraints as CLV, except that constraint 2 ($\sum_{k=1}^K s_{jk} = 1, \forall j$) becomes $\sum_{k=1}^K s_{jk} \leq 1, \forall j$, to allow for the presence of a garbage cluster.

For a given element j in the sum in Eq. (7), only the first or second term inside the curly brackets can be activated at once. The second term is activated if $s_{jk} = 0$ for all $\{1, \dots, K\}$, which means that variable j is assigned to cluster $K + 1$ (the “garbage” cluster). This solution maximizes the constrained problem when, for all k , $\text{cor}(\mathbf{a}_k, \mathbf{x}_j) < \gamma$. The first term in Eq. (7) is only activated if there is a $\{1, \dots, K\}$ for which $\text{cor}(\mathbf{a}_k, \mathbf{x}_j) > \gamma$. If this is the case, the solution s_{jk} maximizing the constrained problem is the same as for CLV.

The hyperparameter γ controls the number of variables that are assigned to the “garbage” cluster: as $\gamma \rightarrow 1$, more variables satisfy the condition $\text{cor}(\mathbf{a}_k, \mathbf{x}_j) < \gamma$ for all k , so more variables are assigned to the garbage cluster. Note that the threshold γ is the same for each cluster k , which is in contrast to AdaCLV, where the membership threshold is different for each cluster.

The CLV $K + 1$ algorithm is the same as the CLV algorithm (see Section 4.1), except that the cluster assignment step (Step 2) is modified and a third

265 step is added for assigning variables to a garbage cluster: $C_{K+1}^{(t)} := \{j \mid j \notin C_1^{(t)}, \dots, C_K^{(t)}\}$. Compared to CLV, fewer variables contribute to the estimation of each latent variable because the conditions for contribution are stricter (see Table 1). As a result, some variables are not used to estimate any of the latent variables.

270 The directional mode of CLV $K + 1$ is explained in Appendix B.

4.3. CLV Sparse LV [9]

Another method proposed by [9] for addressing noise variables is CLV Sparse LV, which combines elements from CLV and CLV $K + 1$ (see Table 1). As with CLV $K + 1$, CLV Sparse LV modifies the estimation of latent variables $\mathbf{a}_k$ with
 275 respect to the original CLV method. However, cluster memberships s_{jk} are estimated using the original CLV approach, meaning that there is no garbage cluster C_{K+1} .

For the estimation of latent variables, a vector of sparse loadings $\mathbf{b}_k$ is introduced so that each latent variable k is estimated using only a subset of the variables in cluster k . Each sparse loading is initialized such that $b_{jk}^{(0)} = \mathbb{I}[j \in C_k^{(0)}]$. Then, for each iteration t , latent variables are updated such that $\mathbf{a}_k^{(t)} = \sqrt{n-1} \mathbf{X} \mathbf{b}_k^{(t-1)} / \|\mathbf{X} \mathbf{b}_k^{(t-1)}\|_2$, and sparse loadings are updated such that

$$b_{jk}^{(t)} = \mathbb{I} \left[k = \arg \max_{\ell} \text{cov}(\mathbf{a}_{\ell}^{(t)}, \mathbf{x}_j) \text{ AND } \text{cor}(\mathbf{a}_k^{(t)}, \mathbf{x}_j) > \gamma \right]. \quad (8)$$

Note that the function in Eq. (8) is the same as the one used for the memberships $s_{jk}^{(t)}$ in CLV $K + 1$. At convergence ($t = T$), the final latent variables are
 280 re-calculated using the final sparse loadings $\mathbf{b}_k^{(T)}$. As with CLV $K + 1$, the threshold parameter γ is the same for each cluster, allowing less flexibility than AdaCLV.

For details concerning the directional mode of CLV Sparse LV, see Appendix C.

	$\mathbf{x}_j$ contributes to $\mathbf{a}_k^{(t)}$ if	$s_{jk}^{(t)} > 0$ if
CLV	$k = \arg \max_{\ell} \text{cov}(\mathbf{a}_{\ell}^{(t-1)}, \mathbf{x}_j)$	$k = \arg \max_{\ell} \text{cov}(\mathbf{a}_{\ell}^{(t)}, \mathbf{x}_j)$
CLV Sparse LV	$k = \arg \max_{\ell} \text{cov}(\mathbf{a}_{\ell}^{(t-1)}, \mathbf{x}_j)$	
CLV $K + 1$	AND $\text{cor}(\mathbf{a}_k^{(t-1)}, \mathbf{x}_j) > \gamma$	$k = \arg \max_{\ell} \text{cov}(\mathbf{a}_{\ell}^{(t)}, \mathbf{x}_j)$ AND $\text{cor}(\mathbf{a}_k^{(t)}, \mathbf{x}_j) > \gamma$
AdaCLV	$\text{cov}(\mathbf{a}_k^{(t-1)}, \mathbf{x}_j) > \tau_k^{(t-1)}$	$\text{cov}(\mathbf{a}_k^{(t)}, \mathbf{x}_j) > \tau_k^{(t)}$

Table 1: Latent variable and cluster membership estimation for AdaCLV and CLV, CLV $K + 1$ and CLV Sparse LV (local mode). CLV Sparse LV is the same as CLV $K + 1$ for latent variable updates (first column) and the same as CLV for cluster membership updates (second column).

285 4.4. Initialization

All existing CLV algorithms require the input of an initial partition of variables, which means that a variable clustering algorithm must be run prior to performing CLV. Vigneau et al. [8] proposed a hierarchical agglomerative clustering algorithm for finding an initialization for the local and directional modes
290 of CLV. At each iteration, this algorithm merges the pair of clusters that incurs the smallest added cost with respect to the CLV objective function.

Let $\mathbf{X}_A$ ($n \times p_A$) (resp. $\mathbf{X}_B$ ($n \times p_B$)) be the matrix of variables in $\mathbf{X}$ that belong to cluster A (resp. B) and $\bar{\mathbf{x}}_A$ (resp. $\bar{\mathbf{x}}_B$) be an $n \times 1$ vector containing the means of each row in $\mathbf{X}_A$ (resp. $\mathbf{X}_B$). For the local mode of CLV, the added cost of merging two clusters A and B , with respect to the CLV objective function, is

$$\Delta = p_A \text{sd}(\bar{\mathbf{x}}_A) + p_B \text{sd}(\bar{\mathbf{x}}_B) - p_{A \cup B} \text{sd}(\bar{\mathbf{x}}_{A \cup B}). \quad (9)$$

Starting from $K = p$ clusters, clusters are merged based on the added cost function Δ until only one cluster remains, resulting in a sequence of partitions with $K = p, \dots, 1$ cluster(s). Each value of K is thus associated with a unique
295 partition from this hierarchical sequence of partitions. The user must choose one of these as the initial partition for CLV, i.e. choose the number K of variable clusters that should be detected in the dataset.

