# Clustering with Statistical Error Control

Michael Vogt[1] Matthias Schmid[2]
University of Bonn University of Bonn

This paper presents a clustering approach that allows for rigorous statistical error control similar to a statistical test. We develop estimators for both the unknown number of clusters and the clusters themselves. The estimators depend on a tuning parameter $\alpha$ which is similar to the significance level of a statistical hypothesis test. By choosing $\alpha$, one can control the probability of overestimating the true number of clusters, while the probability of underestimation is asymptotically negligible. In addition, the probability that the estimated clusters differ from the true ones is controlled. In the theoretical part of the paper, formal versions of these statements on statistical error control are derived in a standard model setting with convex clusters. A simulation study and two applications to temperature and gene expression microarray data complement the theoretical analysis.

**Key words:** Cluster analysis; number of clusters; multiple statistical testing; statistical error control; $k$-means clustering.
**AMS 2010 subject classifications:** 62H30; 62H15; 62E20.

## 1 Introduction

In a wide range of applications, the aim is to cluster a large number of subjects into a small number of groups. Prominent examples are the clustering of genes in microarray analysis (Jiang et al., 2004), the clustering of temperature curves using data recorded on a spatial grid (Fovell and Fovell, 1993; DeGaetano, 2001), and the clustering of consumer profiles on the basis of survey data (Wedel and Kamakura, 2000).

A major challenge in cluster analysis is to estimate the unknown number of groups $K_0$ from a sample of data. A common approach is to compute a criterion function which measures the quality of the clustering for different cluster numbers $K$. An estimator of $K_0$ is then obtained by optimizing the criterion function over $K$. Prominent examples of this approach are the Hartigan index (Hartigan, 1975), the silhouette statistic (Rousseeuw, 1987) and the gap statistic (Tibshirani et al., 2001).

[1] Corresponding author. Address: Department of Economics and Hausdorff Center for Mathematics, University of Bonn, 53113 Bonn, Germany. Email: michael.vogt@uni-bonn.de.

[2] Address: Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, 53105 Bonn, Germany. matthias.schmid@imbie.uni-bonn.de.

Another common way to estimate $K_0$ is based on statistical test theory. Roughly speaking, one can distinguish between two types of test-based procedures: The *first type* relies on a statistical test which either checks whether some clusters can be merged or whether a cluster can be subdivided. Given a set of clusters, the test is repeatedly applied until no clusters can be merged or split any more. The number of remaining clusters serves as an estimator of $K_0$. Classical examples of methods that proceed in this way are discussed in Gordon (1999, Chapter 3.5) who terms them "local methods". Obviously, these methods involve a multiple testing problem. However, the employed critical values do not properly control for the fact that multiple tests are performed. The significance level $\alpha$ used to carry out the tests thus cannot be interpreted strictly. Put differently, the procedures do not allow for rigorous statistical error control.

Test-based approaches of the *second type* proceed by sequentially testing a model with $K$ clusters against one with $K+1$ clusters. The smallest number $K$ for which the test does not reject serves as an estimator of $K_0$. Most work in this direction has been done in the framework of Gaussian mixture models; see McLachlan and Rathnayake (2014) for an overview. However, deriving a general theory for testing a mixture with $K$ components against one with $K' > K$ components has turned out to be a very challenging problem; see Ghosh and Sen (1985) and Hartigan (1985) for a description of the main technical issues involved. Many results are therefore restricted to the special case of testing a homogeneous model against a mixture with $K = 2$ clusters; see Liu and Shao (2004) and Li et al. (2009) among many others. More general test procedures often lack a complete theoretical foundation or are based on very restrictive conditions.

Only recently, there have been some advances in developing a general theory for testing $K$ against $K' > K$ clusters under reasonably weak conditions. In a mixture model setup, Li and Chen (2010) and Chen et al. (2012) have constructed a new expectation-maximization (EM) procedure to approach this testing problem. Outside the mixture model context, Maitra et al. (2012) have developed a bootstrap procedure to test a model with $K$ groups against one with $K' > K$ groups. These papers derive the theoretical properties of the proposed tests under the null hypothesis of $K$ clusters, where $K$ is a pre-specified fixed number. However, they do not formally investigate the properties of a procedure which estimates $K_0$ by sequentially applying the tests. In particular, they do not analyze whether such a sequential procedure may allow for a rigorous interpretation of the significance level $\alpha$ that is used to carry out the tests.

The main contribution of this paper is to construct an estimator $\widehat{K}_0$ of $K_0$ which allows for rigorous statistical error control in the following sense: For any pre-specified significance level $\alpha \in (0,1)$, the proposed estimator $\widehat{K}_0 = \widehat{K}_0(\alpha)$ has the property that

$$\mathbb{P}\big(\widehat{K}_0 > K_0\big) = \alpha + o(1), \tag{1.1}$$

$$\mathbb{P}\big(\widehat{K}_0 < K_0\big) = o(1). \tag{1.2}$$

According to this, the probability of overestimating $K_0$ is controlled by the level $\alpha$, while the probability of underestimating $K_0$ is asymptotically negligible. By picking $\alpha$, we can thus control the probability of choosing too many clusters, while, on the other hand, we can ignore the probability of choosing too few clusters (at least asymptotically).

We show how to construct an estimator $\widehat{K}_0$ with the properties (1.1) and (1.2) in a standard model setting with convex clusters which is introduced in Section 2. Our estimation approach is developed in Section 3. As we will see, the proposed procedure does not only provide us with an estimator of $K_0$. It also yields estimators of the groups themselves which allow for statistical error control similarly to $\widehat{K}_0$. Our approach is based on the following general strategy:

(i) Construct a statistical test which, for any given number $K$, checks the null hypothesis that there are $K$ clusters in the data.

(ii) Starting with $K = 1$, sequentially apply this test until it does not reject the null hypothesis of $K$ clusters any more.

(iii) Define the estimator $\widehat{K}_0$ of $K_0$ as the smallest number $K$ for which the test does not reject the null.

This strategy is discussed in detail in Section 3.1. It is generic in the sense that it can be employed with different test statistics. For our theoretical analysis, we apply it with a specific statistic which is introduced in Section 3.2. For this specific choice, we derive the statements (1.1) and (1.2) on statistical error control under suitable regularity conditions. Some alternative choices of the test statistic are discussed in Section 6. In the following, we refer to our estimation procedure as *CluStErr* ("*Clu*stering with *St*atistical *Err*or Control").

The theoretical properties of our estimators, in particular the statements (1.1) and (1.2), are derived in Section 4. As we will see there, our theory is valid under quite general conditions. First of all, as opposed to many other studies from the clustering literature including those from a Gaussian mixture context, we do not restrict the random variables in our model to be Gaussian. For our theory to work, we merely require them to satisfy a set of moment conditions. Secondly, our approach is essentially free of tuning parameters, the only choice parameter being the significance level $\alpha$. Thirdly, to apply our method, we of course need to compute critical values for the underlying test. However, as opposed to other test-based methods, we do not have to estimate or bootstrap the critical values by a complicated procedure. They can rather be easily computed analytically. This makes our method particularly simple to implement in practice.

We complement the theoretical analysis of the paper by a simulation study and two applications on temperature and microarray data in Section 5. The R code to reproduce the numerical examples is contained in the add-on package **CluStErr** (Lasota et al., 2017), which implements the CluStErr method and which is part of the supplemental materials of the paper.

# 2 Model

Suppose we measure $p$ features on $n$ different subjects. In particular, for each subject $i \in \{1, \ldots, n\}$, we observe the vector $\boldsymbol{Y}_i = (Y_{i1} \ldots, Y_{ip})^\top$, where $Y_{ij}$ denotes the measurement of the $j$-th feature for the $i$-th subject. Our data sample thus has the form $\{\boldsymbol{Y}_i : 1 \leq i \leq n\}$. Both the number of subjects $n$ and the number of features $p$ are assumed to tend to infinity, with $n$ diverging much faster than $p$. This reflects the fact that $n$ is much larger than $p$ in the applications we have in mind. When clustering the genes in a typical microarray data set, for instance, the number of genes $n$ is usually a few thousands, whereas the number of tissue samples $p$ is not more than a few tenths. The exact technical conditions on the sizes of $n$ and $p$ are laid out in Section 4.1.

The data vectors $\boldsymbol{Y}_i$ of the various subjects $i = 1, \ldots, n$ are supposed to satisfy the model

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{e}_i, \tag{2.1}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})^\top$ is a deterministic signal vector and $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{ip})^\top$ is the noise vector. The subjects in our sample are assumed to belong to $K_0$ different classes. More specifically, the set of subjects $\{1, \ldots, n\}$ can be partitioned into $K_0$ groups $G_1, \ldots, G_{K_0}$ such that for each $k = 1, \ldots, K_0$,

$$\boldsymbol{\mu}_i = \boldsymbol{m}_k \quad \text{for all } i \in G_k, \tag{2.2}$$

where $\boldsymbol{m}_k \in \mathbb{R}^p$ are vectors with $\boldsymbol{m}_k \neq \boldsymbol{m}_{k'}$ for $k \neq k'$. Hence, the members of each group $G_k$ all have the same signal vector $\boldsymbol{m}_k$.

Equations (2.1) and (2.2) specify a model with convex spherical clusters which underlies the $k$-means and many other Euclidean distance-based clustering algorithms. This framework has been employed extensively in the literature and is useful in a wide range of applications, which is also illustrated by the examples in

Section 5. It is thus a suitable baseline model for developing our ideas on clustering with statistical error control. We now discuss the two model equations (2.1) and (2.2) in detail.

**Details on equation (2.1).** The noise vector $\boldsymbol{e}_i = (e_{i1}, \dots, e_{ip})^\top$ is assumed to consist of entries $e_{ij}$ with the additive component structure $e_{ij} = \alpha_i + \varepsilon_{ij}$. Equation (2.1) for the $i$-th subject thus writes as

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i, \tag{2.3}$$

where $\boldsymbol{\alpha}_i = (\alpha_i, \dots, \alpha_i)^\top$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{ip})^\top$. Here, $\alpha_i$ is a subject-specific random intercept term. Moreover, the terms $\varepsilon_{ij}$ are standard idiosyncratic noise variables with $\mathbb{E}[\varepsilon_{ij}] = 0$. We assume the error terms $\varepsilon_{ij}$ to be i.i.d. both across $i$ and $j$. The random intercepts $\alpha_i$, in contrast, are allowed to be dependent across subjects $i$ in an arbitrary way.

In general, the components of (2.3) may depend on the sample size $p$. The exact formulation of the model equation (2.3) for the $i$-th subject thus reads $\boldsymbol{Y}_{i,p} = \boldsymbol{\mu}_{i,p} + \boldsymbol{\alpha}_{i,p} + \boldsymbol{\varepsilon}_{i,p}$, where $\boldsymbol{Y}_{i,p} = (Y_{i1,p}, \dots, Y_{ip,p})^\top$, $\boldsymbol{\mu}_{i,p} = (\mu_{i1,p}, \dots, \mu_{ip,p})^\top$, $\boldsymbol{\alpha}_{i,p} = (\alpha_{i,p}, \dots, \alpha_{i,p})^\top$ and $\boldsymbol{\varepsilon}_{i,p} = (\varepsilon_{i1,p}, \dots, \varepsilon_{ip,p})^\top$. However, to keep the notation simple, we suppress this dependence on $p$ and write the model for the $i$-th subject as (2.3).

If we drop the random intercept $\boldsymbol{\alpha}_i$ from (2.3), the signal vector $\boldsymbol{\mu}_i$ is equal to the mean $\mathbb{E}[\boldsymbol{Y}_i]$. In the general equation (2.3) in contrast, $\boldsymbol{\mu}_i$ is only identified up to an additive constant. To identify $\boldsymbol{\mu}_i$ in (2.3), we impose the normalization constraint $p^{-1} \sum_{j=1}^p \mu_{ij} = 0$ for each $i$. We thus normalize the entries of $\boldsymbol{\mu}_i$ to be zero on average for each $i$. Under the technical conditions specified in Section 4.1, the constraint $p^{-1} \sum_{j=1}^p \mu_{ij} = 0$ implies that $\alpha_i = \lim_{p \to \infty} p^{-1} \sum_{j=1}^p Y_{ij}$ almost surely, which in turn identifies the signal vector $\boldsymbol{\mu}_i$.

**Details on equation (2.2).** This equation specifies the group structure in our model. We assume the number of groups $K_0$ to be fixed, implying that the groups $G_k = G_{k,n}$ depend on the sample size $n$. Keeping the number of classes $K_0$ fixed while letting the size of the classes $G_{k,n}$ grow is a reasonable assumption: It reflects the fact that in most applications, we expect the number of groups $K_0$ to be very small as compared to the total number of subjects $n$. To keep the notation simple, we suppress the dependence of the classes $G_{k,n}$ on the sample size $n$ and denote them by $G_k$ throughout the paper.

In the remainder of this section, we discuss two special cases of model (2.1)–(2.2) which are relevant for our applications in Section 5.

**A model for the clustering of time series data.** Suppose we observe time series $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})^\top$ of length $p$ for $n$ different subjects $i$. The time series $\boldsymbol{Y}_i$ of the $i$-th subject is assumed to follow the time trend model

$$Y_{ij} = \mu_i(t_j) + \alpha_i + \varepsilon_{ij} \quad (1 \le j \le p), \tag{2.4}$$

where $\mu_i(\cdot)$ is an unknown nonparametric trend function and $t_1 < \ldots < t_p$ are the observed time points. The deterministic design points $t_j$ are supposed to be the same across subjects $i$ and are normalized to lie in the unit interval. An important example is the equidistant design $t_j = j/p$. However, it is also possible to allow for non-equidistant designs. To identify the trend function $\mu_i(\cdot)$ in (2.4), we suppose that $\int_0^1 \mu_i(w)dw = 0$ for each $i$, which is a slight modification of the identification constraint stipulated in (2.3). Analogous to our general model, we impose a group structure on the observed time series: There are $K_0$ groups of time series $G_1, \ldots, G_{K_0}$ such that $\mu_i(\cdot) = m_k(\cdot)$ for all $i \in G_k$. Hence, the members of each class $G_k$ all have the same time trend function $m_k(\cdot)$.

