TU Clausthal

Master Thesis

by

Bojan Lukić

Benchmarking of Models for Unsupervised Segmentation of Multivariate Time Series With a Novel Homogeneity Metric

1st Supervisor: **Prof. Dr. Andreas Rausch**

2nd Supervisor: **Prof. Dr. Steffen Herbold**

March 24, 2022

Institute for Software and Systems Engineering
Clausthal University of Technology

Declaration of Authorship

I have read and understood the guidelines of the Technical University of Clausthal and those published in the ISSE bulletin, including those regarding the use of literature and other sources. I confirm that I have prepared this thesis independently by myself. Any information taken from other sources and being reproduced in this thesis is clearly referenced.

In terms of the general examination regulations, this work has not yet been submitted to any other examination division.

I hereby agree that my master thesis may be exhibited in the institute's and/or university's library and kept for inspection.

Clausthal-Zellerfeld, March 24, 2022	
Location, Date	Bojan Lukic

Abstract

Multivariate time series segmentation receives an increasing amount of attention from different domains in recent times. An examplary application area for time series segmentation is the analysis of physiological (e.g. electroencephalography) data for classifying sleep stages in patients. Another area of application, which is closely related to this work, is the analysis of process runtimes from software systems for correlations analysis. The clustering of unordered data can be seen as an analogy to time series segmentation, with the main difference being that time series contain the additional dimension of time and are therefore ordered. This makes time series segmentation a more complex discipline, which requires the application of tools other than those usually applied for the segmentation (i.e. clustering) of unordered data. Multivariate time series are a subtype of time series, which consist of multiple single time series, thus adding to the complexity. The main goal of time series segmentation is to divide an ordered dataset into segments with consistent internal behavior. Another term for consistent internal behavior established and used in this work is homogeneity, which quantifies the degree of similarity between the data points inside a segment. On a more abstract level, the consistency of a segment can be defined as the similarity of information that each data point inside of a segment carries.

The problem of segmenting a multivariate time series can be stated as follows: Divide a multivariate data set into an appropriate amount of segments with each segment showing maximum homogeneity. Following subproblems and research questions arise from the problem statement: 1. How can homogeneity in segments of a multivariate time series be defined and measured?, 2. How can the ideal number of segments in a multivariate time series be determined without supervision?, and 3. How can precise segmentation indices for maximizing homogeneity in segments be determined? This work delivers answers to questions one and three with the application of existing models and a newly developed metric. However, question two still remains an open research problem and is discussed towards the end of this work. The complexity of the main problem has yet to be determined, however, it can be assumed that the problem is NP-hard, as there is no a priori knowledge about either the ideal number of segments for a multivariate time series segmentation nor the exact segmentation indices for maximizing the homogeneity of a segmentation.

This work quantifies high homogeneity for a time series segmentation, discusses appropriate methods for dividing a multivariate time series into segments with relatively high homogeneity, and presents procedures for evaluating the segmentation of a multivariate time series. For this purpose, a metric which measures the homogeneity (i.e. the internal consistency) of segmented time series is defined and used for assessment. The time series segmentation is performed with two different time series segmentation models. The tools and models used in this work are examined for effectivity and efficiency by applying them to a synthetic multivariate time series as well as a data set taken from an automotive software system provided by an automotive partner. The findings are positive for applying appropriate models for multivariate time series segmentation and measuring the homogeneity of resulting segmentations. The results also show a high potential for further use of proposed tools and models for subsequent studies in time series analysis, such as classification of segments from time series. The supplementary resources to this work are provided in a GitHub repository[1].

Keywords: multivariate time series, time series segmentation, signal analysis, runtime analysis of automotive system processes, time series analysis.

[1]See `https://github.com/Bojan-Lukic/master-thesis-signal-segmentation` for supplementary material and custom algorithms used in this paper.