More details concerning the local cost function Δ can be found in Appendix D.

300 5. Experiments

Two experiments were performed to compare the performance of the proposed method, AdaCLV, to that of CLV, CLV $K + 1$ and CLV Sparse LV, as well as two baseline methods (K-means and Ward’s hierarchical clustering). In the first experiment, these methods were applied to a real NMR dataset with non-
305 overlapping clusters, “CitrateHippurate” (mean spectrum shown in Figure 1). The estimated cluster memberships were assessed visually. In the second experiment, 500 semi-artificial NMR datasets with overlapping clusters were used. For this experiment, the true cluster memberships and latent variables for the target clusters were known by design, so they were compared to the estimated
310 memberships and latent variables for each method.

In both experiments, the true target clusters were defined as highly positively correlated variables that varied significantly between samples. For the semi-artificial dataset, these clusters were simulated, so they were known by design. For the CitrateHippurate dataset, *a priori* knowledge concerning the peak lo-
315 cations of the varying molecules was used to define the true clusters.

5.1. Data

5.1.1. CitrateHippurate Dataset

The CitrateHippurate dataset used here was drawn from a database of 224 ^1H -NMR spectra of rat urine, which is described in detail in [12] and [13]. The
320 samples measured for this database were collected from a homogeneous pool of rat urine and then spiked with different concentrations of two molecules: citrate and hippurate. All samples were prepared and frozen on the same day but were only thawed and measured later, at different time points. The spectral FID were preprocessed using classical preprocessing tools available in TopSpin from

325 Bruker¹, and then all spectra were normalized to unit sum.

A subset of $n = 34$ observations from this dataset was retained for the present experiment, referred to here as the “CitrateHippurate” dataset. These observations vary according to three factors: 1) the day of measurement (day 2 or day 3), 2) the concentration of added citrate (0, 2 or 4 mM) and 3) the concentration of added hippurate (0, 1 or 2 mM). For each combination of factor levels, two replicates were performed, resulting in 36 spectra. Two observations from day 3 were detected as outliers and removed, as was done in [13]. Variables in low-signal areas and the water zone (variables 1 – 99, 206 – 354 and 501 – 600), as well as all zero-variance variables, were also removed. Finally, 335 the remaining variables were mean-centered. Visualizations of the CitrateHippurate dataset can be found in Figure 2, as well as a table summarizing the experimental design.

Figure 2: **PCA and design for the CitrateHippurate dataset.** The PCA scores (a) and loadings (b) plots show that the first two components represent the concentration of the molecules citrate and hippurate. Moreover, most of the variability between samples (80.29%) emanates from differences in these concentrations. The number of samples for each experimental design condition are presented in (c).

Because citrate and hippurate are the only molecules that vary in concentration between samples, the true number of target variable clusters in this dataset is

¹<https://www.bruker.com/products/mr/nmr/software/topspin.html>

340 defined as $K_{true} = 2$, one cluster for each spiked molecule. The peaks for citrate and hippurate do not overlap each other (see Figure 1).

5.1.2. Semi-Artificial Datasets

The data used to generate semi-artificial NMR spectra came from a real ^1H -NMR database from the COMET project [14] containing over 800 spectral FID
345 of urine from healthy control rats. For the present experiment, 400 of these FID were selected, converted to the frequency domain and preprocessed (water suppression, apodization, baseline correction, peak alignment and bucketing), resulting in a spectral dataset, referred to here as the “COMET dataset,” with $n = 400$ observations and $p = 500$ variables (see [12] for more details). Finally,
350 zero-variance variables were removed, and then variance stabilizing normalization [15, 16] was performed in order to reduce mean-dependent variance in the spectra. The latter transformation makes the variance in low intensity areas more similar to the variance in high intensity areas, making the spectra more homogeneous.

355 Semi-artificial spectra with overlapping peaks were generated by adding artificial peaks to spectra from the COMET dataset, which served as a realistic “background.” The artificial peaks represented $M = 6$ artificial molecules, and the variables in these peaks constituted the $M = K_{true} = 6$ target clusters to be detected. Two factors were varied when generating these artificial datasets:
360 the sample size $n \in \{25, 50, 100, 200, 400\}$ and the proportion of total spectral variability emanating from the artificial peaks (e.g. the target clusters), denoted with the parameter $\theta \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75\}$. This latter factor represented the relative amount of relevant signal in the dataset. For each combination of levels of these two factors, 20 datasets were generated, resulting
365 in $5 \times 5 \times 20 = 500$ total datasets. For each dataset, n spectra were sampled randomly without replacement from the set of 400 real COMET spectra. Then, artificial peaks with randomly generated peak heights were added to each sampled COMET spectrum.

The procedure for generating artificial peaks was the following: first, for each
 370 spectrum i in the real dataset, $M = 6$ molecule concentrations were uniformly
 sampled from an M -dimensional simplex. Next, eight peaks (two tuples and
 four singlets) were generated, with fixed peak width and with peak heights that
 were proportional to the randomly generated concentrations. The peak centers
 were evenly spaced in two regions of the spectrum, one region with five peaks
 375 and the other with three. A fully artificial spectrum $\mathbf{z}_i^{\text{art}}$ was then calculated
 by summing over the spectra for each artificial peak: $\mathbf{z}_i^{\text{art}} = \sum_{\ell=1}^8 \mathbf{z}_{i\ell}^{\text{art}}$, where ℓ
 is the artificial peak index. More details concerning the generation of the fully
 artificial spectra can be found in Appendix E.

Each semi-artificial spectrum $\mathbf{z}_i$ was generated by calculating the weighted sum
 of a real, experimental spectrum ($\mathbf{z}_i^{\text{exp}}$) and an artificial spectrum ($\mathbf{z}_i^{\text{art}}$) con-
 taining synthetic peaks (see Figure 3):

$$\mathbf{z}_i = \alpha_{\text{exp}} \mathbf{z}_i^{\text{exp}} + \alpha_{\text{art}} \mathbf{z}_i^{\text{art}}, \quad (10)$$

where α_{exp} and α_{art} are normalizing constants. These normalizing constants
 380 were used to control the proportion θ of the total spectral variability emanating
 from the artificial peaks (see Appendix E for more details). For all θ values
 tested here, at least half of all inter-sample variability can be attributed to
 the eight simulated peaks, which represent $M = 6$ artificial molecules. The
 $K_{\text{true}} = 6$ variable clusters to detect are thus the groups of variables in these
 385 simulated peaks.

