Clustering with statistical error control

Michael Vogt¹ | Matthias Schmid²

¹Department of Economics and Hausdorff Center for Mathematics, University of Bonn
²Department of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn

Abstract
This article 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 article, formal versions of these statements on statistical error control are derived in a baseline model with convex clusters. A simulation study and two applications to temperature and gene expression microarray data complement the theoretical analysis.

KEYWORDS
k-means clustering, cluster analysis, multiple statistical testing, number of clusters, statistical error control

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 (Frühwirth-Schnatter, Celeux, & Robert, 2018; Jiang, Tang, & Zhang, 2004), the clustering of temperature...
perature curves using data recorded on a spatial grid (DeGaetano, 2001; Finazzi, Haggarty, Miller, Scott, & Fasso, 2015; Fovell & Fovell, 1993), and the clustering of consumer profiles on the basis of survey data (Tripathi, Bhardwaj, & Poovammal, 2018; Wedel & 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 that 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, Walther, & Hastie, 2001).

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 that checks either 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, Chen, and Marriott (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, Li, and Fu (2012) have constructed a new expectation-maximization procedure to tackle this testing problem. Outside the mixture model context, Maitra, Melnykov, and Lahiri (2012) have developed a bootstrap procedure to test a model with $K$ groups against one with $K' > K$ groups. These articles derive the theoretical properties of the proposed tests under the null hypothesis of $K$ clusters, where $K$ is a prespecified fixed number. However, they do not formally investigate the properties of a procedure that 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 aim of this article is to construct an estimator $\hat{K}_0$ of $K_0$, which allows for rigorous statistical error control in the following sense: For any prespecified significance level $\alpha \in (0, 1)$, the proposed estimator $\hat{K}_0 = \hat{K}_0(\alpha)$ has the property that

$$P(\hat{K}_0 > K_0) = \alpha + o(1), \quad (1)$$

$$P(\hat{K}_0 < K_0) = o(1). \quad (2)$$

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

Estimating $K_0$ with statistical error control in the sense of (1) and (2) is a challenging problem. In this article, we make a first step toward solving this problem by providing statistical methodology and rigorous mathematical theory for a simple baseline model. Specifically, we construct an estimator $\hat{K}_0$ with the properties (1) and (2) in a model with convex spherical clusters, which is formally introduced in Section 2. This model is widely used in practice and suitable for a number of applications as discussed in Section 2 and illustrated by our empirical examples in Section 5. It should be noted though that it is not suited for applications where the clusters have general, non-convex shapes. Restricting attention to the simple baseline model of Section 2 allows us to present our main ideas in a clean way and to derive a complete asymptotic theory for our estimators.

Our estimation approach is developed in Section 3. As will be shown there, the proposed procedure does not only provide us with an estimator $\hat{K}_0$ of the unknown number of groups $K_0$. It also yields estimators of the groups themselves, which allow for statistical error control similarly to $\hat{K}_0$. Our approach is based on the following general strategy:

1. Construct a statistical test that, for any given number $K$, checks the null hypothesis that there are $K$ clusters in the data.
2. Starting with $K = 1$, sequentially apply this test until it does not reject the null hypothesis of $K$ clusters any more.
3. Define the estimator $\hat{K}_0$ 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) and (2) on statistical error control under suitable regularity conditions (Section 4). Some alternative choices of the test statistic are discussed in Section 6. In the following, we refer to our estimation procedure as CluStErr (“Clustering with Statistical Error Control”).

We complement the theoretical analysis of the article 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, Vogt, & Schmid, 2019), which implements the CluStErr method and which is part of the supplemental materials of the article.

## 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 $Y_i = (Y_{i1}, \ldots, Y_{ip})^\top$, where $Y_{ij}$ denotes the measurement of the $j$th feature for the $i$th subject. Our data sample thus has the form $\{Y_i : 1 \leq i \leq n\}$. 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 dataset, 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 $Y_i$ of the various subjects $i = 1, \ldots, n$ are supposed to satisfy the model

$$Y_i = \mu_i + e_i,$$  \hspace{1cm} (3)

where $\mu_i = (\mu_{i1}, \ldots, \mu_{ip})^T$ is a deterministic signal vector and $e_i = (e_{i1}, \ldots, e_{ip})^T$ 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 $k = 1, \ldots, K_0$,

$$\mu_i = m_k \text{ for all } i \in G_k,$$  \hspace{1cm} (4)

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

Equations (3) and (4) specify a model with convex spherical clusters, which underlies the $k$-means and many other Euclidean distance-based clustering algorithms. Despite its simplicity, this framework has frequently been employed in the literature and is useful for a number of applications as illustrated by our empirical examples. It is thus a suitable baseline model for developing our ideas on clustering with statistical error control. We now discuss Equations (3) and (4) in detail.

**Details on Equation (3).** The noise vector $e_i = (e_{i1}, \ldots, e_{ip})^T$ is assumed to consist of entries $e_{ij}$ with the additive component structure $e_{ij} = \alpha_i + \varepsilon_{ij}$. Equation (3) for the $i$th subject thus writes as

$$Y_i = \mu_i + \alpha_i + \varepsilon_i,$$  \hspace{1cm} (5)

where $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ip})^T$ and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})^T$. Here, $\alpha_i$ is a subject-specific random intercept term. Moreover, the terms $\varepsilon_{ij}$ are standard idiosyncratic noise variables with $\mathbb{E}[\varepsilon_{ij}] = 0$. For simplicity, 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 (5) may depend on $p$. The exact formulation of Equation (5) for the $i$th subject thus reads $Y_{i,p} = \mu_{i,p} + \alpha_{i,p} + \varepsilon_{i,p}$, where $Y_{i,p} = (Y_{i1,p}, \ldots, Y_{ip,p})^T$, $\mu_{i,p} = (\mu_{i1,p}, \ldots, \mu_{ip,p})^T$, $\alpha_{i,p} = (\alpha_{i1,p}, \ldots, \alpha_{ip,p})^T$, and $\varepsilon_{i,p} = (\varepsilon_{i1,p}, \ldots, \varepsilon_{ip,p})^T$. However, to keep the notation simple, we suppress this dependence on $p$ and write the model for the $i$th subject as (5). If we drop the random intercept $\alpha_i$ from (5), the signal vector $\mu_i$ is equal to the mean $\mathbb{E}[Y_i]$. In the general Equation (5), in contrast, $\mu_i$ is only identified up to an additive constant. To identify $\mu_i$ in (5), we impose the normalization constraint $p^{-1} \sum_{j=1}^p \mu_{ij} = 0$ for each $i$. We thus normalize the entries of $\mu_i$ 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 $\mu_i$.

**Details on Equation (4).** 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 small compared with 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 article.

In the remainder of this section, we discuss two special cases of model (3)–(4), which are relevant for our empirical applications.
A model for the clustering of time series data. Suppose we observe time series \( Y_i = (Y_{i1}, \ldots, Y_{ip})^T \) of length \( p \) for \( n \) different subjects \( i \). The time series \( Y_i \) of the \( i \)th subject is assumed to follow the time trend model

\[
Y_{ij} = \mu_i(t_j) + \alpha_i + \epsilon_{ij} \quad (1 \leq j \leq p),
\]

where \( \mu_i(\cdot) \) is an unknown nonparametric trend function and \( t_1 < \ldots < t_p \) are the time points of the observations. 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 nonequidistant designs. To identify the trend function \( \mu_i(\cdot) \) in (6), 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 (5). 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 \( Y_i = (Y_{i1}, \ldots, Y_{ip})^T \), where \( Y_{ij} \) is the measured expression level of gene \( i \) for tissue sample \( j \). The vector \( Y_i \) of gene \( i \) is supposed to satisfy the model Equation (5), which componentwise reads as

\[
Y_{ij} = \mu_{ij} + \alpha_i + \epsilon_{ij} \quad (1 \leq j \leq p).
\]

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 \( \alpha_i + \epsilon_{ij} \). 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 labeled by \( j_{t-1} \leq j < j_t \), where \( 1 = j_0 < j_1 < \cdots < j_{T-1} < j_T = p + 1 \). 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, that is, \( \mu_{ij} = \mu_{ij'} \) for \( j_{t-1} \leq j, j' < j_t \). The signal vector \( \mu_i \) thus has a piecewise constant structure for each \( i \). As in our general model, we suppose that there are \( K_0 \) groups of genes \( G_1, \ldots, G_{K_0} \) such that \( \mu_i = m_k \) for all \( i \in G_k \) and some vector \( m_k \). The genes of each class \( G_k \) thus have the same (co)expression profile \( m_k \).