**A model for the clustering of genes in microarray experiments.** In a microarray experiment, the expression levels of $n$ different genes are often measured in $p$ different tissue samples (obtained, e.g., from $p$ different patients). For each gene $i$, we observe the vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})^\top$, where $Y_{ij}$ is the measured expression level of gene $i$ for tissue sample $j$. The vector $\boldsymbol{Y}_i$ of gene $i$ is supposed to satisfy the model equation (2.3), which componentwise reads as

$$Y_{ij} = \mu_{ij} + \alpha_i + \varepsilon_{ij} \quad (1 \le j \le p). \tag{2.5}$$

Here, $\mu_{ij}$ can be regarded as the true expression level of gene $i$ for tissue $j$, whereas $Y_{ij}$ is the measured expression level corrupted by the noise term $\alpha_i + \varepsilon_{ij}$.

Most microarray experiments involve different types of tissues, for example tumor "cases" versus healthy "controls", or different tumor (sub)types. We therefore suppose that there are $T$ different types of tissues in our sample and order them according to their type (which is known by experimental design). More specifically, the tissues $j$ of type $t$ are labelled by $j_{t-1} \le j < j_t$, where $1 = j_0 < j_1 < \ldots < j_{T-1} < j_T = p + 1$. If the patients from which tissues are obtained constitute samples of sufficiently homogeneous populations, it is natural to assume that the true expression level $\mu_{ij}$ of gene $i$ is the same for tissues $j$ of the same type, i.e., $\mu_{ij} = \mu_{ij'}$ for $j_{t-1} \le j, j' < j_t$. The signal vector $\boldsymbol{\mu}_i$ thus has a piecewise constant structure for each $i$; see Figures 4 and 6 in Section 5 for an illustration.

As in our general model, we suppose that there are $K_0$ groups of genes $G_1, \ldots, G_{K_0}$ such that $\boldsymbol{\mu}_i = \boldsymbol{m}_k$ for all $i \in G_k$ and some vector $\boldsymbol{m}_k$. The genes of each class $G_k$

thus have the same (co-)expression profile $\boldsymbol{m}_k$.

# 3 Estimation Method

We now present our approach to estimate the unknown groups $G_1, \ldots, G_{K_0}$ and their unknown number $K_0$ in model (2.1)–(2.2). Section 3.1 gives an overview of the general method, while Sections 3.2–3.4 fill in the details.

## 3.1 The general method

To construct our method, we proceed in two steps: In the first step, we specify an algorithm that clusters the set of subjects $\{1, \ldots, n\}$ into $K$ groups for any given number $K$ (which may or may not coincide with the true number of classes $K_0$). Let $\{\widehat{G}_k^{[K]} : 1 \le k \le K\}$ be the $K$ clusters produced by the algorithm when the number of clusters is $K$. For $K = 1$, we trivially set $\widehat{G}_1^{[1]} = \{1, \ldots, n\}$. For our theory to work, we require the clustering algorithm to consistently estimate the class structure $\{G_k : 1 \le k \le K_0\}$ when $K = K_0$. More specifically, we require the estimators $\{\widehat{G}_k^{[K_0]} : 1 \le k \le K_0\}$ to have the property that

$$\mathbb{P}\Big(\big\{\widehat{G}_k^{[K_0]} : 1 \le k \le K_0\big\} = \big\{G_k : 1 \le k \le K_0\big\}\Big) \to 1. \tag{3.1}$$

This is a quite weak restriction which is satisfied by a wide range of clustering algorithms under our regularity conditions. As shown in Section 3.3, it is for example satisfied by a $k$-means type algorithm. Moreover, it can be shown to hold for a number of hierarchical clustering algorithms, in particular for agglomerative algorithms with single, average and complete linkage. Our estimation method can be based on any clustering algorithm that has the consistency property (3.1).

In the second step, we construct a test for each $K$ which checks whether the data can be well described by the $K$ clusters $\widehat{G}_1^{[K]}, \ldots, \widehat{G}_K^{[K]}$. We thereby test whether the number of clusters is equal to $K$. More formally, we use the $K$-cluster partition $\{\widehat{G}_k^{[K]} : 1 \le k \le K\}$ to construct a statistic $\widehat{\mathcal{H}}^{[K]}$ that allows us to test the hypothesis $H_0 : K = K_0$ versus $H_1 : K < K_0$. For any given number of clusters $K$, our test is defined as $T_\alpha^{[K]} = \mathbf{1}(\widehat{\mathcal{H}}^{[K]} > q(\alpha))$, where $q(\alpha)$ is the $(1-\alpha)$-quantile of a known distribution which will be specified later on. We reject $H_0$ at the level $\alpha$ if $T_\alpha^{[K]} = 1$, i.e., if $\widehat{\mathcal{H}}^{[K]} > q(\alpha)$. A detailed construction of the statistic $\widehat{\mathcal{H}}^{[K]}$ along with a precise definition of the quantile $q(\alpha)$ is given in Section 3.2.

To estimate the classes $G_1, \ldots, G_{K_0}$ and their number $K_0$, we proceed as follows: For each $K = 1, 2, \ldots$, we check whether $\widehat{\mathcal{H}}^{[K]} \le q(\alpha)$ and stop as soon as this criterion is satisfied. Put differently, we carry out our test for each $K = 1, 2, \ldots$

until it does not reject $H_0$ any more. Our estimator of $K_0$ is defined as the smallest number $K$ for which $\widehat{\mathcal{H}}^{[K]} \le q(\alpha)$, that is, for which the test does not reject $H_0$. Formally speaking, we define

$$\widehat{K}_0 = \min\big\{K = 1, 2, \ldots \,\big|\, \widehat{\mathcal{H}}^{[K]} \le q(\alpha)\big\}. \tag{3.2}$$

Moreover, we estimate the class structure $\{G_k : 1 \le k \le K_0\}$ by the partition $\{\widehat{G}_k : 1 \le k \le \widehat{K}_0\}$, where we set $\widehat{G}_k = \widehat{G}_k^{[\widehat{K}_0]}$. The definition (3.2) can equivalently be written as

$$\widehat{K}_0 = \min\big\{K = 1, 2, \ldots \,\big|\, \widehat{p}^{[K]} > \alpha\big\}, \tag{3.3}$$

where $\widehat{p}^{[K]}$ is the $p$-value corresponding to the statistic $\widehat{\mathcal{H}}^{[K]}$. The heuristic idea behind (3.3) is as follows: Starting with $K = 1$, we successively test whether the data can be well described by a model with $K$ clusters, in particular by the partition $\{\widehat{G}_k^{[K]} : 1 \le k \le K\}$. For each $K$, we compute the $p$-value $\widehat{p}^{[K]}$ which expresses our confidence in a model with $K$ clusters. We stop as soon as $\widehat{p}^{[K]} > \alpha$, that is, as soon as we have enough statistical confidence in a model with $K$ groups.

As shown in Section 4, under appropriate regularity conditions, our statistic $\widehat{\mathcal{H}}^{[K]}$ has the property that

$$\mathbb{P}\Big(\widehat{\mathcal{H}}^{[K]} \le q(\alpha)\Big) = \begin{cases} o(1) & \text{for } K < K_0 \\ (1-\alpha) + o(1) & \text{for } K = K_0. \end{cases} \tag{3.4}$$

Put differently, $\mathbb{P}(T_\alpha^{[K]} = 0) \to 1 - \alpha$ for $K = K_0$ and $\mathbb{P}(T_\alpha^{[K]} = 1) \to 1$ for $K < K_0$. Hence, our test is asymptotically of level $\alpha$. Moreover, it detects the alternative $H_1 : K < K_0$ with probability tending to 1, that is, its power against $H_1$ is asymptotically equal to 1. From (3.4), it follows that

$$\pi_>(\alpha) := \mathbb{P}\big(\widehat{K}_0 > K_0\big) = \alpha + o(1) \tag{3.5}$$

$$\pi_<(\alpha) := \mathbb{P}\big(\widehat{K}_0 < K_0\big) = o(1). \tag{3.6}$$

Hence, the probability of overestimating $K_0$ is asymptotically bounded by $\alpha$, while the probability of underestimating $K_0$ is asymptotically negligible. By picking $\alpha$, we can thus control the probability of choosing too many clusters similarly to the type-I-error probability of a test. Moreover, we can asymptotically ignore the probability of choosing too few clusters similarly to the type-II-error probability of a test. In finite samples, there is of course a trade-off between the probabilities of under- and overestimating $K_0$: By decreasing the significance level $\alpha$, we can reduce the probability of overestimating $K_0$, since $\alpha' + o(1) = \pi_>(\alpha') \le \pi_>(\alpha) = \alpha + o(1)$ for $\alpha' < \alpha$. However, we pay for this by increasing the probability of underestimating

$K_0$, since $\pi_<(\alpha') \geq \pi_<(\alpha)$ for $\alpha' < \alpha$. This can also be regarded as a trade-off between the size and the power of the test on which $\widehat{K}_0$ is based. Taken together, the two statements (3.5) and (3.6) yield that

$$\mathbb{P}\big(\widehat{K}_0 \neq K_0\big) = \alpha + o(1), \tag{3.7}$$

i.e., the probability that the estimated number of classes $\widehat{K}_0$ differs from the true number of classes $K_0$ is asymptotically equal to $\alpha$. With the help of (3.7) and the consistency property (3.1) of the estimated clusters, we can further show that

$$\mathbb{P}\Big(\big\{\widehat{G}_k : 1 \leq k \leq \widehat{K}_0\big\} \neq \big\{G_k : 1 \leq k \leq K_0\big\}\Big) = \alpha + o(1), \tag{3.8}$$

i.e., the probability of making a classification error is asymptotically equal to $\alpha$ as well. The statements (3.5)–(3.8) give a mathematically precise description of the statistical error control that can be performed by our method.

## 3.2 Construction of the statistic $\widehat{\mathcal{H}}^{[K]}$

To construct the statistic $\widehat{\mathcal{H}}^{[K]}$, we use the following notation:

(i) Let $Y_{ij}^* = Y_{ij} - \alpha_i$ be the observations adjusted for the random intercepts $\alpha_i$ and set $\widehat{Y}_{ij} = Y_{ij} - \overline{Y}_i$ with $\overline{Y}_i = p^{-1}\sum_{j=1}^p Y_{ij}$. The variables $\widehat{Y}_{ij}$ serve as approximations of $Y_{ij}^*$, since under standard regularity conditions

$$\begin{aligned}\widehat{Y}_{ij} &= \mu_{ij} + \varepsilon_{ij} - \frac{1}{p}\sum_{j=1}^p \mu_{ij} - \frac{1}{p}\sum_{j=1}^p \varepsilon_{ij} \\ &= \mu_{ij} + \varepsilon_{ij} + O_p(p^{-1/2}) = Y_{ij}^* + O_p(p^{-1/2}).\end{aligned}$$

(ii) For any set $S \subseteq \{1, \ldots, n\}$, let $m_{j,S} = (\#S)^{-1}\sum_{i \in S} \mu_{ij}$ be the average of the signals $\mu_{ij}$ with $i \in S$ and estimate it by $\widehat{m}_{j,S} = (\#S)^{-1}\sum_{i \in S} \widehat{Y}_{ij}$. We use the notation $m_{j,k}^{[K]} = m_{j,\widehat{G}_k^{[K]}}$ and $\widehat{m}_{j,k}^{[K]} = \widehat{m}_{j,\widehat{G}_k^{[K]}}$ to denote the average of the signals in the cluster $\widehat{G}_k^{[K]}$ and its estimator, respectively.

(iii) For any cluster $\widehat{G}_k^{[K]}$, we define cluster-specific residuals by setting $\widehat{\varepsilon}_{ij}^{[K]} = \widehat{Y}_{ij} - \widehat{m}_{j,k}^{[K]}$ for $i \in \widehat{G}_k^{[K]}$ and $1 \leq j \leq p$.

(iv) Let $\widehat{\sigma}^2$ be an estimator of the error variance $\sigma^2 = \mathbb{E}[\varepsilon_{ij}^2]$. Moreover, let $\widehat{\kappa}$ be an estimator of the parameter $\kappa = (\mathbb{E}[\{(\varepsilon_{ij}/\sigma)^2 - 1\}^2])^{1/2}$, which serves as a normalization constant later on. See Section 3.4 for a detailed construction of the estimators $\widehat{\sigma}^2$ and $\widehat{\kappa}$.

With this notation at hand, we define the statistic

$$\widehat{\Delta}_i^{[K]} = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \Big\{ \Big( \frac{\widehat{\varepsilon}_{ij}^{[K]}}{\widehat{\sigma}} \Big)^2 - 1 \Big\} \Big/ \widehat{\kappa} \tag{3.9}$$

for each subject $i$. This is essentially a scaled version of the residual sum of squares for the $i$-th subject when the number of clusters is $K$. Intuitively, $\widehat{\Delta}_i^{[K]}$ measures how well the data of the $i$-th subject are described when the sample of subjects is partitioned into the $K$ clusters $\widehat{G}_1^{[K]}, \ldots, \widehat{G}_K^{[K]}$. The individual statistics $\widehat{\Delta}_i^{[K]}$ are the building blocks of the overall statistic $\widehat{\mathcal{H}}^{[K]}$.

Before we move on with the construction of $\widehat{\mathcal{H}}^{[K]}$, we have a closer look at the stochastic behaviour of the statistics $\widehat{\Delta}_i^{[K]}$. To do so, we consider the following stylized situation: We assume that the variables $\varepsilon_{ij}$ are i.i.d. normally distributed with mean 0 and variance $\sigma^2$. Moreover, we neglect the estimation error in the expressions $\widehat{Y}_{ij}$, $\widehat{m}_{j,k}^{[K]}$, $\widehat{\sigma}^2$ and $\widehat{\kappa}$. In this situation,

$$\widehat{\Delta}_i^{[K]} = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \Big\{ \frac{(\varepsilon_{ij} + d_{ij})^2}{\sigma^2} - 1 \Big\} \Big/ \kappa$$

for any $i \in S = \widehat{G}_k^{[K]}$, where $d_{ij} = \mu_{ij} - (\#S)^{-1} \sum_{i' \in S} \mu_{i'j}$ is the difference between the signal $\mu_{ij}$ of the $i$-th subject and the average signal in the cluster $S$. We now give a heuristic discussion of the behaviour of $\widehat{\Delta}_i^{[K]}$ in the following two cases:

$K = K_0$: By condition (3.1), $\widehat{G}_k^{[K_0]}$ consistently estimates $G_k$. Neglecting the estimation error in $\widehat{G}_k^{[K_0]}$, we obtain that

$$\widehat{\Delta}_i^{[K_0]} = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \Big\{ \frac{\varepsilon_{ij}^2}{\sigma^2} - 1 \Big\} \Big/ \kappa.$$

Since $\varepsilon_{ij}/\sigma$ is standard normal, $\kappa = \sqrt{2}$ and thus

$$\widehat{\Delta}_i^{[K_0]} \sim \frac{\chi_p^2 - p}{\sqrt{2p}} \tag{3.10}$$

for each $i$. Hence, the individual statistics $\widehat{\Delta}_i^{[K_0]}$ all have a rescaled $\chi^2$-distribution.