Finally, the spectra within each dataset were normalized to unit sum, and then
 their variables were mean-centered. PCA scores and loadings plots of an exam-
 ple dataset are shown in Figure 4.

5.2. Initialization and Hyperparameter Settings for the Clustering Algorithms

390 For K-means and the CLV methods, initial variable clusters were found using the
 local mode of the hierarchical algorithm proposed in [8], for a number of clusters
 K in the range $[2, 12]$. For the semi-artificial datasets, the same set of initial

Figure 3: **Generation of semi-artificial spectra.** The first panel shows the mean spectrum of a real spectral dataset ($n = 400$) and the second represents the mean spectrum of the artificial peaks added to the real spectra. The third panel indicates the true cluster membership degrees for the artificial clusters, where each true membership vector $\mathbf{s}_k$ is plotted in one of six different colors and line types, and line height is proportional to membership degree.

Dataset Type	θ	n	p	K_{true}	Datasets
CitrateHippurate Spectra	-	34	250	2	1
Semi-Artificial Spectra	$[0.5, 0.75]$	$[25, 400]$	424	6	20

Table 2: Characteristics of the datasets used for evaluation. Note that p represents the number of variables after removal of zero-variance variables, and K_{true} is the true number of clusters to detect. For the Semi-Artificial Spectra experiment, 20 datasets were generated for each setting (i.e. each possible combination of the sample sizes $n \in \{25, 50, 100, 200, 400\}$ and levels of signal $\theta \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75\}$).

Figure 4: **PCA of a semi-artificial dataset.** Scores (a-d) and loadings (e-h) plots for one of the semi-artificial datasets with $n = 400$ and $\theta = 0.65$. The scores plots reflect the study design, in which peaks representing $M = 6$ artificial molecules were added to real NMR spectra. The concentrations of these artificial molecules were uniformly sampled from an M -dimensional simplex, which can be seen in the first five PCA dimensions. In the loadings plots, the variables associated with each different artificial molecule are highlighted with different colors, whereas the variables representing background noise are shown in black. The variables in the artificial peaks explain the first five PCA dimensions, which amounts to about 65.68% of the total variability. This is very close to the expected percentage, $\theta \times 100\% = 65\%$.

clusters was used for all datasets that shared the same level of signal θ , regardless of the sample size n . For each level of signal θ , these initial clusters were found
 395 based on one of the largest datasets ($n = 400$). For the CitrateHippurate dataset, all observations were used to generate the initial partition.

The distance measure used for K-means and Ward’s hierarchical clustering algorithm was the Euclidean distance. For CLV $K + 1$, CLV Sparse LV and AdaCLV, a sequence of hyperparameters γ and η were tested for each dataset
 400 and number K . These hyperparameters were proportions (0.05, 0.1, ..., 0.95) of the maximum hyperparameter value, γ_{max} or η_{max} . For CLV $K + 1$ and CLV Sparse LV, $\gamma_{max} = 1$ by definition. For AdaCLV, η_{max} was determined based on the first iteration of the algorithm, as explained in Section 3.

5.3. Calculation of Ground-Truth Matrices $\mathbf{A}$ and $\mathbf{S}$

5.3.1. CitrateHippurate Dataset 405

For the CitrateHippurate dataset, we were interested in detecting the $K_{true} = 2$ molecules that varied significantly between samples. The true latent variable and cluster membership matrices $\mathbf{A}$ ($n \times 2$) and $\mathbf{S}$ ($p \times 2$) were not known by design. Rather, these matrices were approximated based on *a priori* knowledge about the location of citrate and hippurate in the spectra. Let $\tilde{C}_1$ be the set of p_1 variable indexes associated with citrate and $\tilde{C}_2$ the set of p_2 indexes for hippurate. Let $\mathbf{Z}$ ($n \times p$) represent the data matrix before column centering, $\bar{z}_j$ be the sample mean of variable j in $\mathbf{Z}$ and $\bar{\mathbf{z}}_k$ ($p_k \times 1$) be the vector containing means $\bar{z}_j$ for all $j \in \tilde{C}_k$. The approximate ground-truth for cluster membership vector matrix $\mathbf{S}$ was calculated such that

$$s_{jk} = \mathbb{I} \left[j \in \tilde{C}_k \right] \bar{z}_j / \|\bar{\mathbf{z}}_k\|_2. \quad (11)$$

The approximate ground-truth for the latent variable matrix $\mathbf{A}$ was calculated based on the membership matrix $\mathbf{S}$. Let $\mathbf{X}$ be the column-centered data matrix. The true latent variable k was calculated by normalizing the vector $\mathbf{X}\mathbf{s}_k$ ($n \times 1$):

$$\mathbf{a}_k = \frac{\mathbf{X}\mathbf{s}_k - \text{mean}(\mathbf{X}\mathbf{s}_k)}{\text{sd}(\mathbf{X}\mathbf{s}_k)}. \quad (12)$$

5.3.2. Semi-Artificial Datasets

For the Semi-Artificial datasets, we were interested in detecting the synthetic peaks, i.e. the $M = K_{true} = 6$ molecules that contributed a proportion θ of total variability between samples. The true spectral profile matrix $\mathbf{S}$ ($p \times 6$) was defined such that each column k of this matrix was proportional to the mean spectrum for molecule k :

$$\mathbf{s}_k = \frac{\frac{1}{n} \sum_{i=1}^n \mathbf{z}_{ik}^{\text{art}}}{\|\frac{1}{n} \sum_{i=1}^n \mathbf{z}_{ik}^{\text{art}}\|_2} \quad (13)$$

The true latent variables for these datasets were calculated based on the concentration matrix $\mathbf{C}$ ($n \times 6$) used to generate the semi-artificial spectra. Each true latent variable $\mathbf{a}_k$ was found by normalizing the corresponding column $\mathbf{c}_k$ ($n \times 1$) in $\mathbf{C}$ as follows:

$$\mathbf{a}_k = \frac{\mathbf{c}_k - \text{mean}(\mathbf{c}_k)}{\text{sd}(\mathbf{c}_k)}. \quad (14)$$

5.4. Evaluation Criteria

Let $\hat{\mathbf{S}}$ be a matrix of estimated cluster memberships and $\hat{\mathbf{A}}$ be a matrix of estimated latent variables. While both of these matrices are outputs of the CLV algorithms, they are not automatically generated for the baseline methods, K-means and Ward’s hierarchical clustering. Thus, for the latter two methods, $\hat{\mathbf{A}}$ was defined as the matrix of normalized cluster centroids and $\hat{\mathbf{S}}$ as a binary matrix calculated based on the estimated variable partition $\hat{C}_1, \dots, \hat{C}_K$.