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 (3)–(4). 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 \( \{ \hat{G}_k^{[K]} : 1 \leq k \leq K \} \) be the \( K \) clusters produced by the algorithm when the number of clusters is \( K \). For \( K = 1 \), we trivially set \( \hat{G}_1^{[1]} = \{1, \ldots, n\} \). For our theory to work, we require the clustering algorithm to consistently estimate the class structure \( \{G_k : 1 \leq k \leq K_0\} \) when \( K = K_0 \). More specifically, we require the estimators \( \{ \hat{G}_k^{[K]} : 1 \leq k \leq K_0\} \) to have the property that

\[
\mathbb{P}(\{ \hat{G}_k^{[K_0]} : 1 \leq k \leq K_0\} = \{G_k : 1 \leq k \leq K_0\}) \to 1. \tag{8}
\]

This restriction is satisfied by a wide range of clustering algorithms under our regularity conditions. It is, for example, satisfied by a \( k \)-means type algorithm introduced in Section 3.3. 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 (8).

In the second step, we construct a test for each \( K \), which checks whether the data can be well described by the \( K \) clusters \( \hat{G}_1^{[K]}, \ldots, \hat{G}_K^{[K]} \). We thereby test whether the number of clusters is equal to \( K \). More formally, we use the \( K \)-cluster partition \( \{ \hat{G}_k^{[K]} : 1 \leq k \leq K \} \) to construct a statistic \( \hat{H}^{[K]}(\theta) \) 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 \( T_{\alpha}(\theta) = 1(\hat{H}^{[K]} > q(\alpha)) \), 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 \( T_{\alpha}(\theta) = 1 \), that is, if \( \hat{H}^{[K]} > q(\alpha) \). A detailed construction of the statistic \( \hat{H}^{[K]} \) 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 \( \hat{H}^{[K]}(\theta) \leq q(\alpha) \) 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 \( \hat{H}^{[K]}(\theta) \leq q(\alpha) \), that is, for which the test does not reject \( H_0 \). Formally speaking, we define

\[
\hat{K}_0 = \min\{K = 1, 2, \ldots \mid \hat{H}^{[K]}(\theta) \leq q(\alpha)\}. \tag{9}
\]

Moreover, we estimate the class structure \( \{G_k : 1 \leq k \leq K_0\} \) by the partition \( \{\hat{G}_k : 1 \leq k \leq \hat{K}_0\} \), where we set \( \hat{G}_k = \hat{G}_k^{[\hat{K}_0]} \). The definition (9) can equivalently be written as

\[
\hat{K}_0 = \min\{K = 1, 2, \ldots \mid \hat{p}^{[K]} > \alpha\}, \tag{10}
\]

where \( \hat{p}^{[K]} \) is the \( p \)-value corresponding to the statistic \( \hat{H}^{[K]} \). The heuristic idea behind (10) 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 \( \{\hat{G}_k^{[K]} : 1 \leq k \leq K\} \). For each \( K \), we compute the \( p \)-value \( \hat{p}^{[K]} \) that expresses our confidence in a model with \( K \) clusters. We stop as soon as \( \hat{p}^{[K]} > \alpha \), 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 \( \hat{H}^{[K]} \) has the property that

\[
\mathbb{P}(\hat{H}^{[K]} \leq q(\alpha)) = \begin{cases} o(1) & \text{for } K < K_0 \\ (1 - \alpha) + o(1) & \text{for } K = K_0. \end{cases} \tag{11}
\]

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 (11), it follows that

\[
\pi_>(\alpha) := \mathbb{P}(\hat{K}_0 > K_0) = \alpha + o(1) \tag{12}
\]
\[
\pi_<(\alpha) := \mathbb{P}(\hat{K}_0 < K_0) = o(1). \tag{13}
\]

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 \( \alpha' + o(1) = \pi_>(\alpha') \leq \pi_>(\alpha) = \alpha + o(1) \) 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 \( \hat{K}_0 \) is based. Taken together, the two statements (12) and (13) yield that

\[
\mathbb{P}(\hat{K}_0 \neq K_0) = \alpha + o(1), \tag{14}
\]

that is, the probability that the estimated number of classes \( \hat{K}_0 \) differs from the true number of classes \( K_0 \) is asymptotically equal to \( \alpha \). With the help of (14) and the consistency property (8) of the estimated clusters, we can further show that

\[
\mathbb{P}(\{\hat{G}_k : 1 \leq k \leq \hat{K}_0\} \neq \{G_k : 1 \leq k \leq K_0\}) = \alpha + o(1), \tag{15}
\]

that is, the probability of making a classification error is asymptotically equal to \( \alpha \) as well. The statements (12)–(15) give a mathematically precise description of the statistical error control that can be performed by our method.

### 3.2 Construction of the statistic \( \hat{H}^{[K]} \)

To construct the statistic \( \hat{H}^{[K]} \), we use the following notation:

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

\[
\bar{Y}_{ij} = \mu_{ij} + \varepsilon_{ij} - \frac{1}{p} \sum_{j=1}^{p} \mu_{ij} - \frac{1}{p} \sum_{j=1}^{p} \varepsilon_{ij} = \mu_{ij} + \varepsilon_{ij} + O(p^{-1/2}) = Y_{ij}^* + O(p^{-1/2}).
\]

2. For any set \( S \subseteq \{1, \ldots, n\} \), let \( m_{i,S} = (\#S)^{-1} \sum_{i \in S} \mu_{ij} \) be the average of the signals \( \mu_{ij} \) with \( i \in S \) and estimate it by \( \hat{m}_{j,S} = (\#S)^{-1} \sum_{i \in S} \bar{Y}_{ij} \). We use the notation \( m_{j,k}^{[K]} = m_{j,\hat{G}_k^{[K]}} \) and \( \hat{m}_{j,k}^{[K]} = \hat{m}_{j,\hat{G}_k^{[K]}} \) to denote the average of the signals in the cluster \( \hat{G}_k^{[K]} \) and its estimator, respectively.
3. For any cluster \( \hat{G}_k^{[K]} \), we define cluster-specific residuals by setting \( \hat{\varepsilon}_{ij}^{[K]} = \hat{Y}_{ij} - \hat{m}_{j,k}^{[K]} \) for \( i \in \hat{G}_k^{[K]} \) and \( 1 \leq j \leq p \).

4. Let \( \hat{\sigma}^2 \) be an estimator of the error variance \( \sigma^2 = \mathbb{E}[\varepsilon_{ij}^2] \). Moreover, let \( \hat{\kappa} \) 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 \( \hat{\sigma}^2 \) and \( \hat{\kappa} \).

With this notation at hand, we define the statistic

\[
\hat{\Delta}_{i}^{[K]} = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \left\{ \left( \frac{\hat{\varepsilon}_{ij}^{[K]}}{\hat{\sigma}} \right)^2 - 1 \right\} / \hat{\kappa}
\]

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, \( \hat{\Delta}_{i}^{[K]} \) measures how well the data of the \( i \)th subject are described when the sample of subjects is partitioned into the \( K \) clusters \( \hat{G}_1^{[K]}, \ldots, \hat{G}_K^{[K]} \).

The individual statistics \( \hat{\Delta}_{i}^{[K]} \) are the building blocks of the overall statistic \( \hat{H}^{[K]} \).

Before we move on with the construction of \( \hat{H}^{[K]} \), we have a closer look at the stochastic behavior of the statistics \( \hat{\Delta}_{i}^{[K]} \). 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 \( \hat{Y}_{ij}, \hat{m}_{j,k}^{[K]}, \hat{\sigma}^2, \) and \( \hat{\kappa} \). In this situation,

\[
\hat{\Delta}_{i}^{[K]} = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \left\{ \frac{(\varepsilon_{ij} + d_{ij})^2}{\sigma^2} - 1 \right\} / K
\]

for any \( i \in S = \hat{G}_k^{[K]} \), where \( d_{ij} = \mu_{ij} - (\#S)^{-1} \sum_{t \in S} \mu_{t,j} \) 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 behavior of \( \hat{\Delta}_{i}^{[K]} \) in the following two cases:

- \( K = K_0 \): By condition (8), \( \hat{G}_k^{[K_0]} \) consistently estimates \( G_k \). Neglecting the estimation error in \( \hat{G}_k^{[K_0]} \), we obtain that

\[
\hat{\Delta}_{i}^{[K_0]} = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \left\{ \frac{\varepsilon_{ij}^2}{\sigma^2} - 1 \right\} / K.
\]

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

\[
\hat{\Delta}_{i}^{[K_0]} \sim \frac{\chi^2_p - p}{\sqrt{2p}}
\]

for each \( i \). Hence, the individual statistics \( \hat{\Delta}_{i}^{[K_0]} \) 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 \( \{ \hat{G}_k^{[K]} : 1 \leq k \leq K \} \) cannot provide an appropriate approximation of the true class structure \( \{ G_k : 1 \leq k \leq K_0 \} \). In particular, there is always a cluster \( S = \hat{G}_k^{[K]} \) 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

\[
\hat{\Delta}_i^{[K]} = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \left\{ \frac{(e_{ij} + d_{ij})^2}{\sigma^2} - 1 \right\} / k
\]

\[
= \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \frac{d_{ij}^2}{\sigma^2k} + O_p(1)
\]

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 \( i \in S \),

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

that is, the statistic \( \hat{\Delta}_i^{[K]} \) has an explosive behavior. According to these heuristic considerations, the statistics \( \hat{\Delta}_i^{[K]} \) exhibit a quite different behavior depending on whether \( K < K_0 \) or \( K = K_0 \). When \( K < K_0 \), the statistic \( \hat{\Delta}_i^{[K]} \) has an explosive behavior 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 behavior of the (rescaled) residual sum of squares \( \hat{\Delta}_i^{[K]} \) for some subjects \( i \). When \( K = K_0 \) in contrast, \( \{ \hat{\Delta}_i^{[K]} : 1 \leq i \leq n \} \) is a collection of (approximately) independent random variables that (approximately) have a rescaled \( \chi^2 \)-distribution. Hence, all statistics \( \hat{\Delta}_i^{[K]} \) have a stable, nonexplosive behavior. This reflects the fact that the partition \( \{ \hat{G}_k^{[K_0]} : 1 \leq k \leq K_0 \} \) is an accurate estimate of the true class structure and thus yields a moderate residual sum of squares \( \hat{\Delta}_i^{[K]} \) for all subjects \( i \).