Zusammenfassung

Multivariate Zeitreihensegmentierung erhält in letzter Zeit zunehmende Aufmerksamkeit aus verschiedenen Forschungsgebieten. Ein beispielhaftes Anwendungsgebiet für multivariate Zeitreihensegmentierung ist die Analyse von physiologischen (z.B. Elektroenzephalographie) Daten zur Klassifizierung von Schlafstadien bei Patienten. Ein weiterer Bereich, der eng mit dieser Arbeit verwandt ist, ist die Analyse von Prozesslaufzeiten von Softwaresystemen zur Korrelationsanalyse. Das Clustern von ungeordneten Daten kann als Analogie zur Zeitreihensegmentierung angesehen werden, wobei der Hauptunterschied darin besteht, dass Zeitreihen die zusätzliche Dimension der Zeit enthalten und daher geordnet sind. Dies macht die Segmentierung von Zeitreihen zu einer komplexeren Disziplin, die die Anwendung von fortgeschritteneren Werkzeugen als denen erfordert, die üblicherweise für die Segmentierung (d.h. Clustering) von ungeordneten Daten verwendet werden. In allen Bereichen, in denen diese Art der Zeitreihenanalyse angewendet wird, hat die Zeitreihensegmentierung das Hauptziel, einen geordneten Datensatz in Segmente mit konsistentem internen Verhalten zu unterteilen. Ein anderer Begriff für konsistentes internes Verhalten, der in dieser Arbeit verwendet wird, ist Homogenität. Die Homogenität quantifiziert den Grad der Konsistenz zwischen den Datenpunkten innerhalb eines Segments. Auf einer abstrakteren Ebene kann die Konsistenz eines Segments als die Ähnlichkeit von Informationen definiert werden, die jeder Datenpunkt innerhalb eines Segments enthält.

Das Problem der Segmentierung einer multivariaten Zeitreihe kann wie folgt formuliert werden: Es soll ein multivariater Datensatz in eine angemessene Anzahl von Segmenten geteilt werden, sodass jedes Segment eine maximale Homogenität aufweist. Aus der Problemstellung ergeben sich folgende Teilprobleme und Forschungsfragen: 1. Wie kann die Homogenität in Segmenten einer multivariaten Zeitreihe definiert und gemessen werden?, 2. Wie kann die ideale Anzahl von Segmenten in einer multivariaten Zeitreihe ohne Supervision bestimmt werden?, und 3. Wie können genaue Segmentierungsindizes zur Maximierung der Homogenität in Segmenten ermittelt werden? Die vorliegende Arbeit wird Antworten auf die Fragen eins und drei mit der Anwendung bestehender Modelle und einer neu entwickelten Metrik liefern. Jedoch bleibt die Frage zwei ein offenes Forschungsproblem, welches gegen Ende dieser Arbeit diskutiert wird. Die Komplexität des Hauptproblems muss noch ermittelt werden, es kann jedoch davon ausgegangen werden, dass es sich um ein NP-schweres Problem handelt, da a priori weder die ideale Anzahl an Segmenten für eine multivariate Zeitreihensegmentierung noch die exakten Segmentierungsindizes zur Maximierung der Homogenität einer Segmentierung bekannt sind.

In dieser Arbeit wird quantifiziert, was hohe Homogenität für eine Zeitreihensegmentierung bedeutet, es werden geeignete Methoden zur Unterteilung einer multivariaten Zeitreihe in Segmente mit relativ hoher Homogenität diskutiert und Verfahren zur Bewertung der Segmentierung einer multivariaten Zeitreihe vorgestellt. Zu diesem Zweck wird eine Metrik definiert, die die Homogenität segmentierter Zeitreihen (d.h. die interne Konsistenz jedes Segments) misst. Diese Metrik wird verwendet, um die Segmentierungsqualität einer multivariaten Zeitreihe, die mit zwei verschiedenen Modellen der Zeitreihensegmentierung segmentiert wird, zu quantifizieren. Die verwendeten Werkzeuge und Modelle werden auf Effektivität und Effizienz untersucht, indem sie auf eine synthetische multivariate Zeitreihe, sowie einen von einem Automobilpartner bereitgestellten Datensatz angewendet werden. Die Ergebnisse zeigen ein Potential für die Anwendung relevanter Modelle für die multivariate Zeitreihensegmentierung und die Messung der Homogenität der resultierenden Segmentierungen. Die Ergebnisse zeigen auch ein hohes Potenzial für die weitere Verwendung vorgeschlagener Werkzeuge und Modelle für nachfolgende Studien in der Zeitreihenanalyse, wie z.B. die Klassifizierung von Segmenten aus Zeitreihen. Das ergänzende Material zu dieser Arbeit wird auf GitHub zur Verfügung gestellt[2].