For each estimated latent variable $\hat{\mathbf{a}}_k$, a sparse vector $\hat{\mathbf{r}}_k^W$ of within-cluster covariances was also calculated and stored in a $p \times K$ matrix $\hat{\mathbf{R}}^W$, where $\hat{r}_{jk}^W = \text{cov}(\hat{\mathbf{a}}_k, \mathbf{x}_j) \mathbb{I}[\hat{s}_{jk} > 0]$. The matrix $\hat{\mathbf{R}}^W$ summarizes the information contained in $\hat{\mathbf{A}}$ and $\hat{\mathbf{S}}$ and makes it easier to compare AdaCLV to methods that estimate binary cluster memberships. Indeed, cluster assignments ($s_{jk} > 0$ vs

$s_{jk} = 0$), rather than exact membership degrees, are used to calculate this matrix.

430 Let the matrix $\mathbf{Y}$ represent a ground-truth matrix ($\mathbf{A}$, $\mathbf{S}$ or $\mathbf{R}^W$) and $\hat{\mathbf{Y}}$ represent an estimated matrix ($\hat{\mathbf{A}}$, $\hat{\mathbf{S}}$ or $\hat{\mathbf{R}}^W$). In this experiment, we evaluated the *precision* of each method, i.e. the proportion of information in $\hat{\mathbf{Y}}$ that corresponds to information in $\mathbf{Y}$, as well as the *recall* of each method, i.e. the proportion of information in $\mathbf{Y}$ that is represented in $\hat{\mathbf{Y}}$. To appreciate the importance
 435 of both measures, consider the case where the ground-truth $\mathbf{Y} = [\mathbf{a} \ \mathbf{b} \ \mathbf{c}]$, and there are two estimated matrices, $\hat{\mathbf{Y}}_1 = [\mathbf{a} \ \mathbf{a} \ \mathbf{b}]$ and $\hat{\mathbf{Y}}_2 = [\mathbf{a} \ \mathbf{b} \ \mathbf{c} \ \mathbf{d}]$. The estimated matrix $\hat{\mathbf{Y}}_1$ has perfect precision but is missing the vector $\mathbf{c}$ from the ground-truth matrix (imperfect recall). $\hat{\mathbf{Y}}_2$, on the other hand, has perfect recall but includes a vector $\mathbf{d}$ that does not appear in the ground-truth matrix
 440 (imperfect precision).

The precision and recall of an estimated matrix $\hat{\mathbf{Y}}$ was calculated as a function of the Pearson correlation between columns of the estimated and true matrices. Let $\mathbf{y}$ (resp. $\hat{\mathbf{y}}$) be a true (resp. estimated) latent variable, cluster membership or within-cluster covariance vector. A similarity function was defined as follows:

$$\text{Sim}(\mathbf{y}, \hat{\mathbf{y}}) = [\text{cor}(\mathbf{y}, \hat{\mathbf{y}}) + 1] / 2 \in [0, 1] \quad (15)$$

This function evaluates to 1 if the two vectors are maximally similar, 0.5 if they are uncorrelated (perpendicular) and 0 if they are perfectly negatively correlated. If $\hat{\mathbf{y}}$ captures all of the relevant information in $\mathbf{y}$, $\text{Sim}(\mathbf{y}, \hat{\mathbf{y}}) = 1$.

For the precision, each non-zero column k of the estimated matrix $\hat{\mathbf{Y}}$ was matched with the most similar column in the true matrix $\mathbf{Y}$, denoted with the index k^* , based on the similarity function in Eq. (15). The precision was then calculated as the average similarity for the two matrices:

$$\text{Precision}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{K} \sum_{k=1}^K \text{Sim}(\mathbf{y}_{k^*}, \hat{\mathbf{y}}_k), \quad (16)$$

where K is the total number of non-zero columns in $\hat{\mathbf{Y}}$ and

$$k^* = \arg \max_{\ell} \text{Sim}(\mathbf{y}_{\ell}, \hat{\mathbf{y}}_k). \quad (17)$$

445 For matrices $\hat{\mathbf{Y}}$ with no non-zero columns, $K = 0$, so the precision is not defined.

For the recall, the similarity function was also used to match columns in the two matrices, except that each column m of the true matrix $\mathbf{Y}$ was matched with the most similar column in the estimated matrix $\hat{\mathbf{Y}}$, denoted with the index m^* . The recall was then calculated as the average similarity for the two matrices:

$$\text{Recall}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{M} \sum_{m=1}^M \text{Sim}(\mathbf{y}_m, \hat{\mathbf{y}}_{m^*}), \quad (18)$$

where

$$m^* = \arg \max_{\ell} \text{Sim}(\mathbf{y}_m, \hat{\mathbf{y}}_{\ell}). \quad (19)$$

The precision and recall evaluate the relevance of $\hat{\mathbf{Y}}$ by measuring two distinct and potentially diverging qualities. The overall quality of $\hat{\mathbf{Y}}$ can be summarized in a single value with the F_1 Score [17], which is a harmonic mean of the precision
450 and recall, with equal weight given to each measure:

$$F_1(\mathbf{Y}, \hat{\mathbf{Y}}) = \left(\frac{\text{Precision}(\mathbf{Y}, \hat{\mathbf{Y}})^{-1} + \text{Recall}(\mathbf{Y}, \hat{\mathbf{Y}})^{-1}}{2} \right)^{-1}. \quad (20)$$

This score should be maximized, where the maximum value of 1 indicates perfect precision and recall.

5.5. Results

5.5.1. CitrateHippurate Dataset ($K_{\text{true}} = 2$)

455 For the CitrateHippurate dataset, the true latent variable matrix $\mathbf{A}$ and cluster membership matrix $\mathbf{S}$ are not known by design, so they were approximated based on *a priori* knowledge about the location of citrate and hippurate in the spectra (see Section 5.3 for more details). The cluster membership estimates

Figure 5: **Optimal memberships for the CitrateHippurate dataset and for $K = K_{true} = 2$ and $K = 3$.** For binary memberships s_{jk} , the vertical bar at variable index j is shown in color k if $s_{jk} = 1$. White spaces correspond to variables that are assigned to none of the clusters. For non-binary memberships, each membership vector $\mathbf{s}_k$ is plotted in a different color, where line height is proportional to membership degree.

are shown for each method in Figure 5. For the methods that estimate binary
460 cluster memberships, a vertical bar is shown in color k at position j if $s_{jk} = 1$.
For non-binary cluster memberships, a colored line is drawn for each vector
 $\mathbf{s}_k$, where the line height is proportional to membership degree. For AdaCLV,
CLV $K + 1$ and CLV Sparse LV, only results for the optimal hyperparameter
 γ or η are shown here. For each number K of clusters, the optimal γ or η
465 for these methods was chosen based on the F_1 Score of the matrix of within-
cluster covariances, $\hat{\mathbf{R}}^W$. As explained in Section 5.4, the matrix $\hat{\mathbf{R}}^W$ encodes
information contained in both $\hat{\mathbf{A}}$ and $\hat{\mathbf{S}}$.