Since the statistics \( \hat{\Delta}_i^{[K]} \) 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 \( \hat{\Delta}_i^{[K]} \) are i.i.d. variables with a rescaled \( \chi^2 \)-distribution against the alternative that at least one \( \hat{\Delta}_i^{[K]} \) has an explosive behavior. We now construct a statistic \( \hat{\mathcal{H}}^{[K]} \) for this testing problem. A natural approach is to take the maximum of the individual statistics \( \hat{\Delta}_i^{[K]} \): Define

\[
\hat{\mathcal{H}}^{[K]} = \max_{1 \leq i \leq n} \hat{\Delta}_i^{[K]},
\]

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 \((\chi^2_p - p)/\sqrt{2p}\).

Our heuristic discussion from above, in particular formula (17), suggests that for \( K = K_0 \),

\[
\mathbb{P}(\hat{\mathcal{H}}^{[K_0]} \leq q(\alpha)) \approx (1 - \alpha).
\]

Moreover, for \( K < K_0 \), we can show with the help of (18) and some additional considerations that \( \hat{\mathcal{H}}^{[K]} \geq c \sqrt{p} \) 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 formal assumptions on \( n \) and \( p \), \( \hat{\mathcal{H}}^{[K]} \) diverges faster than the quantile \( q(\alpha) \), implying that
\[ \mathbb{P}(\hat{H}^{[K]} \leq q(a)) = o(1) \]

for \( K < K_0 \). This suggests that \( \hat{H}^{[K]} \) has the property (11) and thus is a reasonable statistic to test the hypothesis \( H_0 : K = K_0 \) versus \( H_1 : K < K_0 \).

In this article, we restrict attention to the maximum statistic \( \hat{H}^{[K]} \) defined in (19). In principle though, we may work with any statistic that satisfies the higher order property (11). In Section 6, we discuss some alternative choices of \( \hat{H}^{[K]} \).

### 3.3 A \( k \)-means clustering algorithm

In this section, we construct a \( k \)-means-type clustering algorithm, which has the consistency property (8). 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 differs from the standard one in the choice of the initial values.

To ensure the consistency property (8), we pick initial clusters \( \mathcal{C}_i^{[K]} \) for each given \( K \) as follows: Let \( i_1, \ldots, i_k \) be indices which (with probability tending to 1) belong to \( K \) different classes \( G_{i_1}, \ldots, G_{i_k} \) in the case that \( K \leq K_0 \) and to \( K \) 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 \( \hat{\rho}_k(i) = \hat{\rho}(i_k, i) \) for all \( 1 \leq i \leq n \) and \( 1 \leq k \leq K \), where

\[ \hat{\rho}(i, i') = \frac{1}{p} \sum_{j=1}^{p} (\hat{Y}_{ij} - \hat{Y}_{ij'})^2. \]

The starting values \( \mathcal{C}_i^{[K]} \) are now defined by assigning the index \( i \) to cluster \( \mathcal{C}_{i_k}^{[K]} \) if \( \hat{\rho}_k(i) = \min_{1 \leq k' \leq K} \hat{\rho}_{k'}(i) \). The indices \( i_1, \ldots, i_k \) in this construction are computed as follows: For \( K = 2 \), pick any index \( i_1 \in \{1, \ldots, n\} \) and compute \( i_2 = \arg \max_{1 \leq i \leq n} \hat{\rho}(i_1, i) \). Next suppose we have already constructed the indices \( i_1, \ldots, i_{k-1} \) for the case of \( K - 1 \) clusters and compute the corresponding starting values \( \mathcal{C}_{i_k}^{[K-1]}, \ldots, \mathcal{C}_{i_k}^{[K-1]} \) as described above. Calculate the maximal within-cluster distance \( \hat{\rho}_{\text{max}}(k) = \max_{i \in \mathcal{C}_{i_k}^{[K-1]}} \hat{\rho}_k(i) \) for each \( 1 \leq k \leq K - 1 \) and let \( \mathcal{C}_{i_k}^{[K-1]} \) be a cluster with \( \hat{\rho}_{\text{max}}(k) \geq \hat{\rho}_{\text{max}}(k) \) for all \( k \). Define \( i_K = \arg \max_{i \in \mathcal{C}_{i_k}^{[K-1]}} \hat{\rho}_k(i) \).

To formulate the \( k \)-means algorithm, let the number of clusters \( K \) be given and denote the starting values by \( \mathcal{C}_k^{(0)} = \mathcal{C}_k^{[K]} \) for \( 1 \leq k \leq K \). The \( r \)th iteration of the \( k \)-means algorithm proceeds as follows:

**Step r:** Let \( \mathcal{C}_k^{(r-1)}, \ldots, \mathcal{C}_k^{(r-1)} \) be the clusters from the \( (r-1) \)th iteration step. Compute cluster means \( m_{jk}^{(r)} = (\mathcal{C}_k^{(r-1)})^{-1} \sum_{i \in \mathcal{C}_k^{(r-1)}} \hat{Y}_{ij} \) and calculate the distance measures \( \hat{\rho}_k^{(r)}(i) = p^{-1} \sum_{j=1}^{p} (\hat{Y}_{ij} - m_{jk}^{(r)})^2 \) for all \( 1 \leq i \leq n \) and \( 1 \leq k \leq K \). Define updated groups \( \mathcal{C}_k^{(r)}, \ldots, \mathcal{C}_k^{(r)} \) by assigning the index \( i \) to the cluster \( \mathcal{C}_k^{(r)} \) if \( \hat{\rho}_k^{(r)}(i) = \min_{1 \leq k' \leq K} \hat{\rho}_{k'}^{(r)}(i) \).

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 \( \{ \hat{G}_k^{[K]} : 1 \leq k \leq K \} \). In Section 4, we formally show that these estimators have the consistency property (8) under their 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.

3.4.1 | 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 $\hat{\sigma}^2 = \mathbb{E}[\varepsilon_{ij}^2]$ by

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

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

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

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

$$\hat{\kappa}_{\text{Lip}} = \left(\frac{\hat{\theta}_{\text{Lip}}}{(\hat{\sigma}_{\text{Lip}}^2)^2} - 1\right)^{1/2}.$$

Difference-based estimators of this type have been considered in the context of nonparametric regression by Müller and Stadtmüller (1988), Hall, Kay, and Titterington (1990) and Tecuapetla-Gómez and Munk (2017) among others. Under the technical conditions (C1)–(C3) from Section 4.1, it is straightforward to show that $\hat{\sigma}_{\text{Lip}}^2 = \sigma^2 + O_p((np)^{-1/2} + p^{-2})$ and $\hat{\kappa}_{\text{Lip}} = \kappa + O_p((np)^{-1/2} + p^{-2})$. The estimators $\hat{\sigma}_{\text{Lip}}^2$ and $\hat{\kappa}_{\text{Lip}}$ 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 $\bm{\mu}_i$ have a piecewise constant structure. In particular, $\mu_{ij} = \mu_{ij'}$ for $j_{i-1} \leq j, j' < j_i$, where $j_i$ are known indices with $1 = j_0 < j_1 < \cdots < j_{T-1} < j_T = p + 1$. This implies that $Y_{ij} - Y_{i,j-1} = \varepsilon_{ij} - \varepsilon_{i,j-1}$ for $j_{i-1} < j < j_i$. Similarly as before, we may thus estimate $\sigma^2$, $\theta$, and $\kappa$ by

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

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


3.4.2 Residual-based estimators

Let \( \hat{\sigma}^{[K]}_k \) be the k-means estimators from Section 3.3 for a given \( K \). Moreover, let \( \hat{\varepsilon}^{[K]}_{ij} \) 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 \( \hat{\varepsilon}^{[K]}_i = (\hat{\varepsilon}^{[K]}_{i1}, \ldots, \hat{\varepsilon}^{[K]}_{ip})^T \). With this notation at hand, we define the residual sum of squares for \( K \) clusters by

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

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 \( K \geq K_0 \). The reason is the following: For any \( K \geq K_0 \), the k-means estimators \( \hat{G}_k^{[K]} \) have the property that

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

for \( 1 \leq k \leq K \) under the technical conditions (C1)–(C3). Hence, with probability tending to 1, the estimated clusters \( \hat{G}_k^{[K]} \) contain elements from only one class \( G_k \). The residuals \( \hat{\varepsilon}^{[K]}_{ij} \) should thus give a reasonable approximation to the unknown error terms \( \varepsilon_{ij} \). This in turn suggests that the residual sum of squares 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 \( K_{\text{max}} \). In this situation, we may try to estimate \( \sigma^2 \) by \( \hat{\sigma}^{2}_{\text{RSS}} = \text{RSS}(K_{\text{max}}) \). Even though consistent, this is a very poor estimator of \( \sigma^2 \). The issue is the following: The larger \( K_{\text{max}} \), the smaller the residual sum of squares RSS(\( K_{\text{max}} \)) tends to be. This is a natural consequence of the way in which the k-means algorithm works. Hence, if \( K_{\text{max}} \) is much larger than \( K_0 \), then \( \hat{\sigma}^{2}_{\text{RSS}} = \text{RSS}(K_{\text{max}}) \) tends to strongly underestimate \( \sigma^2 \).