Schlüsselwörter: multivariate time series, time series segmentation, signal analysis, runtime analysis of automotive system processes, homogeneity.

[2]Siehe `https://github.com/Bojan-Lukic/master-thesis-signal-segmentation` für ergänzendes Material mit individuellen Algorithmen, die in dieser Arbeit vorgestellt werden.

Acknowledgements

First and foremost I would like to thank Thorben Knust for his unconditional help not only for this master's thesis but also throughout other projects during my master's program. Thorben, thanks to your expertise, creativity, and continuous support I was able to excel in the area of signal processing and, more generally, data science.

Special thanks go to Prof. Dr. Andreas Rausch for always being at my disposal when it came to interim presentations and expert input on the topic of time series analysis. The detailed discussions and challenges have helped me understand the contexts of the topic in all its details and have made this work possible in the first place.

Lastly, I would like to thank my friends and family for always supporting me mentally and providing me the creative space needed to efficiently finish my study with this work. Without you, writing this thesis would have been considerably more challenging and taxing.

Contents

List of Figures

List of Tables

Abbreviations

EM	expectation-maximization.
LLE	locally-linear embedding.
MHMMR	multiple hidden markov model regression.
PC	principal component.
PCA	principal component analysis.
PCAFC	principal component analysis based fuzzy clustering.

1 Introduction

Nowadays, multivariate time series analysis is an inherent part of many industries. Operations with time ordered data become more prominent and the proper use of it is crucial for research and development in institutions and companies [1, p. 1], [2, p. xxiii]. Effectively dividing a multivariate time series allows the user to observe a smaller subspace of the original data set and therefore presents the opportunity of analyzing information from the data set in more detail. At the same time, multivariate time series are becoming more complex and require increasingly sophisticated tools for analysis. While the research in univariate time series analysis is quite mature and advanced, the research in multivariate time series analysis is, due to the nature of multivariate data, not quite as developed with many research questions still remaining open. Most of the tools used for data analysis (i.e. time series segmentation), are also quite complex and use different techniques from the discipline of statistics, such as dimensionality reduction or, more generally, data transformation [2, p. 25]. Another aspect of multivariate data, and in general the increase of dimensions in data sets, is the difficulty of visualization and feature conception. While univariate time series can easily be plotted in a two dimensional space, multivariate time series with dimensions higher than three can be difficult to visualize and comprehend in their original form.

With the increasing importance in multivariate time series analysis, it can be argued that there is a need for effective and efficient models for multivariate time series segmentation as well as tools for quantifying the quality of multivariate time series segmentations. This work focuses on the application of state-of-the-art statistical models for multivariate time series segmentation and the subsequent quantification of the segmentation quality. For this purpose, two segmentation models are applied to a multivariate time series taken from a real automotive software system. The resulting segmentations are evaluated for their quality using a metric developed for the purpose of this study. The metric measuring the homogeneity of segmented time series is first validated and verified with synthetic multivariate time series before applying it to the multivariate time series taken from the software system.

The remaining work is divided as follows: In chapter 2, the current state of research in the field of multivariate time series segmentation and, more broadly, analysis is discussed. Chapter 3 presents the motivation of this work and the definition of an optimal segmentation using a synthetic multivariate times series. For this purpose, a homogeneity metric is defined, validated and verified. The focus of chapter 4 lies on the tools used for state-of-the-art segmentation models, which are applied to a multivariate time series in chapter 5. Chapter 5 gives an overview on the main data set used with the time series segmentation models and contains

some of the results obtained by the segmentation models. The detailed results from chapter 6 are discussed in chapter 7 before giving a conclusion and future outlook on the work of multivariate time series segmentation in chapter 8.