As seen in Figure 5, the only methods that provide satisfactory estimates of
the true cluster memberships for $K = 2$ are AdaCLV and CLV $K + 1$, the
470 methods that account for “garbage” variables. Both methods incorrectly assign
some noise variables to signal clusters (i.e. generate false positives). However,
the membership degrees for AdaCLV make it possible to interpret these false
positives with caution: small s_{jk} indicate low importance or a low likelihood of

truly belonging to the cluster. In contrast, the binary memberships estimated
 475 by CLV $K + 1$ do not provide this useful information, making misinterpretation
 of false positives more likely.

CLV Sparse LV, CLV, Ward’s Hierarchical clustering and K-means, which do
 not make use of a garbage cluster, achieve better performance when $K = 3$, as
 a third cluster can be used to group noise variables. While Ward’s Hierarchical
 480 clustering and K-means seem to perform better overall than CLV Sparse LV
 and CLV, all of these methods are still subject to error: Ward’s Hierarchical
 clustering and K-means miss one of the peaks for hippurate (in red) and CLV
 Sparse LV and CLV assign many noise variables to the citrate and hippurate
 clusters (in red and green). For $K = 3$, AdaCLV generates the same cluster
 485 memberships as for $K = 2$ (the third cluster is empty), suggesting that AdaCLV
 is relatively robust to the hyperparameter K . In contrast, CLV $K + 1$ erroneously
 assigns noise variables to a third cluster when $K = 3$.

5.5.2. Semi-Artificial Datasets ($K_{true} = 6$)

For the Semi-Artificial spectra, 20 datasets were generated for each sample
 490 size ($n \in \{25, 50, 100, 200, 400\}$) and proportion of artificial variability ($\theta \in$
 $\{0.5, 0.55, 0.6, 0.65, 0.7, 0.75\}$), so the average performance of the different meth-
 ods could be calculated and compared. Moreover, the true latent variable matrix
 $\mathbf{A}$ and cluster membership matrix $\mathbf{S}$ were known by design, making it possible
 to compare results to the ground-truth with numerical criteria.

495 As with the CitrateHippurate dataset, the following results are for the optimal
 hyperparameter settings for each dataset and value K . For all θ values, the
 ranking of the different methods is similar for all evaluation criteria (results
 not shown here). All methods generally benefit from higher values of θ (i.e.
 higher signal), especially for small sample sizes. However, for cluster member-
 500 ship estimation, increasing the amount of signal in the data does not improve
 the precision or recall of CLV or CLV Sparse LV.

For $\theta = 0.65$, the performance results for latent variable estimation ($\hat{\mathbf{A}}$ vs $\mathbf{A}$)

Figure 6: **Performance for latent variable estimation.** Precision($\mathbf{A}, \hat{\mathbf{A}}$) and Recall($\mathbf{A}, \hat{\mathbf{A}}$) for $\theta = 0.65$ and for different n and K , averaged over 20 repetitions.

are presented in Figure 6, for cluster membership estimation ($\hat{\mathbf{S}}$ vs $\mathbf{S}$) in Figure 7 and for within-cluster covariance ($\hat{\mathbf{R}}^W$ vs $\mathbf{R}^W$) in Figure 8. The precision and recall are plotted with respect to K for three of the five sample sizes n , and each method is represented with a different color and line type. The true number of target clusters, $K_{true} = 6$, is indicated with a vertical gray line.

For all K and n tested, AdaCLV has equivalent or better precision than the other methods and comparable recall for latent variable and cluster membership estimation, as well as within-cluster covariance. For $K \geq K_{true} = 6$, AdaCLV has near-perfect precision and recall for all criteria, especially as n increases. As was observed for the CitrateHippurate dataset, the worst precision and recall for cluster membership estimation is obtained by CLV and CLV Sparse LV. CLV and Ward’s Hierarchical clustering have the worst latent variable estimation performance overall.

Examples of cluster membership estimation for the different methods are shown in Figures 9 and 10 for $K = 6, 7$ and 10 and $\theta = 0.65$. For each method, the cluster memberships shown here are the ones that best represent the mean performance of the method in terms of within-cluster covariance. AdaCLV

Figure 7: **Performance for cluster membership estimation.** Precision($\mathbf{S}, \hat{\mathbf{S}}$) and Recall($\mathbf{S}, \hat{\mathbf{S}}$) for $\theta = 0.65$ and for different n and K , averaged over 20 repetitions.

Figure 8: **Performance for within-cluster covariance.** Precision($\mathbf{R}^W, \hat{\mathbf{R}}^W$) and Recall($\mathbf{R}^W, \hat{\mathbf{R}}^W$) for $\theta = 0.65$ and for different n and K , averaged over 20 repetitions.

estimates memberships that are very close to the true memberships for $K = 6, 7$ and 10, further demonstrating its robustness to the choice of the hyperparameter K . Indeed, when $K > K_{true}$, AdaCLV tends to find redundant clusters. For example, for $n = 400$ and $K = 7$, cluster 4 is found twice. For $K = 10$, cluster 3 is found three times and clusters 5 and 6 are found twice. Sometimes
 525 AdaCLV incorrectly assigns variables to signal clusters, but the low membership degrees for these variables indicate that they are less likely to be true members of these clusters. When the sample size n is large, fewer false positives occur and AdaCLV is able to detect more overlap between clusters.

For large sample sizes, CLV $K+1$ and K-means perform comparably with respect
 530 to AdaCLV when K is correctly chosen ($K = 6$ and 7 respectively), but their estimation of cluster memberships varies considerably with K . Compared to AdaCLV, both CLV $K+1$ and K-means are more sensitive to small sample sizes, assigning more noise variables to signal clusters. This is problematic because their memberships are binary, making it impossible to distinguish potential false
 535 positives from true positives. Finally, unlike AdaCLV and CLV $K+1$, K-means and Ward’s Hierarchical clustering do not assign variables in overlapping peak regions to signal clusters. This suggests that these two methods are less well suited to datasets with overlapping clusters.