To avoid this issue, we replace the naive estimator \( \hat{\sigma}^{2}_{\text{RSS}} \) by a refined version:

1. For each \( i \), split the data vector \( Y_i = (Y_{i1}, \ldots, Y_{ip})^T \) into the two parts \( Y^{A}_i = (Y_{ij} : j \in J^A) \) and \( Y^{B}_i = (Y_{ij} : j \notin J^B) \), where \( J^A \) and \( J^B \) are disjoint index sets with \( J^A \cup J^B = \{1, \ldots, p\} \). For \( \ell = A, B \), write \( p^\ell := \#J^\ell \) and assume that \( p^\ell / p \to c^\ell \) with some \( c^\ell \in (0, 1) \). Moreover, for \( \ell = A, B \), define \( \bar{Y}_i = (Y_{ij} - \bar{Y}_i^\ell : j \in J^\ell) \), where \( \bar{Y}_i^\ell \) is the empirical mean of the entries in the vector \( Y_i^\ell \), and set \( Y^\ell = \{ \bar{Y}_i^\ell : 1 \leq i \leq n \} \). Under the technical conditions (C1)–(C3), the random vectors in \( Y^A \) are independent from those in \( Y^B \).

2. Apply the k-means algorithm with \( K = K_{\text{max}} \) to the sample \( Y^A \) and denote the resulting estimators by \( \{ \hat{G}_k^{[K]} : 1 \leq k \leq K_{\text{max}} \} \). These estimators can be shown to have the property (21), provided that we impose the following condition: Let \( m_k \) be the class-specific signal vector of the class \( G_k \) and define the vectors \( m^A_k \) and \( m^B_k \) in the same way as above. Assume that

\[
m^A_k \neq m^A_{k'} \quad \text{for } k \neq k'.
\]

According to this assumption, the signal vectors \( m_k \) and \( m_{k'} \) of two different classes can be distinguished from each other only by looking at \( m^A_k \) and \( m^A_{k'} \). It goes without saying that this is not a very severe restriction.

3. Compute cluster-specific residuals from the data sample \( Y^B \),

\[
\hat{\varepsilon}^B_i = \bar{Y}^B_i - \frac{1}{\#\hat{G}^A_k} \sum_{i' \in \hat{G}^A_k} \hat{Y}^B_{i'} \quad \text{for } i \in \hat{G}^A_k.
\]
and define

\[ \hat{\sigma}_{\text{RSS}}^2 = \frac{1}{np^B} \sum_{k=1}^{K_{\text{max}}} \sum_{i \in \bar{G}_k} \| \hat{\varepsilon}_i^B \|^2. \]

In contrast to the naive estimator \( \hat{\sigma}_{\text{RSS}}^2 \), the refined version \( \hat{\sigma}_{\text{RSS}}^2 \) does not tend to strongly underestimate \( \sigma^2 \). The main reason is that the residuals \( \hat{\varepsilon}_i^B \) are computed from the random vectors \( \hat{Y}_i^B \), which are independent of the estimated clusters \( \bar{G}_k \).

Writing \( \hat{\varepsilon}_i^B = (\hat{\varepsilon}_{ij}^B : j \in J^B) \), we can analogously estimate the fourth error moment \( \theta = \mathbb{E}[\varepsilon_{ij}^4] \) by

\[ \hat{\theta}_{\text{RSS}} = \frac{1}{np^B} \sum_{k=1}^{K_{\text{max}}} \sum_{i \in \bar{G}_k} \sum_{j \in J^B} (\hat{\varepsilon}_{ij}^B)^4 \]

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

\[ \hat{\sigma}_{\text{RSS}}^2 = \sigma^2 + O_p(p^{-1}) \quad \text{and} \quad \hat{\kappa}_{\text{RSS}} = \kappa + O_p(p^{-1}). \tag{23} \]

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

**Remark 1.** The difference- and residual-based estimators of \( \sigma^2 \) and \( \kappa \) constructed above are suited to different situations. While the residual-based estimators \( \hat{\sigma}_{\text{RSS}}^2 \) and \( \hat{\kappa}_{\text{RSS}} \) work under general conditions, the difference-based estimators \( \hat{\sigma}_{\text{Lip}}^2 \) and \( \hat{\kappa}_{\text{Lip}} \) as well as \( \hat{\sigma}_{\text{pc}}^2 \) and \( \hat{\kappa}_{\text{pc}} \) are designed for special model settings where the underlying signal vectors \( \mu_i \) have a specific structure: \( \hat{\sigma}_{\text{Lip}}^2 \) and \( \hat{\kappa}_{\text{Lip}} \) are constructed for applications where the underlying signals \( \mu_i \) are smooth trends, whereas \( \hat{\sigma}_{\text{pc}}^2 \) and \( \hat{\kappa}_{\text{pc}} \) are constructed for the case of piecewise constant signals. We suggest to use the difference-based estimators in the special cases for which they are designed rather than the residual-based estimators, since the latter do not exploit the specific structure of the signal vectors and can thus be expected to be less precise. In the general case where the signal vectors are not known to have any particular structure, in contrast, the residual-based estimators should be used.

**Remark 2.** In this article, we analyze a simple model setting that presupposes that the error variance \( \sigma^2 \) (as well as the parameter \( \kappa \)) is the same across all subjects \( i \). This is of course a quite restrictive assumption that might be violated in applications. In a more general model, it could be assumed that the error variances are group-specific. In particular, for each \( 1 \leq k \leq K_0 \), the subjects \( i \in G_k \) could be assumed to have the error variance \( \sigma_k^2 = \text{Var}(\epsilon_{ij}) \), which may be different across \( k \). Notably, it is in principle possible to generalize the CluStErr method to a model with group-specific error variances. To see this, consider the situation where the signal vectors are piecewise constant. In this case, we can estimate the error variance \( \sigma_i^2 = \text{Var}(\epsilon_{ij}) \) of each subject \( i \) by \( \hat{\sigma}_{i,\text{pc}}^2 = (p - T)^{-1} \sum_{j=1}^p (Y_{ij} - Y_{i,j-1})^2/2 \). If the true groups \( G_k \) were known, we could further estimate the group-specific error variances \( \sigma_k^2 \) by \( \hat{\sigma}_{k,\text{pc}}^2 = \#G_k^{-1} \sum_{i \in G_k} \hat{\sigma}_{i,\text{pc}}^2 \) for \( 1 \leq k \leq K_0 \). These estimators are of course not feasible in practice. Assuming that we know an upper bound \( K_{\text{max}} \) on \( K_0 \), we can replace them by the feasible estimators \( \hat{\sigma}_{k,\text{pc}}^{2|K_{\text{max}}|} = \#G_k^{-1} \sum_{i \in \hat{G}_k^{[K_{\text{max}}]}} \hat{\sigma}_{i,\text{pc}}^2 \) for \( 1 \leq k \leq K_{\text{max}} \). Specifically, we can use \( \hat{\sigma}_{k,\text{pc}}^{2|K_{\text{max}}|} \) as an estimator of the unknown error variance of the subjects \( i \in \hat{G}_k^{[K_{\text{max}}]} \). The other variance estimators defined above (as well as the estimators of \( \kappa \)) could be adapted in a similar vein.
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 (3)–(4), we denote the size, that is, the cardinality of the class $G_k$ by $n_k = |G_k|$. Moreover, we use the shorthand $a \ll b$ to express that $a/ b \leq c \nu^{-\delta}$ for sufficiently large $\nu$ with some $c > 0$ and a small $\delta > 0$. Our assumptions read as follows:

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

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

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

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 $\epsilon_{ij}$ to be i.i.d. Yet, the error terms $e_{ij} = \alpha_i + \epsilon_{ij}$ 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 dataset, 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 compared with $p$. Specifically, it should not grow faster than $p^{(\theta/4) - 1}$, where $\theta$ is the number of existing error moments $\mathbb{E}[|\epsilon_{ij}|^\theta] < \infty$. 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 $\epsilon_{ij}$ 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 $\hat{H}^{[K]} = \max_{1 \leq i \leq n} \hat{\Delta}_i^{[K]}$ has the property (11) and thus is a reasonable statistic to test the hypothesis $H_0 : K = K_0$ versus $H_1 : K < K_0$. 
Theorem 1. Assume that the estimated clusters have the consistency property (8). Moreover, let \( \hat{\sigma}^2 \) and \( \hat{\kappa} \) be any estimators with \( \hat{\sigma}^2 = \sigma^2 + O_p(p^{-(1/2+\delta)}) \) and \( \hat{\kappa} = \kappa + O_p(p^{-\delta}) \) for some \( \delta > 0 \). Under (C1)–(C3), the statistic \( \hat{\mathcal{H}}^{[K]} \) has the property (11), that is,

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

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

Theorem 2. Under the conditions of Theorem 1, it holds that

\[
\mathbb{P}(\hat{K}_0 > K_0) = \alpha + o(1) \quad \text{and} \quad \mathbb{P}(\hat{K}_0 < K_0) = o(1),
\]

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

\[
\mathbb{P}(\{\hat{G}_k : 1 \leq k \leq \hat{K}_0\} \neq \{G_k : 1 \leq k \leq K_0\}) = \alpha + o(1).
\]

Theorem 2 holds true for any clustering algorithm with the consistency property (8). The next result shows that this property is fulfilled, for example, by the \( k \)-means algorithm from Section 3.3.

Theorem 3. Under (C1)–(C3), the \( k \)-means estimators \( \{\hat{G}_k^{[K_0]} : 1 \leq k \leq K_0\} \) from Section 3.3 satisfy (8), that is,

\[
\mathbb{P}(\{\hat{G}_k^{[K_0]} : 1 \leq k \leq K_0\} = \{G_k : 1 \leq k \leq K_0\}) \rightarrow 1.
\]

The proofs of Theorems 1–3 can be found in the Supplementary Material.

5 | SIMULATION STUDY AND APPLICATIONS

5.1 | Simulation study

To explore the properties of the CluStErr method in a systematic way, we carry out a simulation study which splits into three main parts. The first part investigates the finite sample behavior of CluStErr, whereas the second part compares CluStErr with several alternative methods. The third and final part reports the results of some robustness checks concerning the estimation of the error variance \( \sigma^2 \) and the parameter \( \kappa \). The simulation design is inspired by the analysis of the gene expression data in Section 5.3. It is based on model (7) from Section 2. The data vectors \( Y_i \) have the form \( Y_i = \mu_i + \epsilon_i \) with piecewise constant signal profiles \( \mu_i \). We set the number of clusters to \( K_0 = 10 \) and define the cluster-specific signal vectors \( m_k \) by

\[
\begin{align*}
    m_1 &= (1, 0, 0, 0, 0)^T, & m_6 &= (-1, 0, 0, 0, 0)^T, \\
    m_2 &= (0, 1, 0, 0, 0)^T, & m_7 &= (0, -1, 0, 0, 0)^T, \\
    m_3 &= (0, 0, 1, 0, 0)^T, & m_8 &= (0, 0, -1, 0, 0)^T, \\
    m_4 &= (0, 0, 0, 1, 0)^T, & m_9 &= (0, 0, 0, -1, 0)^T, \\
    m_5 &= (0, 0, 0, 0, 1)^T, & m_{10} &= (0, 0, 0, 0, -1)^T.
\end{align*}
\]
where \( \mathbf{1} = (1, \ldots, 1) \) and \( \mathbf{0} = (0, \ldots, 0) \) are vectors of length \( p/5 \). A graphical illustration of the signal vectors \( \mathbf{m}_k \) is provided in Figure 1. The error terms \( \epsilon_{ij} \) 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 \( \mathbf{m}_k \). In particular, we define the noise-to-signal ratios \( \text{NSR}_k(\sigma^2) = \sigma^2 / \text{Var}(\mathbf{m}_k) \), where \( \text{Var}(\mathbf{m}_k) \) denotes the empirical variance of the vector \( \mathbf{m}_k \). Since \( \text{Var}(\mathbf{m}_k) \approx 0.16 \) is the same for all \( k \) in our design, we obtain that \( \text{NSR}_k(\sigma^2) = \text{NSR}(\sigma^2) \approx \sigma^2 / 0.16 \) for all \( k \).

### 5.1.1 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 \((n, p) = (1,000, 30)\). Three different noise-to-signal ratios \( \text{NSR} \) are considered, in particular \( \text{NSR} = 1, 1.5, \) 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 \( \text{NSR} = 1 \) mimics the noise level in the application on gene expression data from Section 5.3, where the estimated noise-to-signal ratios all lie between 0.6 and 1. The ratios \( \text{NSR} = 1.5 \) and \( \text{NSR} = 2 \) 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 \( \hat{\sigma}^2_{pc} \) and \( \hat{\sigma}_{pc} \) from Section 3.4. For each of the three noise-to-signal ratios under consideration, we simulate \( B = 1,000 \) samples and compute the estimate \( \hat{K}_0 \) for each sample.

The simulation results are presented in Figure 2. Each panel shows a histogram of the estimates \( \hat{K}_0 \) for a specific noise-to-signal ratio. For the ratio level \( \text{NSR} = 1 \), the CluStErr method produces very accurate results: About 95% of the estimates are equal to the true value \( K_0 = 10 \).
The probability shows that this is indeed the case for the lowest noise-to-signal ratio \( \text{NSR} = 1 \), underestimating \( P \) that follows sense: It has the property that \( P(K_0 > K_0) = \alpha + o(1) \) and \( P(K_0 < K_0) = o(1) \), implying that \( P(K_0 = K_0) = (1 - \alpha) + o(1) \). Setting \( \alpha \) to 0.05, we should thus observe that \( K_0 \) 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 \( \text{NSR} = 1 \). In this situation, the probability \( P(K_0 > K_0) \) of overestimating \( K_0 \) is around 5%, while the probability \( P(K_0 < K_0) \) of underestimating \( K_0 \) is 0%, implying that \( P(K_0 = K_0) \) is about 95%. For the two higher ratio levels \( \text{NSR} = 1.5 \) and 2, in contrast, the estimated values of the probabilities \( P(K_0 < K_0), P(K_0 = K_0), \) and \( P(K_0 > K_0) \) 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 \( P(K_0 < K_0), P(K_0 = K_0), \) and \( P(K_0 > K_0) \) for the noise-to-signal ratio \( \text{NSR} = 1.5 \) and various sample sizes \( (n, p) = (1,000, 30), (1,500, 40), (2,000, 50), (2,500, 60), (3,000, 70) \). As one can clearly see, the estimated probabilities gradually approach the values predicted by the theory as the sample size increases.

To summarize, our simulations on the finite sample behavior 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-
TABLE 2  Estimates of the probabilities $P(\hat{K}_0 < K_0)$, $P(\hat{K}_0 = K_0)$, and $P(\hat{K}_0 > K_0)$ in the simulation scenario with $\text{NSR} = 1.5$ and five different sample sizes $(n, p)$

| $(n, p)$     | (1,000,30) | (1,500,40) | (2,000,50) | (2,500,50) | (3,000,60) |
|-------------|------------|------------|------------|------------|------------|
| $P(\hat{K}_0 < K_0)$ | .274 | .037 | .002 | .000 | .000 |
| $P(\hat{K}_0 = K_0)$ | .549 | .795 | .893 | .918 | .956 |
| $P(\hat{K}_0 > K_0)$ | .177 | .168 | .105 | .082 | .044 |

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 major 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 2 with $\text{NSR} = 1.5$, for example, the estimated probability $P(\hat{K}_0 = K_0)$ is seen to be only around 55% rather than 95%. Nevertheless, 904 out of 1,000 estimates take a value between 9 and 11. Hence, in a large number of 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 $\hat{H}_k^{[K]}$ 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 $\alpha = .05$ 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 $K_0$.

5.1.2 | Comparison of CluStErr with alternative 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), (iii) the Hartigan index (Hartigan, 1975), and (d) a BIC based method for Gaussian mixtures (Fraley & Raftery, 2002). As before, we set the sample size to $(n, p) = (1,000,30)$ and consider the three noise-to-signal ratios $\text{NSR} = 1, 1.5, \text{and } 2$. 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 compare the estimated cluster numbers obtained from the CluStErr method with those produced by the alternative algorithms.