$K < K_0$: If we pick $K$ smaller than the true number of classes $K_0$, the clusters $\{\widehat{G}_k^{[K]} : 1 \le k \le K\}$ cannot provide an appropriate approximation of the true class structure $\{G_k : 1 \le k \le K_0\}$. In particular, there is always a cluster $S = \widehat{G}_k^{[K]}$ which contains subjects from at least two different classes. For simplicity, let $S = G_{k_1} \cup G_{k_2}$. For any $i \in S$, it holds that

$$\begin{aligned} \widehat{\Delta}_i^{[K]} &= \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \Big\{ \frac{(\varepsilon_{ij} + d_{ij})^2}{\sigma^2} - 1 \Big\} \Big/ \kappa \\ &= \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \frac{d_{ij}^2}{\sigma^2 \kappa} + O_p(1) \end{aligned}$$

under our regularity conditions from Section 4.1. Moreover, it is not difficult to see that for at least one $i \in S$, $p^{-1/2} \sum_{j=1}^{p} d_{ij}^2 \geq c\sqrt{p}$ for some small constant $c > 0$. This implies that for some $i \in S$,

$$\widehat{\Delta}_i^{[K]} \geq c\sqrt{p} \quad \text{for some } c > 0 \text{ with prob. tending to 1,} \tag{3.11}$$

i.e., the statistic $\widehat{\Delta}_i^{[K]}$ has an explosive behaviour.

According to these heuristic considerations, the statistics $\widehat{\Delta}_i^{[K]}$ exhibit a quite different behaviour depending on whether $K < K_0$ or $K = K_0$. When $K < K_0$, the statistic $\widehat{\Delta}_i^{[K]}$ has an explosive behaviour at least for some subjects $i$. This mirrors the fact that a partition with $K < K_0$ clusters cannot give a reasonable approximation to the true class structure. In particular, it cannot describe the data of all subjects $i$ in an accurate way, resulting in an explosive behaviour of the (rescaled) residual sum of squares $\widehat{\Delta}_i^{[K]}$ for some subjects $i$. When $K = K_0$ in contrast, $\{\widehat{\Delta}_i^{[K_0]} : 1 \leq i \leq n\}$ is a collection of (approximately) independent random variables that (approximately) have a rescaled $\chi^2$-distribution. Hence, all statistics $\widehat{\Delta}_i^{[K_0]}$ have a stable, non-explosive behaviour. This reflects the fact that the partition $\{\widehat{G}_k^{[K_0]} : 1 \leq k \leq K_0\}$ is an accurate estimate of the true class structure and thus yields a moderate residual sum of squares $\widehat{\Delta}_i^{[K_0]}$ for all subjects $i$.

Since the statistics $\widehat{\Delta}_i^{[K]}$ behave quite differently depending on whether $K = K_0$ or $K < K_0$, they can be used to test $H_0 : K = K_0$ versus $H_1 : K < K_0$. In particular, testing $H_0$ versus $H_1$ can be achieved by testing the hypothesis that $\widehat{\Delta}_i^{[K]}$ are i.i.d. variables with a rescaled $\chi^2$-distribution against the alternative that at least one $\widehat{\Delta}_i^{[K]}$ has an explosive behaviour. We now construct a statistic $\widehat{\mathcal{H}}^{[K]}$ for this testing problem. A natural approach is to take the maximum of the individual statistics $\widehat{\Delta}_i^{[K]}$: Define

$$\widehat{\mathcal{H}}^{[K]} = \max_{1 \leq i \leq n} \widehat{\Delta}_i^{[K]} \tag{3.12}$$

and let $q(\alpha)$ be the $(1-\alpha)$-quantile of $\mathcal{H} = \max_{1 \leq i \leq n} Z_i$, where $Z_i$ are independent random variables with the distribution $(\chi_p^2 - p)/\sqrt{2p}$.

Our heuristic discussion from above, in particular formula (3.10), suggests that for $K = K_0$,

$$\mathbb{P}\Big(\widehat{\mathcal{H}}^{[K_0]} \le q(\alpha)\Big) \approx (1-\alpha).$$

Moreover, for $K < K_0$, we can show with the help of (3.11) and some additional considerations that $\widehat{\mathcal{H}}^{[K]} \ge c\sqrt{p}$ for some $c > 0$ with probability tending to 1. The quantile $q(\alpha)$, in contrast, can be shown to grow at the rate $\sqrt{\log n}$. Since $\sqrt{\log n} = o(\sqrt{p})$ under our conditions from Section 4.1, $\widehat{\mathcal{H}}^{[K]}$ diverges faster than the quantile $q(\alpha)$, implying that

$$\mathbb{P}\Big(\widehat{\mathcal{H}}^{[K]} \le q(\alpha)\Big) = o(1)$$

for $K < K_0$. This suggests that $\widehat{\mathcal{H}}^{[K]}$ has the property (3.4) and thus is a reasonable statistic to test the hypothesis $H_0 : K = K_0$ versus $H_1 : K < K_0$.

In this paper, we restrict attention to the maximum statistic $\widehat{\mathcal{H}}^{[K]}$ defined in (3.12). In principle though, we may work with any statistic that satisfies the higher-order property (3.4). In Section 6, we discuss some alternative choices of $\widehat{\mathcal{H}}^{[K]}$.

## 3.3 A *k*-means clustering algorithm

We now construct a $k$-means type clustering algorithm which has the consistency property (3.1). Since its introduction by Cox (1957) and Fisher (1958), the $k$-means algorithm has become one of the most popular tools in cluster analysis. Our version of the algorithm mainly differs from the standard one in the choice of the initial values. To ensure the consistency property (3.1), we pick initial clusters $\mathscr{C}_1^{[K]}, \dots, \mathscr{C}_K^{[K]}$ for each given $K$ as follows:

**Choice of the starting values.** Let $i_1, \dots, i_K$ be indices which (with probability tending to 1) belong to $K$ different classes $G_{k_1}, \dots, G_{k_K}$ in the case that $K \le K_0$ and to $K_0$ different classes in the case that $K > K_0$. We explain how to obtain such indices below. With these indices at hand, we compute the distance measures $\widehat{\rho}_k(i) = \widehat{\rho}(i_k, i)$ for all $1 \le i \le n$ and $1 \le k \le K$, where

$$\widehat{\rho}(i, i') = \frac{1}{p} \sum_{j=1}^{p} \big(\widehat{Y}_{ij} - \widehat{Y}_{i'j}\big)^2.$$

The starting values $\mathscr{C}_1^{[K]}, \dots, \mathscr{C}_K^{[K]}$ are now defined by assigning the index $i$ to cluster $\mathscr{C}_k^{[K]}$ if $\widehat{\rho}_k(i) = \min_{1 \le k' \le K} \widehat{\rho}_{k'}(i)$.

The indices $i_1, \dots, i_K$ in this construction are computed as follows: For $K = 2$, pick any index $i_1 \in \{1, \dots, n\}$ and calculate $i_2 = \arg\max_{1 \le i \le n} \widehat{\rho}(i_1, i)$. Next suppose we have already constructed the indices $i_1, \dots, i_{K-1}$ for the case of $K-1$ clusters and compute the corresponding starting values $\mathscr{C}_1^{[K-1]}, \dots, \mathscr{C}_{K-1}^{[K-1]}$ as described above. Calculate the maximal within-cluster distance $\widehat{\rho}_{\max}(k) = \max_{i \in \mathscr{C}_k^{[K-1]}} \widehat{\rho}_k(i)$ for each $1 \le k \le K-1$ and let $\mathscr{C}_{k^*}^{[K-1]}$ be a cluster with $\widehat{\rho}_{\max}(k^*) \ge \widehat{\rho}_{\max}(k)$ for all $k$. Define $i_K = \arg\max_{i \in \mathscr{C}_{k^*}^{[K-1]}} \widehat{\rho}_{k^*}(i)$.

**The *k*-means algorithm.** Let the number of clusters $K$ be given and denote the starting values by $C_k^{(0)} := \mathscr{C}_k^{[K]}$ for $1 \le k \le K$. The $r$-th iteration of our $k$-means algorithm proceeds as follows:

Step $r$: Let $C_1^{(r-1)}, \dots, C_K^{(r-1)}$ be the clusters from the $(r-1)$-th iteration step. Compute cluster means $m_{j,k}^{(r)} = (\#C_k^{(r-1)})^{-1} \sum_{i \in C_k^{(r-1)}} \widehat{Y}_{ij}$ and calculate the distance measures $\widehat{\rho}_k^{(r)}(i) = p^{-1} \sum_{j=1}^p (\widehat{Y}_{ij} - m_{j,k}^{(r)})^2$ for all $1 \le i \le n$ and $1 \le k \le K$. Define updated groups $C_1^{(r)}, \dots, C_K^{(r)}$ by assigning the index $i$ to the cluster $C_k^{(r)}$ if $\widehat{\rho}_k^{(r)}(i) = \min_{1 \le k' \le K} \widehat{\rho}_{k'}^{(r)}(i)$.

Repeat this algorithm until the estimated groups do not change any more. For a given sample of data, this is guaranteed to happen after finitely many steps. The resulting $k$-means estimators are denoted by $\{\widehat{G}_k^{[K]} : 1 \le k \le K\}$. In Section 4, we formally show that these estimators have the consistency property (3.1) under our regularity conditions.

### 3.4 Estimation of $\sigma^2$ and $\kappa$

In practice, the error variance $\sigma^2$ and the normalization constant $\kappa$ are unknown and need to be estimated from the data at hand. We distinguish between two different estimation approaches, namely a difference- and a residual-based approach.

**Difference-based estimators.** To start with, consider the time trend model from Section 2, where $\mu_{ij} = \mu_i(j/p)$ with some trend function $\mu_i(\cdot)$. Supposing that the functions $\mu_i(\cdot)$ are Lipschitz continuous, we get that $Y_{ij} - Y_{i,j-1} = \{\varepsilon_{ij} - \varepsilon_{i,j-1}\} + \{\mu_i(j/p) - \mu_i((j-1)/p)\} = \{\varepsilon_{ij} - \varepsilon_{i,j-1}\} + O(p^{-1})$. This motivates to estimate the error variance $\sigma^2 = \mathbb{E}[\varepsilon_{ij}^2]$ by

$$\widehat{\sigma}^2_{\text{Lip}} = \frac{1}{n(p-1)} \sum_{i=1}^n \sum_{j=2}^p \frac{(Y_{ij} - Y_{i,j-1})^2}{2}.$$

Similarly, the fourth moment $\vartheta = \mathbb{E}[\varepsilon_{ij}^4]$ can be estimated by

$$\widehat{\vartheta}_{\text{Lip}} = \frac{1}{n(p-1)} \sum_{i=1}^n \sum_{j=2}^p \frac{(Y_{ij} - Y_{i,j-1})^4}{2} - 3(\widehat{\sigma}^2_{\text{Lip}})^2,$$

which in turn allows us to estimate the parameter $\kappa$ by

$$\widehat{\kappa}_{\text{Lip}} = \Big( \frac{\widehat{\vartheta}_{\text{Lip}}}{(\widehat{\sigma}^2_{\text{Lip}})^2} - 1 \Big)^{1/2}.$$

Difference-based estimators of this type have been considered in the context of nonparametric regression by Müller et al. (1988) and Hall et al. (1990) among others. Under the technical conditions (C1)–(C3) from Section 4.1, it is straightforward to show that $\widehat{\sigma}^2_{\text{Lip}} = \sigma^2 + O_p((np)^{-1/2} + p^{-2})$ and $\widehat{\kappa}_{\text{Lip}} = \kappa + O_p((np)^{-1/2} + p^{-2})$. The estimators $\widehat{\sigma}^2_{\text{Lip}}$ and $\widehat{\kappa}_{\text{Lip}}$ are particularly suited for applications where the trend functions $\mu_i(\cdot)$ can be expected to be fairly smooth. This ensures that the unknown first differences $(\varepsilon_{ij} - \varepsilon_{i,j-1})$ can be sufficiently well approximated by the terms $(Y_{ij} - Y_{i,j-1})$.

A similar difference-based estimation strategy can be used in the model for gene expression microarray data from Section 2. In this setting, the signal vectors $\boldsymbol{\mu}_i$ have a piecewise constant structure. In particular, $\mu_{ij} = \mu_{ij'}$ for $j_{t-1} \le j, j' < j_t$, where $j_t$ are known indices with $1 = j_0 < j_1 < \ldots < j_{T-1} < j_T = p+1$. This implies that $Y_{ij} - Y_{i,j-1} = \varepsilon_{ij} - \varepsilon_{i,j-1}$ for $j_{t-1} < j < j_t$. Similarly as before, we may thus estimate $\sigma^2$, $\vartheta$ and $\kappa$ by

$$\widehat{\sigma}^2_{\text{pc}} = \frac{1}{n(p-T)} \sum_{i=1}^n \sum_{j=1}^p \mathbf{1}_j \frac{(Y_{ij} - Y_{i,j-1})^2}{2}$$

$$\widehat{\vartheta}_{\text{pc}} = \frac{1}{n(p-T)} \sum_{i=1}^n \sum_{j=1}^p \mathbf{1}_j \frac{(Y_{ij} - Y_{i,j-1})^4}{2} - 3(\widehat{\sigma}^2_{\text{pc}})^2$$

and $\widehat{\kappa}_{\text{pc}} = (\widehat{\vartheta}_{\text{pc}}/(\widehat{\sigma}^2_{\text{pc}})^2 - 1)^{1/2}$, where $\mathbf{1}_j = \mathbf{1}(j \notin \{j_0, j_1, \ldots, j_T\})$. It is not difficult to see that under the conditions (C1)–(C3), $\widehat{\sigma}^2_{\text{pc}} = \sigma^2 + O_p((np)^{-1/2})$ and $\widehat{\kappa}_{\text{pc}} = \kappa + O_p((np)^{-1/2})$.