For CLV $K+1$, CLV Sparse LV and AdaCLV, the results presented above were
 540 for the optimal choice of hyperparameters γ and η . To explore the effect of these hyperparameters on performance, the average F_1 scores for latent variable, cluster membership and within-cluster covariance estimation were plotted with respect to these hyperparameters (see the figures in Appendix F, which represent the datasets with $\theta = 0.65$). Overall, AdaCLV is least sensitive to
 545 η for cluster membership estimation. For all n and $K \in [4, 8]$, η values in the range $[0.05, 0.6] \times \eta_{max}$ yield cluster membership estimates that are better than those for the other two methods. AdaCLV is most sensitive to the choice of η for within-cluster covariance, but this sensitivity decreases as n increases. Even in the worst case scenario, 30% of the hyperparameter range for AdaCLV

Figure 9: **Examples of optimal memberships for semi-artificial datasets with $\theta = 0.65$ ($K = 6$ and $K = 7$).** Cluster membership estimation for $K = K_{true} = 6$ (top) and $K = 7$ (bottom) for $n = 25$ (left) and $n = 400$ (right). These plots may be interpreted in the same way as for those in Figure 5.

Figure 10: **Examples of optimal cluster memberships for semi-artificial datasets with $\theta = 0.65$ ($K = 10$).** Cluster membership estimation for $K = 10$ and $n = 25$ (left) and $n = 400$ (right). These plots may be interpreted in the same way as for those in Figure 5.

550 yields better within-cluster covariance estimates than the other methods, suggesting that AdaCLV’s superior performance is relatively robust to the choice of η .

6. Discussion and Conclusions

This paper presented a new method for simultaneous dimensionality reduction
555 and variable clustering, Adaptive Clustering around Latent Variables (AdaCLV). This method can facilitate the analysis of spectroscopic datasets, such as Nuclear Magnetic Resonance (NMR) spectra, which suffer from high dimensionality and complex correlation structures. Using a data-adaptive threshold, AdaCLV filters out variables that do not vary significantly between samples (i.e.
560 noise variables). For the remaining variables, cluster membership degrees are estimated, making it possible to identify potentially overlapping clusters and to rank the variables within a cluster in order of importance. Latent variables are estimated based on these membership degrees, summarizing the clusters in a low-dimensional representation that can be used for other tasks, such as classi-

565 fication. AdaCLV’s ability to adapt to clustering structures present in the data makes it a candidate approach for the discovery of new biomarkers.

AdaCLV is closely related to a group of multivariate methods in the Clustering around Latent Variables (CLV) family. However, existing CLV methods cannot find overlapping clusters, and they treat cluster membership as binary. 570 These memberships are less informative than membership degrees, making it difficult to assess the likelihood that a variable truly belongs to a given cluster. Moreover, the latent variables estimated based on these binary memberships are less representative of the true clusters. Indeed, the true clusters in NMR datasets are often overlapping, with some variables being more important than 575 others.

Several numerical experiments were conducted to compare AdaCLV to existing CLV methods (CLV, CLV $K + 1$ and CLV Sparse LV), as well as two baseline clustering methods (K-means and Ward’s hierarchical clustering). For NMR spectra with overlapping and non-overlapping clusters, and for all sample sizes 580 and levels of signal tested, AdaCLV estimated latent variables and cluster memberships with greater or equivalent precision with respect to the other methods. This was the case for all values tested for the hyperparameter K , the number of clusters to detect. In terms of recall, AdaCLV performed comparably to the other methods for all datasets. In contrast to the CLV methods, neither of the 585 baseline clustering methods assigned variables in regions of peak overlap to signal clusters, making these methods less attractive for datasets with overlapping clusters. While CLV $K + 1$ and K-means were able to estimate fairly accurate cluster memberships for some datasets, this depended on the sample size and the careful selection of hyperparameters. AdaCLV, on the other hand, was able 590 to estimate more interpretable and representative cluster memberships for a variety of hyperparameter choices, sample sizes and levels of signal, suggesting that it is more robust in practice.

While the choice of AdaCLV’s hyperparameter η had little impact on its performance, future works could nonetheless investigate the data characteristics that

595 make certain choices better than others. In addition, it would be interesting
to identify a data-dependent criterion for choosing AdaCLV’s hyperparameters
when the ground-truth is not available. The optimization of hyperparameters
was beyond the scope of this paper, but many candidate criteria for hyperparam-
eter optimization have been evaluated in the clustering literature [18, 19, 20].
600 These criteria measure clustering quality based on data characteristics, without
requiring knowledge of the ground-truth. Finally, another avenue for future
work would be to modify AdaCLV and the other CLV methods so that they no
longer depend on an initial partition of variables. AdaCLV would be a particu-
larly good candidate for this modification, as it is less sensitive to the number
605 of initial clusters provided.

Funding

The first author’s work was supported by the Belgian Fund for Scientific Re-
search (F.R.S.-FNRS, FRIA grant).

Appendices

610 Appendix A. CLV Directional Mode

The local mode for CLV was explained in Section 4.1. This appendix summa-
rizes the directional mode, where covariances are replaced with squared covari-
ances.

The matrices $\mathbf{A}$ and $\mathbf{S}$ are found by maximizing the function

$$\begin{aligned}
J(\mathbf{A}, \mathbf{S}) &= \sum_{k=1}^K \sum_{j=1}^p s_{jk} \text{cov}^2(\mathbf{a}_k, \mathbf{x}_j) \\
&= \frac{1}{(n-1)^2} \sum_{k=1}^K (\mathbf{a}_k^\top \mathbf{X} \mathbf{s}_k)^2
\end{aligned} \tag{A.1}$$

with respect to $\mathbf{A}$ and $\mathbf{S}$, subject to the same constraints as the local mode.

615 The solution for each latent variable $\mathbf{a}_k$ is the first standardized principal com-

ponent of $\mathbf{X}\text{diag}(\mathbf{s}_k)$, and the optimal cluster k is now defined based on squared covariances: $C_k := \{j \mid k = \arg \max_{\ell} \text{cov}^2(\mathbf{a}_{\ell}, \mathbf{x}_j)\}$.

Appendix B. CLV $K + 1$ Directional Mode

The local mode for CLV $K + 1$ was described in Section 4.2. This appendix
620 summarizes the directional mode, where covariances are replaced with squared covariances and $\sqrt{\text{var}(\mathbf{x}_j)}$ and the hyperparameter γ are squared.