The CluStErr estimates are computed as described in the first part of the simulation study. The gap, silhouette, and Hartigan statistics 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, Rousseeuw, Struyf, & Hubert, 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, Ghazzali, Boiteau, & Niknafs, 2015). The BIC based method for Gaussian mixtures (GM) is computed using the R package mclust (Fraley, Raftery, & Scrucca, 2019).
| NSR | Method      | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | ≥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  | 5  | 19 | 43 | 20 | 11 | 1  | 0  | 1  |    |
|     | Hartigan    | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 4  | 40 | 31 | 10 | 5  | 6  | 3  |    |
|     | BIC for GM  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 100| 0  | 0  | 0  | 0  | 0  |    |
| 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  |    |
|     | Silhouette  | 0  | 0  | 0  | 0  | 0  | 0  | 2  | 12 | 54 | 23 | 8  | 1  | 0  | 0  |    |
|     | Hartigan    | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 3  | 61 | 18 | 9  | 2  | 4  | 2  |    |
|     | BIC for GM  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 100| 0  | 0  | 0  | 0  | 0  |    |
| 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  |    |
|     | Silhouette  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 6  | 66 | 26 | 2  | 0  | 0  | 0  |    |
|     | Hartigan    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 10 | 68 | 18 | 2  | 2  | 0  | 0  |    |
|     | BIC for GM  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 98 | 2  | 0  | 0  | 0  | 0  |    |

(a) Balanced scenario

(Continues)
TABLE 3 (Continued)

|   | Estimated number of clusters | (b) Unbalanced scenario |
|---|-------------------------------|-------------------------|
|   | CluStErr                      | 0 0 0 0 0 0 0 0 0 94 4 2 0 0 0 0 0 |
|   | Gap                           | 0 1 4 9 17 11 21 20 16 1 0 0 0 0 0 0 |
|   | Silhouette                    | 0 0 0 0 0 0 10 26 37 19 6 2 0 0 0 0 |
|   | Hartigan                      | 0 0 1 2 7 8 21 21 12 11 7 2 2 1 5 |
|   | BIC for GM                    | 0 0 0 0 0 0 0 0 0 98 2 0 0 0 0 0 |
| 1  | CluStErr                      | 0 0 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 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 |
|   | BIC for GM                    | 0 0 0 0 0 0 0 1 4 12 65 12 3 1 0 2 |
| 1.5| CluStErr                      | 0 0 0 0 0 0 3 17 25 31 15 3 6 0 0 0 |
|   | Gap                           | 14 4 5 1 6 6 19 26 19 0 0 0 0 0 0 0 |
|   | Silhouette                    | 0 0 0 0 0 0 5 28 48 19 0 0 0 0 0 0 |
|   | Hartigan                      | 0 0 1 4 15 15 15 18 17 7 5 0 0 2 1 1 |
|   | BIC for GM                    | 0 0 0 0 0 0 2 4 17 33 25 9 4 2 1 3 |

Note: 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 $\hat{K}_0$ 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.
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, CluStErr clearly outperforms the gap, silhouette and Hartigan statistics in the setting with NSR = 1. In the scenario with NSR = 1.5, it also performs well in comparison with these three 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. The simulation results further show that the BIC method for Gaussian mixtures outperforms all other methods including CluStErr. This result is most likely due to the fact that the simulation setup under consideration is a Gaussian mixture model, so that methods tailored to this class of models (such as the BIC based method) can be expected to perform better than the more general CluStErr, gap, 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, the gap, silhouette and Hartigan statistics substantially underestimate the number of clusters. CluStErr, 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 the gap, silhouette and Hartigan statistics. As can further be seen, CluStErr performs quite similar to the BIC method for Gaussian mixtures. Hence, in the unbalanced setting, CluStErr is competitive with the BIC procedure, even though it is not specially constructed for Gaussian mixture models.

To summarize, the main findings of our comparison study are as follows: (i) Overall, CluStErr performs well in comparison with the three competing methods that are not specifically designed for Gaussian mixture models (that is, in comparison with the gap, silhouette and Hartigan statistics). On the other hand, as expected, it is outperformed by the BIC method for Gaussian mixtures. (ii) CluStErr outperforms the gap, silhouette and Hartigan statistics as long as the noise-to-signal ratio is not too high. It is, however, outperformed by part of them in the balanced scenario when the noise level is high. In Section 6, we discuss some modifications of the CluStErr method that are designed to improve its behavior in the case of high noise. (iii) CluStErr is able to deal with both balanced and unbalanced cluster sizes, whereas the gap, silhouette and Hartigan statistics perform less adequately in unbalanced settings.

The findings (ii) and (iii) can heuristically be explained as follows: The CluStErr method is based on the test statistic \( \hat{H}^{[K]} = \max_{1 \leq i \leq n} \hat{\Delta}_i^{[K]} \), which is essentially the maximum over the residual sums of squares of the various individuals \( i \). The gap, silhouette and Hartigan approaches, 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 (20) 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.

5.1.3 | Robustness checks

In order to run CluStErr, one needs to estimate the error variance \( \sigma^2 \) (as well as the parameter \( \kappa \)). We now present the results of some additional simulations that were carried out to explore in more
TABLE 4 Comparison of CluStErr implemented with the difference-based estimators \( \hat{\sigma}_2^2 \) and \( \hat{\kappa}_{pc} \) (CluStErr-pc) and CluStErr implemented with the residual-based estimators \( \hat{\sigma}_2^2 \) and \( \hat{\kappa}_{RSS} \) (CluStErr-RSS)

| NSR | Method      | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | ≥15 |
|-----|-------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|-----|
| 1   | CluStErr-pc | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 95| 2  | 1  | 2  | 0  | 0  |     |
|     | CluStErr-RSS| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 93| 1  | 3  | 2  | 0  | 1  |     |
| 1.5 | CluStErr-pc | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22| 28 | 53 | 10 | 3  | 3  | 1  |
|     | CluStErr-RSS| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2  | 31 | 32 | 0  | 1  | 0  |     |
| 2   | CluStErr-pc | 0 | 0 | 0 | 0 | 1 | 2 | 10| 13| 32| 31 | 7  | 2  | 1  | 0  | 1  |
|     | CluStErr-RSS| 0 | 0 | 0 | 1 | 5 | 2 | 17| 19| 34| 18 | 4  | 0  | 0  | 0  |     |

Note: The format of the table is the same as that of Table 3.

detail how CluStErr is affected by the estimation of the error variance \( \sigma^2 \) (and the parameter \( \kappa \)). To do so, we restrict attention to the balanced scenario with clusters of the same size, where we set the sample size to \( (n, p) = (1,000, 30) \) and consider the three noise-to-signal ratios NSR = 1, 1.5, and 2.

(i) So far, we have implemented CluStErr with the difference-based estimators \( \hat{\sigma}_2^2 \) and \( \hat{\kappa}_{pc} \), which are specifically designed for piecewise constant signals. Alternatively, we can use the residual-based estimators \( \hat{\sigma}_2^2 \) and \( \hat{\kappa}_{RSS} \). Since these are not specifically constructed for piecewise constant signals, we expect CluStErr to produce somewhat less precise results when implemented with these. Table 4 shows how much precision is lost when CluStErr is run with the residual-based estimators \( \hat{\sigma}_2^2 \) and \( \hat{\kappa}_{RSS} \), where we set \( K_{max} = 20, J_B = \{5, 10, 15, \ldots, p\} \), and \( J_A = \{1, \ldots, p\} \setminus J_B \). As can be seen, the loss of precision is fairly moderate, at least in the simulation scenario at hand.

(ii) Throughout the article, we have assumed that the error variance \( \sigma^2 \) (as well as the parameter \( \kappa \)) is the same across groups. To investigate the sensitivity of CluStErr to violations of this assumption, we simulate data from the same setup as above, the only difference being that the error variances are group-specific. In particular, for each \( 1 \leq k \leq K_0 \), the error variance is \( \sigma^2_k = c_k \sigma^2 \) for all \( i \in G_k \), where \( \sigma^2 \approx 0.16, 0.25, 0.32 \) as before and \( c_k = 0.5 + 0.05k \) for \( 1 \leq k \leq K_0 \). In this scenario, the error variances \( \sigma^2_k \) range between 0.55\( \sigma^2 \) and \( \sigma^2 \), implying that the largest variance is about double the smallest one. Table 5 compares the performance of CluStErr in this setting and in the homogeneous setting, where the error variances are all the same, that is, where

TABLE 5 Performance of CluStErr in the simulation settings with homogeneous and heterogeneous error variances

| \( \sigma^2 \) | Setting      | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | ≥15 |
|--------------|--------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|-----|
| 0.16         | Homogeneous  | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 93| 4  | 0  | 0  | 0  | 3  |     |
|              | Heterogeneous| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1  | 60| 11 | 7  | 4  | 1  | 16 |     |
| 0.25         | Homogeneous  | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1  | 25| 50 | 13 | 4  | 1  | 0  | 4   |
|              | Heterogeneous| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5  | 15| 35 | 16 | 6  | 4  | 8  | 11  |
| 0.32         | Homogeneous  | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1  | 33| 31 | 3  | 2  | 2  | 0  |     |
|              | Heterogeneous| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17| 19| 17 | 19 | 12 | 2  | 2  | 11  |

Note: The format of the table is the same as that of Table 3.
\[ \sigma_k^2 = \sigma^2 \text{ for all } k. \]

As can be seen, CluStErr tends to overestimate the true number of groups \( K_0 \) much more strongly in the heterogeneous than in the homogeneous setting. This can be explained as follows: When the error variances are heterogeneous, the statistic \( \hat{\sigma}^2 \) estimates a weighted average of them. Hence, it tends to lie between the smallest and the largest error variance, that is, between \( 0.55\sigma^2 \) and \( \sigma^2 \). This implies that the statistics \( \hat{\Delta}^{[K_0]} \) are not normalized appropriately. In particular, since \( \hat{\sigma}^2 \) tends to be smaller than \( \sigma^2 \), the statistics \( \hat{\Delta}^{[K_0]}_i \) tend to be too large for subjects \( i \) in the group with the error variance \( \sigma^2 \). As a consequence, CluStErr tends to overestimate \( K_0 \). This tendency will be the stronger, the more heterogeneous the error variances are. Hence, in applications where the noise level in the data is expected to be strongly heterogeneous across groups, one should work with an extended version of CluStErr which can handle group-specific error variances (see Remark 1).