**Residual-based estimators.** Let $\{\widehat{G}^{[K]}_k : 1 \le k \le K\}$ be the $k$-means estimators from Section 3.3 for a given $K$. Moreover, let $\widehat{\varepsilon}^{[K]}_{ij}$ be the cluster-specific residuals introduced at the beginning of Section 3.2 and denote the vector of residuals for the $i$-th subject by $\widehat{\boldsymbol{\varepsilon}}^{[K]}_i = (\widehat{\varepsilon}^{[K]}_{i1}, \ldots, \widehat{\varepsilon}^{[K]}_{ip})^\top$. With this notation at hand, we define the residual sum of squares for $K$ clusters by

$$\text{RSS}(K) = \frac{1}{np} \sum_{k=1}^K \sum_{i \in \widehat{G}^{[K]}_k} \|\widehat{\boldsymbol{\varepsilon}}^{[K]}_i\|^2, \tag{3.13}$$

where $\|\cdot\|$ denotes the usual Euclidean norm for vectors. RSS$(K)$ can be shown to be a consistent estimator of $\sigma^2$ for any fixed $K \ge K_0$. The reason is the following:

For any $K \geq K_0$, the $k$-means estimators $\widehat{G}_k^{[K]}$ have the property that

$$\mathbb{P}\Big(\widehat{G}_k^{[K]} \subseteq G_{k'} \text{ for some } 1 \leq k' \leq K_0\Big) \to 1 \tag{3.14}$$

for $1 \leq k \leq K$ under the technical conditions (C1)–(C3) from Section 4.1. Hence, with probability tending to 1, the estimated clusters $\widehat{G}_k^{[K]}$ contain elements from only one class $G_{k'}$. The residuals $\widehat{\varepsilon}_{ij}^{[K]}$ should thus give a reasonable approximation to the unknown error terms $\varepsilon_{ij}$. This in turn suggests that the residual sum of squares $\text{RSS}(K)$ should be a consistent estimator of $\sigma^2$ for $K \geq K_0$.

Now suppose we know that $K_0$ is not larger than some upper bound $K_{\max}$. In this situation, we may try to estimate $\sigma^2$ by $\widetilde{\sigma}^2_{\text{RSS}} = \text{RSS}(K_{\max})$. Even though consistent, this is a very poor estimator of $\sigma^2$. The issue is the following: The larger $K_{\max}$, the smaller the residual sum of squares $\text{RSS}(K_{\max})$ tends to be. This is a natural consequence of the way in which the $k$-means algorithm works. Hence, if $K_{\max}$ is much larger than $K_0$, then $\widetilde{\sigma}^2_{\text{RSS}} = \text{RSS}(K_{\max})$ tends to strongly underestimate $\sigma^2$. To avoid this issue, we replace the naive estimator $\widetilde{\sigma}^2_{\text{RSS}}$ by a refined version:

(i) Split the data vector $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{ip})^\top$ into the two parts $\boldsymbol{Y}_i^A = (Y_{i1}, Y_{i3}, \dots)^\top$ and $\boldsymbol{Y}_i^B = (Y_{i2}, Y_{i4}, \dots)^\top$. Moreover, let $\overline{Y}_i^A$ be the empirical mean of the entries in the vector $\boldsymbol{Y}_i^A$ and define $\widehat{\boldsymbol{Y}}_i^A = (Y_{i1} - \overline{Y}_i^A, Y_{i3} - \overline{Y}_i^A, \dots)^\top$. Finally, set $\mathcal{Y}^A = \{\widehat{\boldsymbol{Y}}_i^A : 1 \leq i \leq n\}$ and analogously define $\mathcal{Y}^B = \{\widehat{\boldsymbol{Y}}_i^B : 1 \leq i \leq n\}$. Importantly, under our technical conditions, the random vectors in $\mathcal{Y}^A$ are independent from those in $\mathcal{Y}^B$.

(ii) Apply the $k$-means algorithm with $K = K_{\max}$ to the sample $\mathcal{Y}^A$ and denote the resulting estimators by $\{\widehat{G}_k^A : 1 \leq k \leq K_{\max}\}$. These estimators can be shown to have the property (3.14), provided that we impose the following condition: Let $\boldsymbol{m}_k$ be the class-specific signal vector of the class $G_k$ and define the vectors $\boldsymbol{m}_k^A$ and $\boldsymbol{m}_k^B$ in the same way as above. Assume that

$$\boldsymbol{m}_k^A \neq \boldsymbol{m}_{k'}^A \text{ for } k \neq k'. \tag{3.15}$$

According to this assumption, the signal vectors $\boldsymbol{m}_k$ and $\boldsymbol{m}_{k'}$ of two different classes can be distinguished from each other only by looking at their odd entries $\boldsymbol{m}_k^A$ and $\boldsymbol{m}_{k'}^A$. It goes without saying that this is not a very severe restriction.

(iii) Compute cluster-specific residuals from the data sample $\mathcal{Y}^B$,

$$\widehat{\boldsymbol{\varepsilon}}_i^B = \widehat{\boldsymbol{Y}}_i^B - \frac{1}{\#\widehat{G}_k^A} \sum_{i' \in \widehat{G}_k^A} \widehat{\boldsymbol{Y}}_{i'}^B \quad \text{for } i \in \widehat{G}_k^A,$$

and define

$$\widehat{\sigma}^2_{\mathrm{RSS}} = \frac{1}{n\lfloor p/2 \rfloor} \sum_{k=1}^{K_{\max}} \sum_{i \in \widehat{G}^A_k} \|\widehat{\boldsymbol{\varepsilon}}^B_i\|^2.$$

In contrast to the naive estimator $\widetilde{\sigma}^2_{\mathrm{RSS}}$, the refined version $\widehat{\sigma}^2_{\mathrm{RSS}}$ does not tend to strongly underestimate $\sigma^2$. The main reason is that the residuals $\widehat{\boldsymbol{\varepsilon}}^B_i$ are computed from the random vectors $\widehat{\boldsymbol{Y}}^B_i$ which are independent of the estimated clusters $\widehat{G}^A_k$.

Writing $\widehat{\boldsymbol{\varepsilon}}^B_i = (\widehat{\varepsilon}^B_{i1}, \dots, \widehat{\varepsilon}^B_{i\lfloor p/2 \rfloor})^\top$, we can analogously estimate the fourth error moment $\vartheta = \mathbb{E}[\varepsilon^4_{ij}]$ by

$$\widehat{\vartheta}_{\mathrm{RSS}} = \frac{1}{n\lfloor p/2 \rfloor} \sum_{k=1}^{K_{\max}} \sum_{i \in \widehat{G}^A_k} \sum_{j=1}^{\lfloor p/2 \rfloor} \big(\widehat{\varepsilon}^B_{ij}\big)^4$$

and set $\widehat{\kappa}_{\mathrm{RSS}} = (\widehat{\vartheta}_{\mathrm{RSS}}/(\widehat{\sigma}^2_{\mathrm{RSS}})^2 - 1)^{1/2}$. Under the conditions (C1)–(C3), it can be shown that

$$\widehat{\sigma}^2_{\mathrm{RSS}} = \sigma^2 + O_p(p^{-1}) \quad \text{and} \quad \widehat{\kappa}_{\mathrm{RSS}} = \kappa + O_p(p^{-1}). \tag{3.16}$$

A sketch of the proof is given in the Supplementary Material.

# 4 Asymptotics

In this section, we investigate the asymptotic properties of our estimators. We first list the assumptions needed for the analysis and then summarize the main results.

## 4.1 Assumptions

To formulate the technical conditions that we impose on model (2.1)–(2.2), we denote the size, i.e., the cardinality of the class $G_k$ by $n_k = \#G_k$. Moreover, we use the shorthand $a_\nu \ll b_\nu$ to express that $a_\nu / b_\nu \le c\nu^{-\delta}$ for sufficiently large $\nu$ with some $c > 0$ and a small $\delta > 0$. Our assumptions read as follows:

(C1) The errors $\varepsilon_{ij}$ are identically distributed and independent across both $i$ and $j$ with $\mathbb{E}[\varepsilon_{ij}] = 0$ and $\mathbb{E}[|\varepsilon_{ij}|^\theta] \le C < \infty$ for some $\theta > 8$.

(C2) The class-specific signal vectors $\boldsymbol{m}_k = (m_{1,k}, \dots, m_{p,k})^\top$ differ across groups in the following sense: There exists a constant $\delta_0 > 0$ such that

$$\frac{1}{p} \sum_{j=1}^{p} \big(m_{j,k} - m_{j,k'}\big)^2 \ge \delta_0$$

for any pair of groups $G_k$ and $G_{k'}$ with $k \neq k'$. Moreover, $|m_{j,k}| \leq C$ for all $k$ and $j$, where $C > 0$ is a sufficiently large constant.

(C3) Both $n$ and $p$ tend to infinity. The group sizes $n_k = \#G_k$ are such that $p \ll n_k \ll p^{(\theta/4)-1}$ for all $1 \leq k \leq K_0$, implying that $p \ll n \ll p^{(\theta/4)-1}$.

We briefly comment on the above conditions. By imposing (C1), we restrict the noise terms $\varepsilon_{ij}$ to be i.i.d. Yet the error terms $e_{ij} = \alpha_i + \varepsilon_{ij}$ of our model may be dependent across subjects $i$, as we do not impose any restrictions on the random intercepts $\alpha_i$. This is important, for instance, when clustering the genes in a microarray data set, where we may expect different genes $i$ to be correlated. (C2) is a fairly harmless condition, which requires the signal vectors to differ in an $L_2$-sense across groups. By (C3), the group sizes $n_k$ and thus the total number of subjects $n$ are supposed to grow faster than the number of features $p$. We thus focus attention on applications where $n$ is (much) larger than $p$. However, $n$ should not grow too quickly as compared to $p$. Specifically, it should not grow faster than $p^{(\theta/4)-1}$, where $\theta$ is the number of existing error moments $\mathbb{E}[|\varepsilon_{ij}|^\theta] < \infty$. As can be seen, the bound $p^{(\theta/4)-1}$ on the growth rate of $n$ gets larger with increasing $\theta$. In particular, if all moments of $\varepsilon_{ij}$ exist, $n$ may grow as quickly as any polynomial of $p$. Importantly, (C3) allows the group sizes $n_k$ to grow at different rates (between $p$ and $p^{(\theta/4)-1}$). Put differently, it allows for strongly heterogeneous group sizes. Our estimation methods are thus able to deal with situations where some groups are much smaller than others.

## 4.2 Main results

Our first result shows that the maximum statistic $\widehat{\mathcal{H}}^{[K]} = \max_{1 \leq i \leq n} \widehat{\Delta}_i^{[K]}$ has the property (3.4) and thus is a reasonable statistic to test the hypothesis $H_0 : K = K_0$ versus $H_1 : K < K_0$.

**Theorem 4.1.** *Assume that the estimated clusters have the consistency property* (3.1)*. Moreover, let $\widehat{\sigma}^2$ and $\widehat{\kappa}$ be any estimators with $\widehat{\sigma}^2 = \sigma^2 + O_p(p^{-(1/2+\delta)})$ and $\widehat{\kappa} = \kappa + O_p(p^{-\delta})$ for some $\delta > 0$. Under (C1)–(C3), the statistic $\widehat{\mathcal{H}}^{[K]}$ has the property* (3.4)*, that is,*

$$\mathbb{P}\Big(\widehat{\mathcal{H}}^{[K]} \leq q(\alpha)\Big) = \begin{cases} o(1) & \text{for } K < K_0 \\ (1-\alpha) + o(1) & \text{for } K = K_0. \end{cases}$$

This theorem is the main stepping stone to derive the central result of the paper, which describes the asymptotic properties of the estimators $\widehat{K}_0$ and $\{\widehat{G}_k : 1 \leq k \leq \widehat{K}_0\}$.

**Theorem 4.2.** *Under the conditions of Theorem 4.1, it holds that*

$$\mathbb{P}\big(\widehat{K}_0 > K_0\big) = \alpha + o(1) \quad \textit{and} \quad \mathbb{P}\big(\widehat{K}_0 < K_0\big) = o(1),$$

*implying that* $\mathbb{P}(\widehat{K}_0 \neq K_0) = \alpha + o(1)$. *Moreover,*

$$\mathbb{P}\Big(\big\{\widehat{G}_k : 1 \leq k \leq \widehat{K}_0\big\} \neq \big\{G_k : 1 \leq k \leq K_0\big\}\Big) = \alpha + o(1).$$

Theorem 4.2 holds true for any clustering algorithm with the consistency property (3.1). The next result shows that this property is fulfilled, for example, by the $k$-means algorithm from Section 3.3.

**Theorem 4.3.** *Under (C1)–(C3), the k-means estimators* $\{\widehat{G}_k^{[K_0]} : 1 \leq k \leq K_0\}$ *from Section 3.3 satisfy* (3.1)*, that is,*

$$\mathbb{P}\Big(\big\{\widehat{G}_k^{[K_0]} : 1 \leq k \leq K_0\big\} = \big\{G_k : 1 \leq k \leq K_0\big\}\Big) \to 1.$$

The proofs of Theorems 4.1–4.3 can be found in the Supplementary Material.

# 5 Applications and Simulation Study

## 5.1 Clustering of temperature curves

Our first application is concerned with the analysis of a data set on land surface temperatures that was collected by the investigators of the Berkeley Earth project (Rohde et al., 2013). The data, which are publicly available at `http://berkeleyearth.org/data`, contain measurements on a grid of worldwide locations that is defined on a one degree (longitude) by one degree (latitude) basis. For each grid point, the data set contains a monthly land surface temperature profile. This profile is a vector with twelve entries, the first entry specifying the average temperature of all Januaries from 1951 to 1980, the second entry specifying the average temperature of all Februaries from 1951 to 1980, and so on. The temperature profiles at various example locations on earth are shown in Figure 1. As grid points containing 100% sea surface are not taken into account, the overall number of grid points is equal to $n$ = 24,311. A detailed description of the derivation of the data can be found in Rohde et al. (2013). Our analysis is based on the Berkeley Earth source file from April 19, 2016.