The matrices $\mathbf{A}$ and $\mathbf{S}$ are found by maximizing the function

$$J(\mathbf{A}, \mathbf{S}) = \sum_{j=1}^p \left\{ \sum_{k=1}^K s_{jk} \text{cov}^2(\mathbf{a}_k, \mathbf{x}_j) + \gamma^2 \left(1 - \sum_{k=1}^K s_{jk} \right) \text{var}(\mathbf{x}_j) \right\} \quad (\text{B.1})$$

with respect to $\mathbf{A}$ and $\mathbf{S}$, subject to the same constraints as the local mode. The solution for each latent variable $\mathbf{a}_k$ is the same as for CLV (see Appendix A), and the optimal cluster k is now defined based on squared covariances
625 and correlations: $C_k := \{j \mid k = \arg \max_{\ell} \text{cov}^2(\mathbf{a}_{\ell}, \mathbf{x}_j) \text{ AND } \text{cor}^2(\mathbf{a}_k, \mathbf{x}_j) > \gamma^2\}$.

Appendix C. CLV Sparse LV Directional Mode

The local mode for CLV Sparse LV was summarized in Section 4.3. For the directional mode and for fixed $\mathbf{A}$, the matrix $\mathbf{S}$ is found by maximizing the
630 original CLV function with respect to $\mathbf{S}$, subject to the same constraints as the local mode. For fixed $\mathbf{S}$, the optimal latent variable $\mathbf{a}_k$ is estimated using a sparse loadings vector $\mathbf{b}_k$, which is calculated according to a soft-thresholding function. For more details about this procedure, see [9].

Appendix D. Initialization for CLV Local Mode

635 The initialization procedure proposed by [8] for the local mode of CLV was introduced in Section 4.4. This appendix provides theoretical justifications for the cost function Δ in Eq (9).

Let $\mathbf{X}_k$ ($n \times p_k$) be the matrix of variables in $\mathbf{X}$ that belong to cluster k , and let $\mathbf{x}_{jk}$ be column j in $\mathbf{X}_k$. Moreover, let $\bar{\mathbf{x}}_k$ ($n \times 1$) be a vector containing the means of each row in $\mathbf{X}_k$. For fixed clusters $C_1, \dots, C_K$, the local criterion in Eq. (6) can be rewritten as follows:

$$\begin{aligned} J(\mathbf{A}) &= \frac{1}{n-1} \sum_{k=1}^K \sum_{j=1}^{p_k} \mathbf{a}_k^\top \mathbf{x}_{jk} \\ &= \frac{1}{n-1} \sum_{k=1}^K p_k \mathbf{a}_k^\top \left(\frac{1}{p_k} \sum_{j=1}^{p_k} \mathbf{x}_{jk} \right) \\ &= \frac{1}{n-1} \sum_{k=1}^K p_k \mathbf{a}_k^\top \bar{\mathbf{x}}_k. \end{aligned} \quad (\text{D.1})$$

For a given k , the unit-variance latent variable $\mathbf{a}_k$ that maximizes this function is

$$\mathbf{a}_k = \sqrt{n-1} \frac{\bar{\mathbf{x}}_k}{\|\bar{\mathbf{x}}_k\|_2}. \quad (\text{D.2})$$

Thus, the maximum of the local criterion can be rewritten as follows:

$$\begin{aligned} \max_{\mathbf{A}} J(\mathbf{A}) &= \frac{1}{n-1} \sum_{k=1}^K p_k \left(\sqrt{n-1} \frac{\bar{\mathbf{x}}_k}{\|\bar{\mathbf{x}}_k\|_2} \right)^\top \bar{\mathbf{x}}_k \\ &= \sum_{k=1}^K \frac{p_k}{\sqrt{n-1}} \|\bar{\mathbf{x}}_k\|_2 \\ &= \sum_{k=1}^K p_k \text{sd}(\bar{\mathbf{x}}_k). \end{aligned} \quad (\text{D.3})$$

The added cost of merging two clusters A and B , with respect to this criterion, is then

$$\Delta = p_A \text{sd}(\bar{\mathbf{x}}_A) + p_B \text{sd}(\bar{\mathbf{x}}_B) - p_{A \cup B} \text{sd}(\bar{\mathbf{x}}_{A \cup B}). \quad (\text{D.4})$$

Appendix E. Data Generation for Semi-Artificial Spectra

640 The procedure for generating the Semi-Artificial datasets was briefly outlined in Section 5.1.2. This appendix provides further details about this procedure.

For each repetition $r : 1, \dots, 20$, sample size $n : 25, 50, 100, 200, 400$ and signal level $\theta : 0.5, 0.55, 0.6, 0.65, 0.7, 0.75$, spectral vectors $\mathbf{z}_i^{\text{art}}$ with artificial peaks were generated as follows. $M = 6$ molecule concentrations c_{im} were uniformly sampled from an M -dimensional simplex, then 8 peaks were generated based on these concentrations (2 tuples and 4 singlets). The peak heights were set to

$$\mathbf{h}_i = (c_{i1}, c_{i2}, c_{i1} \times 0.75, c_{i2} \times 0.75, c_{i3}, c_{i4}, c_{i5}, c_{i6}).$$

Peak centers $\nu_{0i\ell}$ were chosen so that evenly-spaced peaks appeared in two regions with five and three peaks respectively. Intensities $z_{ij\ell}^{\text{art}}$ for each peak ℓ , sample i and frequency ν_j were calculated using a Lorentzian function, as was done in [21]:

$$z_{ij\ell}^{\text{art}} = \frac{h_{i\ell} w^2}{w^2 + 4(\nu_j - \nu_{0i\ell})^2}, \quad (\text{E.1})$$

where w is a fixed peak width. Small intensities ($z_{ij\ell}^{\text{art}} < 0.01 h_{i\ell}$) were then set to zero. The pure spectrum for a given molecule m , $\mathbf{z}_{im}^{\text{art}}$, was calculated as the sum of all peak spectra associated with molecule m :

$$\mathbf{z}_{im}^{\text{art}} = \sum_{\ell \in \text{mol } m} \mathbf{z}_{i\ell}^{\text{art}}. \quad (\text{E.2})$$

The pure spectrum $\mathbf{z}_i^{\text{art}}$ was calculated as a sum over all molecules m :

$$\mathbf{z}_i^{\text{art}} = \sum_{m=1}^M \mathbf{z}_{im}^{\text{art}}. \quad (\text{E.3})$$