### 5.2 Clustering of temperature curves

Our first application is concerned with the analysis of a dataset 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](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 dataset contains a monthly land surface temperature profile. This profile is a vector with 12 entries, the first entry specifying the average temperature of all Januarys from 1951 to 1980, the second entry specifying the average temperature of all Februarys from 1951 to 1980, and so on. The temperature profiles at various example locations on earth are shown in Figure 3. 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).

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 (6) on the data and apply the CluStErr method to them, setting \( n = 24,311, \ p = 12, \) and \( \alpha = .05 \). To estimate the error variance \( \sigma^2 \) and the normalization constant \( \kappa \), we apply the difference-based estimators \( \hat{\sigma}^2_{\text{Lip}} \) and \( \hat{\kappa}_{\text{Lip}} \) from Section 3.4, thus making use of the smoothness of the temperature curves illustrated in Figure 3.

The estimation results are presented in Figures 4 and 5. Figure 4 depicts the \( p \)-values \( \hat{p}^{[K]} \) corresponding to the test statistics \( \hat{\alpha}^{[K]} \) for different numbers of clusters \( K \). It shows that the \( p \)-value \( \hat{p}^{[K]} \)
FIGURE 4  Analysis of the Berkeley Earth temperature data. The plot depicts the $p$-values $\hat{\beta}^{[K]}$ corresponding to the test statistics $\hat{Y}^{[K]}$ as a function of $K$. The horizontal dashed line specifies the significance level $\alpha = .05$, and the vertical dashed line indicates that the estimated number of clusters is $\hat{K}_0 = 26$.

FIGURE 5  Visualization of the $\hat{K}_0 = 26$ clusters obtained from the analysis of the Berkeley Earth temperature data. Each shade of color refers to one cluster.

remains below the $\alpha = .05$ 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 $\hat{K}_0 = 26$, 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 $\hat{\sigma}^2 = 16.25$. Figure 5 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 5 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, Finlayson, & McMahon, 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 $\hat{K}_0 = 26$ 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 5 confirms the usefulness of the CluStErr method in this setting.
5.3 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 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 “coexpressed” genes showing similar expression patterns across clinical conditions (Chipman, Hastie, & Tibshirani, 2003; D’haeseleer, 2005; Jiang et al., 2004).

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 dataset 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 genewise standardization of the expression levels. For details on data collection and preprocessing, 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 who 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 chronic myeloid leukemia (CML, 12 patients). The other 12 samples were obtained from nonleukemia (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 \( \geq .01 \) 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 (7) from Section 2 on the data. The measured expression profiles \( Y_i = (Y_{i1}, \ldots, Y_{ip})^T \) of the various genes \( i = 1, \ldots, n \) are thus assumed to follow the model equation \( Y_i = \mu_i + \alpha_i + \epsilon_i \). The signal vectors \( \mu_i \) are supposed to have a piecewise constant structure after the patients have been ordered according to their tissue type (ALL, AML, CLL, CML, NoL). To illustrate, the expression profile \( Y_i \) of a randomly selected gene is plotted in Figure 6.

To cluster the genes, we apply the CluStErr algorithm with the significance level \( \alpha = .05 \) and the difference-based estimators \( \hat{\sigma}^2_{pc} \) and \( \hat{\kappa}_{pc} \) from Section 3.4, thus exploiting the piecewise constant structure of the signal vectors. The estimation results are presented in Figures 7 and 8. The plot in Figure 7 depicts the \( p \)-values \( \hat{p}^{[K]} \) corresponding to the test statistics \( \hat{H}^{[K]} \) as a function of the cluster number \( K \). It shows that the estimated number of clusters is \( \hat{K}_0 = 14 \). The estimated sizes of the 14 clusters range between 58 and 469. Moreover, the estimated error variance is \( \hat{\sigma}^2 = 0.442 \).

Inspecting Figure 7, the \( p \)-values \( \hat{p}^{[K]} \) can be seen to be fairly small for any \( K < 14 \), which suggests that the structure in the data is not well described by a model with \( K < 14 \) clusters. A first larger upward jump occurs at \( K = 14 \), indicating that the data are much better described
**FIGURE 6** 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, Fontanillo, & De Las Rivas, 2013). Horizontal lines represent the average gene expression levels across the five tissue types.

**FIGURE 7** Analysis of the MILE study gene expression data. The plot depicts the $p$-values $\hat{p}^{[K]}$ corresponding to the test statistics $\hat{H}^{[K]}$ as a function of $K$. The dashed vertical line indicates that the number of clusters is estimated to be $\hat{K}_0 = 14$. 

by a model with $K = 14$ clusters. This large jump or “elbow” in the $p$-value plot indicates that $\hat{K}_0 = 14$ is a reasonable estimate of the number of clusters. The $p$-value plot thus lends credibility to the estimate $\hat{K}_0 = 14$. Generally speaking, the $p$-value plot is a useful diagnostic tool in practice.

In Figure 8, the cluster centers $\hat{m}_k = (\#\hat{G}_k)^{-1} \sum_{i \in \hat{G}_k} \hat{Y}_i$ are presented, which estimate the cluster-specific signal vectors $m_k = (\#G_k)^{-1} \sum_{i \in G_k} \mu_i$. All clusters show a distinct separation of at least one tissue type, supporting the assumption of piecewise constant signals $m_k$ and indicating that the genes contained in the clusters are coexpressed differently across the five tissue types. 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 coexpressed gene profiles.
FIGURE 8  Visualization of the cluster centers obtained from the analysis of the MILE study gene expression data. The dots represent the cluster centers $\hat{m}_k = (\#G_k)^{-1} \sum_{i \in G_k} \bar{Y}_i$, which estimate the cluster-specific signal vectors $m_k = (\#G_k)^{-1} \sum_{i \in G_k} \mu_i$. Gene ENSG00000002834, whose expression profile is visualized in Figure 6, is an element of cluster #13. Note the similarity of the patterns in cluster #13 and Figure 6.
6 | EXTENSIONS

In this article, we have developed an approach for estimating the number of clusters with statistical error control. We have derived rigorous statistical theory for a simple model with convex spherical clusters. An interesting question is how to extend our ideas to the case of general, potentially nonconvex clusters. Developing theory for this general case, however, is a very challenging problem and goes beyond the scope of this article. In what follows, we discuss some modifications and extensions of our methods within the limits of the model setting (3)–(4).

So far, we have based our estimation methods on the maximum statistic \( \hat{H}^{[K]} = \max_{1 \leq i \leq n} \hat{\Delta}_i^{[K]} \). However, we are not bound to this choice. Our approach can be based on any test statistic \( \hat{H}^{[K]} \) that fulfills the higher order property (11). 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 \( \hat{\Delta}_i^{[K]} \). Their stochastic behavior has been analyzed in detail in Section 3.2. Under the null hypothesis \( H_0 : K = K_0 \), the statistics \( \hat{\Delta}_i^{[K]} \) are approximately independent and distributed as \( (\chi^2_p - p)/\sqrt{2p} \) variables. Under the alternative \( H_1 : K < K_0 \) in contrast, they have an explosive behavior at least for some \( i \). This difference in behavior suggests to test \( H_0 \) by checking whether the hypothesis

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

holds for all subjects \( i = 1, \ldots, n \). 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 \( \hat{H}^{[K]} = \max_{1 \leq i \leq n} \hat{\Delta}_i^{[K]} \) 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 \( \hat{K}_0 \) 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 5.1. We now present two ways how to construct a statistic \( \hat{H}^{[K]} \) with better power properties.