The aim of our analysis is to cluster the 24,311 grid points in order to obtain a set of climate regions characterized by distinct temperature profiles. For this purpose, we impose the time trend model (2.4) on the data and apply the CluStErr method

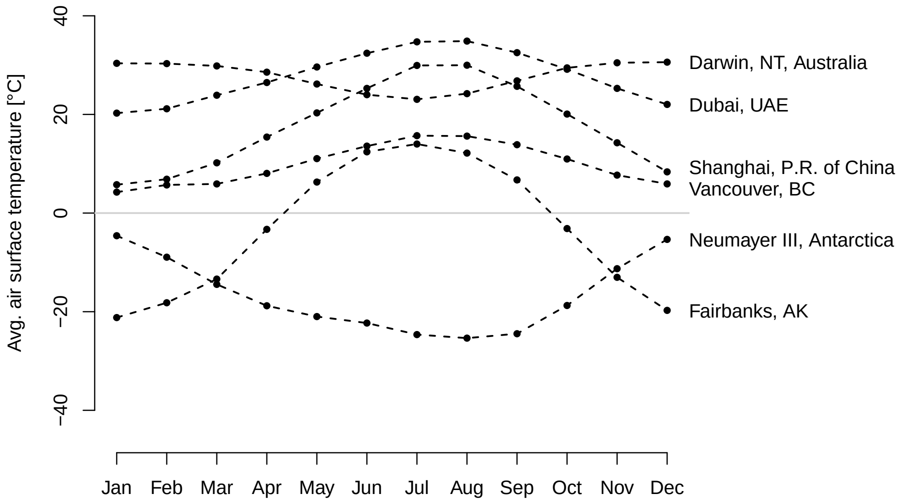

Figure 1: Analysis of the Berkeley Earth temperature data. The plot depicts the average land surface temperature curves at various example locations on earth.

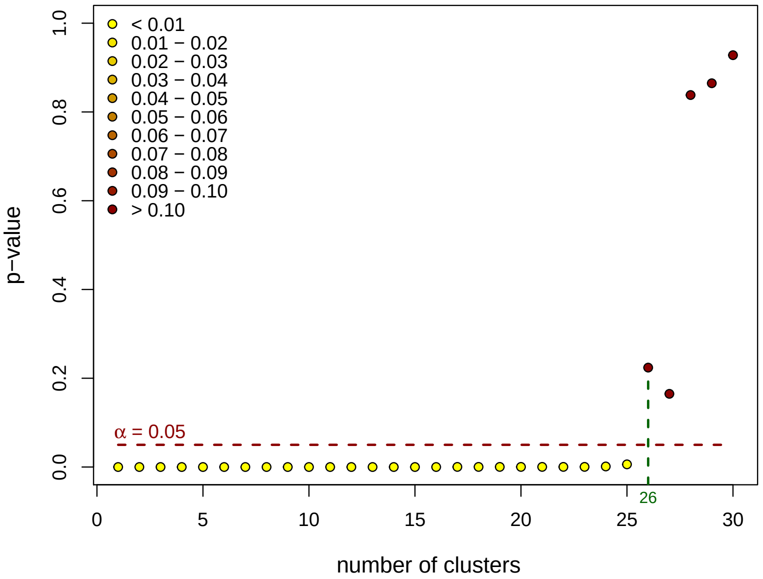

Figure 2: Analysis of the Berkeley Earth temperature data. The plot depicts the $p$-values $\widehat{p}^{[K]}$ corresponding to the test statistics $\widehat{\mathcal{H}}^{[K]}$ as a function of $K$. The horizontal dashed line specifies the significance level $\alpha = 0.05$, and the vertical dashed line indicates that the estimated number of clusters is $\widehat{K}_0 = 26$.

to them, setting $n = 24{,}311$, $p = 12$ and $\alpha = 0.05$. To estimate the error variance $\sigma^2$ and the normalization constant $\kappa$, we apply the difference-based estimators $\widehat{\sigma}^2_{\text{Lip}}$ and $\widehat{\kappa}_{\text{Lip}}$ from Section 3.4, thus making use of the smoothness of the temperature curves illustrated in Figure 1.

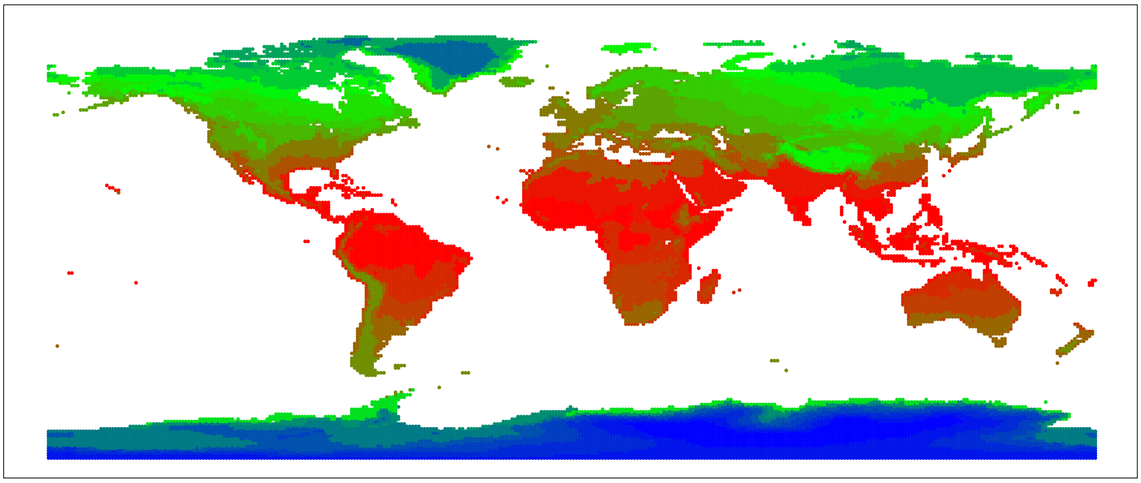

Figure 3: Visualization of the $\widehat{K}_0 = 26$ clusters obtained from the analysis of the Berkeley Earth temperature data. Each shade of color refers to one cluster.

The estimation results are presented in Figures 2 and 3. Figure 2 depicts the $p$-values $\widehat{p}^{[K]}$ corresponding to the test statistics $\widehat{\mathcal{H}}^{[K]}$ for different numbers of clusters $K$. It shows that the $p$-value $\widehat{p}^{[K]}$ remains below the $\alpha = 0.05$ threshold for any $K < 26$ but jumps across this threshold for $K = 26$. The CluStErr algorithm thus estimates the number of clusters to be equal to $\widehat{K}_0 = 26$, suggesting that there are 26 distinct climate regions. The sizes of the estimated clusters range between 244 and 2,110; the error variance is estimated to be $\widehat{\sigma}^2 = 16.25$. Figure 3 uses a spatial grid to visualize the 26 regions and demonstrates the plausibility of the obtained results. For example, mountain ranges such as the Himalayas and the South American Andes, but also tropical climates in Africa, South America and Indonesia are easily identified from the plot. Of note, the results presented in Figure 3 show a remarkable similarity to the most recent modification of the Köppen-Geiger classification, which is one of the most widely used classification systems in environmental research (Peel et al., 2007). In particular, the overall number of climate regions defined in Peel et al. (2007) is equal to 29, which is similar to the cluster number $\widehat{K}_0 = 26$ identified by the CluStErr algorithm. Thus, although our example is purely illustrative, and although expert classification systems account for additional characteristics such as precipitation and vegetation, Figure 3 confirms the usefulness of the CluStErr method.

## 5.2 Clustering of gene expression data

Our second application is concerned with the analysis of gene expression data, which has become a powerful tool for the understanding of disease processes in biomedical research (Jiang et al., 2004). A popular approach to measure gene expression is

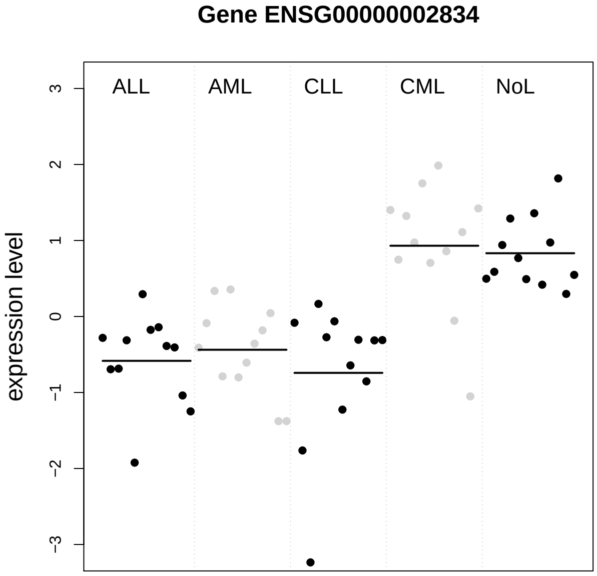

Figure 4: Analysis of the MILE study gene expression data. The plot depicts the expression levels of a randomly selected gene after normalization and standardization. The label of the gene ("ENSG00000002834") refers to its *Ensembl* gene ID to which the original Affymetrix probesets were mapped (Aibar et al., 2013). Horizontal lines represent the average gene expression levels across the five tissue types.

to carry out microarray experiments. These experiments simultaneously quantify the expression levels of $n$ genes across $p$ samples of patients with different clinical conditions, such as tumor stages or disease subtypes. In the analysis of microarray data, clustering of the $n$ genes is frequently used to detect genes with similar cellular function and to discover groups of "co-expressed" genes showing similar expression patterns across clinical conditions (Chipman et al., 2003; Jiang et al., 2004; D'haeseleer, 2005).

In what follows, we analyze a set of gene expression data that was collected during the first stage of the Microarray Innovations in Leukemia (MILE) study (Haferlach et al., 2010). The data set contains expression level measurements for 20,172 genes and is publicly available as part of the Bioconductor package **leukemiasEset** (Aibar et al., 2013). The gene expression levels were measured using Affymetrix HG-U133 Plus 2.0 microarrays. For statistical analysis, the raw expression data were normalized using the Robust Multichip Average (RMA) method, followed by an additional gene-wise standardization of the expression levels. For details on data collection and pre-processing, we refer to Haferlach et al. (2010) and Aibar et al. (2013).

The data of the MILE study were obtained from $p = 60$ bone marrow samples of patients that were untreated at the time of diagnosis. Of these patients, 48 were either diagnosed with acute lymphoblastic leukemia (ALL, 12 patients), acute myeloid leukemia (AML, 12 patients), chronic lymphocytic leukemia (CLL, 12 patients), or

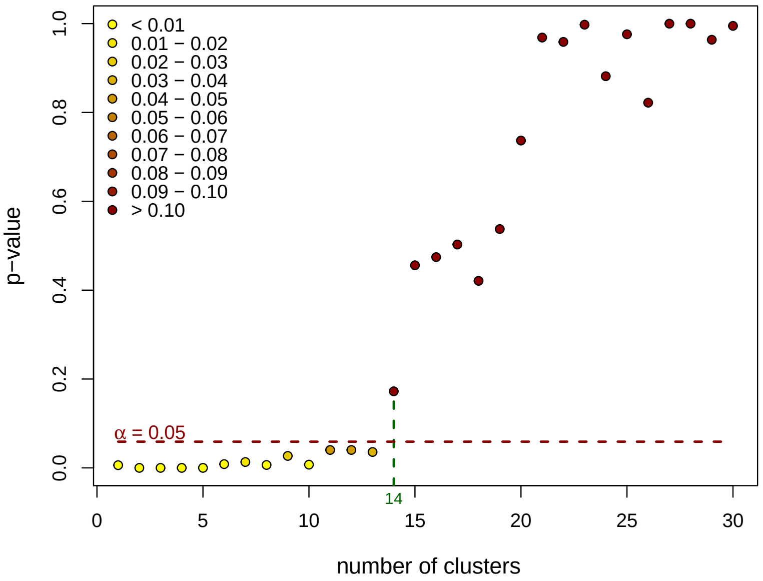

Figure 5: Analysis of the MILE study gene expression data. The plot depicts the $p$-values $\widehat{p}^{[K]}$ corresponding to the test statistics $\widehat{\mathcal{H}}^{[K]}$ as a function of $K$. The dashed vertical line indicates that the number of clusters is estimated to be $\widehat{K}_0 = 14$.

chronic myeloid leukemia (CML, 12 patients). The other 12 samples were obtained from non-leukemia (NoL) patients. From a biomedical point of view, the main interest focuses on the set of “differentially expressed” genes, that is, on those genes that show a sufficient amount of variation in their expression levels across the five tissue types (ALL, AML, CLL, CML, NoL). To identify the set of these genes, we run a univariate ANOVA for each gene and discard those with Bonferroni-corrected $p$-values $\geq 0.01$ in the respective overall $F$-tests. Application of this procedure results in a sample of $n = 3{,}167$ univariately significant genes.

The aim of our analysis is to cluster the $n = 3{,}167$ genes into groups whose members have similar expression patterns across the five tissue types (ALL, AML, CLL, CML, NoL). To do so, we impose model (2.5) from Section 2 on the data. The measured expression profiles $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{ip})^\top$ of the various genes $i = 1, \dots, n$ are thus assumed to follow the model equation $\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i$. The signal vectors $\boldsymbol{\mu}_i$ are supposed to have a piecewise constant structure after the patients have been ordered according to their tissue type (ALL, AML, CLL, CML, NoL). For illustration, the expression profile $\boldsymbol{Y}_i$ of a randomly selected gene is plotted in Figure 4.