2 Status Quo and Related Work

In this chapter the current state of research in the field of multivariate time series segmentation along with homogeneity metrics is presented. For this purpose, the chapter is divided into related works from the three subdomains, namely time series segmentation, homogeneity metrics for time series, and evaluation of methods for time series segmentation.

2.1 Time Series Segmentation

In their book [3] *Cluster Analysis for Data Mining and System Identification* as well as in [4] and [5] János Abonyi and Balázs Feil present a novel principal component analysis based fuzzy clustering (PCAFC) algorithm for unsupervised segmentation of multivariate time series into homogeneous segments. Their approach of bottom-up clustering incorporates the comparison and merging of adjacent segments with principal component analysis (PCA) based similarity measures and weight matrices for a fuzzy segmentation. Besides a cost function as a similarity metric and the definition of homogeneity in segments, the authors propose a new technique for estimating an appropriate amount of segments in a data set without a priori knowledge about the features of the data set. The authors conclude that their novel method of PCAFC segmentation is an efficient and effective way of dividing a multivariate time series into an appropriate amount of segments with high homogeneity. In works such as [6], [7], [8], and [9], different authors apply multivariate time series segmentation with approaches closely related to the one of Abonyi and Feil. Across these works, tools for dimensionality reduction in combination with similarity measures for merging adjacent segments of time series in a bottom-up approach are applied.

Another method, the segmentation of multivariate time series with convolutional neural networks, is described in [10]. The authors Yu et al. choose local boundaries for segmentation of multi-channel signals with a convolutional neural network. In total, three convolutional layers and a fully connected layer are used to create a precise segmentation of single segments. The authors state a precision of 97 % for their approach of finding boundaries in a time series with a neural network. In their paper [11], Murphey et al. present a supervised learning algorithm based on fuzzy segmentation for automotive fault diagnosis. Their distributed intelligent-agent system provides information for diagnostics decision in form of signal-segment fault, signal fault, and vehicle fault. This work takes the direction of classifying the segments found in a multivariate time series, which is also of significance in the underlying work. David Guijo-Rubio shows [12] an approach of segmenting multivariate time series with varying lengths per signal by

using a combination of least squares polynomial segmentation, feature extraction for segments, and clustering of features. The method outperforms two state-of-the-art methods (namely hierarchical clustering using the DDDTW distance measure as well as the K-spectral centroid clustering algorithm) and provides a way of segmenting multivariate time series independent of the lengths of the signals contained in the time series.

2.2 Internal Homogeneity Measures for Time Series

As much as methods for multivariate time series segmentation are important for this work, an integral part of the topic is the definition of precise segmentation, or rather the homogeneity of segments. It is important to note that the measure of homogeneity can be used as an internal measure for the procedure of dividing a data set into homogeneous segments (e.g. with a cost function), as well as an external measure for quantifying (e.g. as a benchmark metric) the segmentation quality of an already performed segmentation. This subsection discusses the progress in the field of homogeneity criteria used as an internal measure during the procedure of dividing a data set precisely into homogeneous segments. The metrics used for quantifying segmentation quality are discussed in chapter 3.

Several papers discuss the subtopic of homogeneity for segmentation. In their publication [13], Giannakopoulos and Pikrakis define various abstraction layers for homogeneity in the context of audio segmentation. Depending on the application and domain, a homogeneous segment, in the context of dividing an audio signal, could be containing either speech, classical music, or silence. Therefore, homogeneity can be based on the type of information a segment carries. In [10] it is suggested that an advanced way of evaluating homogeneity in segments of times series is PCA, which is backed by [3]. The latter source states that similarity measures can be used to compare the homogeneity between adjacent segments before applying bottom-up clustering (i.e. segmentation). In additional papers [4], [5], [8], cost functions are proposed to measure the internal homogeneity of individual segments. These cost functions can incorporate anything from PCA based similarity measures (i.e. parallelism of segments) to the closeness of the centroids of segments for deciding on precise segmentation indices for all segments.

The cost functions for homogeneity along with cluster purity measures are analyzed and presented in chapter 3.

2.3 Evaluation of Time Series Segmentations

Evaluating time series segmentations is a relatively new field, which has only been researched on a limited scale. While the term clustering strictly speaking refers to unordered data sets, the evaluation of data clustering is nonetheless discussed in this section due to its significant similarities to (time series) segmentation.