Let $\mathbf{z}_i^{\text{exp}}$ be the experimental spectrum for a given sample i and $\mathbf{z}_i^{\text{art}}$ be a spectrum containing artificial peaks. Each final spectrum $\mathbf{z}_i$ was generated by adding these two spectra,

$$\mathbf{z}_i = \alpha_{\text{exp}} \mathbf{z}_i^{\text{exp}} + \alpha_{\text{art}} \mathbf{z}_i^{\text{art}}, \quad (\text{E.4})$$

where α_{exp} and α_{art} are normalizing constants. These normalizing constants were used to control the proportion θ of the total spectral variability emanating from the artificial peaks. If $\alpha_{\text{exp}} = \sqrt{1 - \theta} / \sqrt{\sum_{j=1}^p \sigma_{j,\text{exp}}^2}$ and $\alpha_{\text{art}} = \sqrt{\theta} / \sqrt{\sum_{j=1}^p \sigma_{j,\text{art}}^2}$, then the total population variance of the final set of spectra, $\sum_{j=1}^p \sigma_j^2$, is equal to 1, where a fraction θ of this variance comes from the

artificial spectra.

$$\begin{aligned}
\sum_{j=1}^p \sigma_j^2 &= \alpha_{\text{exp}}^2 \sum_{j=1}^p \sigma_{j,\text{exp}}^2 + \alpha_{\text{art}}^2 \sum_{j=1}^p \sigma_{j,\text{art}}^2 \\
&= 1 - \theta + \theta \\
&= 1
\end{aligned} \tag{E.5}$$

In practice, α_{exp} and α_{art} were calculated using sample variances $\hat{\sigma}_{j,\text{exp}}^2$ and $\hat{\sigma}_{j,\text{art}}^2$.

645 Appendix F. Sensitivity Analysis for the Semi-Artificial Datasets

The overall performance (F_1 Score) of AdaCLV, CLV $K + 1$ and CLV Sparse LV for latent variable estimation (Figure F.11), cluster membership estimation (Figure F.12) and within-cluster covariance (Figure F.13) is presented here for $\theta = 0.65$. The mean performance for each method is plotted with respect to the value of the hyperparameter γ (for CLV $K + 1$ and CLV Sparse LV) or η (for
650 AdaCLV). When no clusters are found for a given hyperparameter setting, the corresponding F_1 score is undefined and thus does not appear in the plot.

References

- [1] K. Bingol, L. Bruschweiler-Li, D. Li, B. Zhang, M. Xie, R. Bruschweiler,
655 Emerging new strategies for successful metabolite identification in metabolomics, *Bioanalysis* 8 (6) (2016) 557–573.
- [2] J. Boccard, S. Rudaz, Harnessing the complexity of metabolomic data with chemometrics, *Journal of Chemometrics* 28 (1) (2014) 1–9.
- [3] J. L. Markley, R. Bruschweiler, A. S. Edison, H. R. Eghbalnia, R. Powers, D. Raftery, D. S. Wishart, The future of NMR-based metabolomics,
660 *Current opinion in biotechnology* 43 (2017) 34–40.
- [4] H. J. Muncey, R. Jones, M. De Iorio, T. M. Ebbels, Metassimulo: simulation of realistic NMR metabolic profiles, *BMC bioinformatics* 11 (1) (2010) 496.

Figure F.11: **Sensitivity for latent variable estimation.** F_1 score for $\theta = 0.65$ and for different n and hyperparameter values, averaged over 20 repetitions.

Figure F.12: **Sensitivity for cluster membership estimation.** F_1 score for $\theta = 0.65$ and for different n and hyperparameter values, averaged over 20 repetitions.

Figure F.13: **Sensitivity for within-cluster covariance.** F_1 score for $\theta = 0.65$ and for different n and hyperparameter values, averaged over 20 repetitions.

- 665 [5] J. C. Lindon, J. K. Nicholson, E. Holmes, J. R. Everett, Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids, Concepts in Magnetic Resonance: An Educational Journal 12 (5) (2000) 289–320.
- [6] K. H. Liland, Multivariate methods in metabolomics—from pre-processing to dimension reduction and statistical analysis, TrAC Trends in Analytical
670 Chemistry 30 (6) (2011) 827–841.
- [7] P. Bühlmann, P. Rütimann, S. van de Geer, C.-H. Zhang, Correlated variables in regression: clustering and sparse estimation, Journal of Statistical Planning and Inference 143 (11) (2013) 1835–1858.
- [8] E. Vigneau, E. Qannari, Clustering of variables around latent components,
675 Communications in Statistics-Simulation and Computation 32 (4) (2003) 1131–1150.
- [9] E. Vigneau, Dimensionality reduction by clustering of variables while set-

ting aside atypical variables, *Electronic Journal of Applied Statistical Analysis* 9 (1) (2016) 134–153.

- 680 [10] J. Nayak, B. Naik, H. Behera, Fuzzy c-means (FCM) clustering algorithm: a decade review from 2000 to 2014, in: *Computational intelligence in data mining-volume 2*, Springer, 2015, pp. 133–149.
- [11] S. Golotvin, A. Williams, Improved baseline recognition and modeling of FT NMR spectra, *Journal of Magnetic Resonance* 146 (1) (2000) 122–125.
- 685 [12] R. Rousseau, Statistical contribution to the analysis of metabonomics data in ¹H NMR spectroscopy, Université catholique de Louvain (2011).
- [13] S. Guisset, M. Martin, B. Govaerts, Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs, *Chemometrics and Intelligent Laboratory Systems* 184 (2019) 44–63.
- 690 [14] J. C. Lindon, J. K. Nicholson, E. Holmes, H. Antti, M. E. Bollard, H. Keun, O. Beckonert, T. M. Ebbels, M. D. Reily, D. Robertson, et al., Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project, *Toxicology and applied pharmacology* 187 (3) (2003) 137–146.
- 695 [15] B. P. Durbin, J. S. Hardin, D. M. Hawkins, D. M. Rocke, A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics* 18 (suppl_1) (2002) S105–S110.
- [16] H. M. Parsons, C. Ludwig, U. L. Günther, M. R. Viant, Improved classification accuracy in 1-and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation, *BMC bioinformatics* 8 (1) (2007) 234.
- 700 [17] N. Jardine, C. J. van Rijsbergen, The use of hierarchic clustering in information retrieval, *Information storage and retrieval* 7 (5) (1971) 217–240.

- 705 [18] M. Chiang, B. Mirkin, Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads, *Journal of Classification* 27 (1) (2010) 3–40.
- [19] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, I. Perona, An extensive comparative study of cluster validity indices, *Pattern Recognition* 710 46 (1) (2013) 243–256.
- [20] S. Kaczyńska, R. Marion, R. von Sachs, Comparison of cluster validity indices and decision rules for different degrees of cluster separation, in: *ESANN*, forthcoming.
- [21] O. Cloarec, M. E. Dumas, J. Trygg, A. Craig, R. H. Barton, J. C. Lindon, J. K. Nicholson, E. Holmes, Evaluation of the orthogonal projection 715 on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in ^1H NMR spectroscopic metabonomic studies, *Analytical chemistry* 77 (2) (2005) 517–526.