To cluster the genes, we apply the CluStErr algorithm with the significance level $\alpha = 0.05$ and the difference-based estimators $\widehat{\sigma}^2_{\text{pc}}$ and $\widehat{\kappa}_{\text{pc}}$ from Section 3.4, thus exploiting the piecewise constant structure of the signal vectors. The estimation results are presented in Figures 5 and 6. The plot in Figure 5 depicts the $p$-values

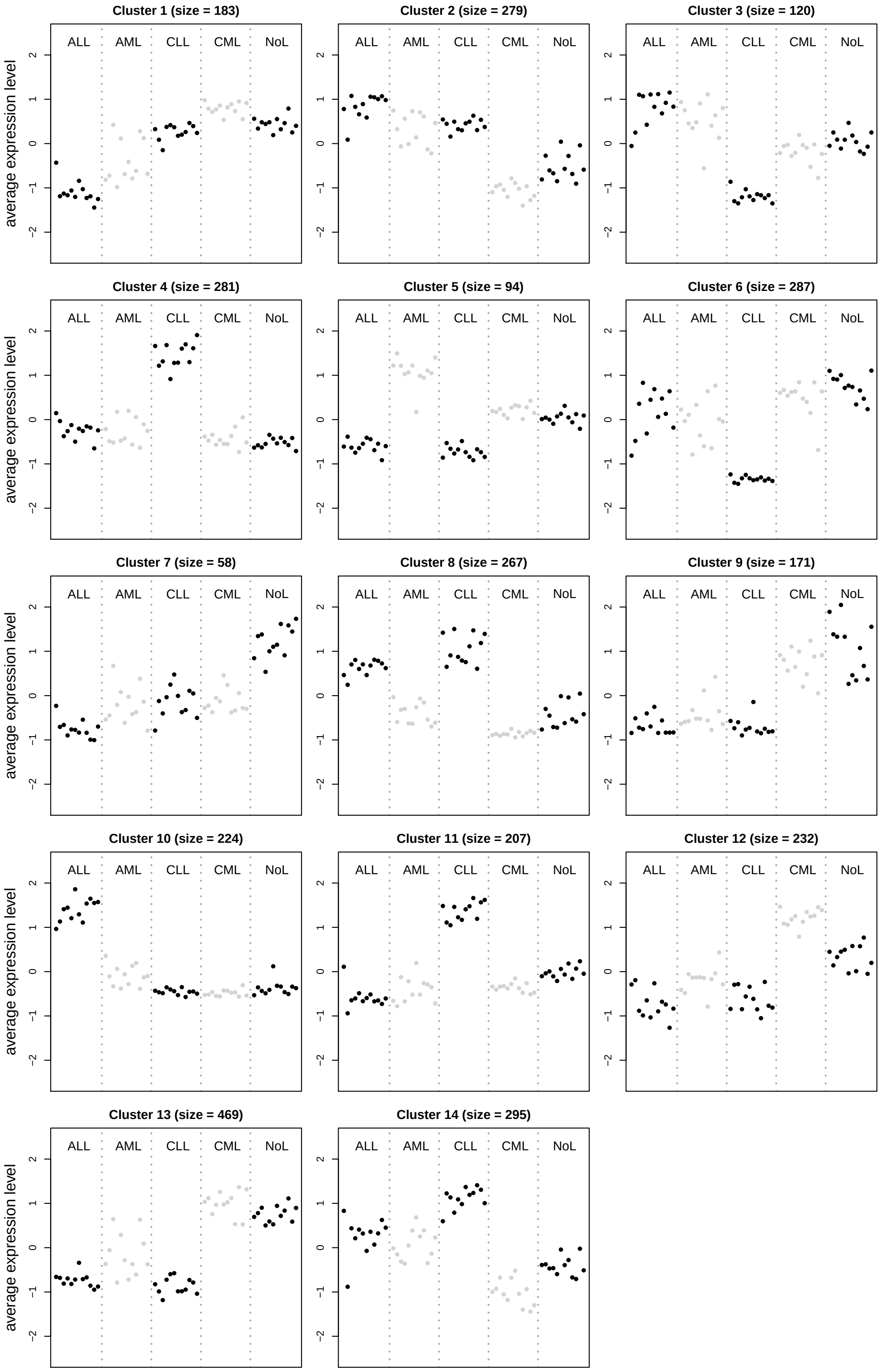

Figure 6: Visualization of the cluster centres obtained from the analysis of the MILE study gene expression data. The dots represent the cluster centres $\widehat{\boldsymbol{m}}_k = (\#\widehat{G}_k)^{-1} \sum_{i \in \widehat{G}_k} \widehat{\boldsymbol{Y}}_i$, which estimate the cluster-specific signal vectors $\boldsymbol{m}_k = (\#G_k)^{-1} \sum_{i \in G_k} \boldsymbol{\mu}_i$. Gene ENSG00000002834, whose expression profile is visualized in Figure 4, is an element of cluster #13. Note the similarity of the patterns in cluster #13 and Figure 4.

$\widehat{p}^{[K]}$ corresponding to the test statistics $\widehat{\mathcal{H}}^{[K]}$ as a function of the cluster number $K$. It shows that the estimated number of clusters is $\widehat{K}_0 = 14$. The estimated sizes of the 14 clusters range between 58 and 469. Moreover, the estimated error variance is $\widehat{\sigma}^2 = 0.442$. In Figure 6, the cluster centres $\widehat{\boldsymbol{m}}_k = (\#\widehat{G}_k)^{-1} \sum_{i \in \widehat{G}_k} \widehat{\boldsymbol{Y}}_i$ are presented, which estimate the cluster-specific signal vectors $\boldsymbol{m}_k = (\#G_k)^{-1} \sum_{i \in G_k} \boldsymbol{\mu}_i$. All clusters show a distinct separation of at least one tissue type, supporting the assumption of piecewise constant signals $\boldsymbol{m}_k$ and indicating that the genes contained in the clusters are co-expressed differently across the five groups. For example, cluster #2 separates CML and NoL samples from ALL, AML and CLL samples, whereas cluster #4 separates CLL samples from the other tissue types. Thus, each of the 14 clusters represents a specific pattern of co-expressed gene profiles.

## 5.3 Simulation study

To explore the properties of the CluStErr method more systematically, we carry out a simulation study which splits into two main parts. The first part investigates the finite sample behaviour of CluStErr, whereas the second part compares CluStErr with several competing methods. The simulation design is inspired by the analysis of the gene expression data in Section 5.2. It is based on model (2.5) from Section 2. The data vectors $\boldsymbol{Y}_i$ have the form $\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$ with piecewise constant signal profiles $\boldsymbol{\mu}_i$. We set the number of clusters to $K_0 = 10$ and define the cluster-specific signal vectors $\boldsymbol{m}_k$ by

$$\begin{aligned}
\boldsymbol{m}_1 &= (\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})^\top, & \boldsymbol{m}_6 &= (-\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})^\top, \\
\boldsymbol{m}_2 &= (\mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0})^\top, & \boldsymbol{m}_7 &= (\mathbf{0}, -\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0})^\top, \\
\boldsymbol{m}_3 &= (\mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0})^\top, & \boldsymbol{m}_8 &= (\mathbf{0}, \mathbf{0}, -\mathbf{1}, \mathbf{0}, \mathbf{0})^\top, \\
\boldsymbol{m}_4 &= (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0})^\top, & \boldsymbol{m}_9 &= (\mathbf{0}, \mathbf{0}, \mathbf{0}, -\mathbf{1}, \mathbf{0})^\top, \\
\boldsymbol{m}_5 &= (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1})^\top, & \boldsymbol{m}_{10} &= (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, -\mathbf{1})^\top,
\end{aligned}$$

where $\mathbf{1} = (1, \dots, 1)$ and $\mathbf{0} = (0, \dots, 0)$ are vectors of length $p/5$. A graphical illustration of the signal vectors $\boldsymbol{m}_k$ is provided in Figure 7. The error terms $\varepsilon_{ij}$ are assumed to be i.i.d. normally distributed with mean 0 and variance $\sigma^2$. In the course of the simulation study, we consider different values of $n$, $p$ and $\sigma^2$ as well as different cluster sizes. To assess the noise level in the simulated data, we consider the ratios between the error variance $\sigma^2$ and the "variances" of the signals $\boldsymbol{m}_k$. In particular, we define the noise-to-signal ratios $\mathrm{NSR}_k(\sigma^2) = \sigma^2/\mathrm{Var}(\boldsymbol{m}_k)$, where $\mathrm{Var}(\boldsymbol{m}_k)$ denotes the empirical variance of the vector $\boldsymbol{m}_k$. Since $\mathrm{Var}(\boldsymbol{m}_k) \approx 0.16$ is the same for all $k$ in our design, we obtain that $\mathrm{NSR}_k(\sigma^2) = \mathrm{NSR}(\sigma^2) \approx \sigma^2/0.16$ for all $k$.

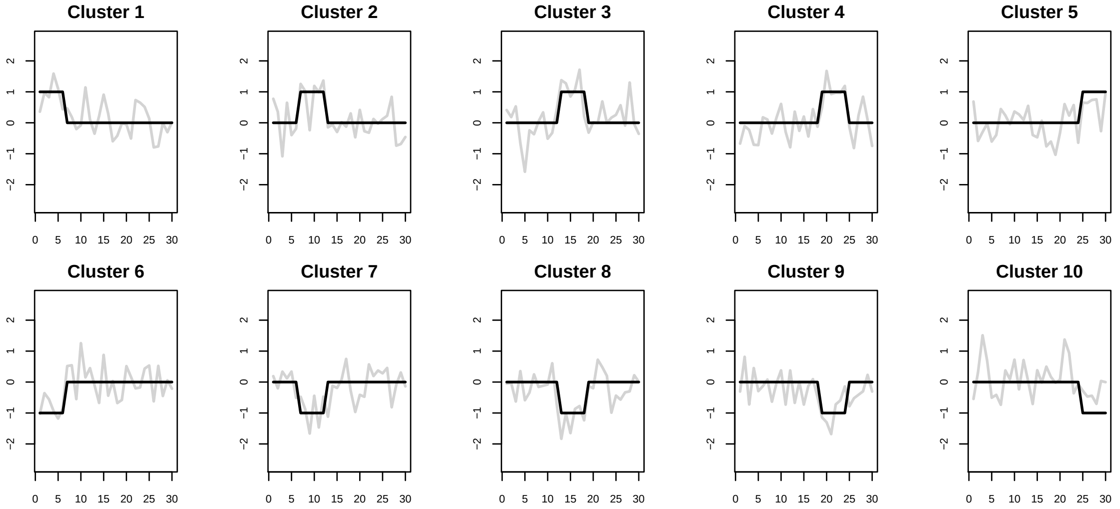

Figure 7: Visualization of the cluster-specific signal vectors $\boldsymbol{m}_k$ for the simulation study. The black lines represent the signal vectors $\boldsymbol{m}_k$ for $k = 1, \ldots, K_0 = 10$. The gray lines depict the data vectors $\boldsymbol{Y}_i = \boldsymbol{m}_k + \boldsymbol{\varepsilon}_i$ of a randomly selected member $i$ of the $k$-th cluster for each $k$. All plots are based on a setting with $p = 30$ and noise-to-signal ratio NSR $= 1.5$.

**Finite sample properties of CluStErr.** In this part of the simulation study, we analyze a design with equally sized clusters and set the sample size to $(n, p) = (1000, 30)$. Three different noise-to-signal ratios NSR are considered, in particular NSR $= 1$, 1.5 and 2. Since $\sigma^2 \approx 0.16\,\text{NSR}$, the corresponding error variances amount to $\sigma^2 \approx 0.16$, 0.25 and 0.32, respectively. The noise-to-signal ratio NSR $= 1$ mimics the noise level in the application on gene expression data from Section 5.2, where the estimated noise-to-signal ratios all lie between 0.6 and 1. The ratios NSR $= 1.5$ and NSR $= 2$ are used to investigate how the CluStErr method behaves when the noise level increases. We implement the CluStErr algorithm with $\alpha = 0.05$ and the difference-based estimators $\widehat{\sigma}^2_{\text{pc}}$ and $\widehat{\kappa}_{\text{pc}}$ from Section 3.4. For each of the three noise-to-signal ratios under consideration, we simulate $B = 1000$ samples and compute the estimate $\widehat{K}_0$ for each sample.

The simulation results are presented in Figure 8. Each panel shows a histogram of the estimates $\widehat{K}_0$ for a specific noise-to-signal ratio. For the ratio level NSR $= 1$, the CluStErr method produces very accurate results: About 95% of the estimates are equal to the true value $K_0 = 10$ and most of the remaining estimates take the value 11. For the ratio level NSR $= 1.5$, the estimation results are also quite precise: Most of the estimates take a value between 9 and 11 with around 55% of them being equal to the true value $K_0 = 10$. Only for the highest noise-to-signal ratio NSR $= 2$, the estimation results are less accurate. In this case, the noise level in the data is too high for the method to produce precise results. As one can see, the estimates have a strong downward bias, which can be explained as follows: When there is too much noise in the data, the test procedure on which the estimator $\widehat{K}_0$ is based does

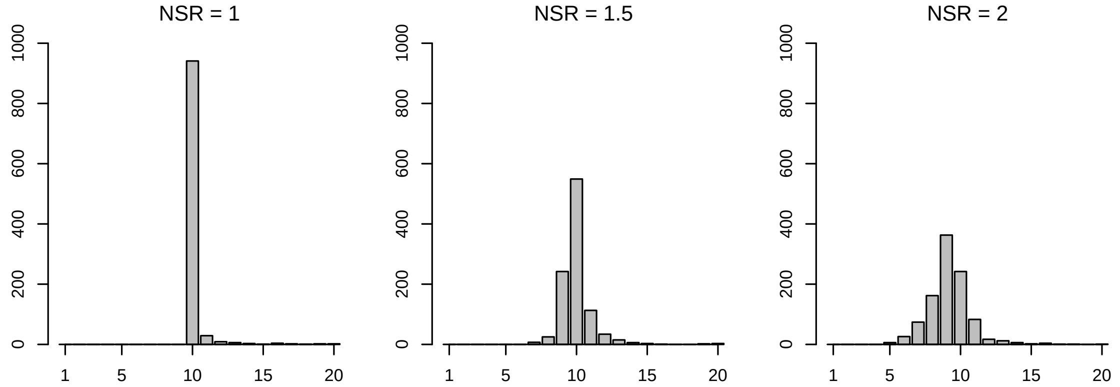

Figure 8: Histograms of the estimates $\widehat{K}_0$ in the three simulation scenarios with the noise-to-signal ratios NSR = 1, 1.5 and 2.

not have enough power to detect the alternative $H_1 : K < K_0$. As a result, our repeated test procedure stops too soon, thus underestimating the true number of clusters.