According to [14], retrieving features from an unlabeled data set with unsupervised approaches is relatively challenging. The evaluation of the clustering (i.e. segmentation) quality has thus far been done by using classified datasets and by comparing the clustering precision of custom algorithms with some ground truth. In 2015, Aghabozorgi, Shirkhorshidi, and Wah discussed in their work [15] that the evaluation of clusters in data sets without classes is difficult and still an open problem. Similarly to [13] the authors say that the definition of precise clusters (i.e. segments) in a data set depends on the preferences of the user and the area of application. Further, the authors propose different measures for evaluating the quality of clusters, such as Rand index, cluster purity, Jaccard measure, and F-measure. They come to the conclusion that in unsupervised clustering tasks, some of the measures (belonging to the group of external indices) are not applicable for measuring the quality of a segmentation and that measures belonging to the group of internal indices should be used. An auspicious measure, which the authors mention in their work, is the homogeneity index which is discussed and used as an evaluation measure of clustering methods in the recent publication [16] of Javed, Lee, and Rizzo. In their work, the authors discuss the current state for evaluating time series segmentation methods. For their discussed measure, the authors reference the work [17] of Andrew Rosenberg and Julia Hirschberg and just like in [15], they differentiate between internal and external measures. The conclusion is that external measures can be used when class labels for the data points of a data set are present and internal measures should be used otherwise. Due to the data points of all time series used in this work not being labeled, solely internal measures for quantifying the segmentation quality of multivariate time series are discussed and evaluated.

3 Motivation

Before presenting some tools, which are used in state-of-the-art models for multivariate times series segmentation in chapter 4, an overview of the problem statement and the most important definitions in the domain of univariate and multivariate time series analysis are provided in the next two sections.

3.1 Fundamental Problem Statement

Sought-after is a method that can divide a multivariate time series consisting of multiple single time series into an adequate amount of non-intersecting segments, with each segment being homogeneous (i.e. having consistent inner behavior). More formally, a multivariate time series $M = (T_1, ..., T_m)$ consisting of single time series $T = (x_1, ..., x_n)$ should be divided into k segments $S_1, ..., S_k$ of consecutive vectors $S_i(a, b) = (X_a, ..., X_b)$ such that the segments do not overlap each other. Therefore, a c-segmentation of the multivariate time series is a partition of M into c segments, such that $a_1 = 1$, $a_i = b_{i-1} + 1$, and $b_c = n$. In other words, the c-segmentation shall produce the best representation of M using only c disjoint time intervals by segment boundaries, with each segment being homogeneous.

As there is no formal definition of homogeneity for segments in time series at this point, a theoretical quantification of the segmentation quality is defined. For better understanding, a univariate time series consisting of one synthetic signal is presented in 3.1. Hereinafter, all time series presented for this work share the same structure and definition: The time series are treated as software signals with the signal range on the y-axis and the reading point (i.e. the time) on the x-axis. This has the advantage of a uniform analysis and follows the structure of the multivariate time series provided by the automotive partner. For below example, a possible c-segmentation producing the best representation of T could be one with three segments.

In figure 3.1 there are three segments defined by segmentation lines between the beginning of the signal and its end. Using the definition above, the time series T shown in figure 3.1 contains the three segments $S_1(0, 19)$, $S_2(20, 79)$, and $S_3(80, 100)$. The idea behind this segmentation is that the first and third segment contain parts of the signal which have a relatively low fluctuation and have values in the range of 20. The second segment contains a part of the signal which first rises until reading point 50 and then drops smoothly.