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

\[
\hat{G}_1^{[K]} = \{ i_1, \ldots, i_{w_1} \}, \quad \hat{G}_2^{[K]} = \{ i_{w_1+1}, \ldots, i_{w_2} \}, \quad \ldots \quad \hat{G}_K^{[K]} = \{ i_{w_{K-1}+1}, \ldots, i_{w_K} \}.
\]

We next partition the ordered indices into blocks

\[
B_{\ell}^{[K]} = \{ i_1, \ldots, i_{w_{\ell N}} \} \quad \text{for } 1 \leq \ell \leq L,
\]

where \( N \) is the block length and \( L = \lfloor n/N \rfloor \) is the number of blocks. With this notation at hand, we construct blockwise averages

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

of the individual statistics \( \hat{\Delta}_i^{[K]} \) and aggregate them by taking their maximum, thus defining
\[ \hat{H}_B^{[K]} = \max_{1 \leq \ell \leq L} \hat{\Lambda}_\ell^{[K]} . \]

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

\[ H_B = \max_{1 \leq \ell \leq L} \Lambda_\ell \quad \text{with} \quad \Lambda_\ell = \frac{1}{\sqrt{N}} \sum_{i=1}^{\ell} Z_i, \]

where \( Z_i \) are i.i.d. variables with the distribution \((\chi^2_p - p)/\sqrt{2p}\). Note that this definition of \( \hat{H}_B^{[K]} \) nests the maximum statistic \( \hat{H}_B^{[K]} = \max_{1 \leq \ell \leq L} \hat{\Delta}_\ell^{[K]} \) 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 \( \hat{H}_B^{[K]} \) can be shown to have the theoretical properties stated in Theorems 1–3. More specifically, Theorems 1–3 can be shown to hold true for the blocked statistic \( \hat{H}_B^{[K]} \) 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 \( \hat{H}_B^{[K]} \), we average over the individual statistics \( \hat{\Delta}_i^{[K]} \). As already discussed in the context of the simulation study of Section 5.1, averaging has the effect that small clusters are effectively ignored. Hence, in contrast to the maximum statistic \( \hat{H}_B^{[K]} = \max_{1 \leq \ell \leq L} \hat{\Delta}_\ell^{[K]} \), the blocked statistic \( \hat{H}_B^{[K]} \) with a large block size \( N \) can be expected not to perform adequately when the cluster sizes are highly unbalanced.

In balanced settings, however, the blocked statistic \( \hat{H}_B^{[K]} \) 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 behavior of \( \hat{H}_B^{[K]} \) for different block lengths \( N \). Our heuristic discussion of the individual statistics \( \hat{\Delta}_i^{[K]} \) from Section 3.2 directly carries over to the blocked versions \( \hat{\Lambda}_\ell^{[K]} \). With the help of (17), it is easy to see that

\[ \mathbb{P}(\hat{H}_B^{[K]} \leq q_B(\alpha)) \approx (1 - \alpha) . \]

Moreover, (18) together with some additional arguments suggests that \( \hat{H}_B^{[K]} \) has an explosive behavior 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

\[ \hat{H}_B^{[K]} \geq c\sqrt{Np} \quad \text{for some} \quad c > 0 \quad \text{with prob. tending to} \quad 1 . \quad (24) \]

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

\[ \mathbb{P}(\hat{H}_B^{[K]} \leq q_B(\alpha)) = o(1) \]

for \( K < K_0 \). As a result, \( \hat{H}_B^{[K]} \) should satisfy the higher order property (11). Moreover, according to (24), the statistic \( \hat{H}_B^{[K]} \) explodes at the rate \( \sqrt{Np} \) for \( K < K_0 \). Hence, the faster the block size \( N \) grows, the faster \( \hat{H}_B^{[K]} \) diverges to infinity. Put differently, the larger \( N \), the more power we have.
Table 6 Simulation results for the blocked maximum statistic

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

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 $\hat{H}^{[K]}_B$ with block length $N = p$ should be more powerful than the maximum statistic $\hat{H}^{[K]} = \max_{1 \leq i \leq n} \hat{\Delta}^{[K]}_i$ 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 5.1. 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 $\hat{H}^{[K]}_B$ with $N = p$. Table 6 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 with those in Table 3a, the blocked method can be seen to perform very well in comparison with the gap, silhouette and Hartigan statistics 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 $\hat{H}^{[K]}$. As an example, we set up a test statistic that uses ideas from the FDR literature: Rather than only taking into account the maximum of the statistics $\hat{\Delta}^{[K]}_i$, we may try to exploit the information in all of the ordered statistics $\hat{\Delta}^{[K]}_i \geq \ldots \geq \hat{\Delta}^{[K]}_n$. 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

$$\hat{\Delta}^{[K]}_{(i)} > q_{\chi^2} \left( \frac{i}{n} \alpha \right) \text{ for some } i \in \{1, \ldots, n\},$$

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

$$\hat{H}^{[K]}_{\text{FDR}} = \max_{1 \leq i \leq n} \frac{\hat{\Delta}^{[K]}_{(i)}}{q_{\chi^2}(i\alpha/n)}$$

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

The authors thank the associate editor and two referees for their constructive and helpful comments on an earlier version of the article. M.V. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy, GZ 2047/1, Project-ID 390685813. M.S. was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Grant SCHM 2966/2-1.

ORCID

Michael Vogt https://orcid.org/0000-0002-5885-4303

REFERENCES

Aibar, S., Fontanillo, C. & De Las Rivas, J. (2013). LeukemiasEset: Leukemia’s microarray gene expression data (expressionSet) (R Package Version 1.8). Retrieved from https://bioconductor.org/packages/release/data/experiment/html/leukemiasEset.html.

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

Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. (2015). NbClust: Determining the best number of clusters in a data set (R Package Version 3). Retrieved from https://cran.r-project.org/web/packages/NbClust/.

Chen, J., Li, P., & 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., & Tibshirani, R. (2003). Clustering microarray data. In T. Speed (Ed.), Statistical analysis of gene expression microarray data (pp. 161–203). Boca Raton, FL: Chapman & Hall / CRC.

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

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

Degaetano, A. T. (2001). Spatial grouping of United States climate stations using a hybrid clustering approach. International Journal of Climatology: A Journal of the Royal Meteorological Society, 21, 791–807.

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

Finazzi, F., Haggarty, R., Miller, C., Scott, M., & Fasso, A. (2015). A comparison of clustering approaches for the study of the temporal coherence of multiple time series. Stochastic Environmental Research and Risk Assessment, 29, 463–475.

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

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

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97, 611–631.

Fraley, C., Raftery, A. E. & Scrucca, L. (2019). mclust: Gaussian mixture modelling for model-based clustering, classification, and density estimation. (R package version 5.4.5). Retrieved from https://cran.r-project.org/web/packages/mclust/.

Frühwirth-Schnatter, S., Celeux, G., & Robert, C. P. (2018). Handbook of mixture analysis. New York: Chapman and Hall/CRC.

Ghosh, J. K., & Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In L. M. Le Cam & R. A. Olshen (Eds.), Proceedings of Berkeley conference in Honor of Jerzy Neyman and Jack Kiefer (Vol. 2, pp. 789–806). Belmont, CA: Wadsworth.

Gordon, A. (1999). Classification. New York, NY: Chapman & Hall/CRC.

Haferlach, T., Kohlmann, A., Wieczorek, L., Basso, G., Te Kronnie, G., Bene, M.-C., … 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., & Titterington, D. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika, 77*, 521–528.

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

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In L. M. Le Cam & R. A. Olshen (Eds.), *Proceedings of Berkeley conference in Honor of Jerzy Neyman and Jack Kiefer* (Vol. 2, pp. 807–810). Belmont, CA: Wadsworth.

Jiang, D., Tang, C., & 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., & Schmid, M. (2019). CluStErr: Clustering with statistical error control. *R Package Version 1.*

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

Liu, X., & 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. & Hubert, M. (2016). Finding groups in data: Cluster analysis extended *R* Package Version 2.4. Retrieved from https://cran.r-project.org/web/packages/cluster/.

Maitra, R., Melnykov, V., & Lahiri, S. N. (2012). Bootstrapping for significance of compact clusters in multidimensional datasets. *J Am Stat Assoc, 107*, 378–392.

McLachlan, G. J., & 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., & Stadtmüller, U. (1988). Detecting dependencies in smooth regression models. *Biometrika, 75*, 639–650.

Peel, M. C., Finlayson, B. L., & 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., … 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.

Tecuapetla-Gómez, I., & Munk, A. (2017). Autocovariance estimation in regression with a discontinuous signal and m-dependent errors: A difference-based approach. *Scandinavian Journal of Statistics, 44*, 346–368.

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

Tripathi, S., Bhardwaj, A., & Poovammal, E. (2018). Approaches to clustering in customer segmentation. *International Journal of Engineering & Technology, 7*, 802–807.

Wedel, M., & Kamakura, W. (2000). *Market segmentation: Conceptual and methodological foundations*. New York, NY: Springer.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

---

**How to cite this article:** Vogt M, Schmid M. Clustering with statistical error control. *Scand J Statist.* 2021;48:729–760. https://doi.org/10.1111/sjos.12450