According to our theoretical results, the estimator $\widehat{K}_0$ allows for statistical error control in the following sense: It has the property that $\mathbb{P}(\widehat{K}_0 > K_0) = \alpha + o(1)$ and $\mathbb{P}(\widehat{K}_0 < K_0) = o(1)$, implying that $\mathbb{P}(\widehat{K}_0 = K_0) = (1 - \alpha) + o(1)$. Setting $\alpha$ to 0.05, we should thus observe that $\widehat{K}_0$ equals $K_0 = 10$ in approximately 95% of the simulations and overestimates $K_0$ in about 5% of them. Table 1 shows that this is indeed the case for the lowest noise-to-signal ratio NSR = 1. In this situation, the probability $\mathbb{P}(\widehat{K}_0 > K_0)$ of overestimating $K_0$ is around 5%, while the probability $\mathbb{P}(\widehat{K}_0 < K_0)$ of underestimating $K_0$ is 0%, implying that $\mathbb{P}(\widehat{K}_0 = K_0)$ is about 95%. For the two higher ratio levels NSR = 1.5 and 2, in contrast, the estimated values of the probabilities $\mathbb{P}(\widehat{K}_0 < K_0)$, $\mathbb{P}(\widehat{K}_0 = K_0)$ and $\mathbb{P}(\widehat{K}_0 > K_0)$ do not accurately match the values predicted by the theory. This is due to the fact that the statistical error control of the CluStErr method is asymptotic in nature. Table 2 illustrates this fact by reporting the estimated values of the probabilities $\mathbb{P}(\widehat{K}_0 < K_0)$, $\mathbb{P}(\widehat{K}_0 = K_0)$ and $\mathbb{P}(\widehat{K}_0 > K_0)$ for the noise-to-signal ratio NSR = 1.5 and various sample sizes $(n, p) = (1000, 30)$, $(1500, 40)$, $(2000, 50)$, $(2500, 60)$, $(3000, 70)$. As one can clearly see, the estimated probabilities gradually approach the values predicted by the theory as the sample size increases.

Table 1: Estimates of the probabilities $\mathbb{P}(\widehat{K}_0 < K_0)$, $\mathbb{P}(\widehat{K}_0 = K_0)$ and $\mathbb{P}(\widehat{K}_0 > K_0)$ in the three simulation scenarios with the noise-to-signal ratios NSR = 1, 1.5 and 2.

| | NSR = 1 | NSR = 1.5 | NSR = 2 |
|---|---|---|---|
| $\mathbb{P}(\widehat{K}_0 < K_0)$ | 0.000 | 0.274 | 0.631 |
| $\mathbb{P}(\widehat{K}_0 = K_0)$ | 0.941 | 0.549 | 0.242 |
| $\mathbb{P}(\widehat{K}_0 > K_0)$ | 0.059 | 0.177 | 0.127 |

Table 2: Estimates of the probabilities $\mathbb{P}(\widehat{K}_0 < K_0)$, $\mathbb{P}(\widehat{K}_0 = K_0)$ and $\mathbb{P}(\widehat{K}_0 > K_0)$ in the simulation scenario with NSR $= 1.5$ and five different sample sizes $(n, p)$.

| $(n,p)$ | $(1000,30)$ | $(1500,40)$ | $(2000,50)$ | $(2500,50)$ | $(3000,60)$ |
|---|---|---|---|---|---|
| $\mathbb{P}(\widehat{K}_0 < K_0)$ | 0.274 | 0.037 | 0.002 | 0.000 | 0.000 |
| $\mathbb{P}(\widehat{K}_0 = K_0)$ | 0.549 | 0.795 | 0.893 | 0.918 | 0.956 |
| $\mathbb{P}(\widehat{K}_0 > K_0)$ | 0.177 | 0.168 | 0.105 | 0.082 | 0.044 |

To summarize, our simulations on the finite sample behaviour of the CluStErr method indicate the following: (i) The method produces accurate estimates of $K_0$ as long as the noise level in the data is not too high. (ii) For sufficiently large sample sizes, it controls the probability of under- and overestimating the number of clusters $K_0$ as predicted by the theory. (iii) For smaller sample sizes, however, the error control is not fully accurate.

It is important to note that (iii) is not a big issue: Even in situations where the error control is not very precise, the CluStErr method may still produce accurate estimates of $K_0$. This is illustrated by our simulations. Inspecting the histogram of Figure 8 with NSR $= 1.5$, for example, the estimated probability $\mathbb{P}(\widehat{K}_0 = K_0)$ is seen to be only around 55% rather than 95%. Nevertheless, most of the estimates take a value between 9 and 11. Hence, in most of the simulations, the CluStErr method yields a reasonable approximation to the true number of clusters. From a heuristic perspective, the CluStErr method can indeed be expected to produce satisfying estimation results even in smaller samples when the error control is not very precise. This becomes clear when regarding CluStErr as a thresholding procedure. For $K = 1, 2, \ldots$, it checks whether the statistic $\widehat{\mathcal{H}}^{[K]}$ is below a certain threshold level $q$ and stops as soon as this is the case. For this approach to work, it is crucial to pick the threshold level $q$ appropriately. Our theoretical results suggest that the choice $q = q(\alpha)$ for a common significance level such as $\alpha = 0.05$ should be appropriate. Of course, this choice guarantees precise error control only for sufficiently large sample sizes. Nevertheless, in smaller samples, the threshold level $q = q(\alpha)$ can still be expected to be of the right order of magnitude, thus resulting in reasonable estimates of $K_0$.

**Comparison of CluStErr with competing methods.** We now compare the CluStErr method to other criteria for selecting the number of clusters $K_0$, in particular to (i) the gap statistic (Tibshirani et al., 2001), (ii) the silhouette statistic (Rousseeuw, 1987), and (iii) the Hartigan index (Hartigan, 1975). As before, we set the sample size to $(n, p) = (1000, 30)$ and consider the three noise-to-signal ratios NSR $= 1$, 1.5 and 2. In addition to a "balanced" scenario with clusters of the same size $n/K_0$ each, we also consider an "unbalanced" scenario with clusters of sizes $1 + 18k$ for $k = 1, \ldots, K_0$. For each design, we simulate $B = 100$ samples and

Table 3: Results of the comparison study. The entries of the table give the numbers of simulations (out of a total of 100) for which a certain estimate of $K_0$ is obtained. The first line in part (a) of the table, for example, has to be read as follows: The CluStErr estimate $\widehat{K}_0$ is equal to the true $K_0 = 10$ in 95 out of 100 simulations, and it is equal to $K = 11, 12, 13$ in $2, 1, 2$ simulations, respectively.

(a) balanced scenario

| | | Estimated number of clusters | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NSR | Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | $\geq 15$ |
| 1 | CluStErr | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 95 | 2 | 1 | 2 | 0 | 0 |
| | Gap | 3 | 5 | 0 | 0 | 1 | 3 | 11 | 25 | 36 | 16 | 0 | 0 | 0 | 0 | 0 |
| | Silhouette | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 19 | 43 | 20 | 11 | 1 | 0 | 1 |
| | Hartigan | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4 | 40 | 31 | 10 | 5 | 6 | 3 |
| 1.5 | CluStErr | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 22 | 59 | 10 | 3 | 3 | 1 | 1 |
| | Gap | 13 | 26 | 0 | 0 | 0 | 0 | 2 | 13 | 17 | 29 | 0 | 0 | 0 | 0 | 0 |
| | Silhouette | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 12 | 54 | 23 | 8 | 1 | 0 | 0 |
| | Hartigan | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 61 | 18 | 9 | 2 | 4 | 2 |
| 2 | CluStErr | 0 | 0 | 0 | 0 | 1 | 2 | 10 | 13 | 32 | 31 | 7 | 2 | 1 | 0 | 1 |
| | Gap | 22 | 26 | 2 | 0 | 0 | 0 | 1 | 6 | 9 | 34 | 0 | 0 | 0 | 0 | 0 |
| | Silhouette | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 66 | 26 | 2 | 0 | 0 | 0 |
| | Hartigan | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 68 | 18 | 2 | 2 | 0 | 0 |

(b) unbalanced scenario

| | | Estimated number of clusters | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NSR | Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | $\geq 15$ |
| 1 | CluStErr | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 94 | 4 | 2 | 0 | 0 | 0 |
| | Gap | 0 | 1 | 4 | 9 | 17 | 11 | 21 | 20 | 16 | 1 | 0 | 0 | 0 | 0 | 0 |
| | Silhouette | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 26 | 37 | 19 | 6 | 2 | 0 | 0 | 0 |
| | Hartigan | 0 | 0 | 1 | 2 | 7 | 8 | 21 | 21 | 12 | 11 | 7 | 2 | 2 | 1 | 5 |
| 1.5 | CluStErr | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 15 | 55 | 20 | 5 | 2 | 0 | 0 |
| | Gap | 2 | 4 | 4 | 3 | 8 | 15 | 21 | 21 | 22 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Silhouette | 0 | 0 | 0 | 0 | 0 | 1 | 7 | 20 | 47 | 24 | 1 | 0 | 0 | 0 | 0 |
| | Hartigan | 0 | 0 | 0 | 1 | 14 | 20 | 12 | 14 | 20 | 8 | 7 | 4 | 0 | 0 | 0 |
| 2 | CluStErr | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 17 | 25 | 31 | 15 | 3 | 6 | 0 | 0 |
| | Gap | 14 | 4 | 5 | 1 | 6 | 6 | 19 | 26 | 19 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Silhouette | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 28 | 48 | 19 | 0 | 0 | 0 | 0 | 0 |
| | Hartigan | 0 | 0 | 1 | 4 | 15 | 15 | 15 | 18 | 17 | 7 | 5 | 0 | 0 | 2 | 1 |

compare the estimated cluster numbers obtained from the CluStErr method with those produced by the three competing algorithms.

The CluStErr estimates are computed as described in the first part of the simulation study. The three competing methods are implemented with a $k$-means algorithm as the underlying clustering method. To compute the values of the gap statistic, we employ the `clusGap` function contained in the R package **cluster** (Maechler et al., 2016). The number of clusters is estimated by the function `maxSE` with the option `Tibs2001SEmax`. We thus determine the number of clusters as suggested in Tibshirani et al. (2001). To compute the silhouette and Hartigan statistics, we apply the R package **NbClust** (Charrad et al., 2015).

The results of the comparison study are presented in Table 3. Part (a) of the table provides the results for the balanced scenario with equal cluster sizes. As can be seen, the CluStErr method clearly outperforms its competitors in the setting with NSR $= 1$. In the scenario with NSR $= 1.5$, it also performs well in comparison to the other methods. Only for the highest noise-to-signal ratio NSR $= 2$, it produces estimates of $K_0$ with a strong downward bias and is outperformed by the silhouette and Hartigan statistics. Part (b) of Table 3 presents the results for the unbalanced scenario where the clusters strongly differ in size. In this scenario, all of the three competing methods substantially underestimate the number of clusters. The CluStErr method, in contrast, provides accurate estimates of $K_0$ in the two designs with NSR $= 1$ and NSR $= 1.5$. Only in the high-noise design with NSR $= 2$, it produces estimates with a substantial downward bias, which nevertheless is much less pronounced than that of its competitors.

To summarize, the main findings of our comparison study are as follows: (i) The CluStErr method performs well in comparison to its competitors as long as the noise-to-signal ratio is not too high. It is however outperformed by its competitors in a balanced setting when the noise level is high. In Section 6, we discuss some modifications of the CluStErr method to improve its behaviour in the case of high noise. (ii) The CluStErr method is able to deal with both balanced and unbalanced cluster sizes, whereas its competitors perform less adequately in unbalanced settings.

The findings (i) and (ii) can heuristically be explained as follows: The CluStErr method is based on the test statistic $\widehat{\mathcal{H}}^{[K]} = \max_{1 \le i \le n} \widehat{\Delta}_i^{[K]}$, which is essentially the maximum over the residual sums of squares of the various individuals $i$. Its competitors, in contrast, are based on statistics which evaluate averages rather than maxima. Hartigan's rule, for instance, relies on a statistic which is essentially a scaled version of the ratio $\text{RSS}(K)/\text{RSS}(K+1)$, where $\text{RSS}(K)$ is defined as in (3.13) and denotes the average residual sum of squares for a partition with $K$ clusters. Averaging the residual sums of squares reduces the noise in the data much more strongly than taking the maximum. This is the reason why Hartigan's rule tends to perform better than the CluStErr method in a balanced setting with high noise. On the other hand, the average residual sum of squares hardly reacts to changes in the residual sums of squares of a few individuals that form a small cluster. Hence, small clusters are effectively ignored when taking the average of the residual sums of squares. This is the reason why Hartigan's statistic is not able to deal adequately with unbalanced settings. Taking the maximum of the residual sums of squares instead allows us to handle even highly unbalanced cluster sizes.

# 6 Extensions

In this paper, we have developed an approach for estimating the number of clusters with statistical error control. We have derived a rigorous mathematical theory for a model with convex spherical clusters. This model is widely used in practice and is suitable for a large number of applications. Nevertheless, it of course has some limitations. In particular, it is not suitable for applications where the clusters have non-convex shapes. An interesting question is how to extend our ideas to the case of general, potentially non-convex clusters. Developing theory for this general case is a very challenging problem. We have made a first step into this direction by providing theory for the case of spherical clusters.

There are several ways to modify and extend our estimation methods in the model setting (2.1)–(2.2). So far, we have based our methods on the maximum statistic $\widehat{\mathcal{H}}^{[K]} = \max_{1 \le i \le n} \widehat{\Delta}_i^{[K]}$. However, we are not bound to this choice. Our approach can be based on any test statistic $\widehat{\mathcal{H}}^{[K]}$ that fulfills the higher-order property (3.4). The maximum statistic serves as a baseline which may be modified and improved in several directions. The building blocks of the maximum statistic are the individual statistics $\widehat{\Delta}_i^{[K]}$. Their stochastic behaviour has been analyzed in detail in Section 3.2. Under the null hypothesis $H_0 : K = K_0$, the statistics $\widehat{\Delta}_i^{[K]}$ are approximately independent and distributed as $(\chi_p^2 - p)/\sqrt{2p}$ variables. Under the alternative $H_1 : K < K_0$ in contrast, they have an explosive behaviour at least for some $i$. This difference in behaviour suggests to test $H_0$ by checking whether the hypothesis

$$H_{0,i} : \widehat{\Delta}_i^{[K]} \text{ has a } (\chi_p^2 - p)/\sqrt{2p} \text{ distribution}$$

holds for all subjects $i = 1, \dots, n$. We are thus faced with a multiple testing problem. A maximum statistic is a classical tool to tackle this problem. However, as is well known from the field of multiple testing, maximum statistics tend to be fairly conservative. When the noise level in the data is high, a test based on the maximum statistic $\widehat{\mathcal{H}}^{[K]} = \max_{1 \le i \le n} \widehat{\Delta}_i^{[K]}$ can thus be expected to have low power against the alternative $H_1 : K < K_0$. As a consequence, the repeated test procedure on which the estimator $\widehat{K}_0$ is based tends to stop too soon, thus underestimating the true number of clusters. This is exactly what we have seen in the high-noise scenarios of the simulation study from Section 5.3. We now present two ways how to construct a statistic $\widehat{\mathcal{H}}^{[K]}$ with better power properties.