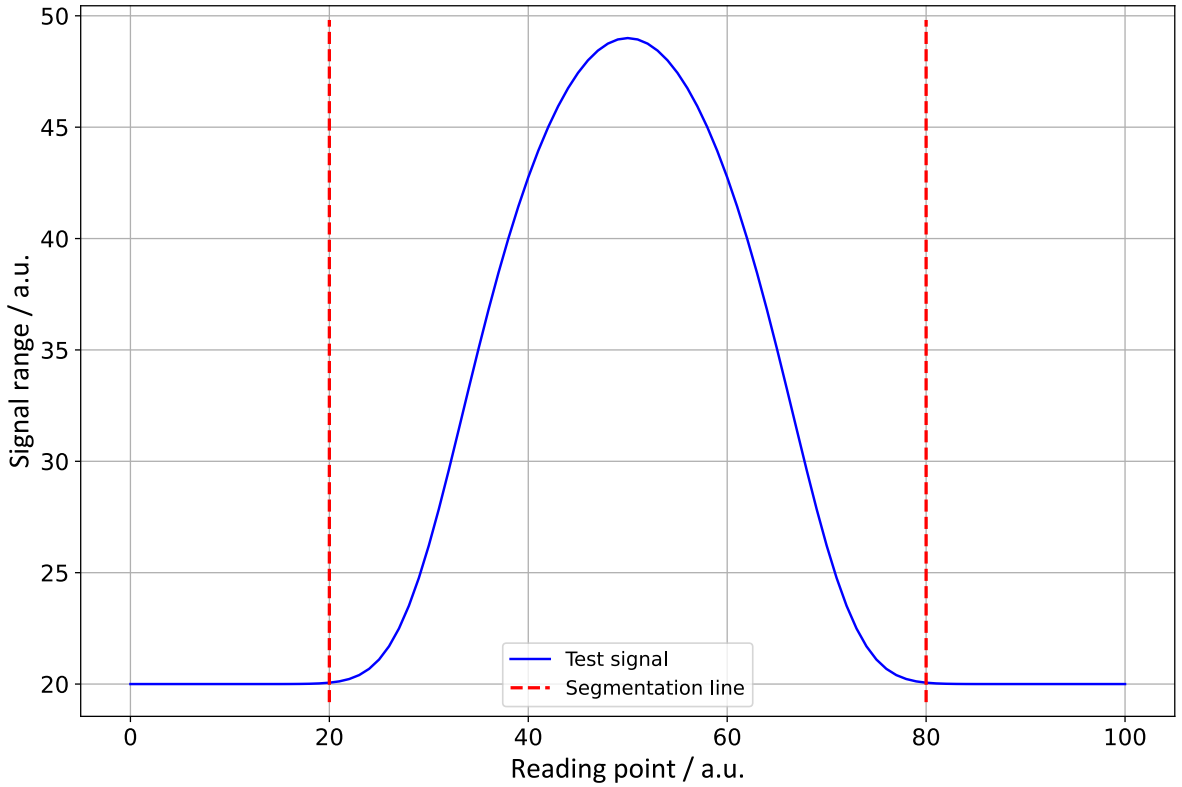

Figure 3.1: Intuitive solution for a c = 3 segmentation of the test signal.

Another possible segmentation can be seen in figure 3.2. Here, there are four underlying segments. The second segment presented in figure 3.1 has been divided into two smaller segments with the thought that the rise and fall of a signal can each be seen as individual and independent behaviors (i.e. states), justifying two separate segments.

The problem of dividing a time series into homogeneous segments can be split into two subproblems and expressed as an optimization problem:

> What is an appropriate number of segments for dividing a time series into segments with relatively high homogeneity?

It can be argued that it is preferred to choose a segmentation with fewer segments if with a decreasing number of segments the homogeneity in each segment stays the same or increases. This follows the principal of Occam's Razor[1], which in this context translates to minimizing the total amount of segments in a time series.

[1]See CE. Rasmussen and Z. Ghahramani. "Occam's Razor". In: *Advances in Neural Information Processing Systems 13*.

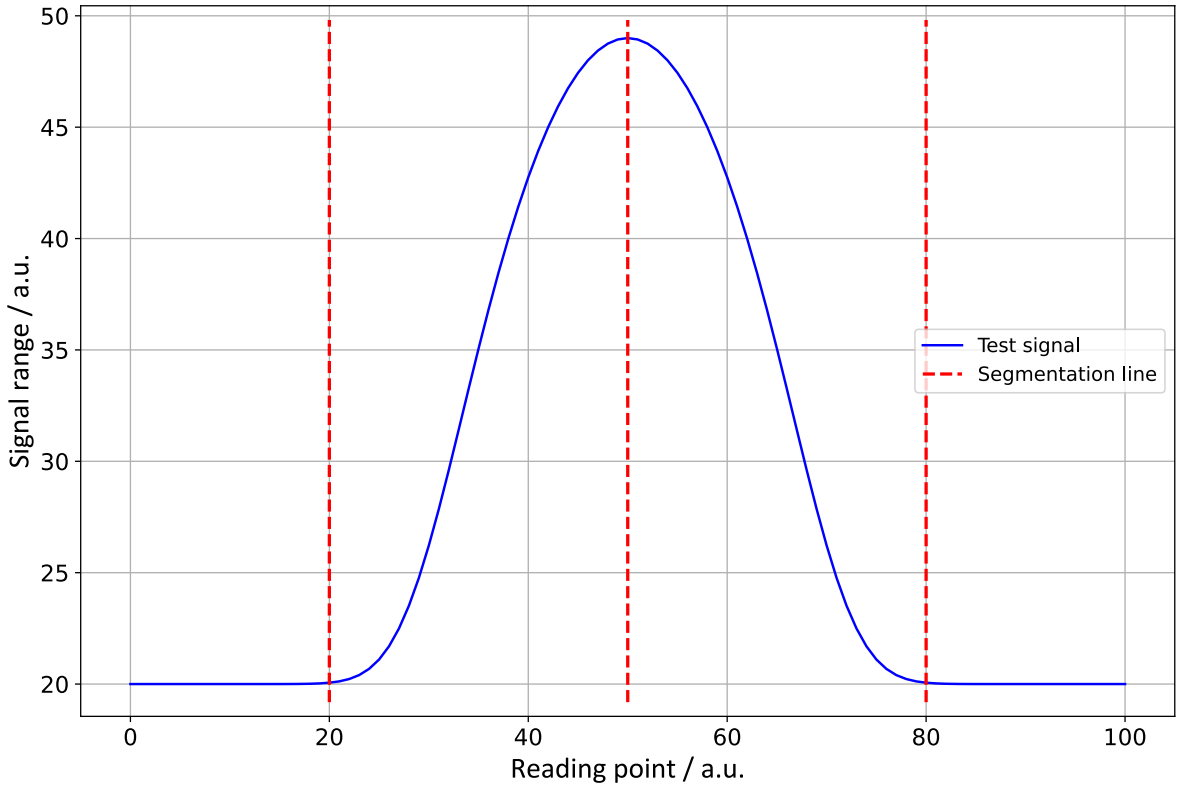

Figure 3.2: Intuitive solution for a c = 4 segmentation of the test signal.

Getting back to the examples shown above, if the homogeneity of segments in both segmentation scenarios is the same, then the c = 3 segmentation shown in 3.1 is preferred over the segmentation in 3.2.

> With a given number of segments, what are the ideal segmentation indices for each segment, yielding the maximum homogeneity for the segmentation?

After answering the question to the first subproblem of an appropriate number of segments in a time series, the segmentation indices found during the time series segmentation shall be adjusted and optimized to maximize the homogeneity in each segment.

For above signal, an example of a disadvantageous c = 3 segmentation could be one with the first segment having an upper segmentation index of 10 and the second segment an index of 60. Figure 3.3 shows this segmentation. The disadvantageous segmentation leads to the first segment still being relatively homogeneous, while the second and third segment contain parts of the signal which should intuitively belong to one of the other segments.