**A blocked maximum statistic.** Let $w_0 = 0$ and define $w_k = \sum_{r=1}^k \#\widehat{G}_r^{[K]}$ for $1 \le k \le K$. Moreover, write $\widehat{G}_k^{[K]} = \{i_{w_{k-1}+1}, \dots, i_{w_k}\}$ with $i_{w_{k-1}+1} < \ldots < i_{w_k}$ for any $k$. To start with, we order the indices $\{1, \dots, n\}$ clusterwise. In particular, we write them as $\{i_1, i_2, \ldots, i_n\}$, which yields the ordering

$$\overbrace{i_1 < \ldots < i_{w_1}}^{\widehat{G}_1^{[K]}} \quad \overbrace{i_{w_1+1} < \ldots < i_{w_2}}^{\widehat{G}_2^{[K]}} \quad \ldots\ldots \quad \overbrace{i_{w_{K-1}+1} < \ldots < i_{w_K}}^{\widehat{G}_K^{[K]}} .$$

We next partition the ordered indices into blocks

$$B_\ell^{[K]} = \big\{ i_{(\ell-1)N+1}, \ldots, i_{\ell N \wedge n} \big\} \quad \text{for } 1 \le \ell \le L,$$

where $N$ is the block length and $L = \lceil n/N \rceil$ is the number of blocks. With this notation at hand, we construct blockwise averages

$$\widehat{\Lambda}_\ell^{[K]} = \frac{1}{\sqrt{N}} \sum_{i \in B_\ell^{[K]}} \widehat{\Delta}_i^{[K]}$$

of the individual statistics $\widehat{\Delta}_i^{[K]}$ and aggregate them by taking their maximum, thus defining

$$\widehat{\mathcal{H}}_B^{[K]} = \max_{1 \le \ell \le L} \widehat{\Lambda}_\ell^{[K]}.$$

In addition, we let $q_B(\alpha)$ be the $(1-\alpha)$-quantile of

$$\mathcal{H}_B = \max_{1 \le \ell \le L} \Lambda_\ell \quad \text{with} \quad \Lambda_\ell = \frac{1}{\sqrt{N}} \sum_{i=(\ell-1)N+1}^{\ell N} Z_i,$$

where $Z_i$ are i.i.d. variables with the distribution $(\chi_p^2 - p)/\sqrt{2p}$. Note that this definition of $\widehat{\mathcal{H}}_B^{[K]}$ nests the maximum statistic $\widehat{\mathcal{H}}^{[K]} = \max_{1 \le i \le n} \widehat{\Delta}_i^{[K]}$ as a special case with the block length $N = 1$.

Under appropriate restrictions on the block length $N$, the estimators that result from applying the CluStErr method with the blocked statistic $\widehat{\mathcal{H}}_B^{[K]}$ can be shown to have the theoretical properties stated in Theorems 4.1–4.3. More specifically, Theorems 4.1–4.3 can be shown to hold true for the blocked statistic $\widehat{\mathcal{H}}_B^{[K]}$ if the following two restrictions are satisfied: (i) $N/p^{1-\eta} = O(1)$ for some small $\eta > 0$, that is, the block length $N$ diverges more slowly than $p$. (ii) $\#G_k/n \to c_k > 0$ for all $k$, that is, the cluster sizes $\#G_k$ all grow at the same rate. Condition (ii) essentially rules out strongly differing cluster sizes. It is not surprising that we require such a restriction: To construct the blocked statistic $\widehat{\mathcal{H}}_B^{[K]}$, we average over the individual statistics $\widehat{\Delta}_i^{[K]}$. As already discussed in the context of the simulation study of Section 5.3, averaging has the effect that small clusters are effectively ignored. Hence, in contrast to the maximum statistic $\widehat{\mathcal{H}}^{[K]} = \max_{1 \le i \le n} \widehat{\Delta}_i^{[K]}$, the blocked statistic $\widehat{\mathcal{H}}_B^{[K]}$ with a large block size $N$ can be expected not to perform adequately when the

Table 4: Simulation results for the blocked maximum statistic.

| | Estimated number of clusters | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NSR | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 98 | 1 | 1 | 0 | 0 |
| 1.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 79 | 12 | 5 | 2 | 2 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 73 | 21 | 2 | 3 | 1 |

cluster sizes are highly unbalanced.

In balanced settings, however, the blocked statistic $\widehat{\mathcal{H}}_B^{[K]}$ can be shown to have better power properties than the maximum statistic when the block size $N$ is chosen sufficiently large. To see this, we examine the behaviour of $\widehat{\mathcal{H}}_B^{[K]}$ for different block lengths $N$. Our heuristic discussion of the individual statistics $\widehat{\Delta}_i^{[K]}$ from Section 3.2 directly carries over to the blocked versions $\widehat{\Lambda}_\ell^{[K]}$: With the help of (3.10), it is easy to see that

$$\mathbb{P}\Big(\widehat{\mathcal{H}}_B^{[K_0]} \le q_B(\alpha)\Big) \approx (1-\alpha).$$

Moreover, (3.11) together with some additional arguments suggests that $\widehat{\mathcal{H}}_B^{[K]}$ has an explosive behaviour for $K < K_0$. Specifically, under our technical conditions from Section 4.1 and the two additional restrictions (i) and (ii) from above, we can show that

$$\widehat{\mathcal{H}}_B^{[K]} \ge c\sqrt{Np} \quad \text{for some } c > 0 \text{ with prob. tending to 1.} \tag{6.1}$$

As the quantile $q_B(\alpha)$ grows at the slower rate $\sqrt{\log L}$ ($\le \sqrt{\log n}$), we can conclude that

$$\mathbb{P}\Big(\widehat{\mathcal{H}}_B^{[K]} \le q_B(\alpha)\Big) = o(1)$$

for $K < K_0$. As a result, $\widehat{\mathcal{H}}_B^{[K]}$ should satisfy the higher-order property (3.4). Moreover, according to (6.1), the statistic $\widehat{\mathcal{H}}_B^{[K]}$ explodes at the rate $\sqrt{Np}$ for $K < K_0$. Hence, the faster the block size $N$ grows, the faster $\widehat{\mathcal{H}}_B^{[K]}$ diverges to infinity. Put differently, the larger $N$, the more power we have to detect that $K < K_0$. This suggests to select $N$ as large as possible. According to restriction (i) from above, we may choose any $N$ with $N/p^{1-\eta} = O(1)$ for some small $\eta > 0$. Ideally, we would thus like to pick $N$ so large that it grows at the same rate as $p^{1-\eta}$. In practice, we neglect the small constant $\eta > 0$ and set $N = p$ as a simple rule of thumb.

According to the heuristic arguments from above, the blocked maximum statistic $\widehat{\mathcal{H}}_B^{[K]}$ with block length $N = p$ should be more powerful than the maximum statistic $\widehat{\mathcal{H}}^{[K]} = \max_{1\le i\le n} \widehat{\Delta}_i^{[K]}$ in settings with balanced cluster sizes. We examine this claim with the help of some simulations. To do so, we return to the balanced scenario of the comparison study in Section 5.3. For each of the data samples that were simulated for this scenario, we compute the CluStErr estimate of $K_0$ based on the blocked statistic $\widehat{\mathcal{H}}_B^{[K]}$ with $N = p$. Table 4 presents the results. It shows that the blocked CluStErr method yields accurate estimates of $K_0$ for all three noise-to-signal ratios. Comparing the results to those in Table 3(a), the blocked method can be seen to perform very well in comparison to the other procedures even in the high-noise setting with NSR $= 2$. This clearly shows the gain in power induced by the block structure of the statistic.

**An FDR-based statistic.** There are several approaches in the literature how to construct multiple testing procedures that have better power properties than the classical maximum statistic. Prominent examples are methods that control the false discovery rate (FDR) or the higher criticism procedure by Donoho and Jin (2004). We may try to exploit ideas from these approaches to construct a more powerful statistic $\widehat{\mathcal{H}}^{[K]}$. As an example, we set up a test statistic which uses ideas from the FDR literature: Rather than only taking into account the maximum of the statistics $\widehat{\Delta}_i^{[K]}$, we may try to exploit the information in all of the ordered statistics $\widehat{\Delta}_{(1)}^{[K]} \ge \ldots \ge \widehat{\Delta}_{(n)}^{[K]}$. In particular, following Simes (1986) and Benjamini and Hochberg (1995), we may set up our test for a given number of clusters $K$ as follows: Reject $H_0 : K = K_0$ if

$$\widehat{\Delta}_{(i)}^{[K]} > q_\chi\Big(\frac{i}{n}\alpha\Big) \quad \text{for some } i \in \{1, \ldots, n\},$$

where $q_\chi(\beta)$ is the $(1-\beta)$-quantile of the distribution $(\chi_p^2 - p)/\sqrt{2p}$. This procedure can be rephrased as follows: Define the statistic

$$\widehat{\mathcal{H}}_{\text{FDR}}^{[K]} = \max_{1 \le i \le n} \frac{\widehat{\Delta}_{(i)}^{[K]}}{q_\chi(i\alpha/n)}$$

and reject $H_0$ if $\widehat{\mathcal{H}}_{\text{FDR}}^{[K]} > 1$. Instead of taking the maximum over the original statistics $\widehat{\Delta}_i^{[K]}$, we thus take the maximum over rescaled versions of the ordered statistics $\widehat{\Delta}_{(i)}^{[K]}$. Developing theory for the FDR-type statistic $\widehat{\mathcal{H}}_{\text{FDR}}^{[K]}$ is a very interesting topic which is however far from trivial.

# References

Aibar, S., Fontanillo, C. and De Las Rivas, J. (2013). *LeukemiasEset: Leukemia's microarray gene expression data (expressionSet).* R package version 1.8.0. https://bioconductor.org/packages/release/data/experiment/html/leukemiasEset.html.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, **57** 289–300.

Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A. (2015). *NbClust: determining the best number of clusters in a data set.* R package version 3.0. https://cran.r-project.org/web/packages/NbClust/.

Chen, J., Li, P. and Fu, Y. (2012). Inference on the order of a normal mixture. *Journal of the American Statistical Association*, **107** 1096–1105.

Chipman, H., Hastie, T. J. and Tibshirani, R. (2003). Clustering microarray data. In *Statistical Analysis of Gene Expression Microarray Data* (T. Speed, ed.). Chapman & Hall / CRC, Boca Raton, 161–203.

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, **52** 543–547.

DeGaetano, A. T. (2001). Spatial grouping of United States climate stations using a hybrid clustering approach. *International Journal of Climatology*, **21** 791–807.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, **23** 1499–1501.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, **32** 962–994.

Fisher, D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, **53** 789–798.

Fovell, R. G. and Fovell, M.-Y. C. (1993). Climate zones of the conterminous United States defined using cluster analysis. *Journal of Climate*, **6** 2103–2135.

Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer Vol. II.* (L. M. Le Cam and R. A. Olshen, eds.). Wadsworth, 789–806.

Gordon, A. (1999). *Classification.* Chapman & Hall.

Haferlach, T., Kohlmann, A., Wieczorek, L., Basso, G., Te Kronnie, G., Bene, M.-C., De Vos, J., Hernandez, J. M., Hofmann, W.-K., Mills, K. I., Gilkes, A., Chiaretti, S., Shurtleff, S. A., Kipps, T. J., Rassenti, L. Z., Yeoh, A. E., Papenhausen, P. R., Liu, W.-M., Williams, P. M. and Foa, R. (2010). Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the International Microarray Innovations in Leukemia Study Group. *Journal of Clinical Oncology*, **28** 2529–2537.

Hall, P., Kay, J. and Titterington, D. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77** 521–528.

Hartigan, J. A. (1975). *Clustering Algorithms.* John Wiley & Sons, New York.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer Vol. II.*

(L. M. Le Cam and R. A. Olshen, eds.). Wadsworth, 807–810.

Jiang, D., Tang, C. and Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, **16** 1370–1386.

Lasota, L., Vogt, M. and Schmid, M. (2017). *CluStErr: Clustering with Statistical Error Control.* R package version 1.0.

Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*, **105** 1084–1092.

Li, P., Chen, J. and Marriott, P. (2009). Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, **96** 411–426.

Liu, X. and Shao, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference*, **123** 61–81.

Maechler, M., Rousseeuw, P., Struyf, A. and Hubert, M. (2016). *Finding groups in data: cluster analysis extended Rousseeuw et al.* R package version 2.0.4. https://cran.r-project.org/web/packages/cluster/.

Maitra, R., Melnykov, V. and Lahiri, S. N. (2012). Bootstrapping for significance of compact clusters in multidimensional datasets. *Journal of the American Statistical Association*, **107** 378–392.

McLachlan, G. J. and Rathnayake, S. (2014). On the number of components in a Gaussian mixture model. *WIREs Data Mining and Knowledge Discovery*, **4** 341–355.

Müller, H.-G., and Stadtmüller, U. (1988). Detecting dependencies in smooth regression models. *Biometrika*, **75** 639–650.

Peel, M. C., Finlayson, B. L. and McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, **11** 1633–1644.

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickham, C. and Mosher, S. (2013). Berkeley Earth Temperature averaging process. *Geoinformatics & Geostatistics: An Overview*, **1:2**.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20** 53–65.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73** 751–754.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, **63** 411–423.

Wedel, M. and Kamakura, W. (2000). *Market segmentation: conceptual and methodological foundations.* Springer.