According to this work's interpretation, the two subproblems of minimizing the number of segments and finding segments with high homogeneity in a time series segmentation are

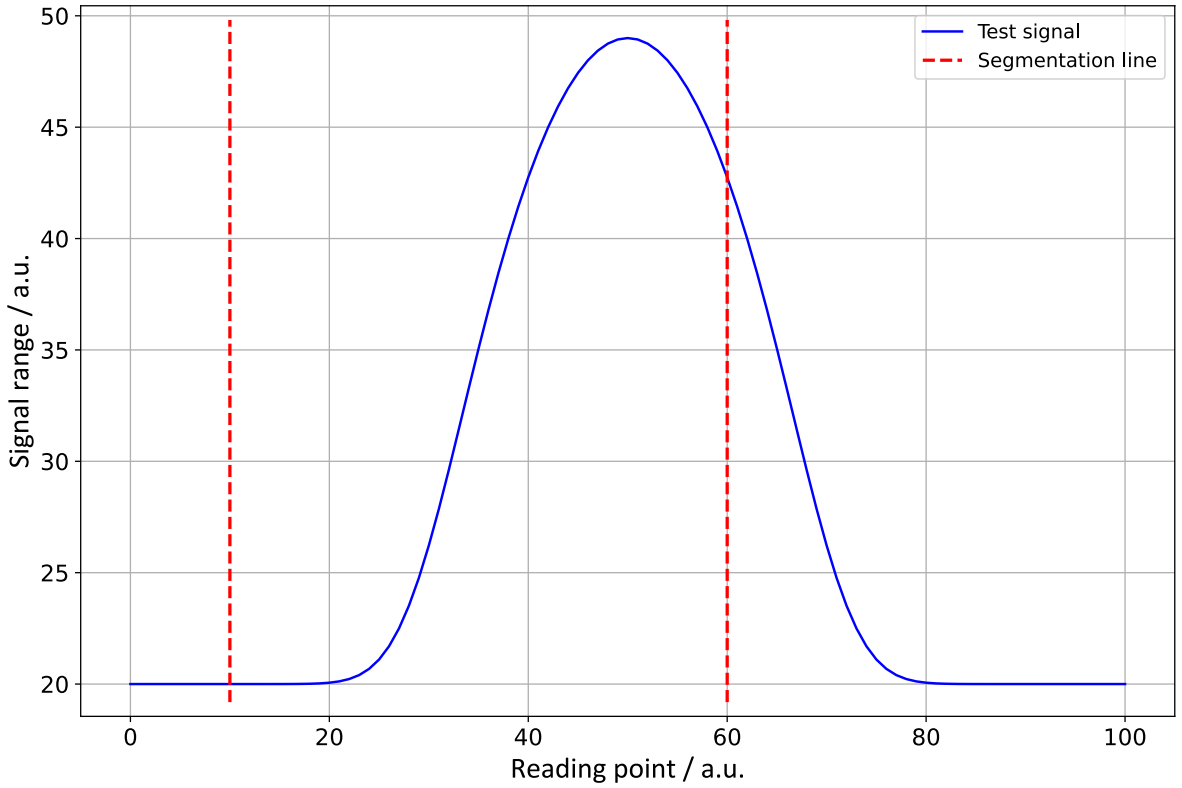

Figure 3.3: Disadvantageous segmentation of the test signal.

working against each other: Purely minimizing the number of segments in a time series would result in a single segment over all data points. Segmenting with the aim of increasing the homogeneity of each segment leads to an overly fine segmentation with segments containing only a few data points. Eventually this results in a segmentation with one segment per data point. The optimal solution to the main problem of segmenting a time series is therefore a tradeoff between these two subproblems. A central research problem is how this tradeoff can be achieved. This question remains open, as for now no algorithm has been found which can optimally segment a data set with above criteria. This leads to the search for algorithms which deliver an approximate solution to the problem of segmenting multivariate time series. Heuristics for evaluating the efficiency and effectiveness of these algorithms to approximate optimal solutions are further investigated.

3.2 Defining an Optimal Segmentation with a Novel Homogeneity Metric on a Synthetic Signal

In this unsupervised approach, it is not apparent if the c = 3 or the c = 4 segmentation of the signal shown in figure 3.1 is preferable. The current advancements in defining homogeneity criteria for segments in time series are not far, with only some papers recently presenting possible metrics for measuring homogeneity in the field of time series segmentation. Some of these metrics are discussed in the next chapter and applied to new test signals. The new test signals relate more to the multivariate time series provided by the automotive partner that is used for segmentation in chapter 5. Figure 3.4 shows a new multivariate time series, which is used as a sample for further definitions and tests. The new multivariate time series consists of three univariate time series (i.e. signals).

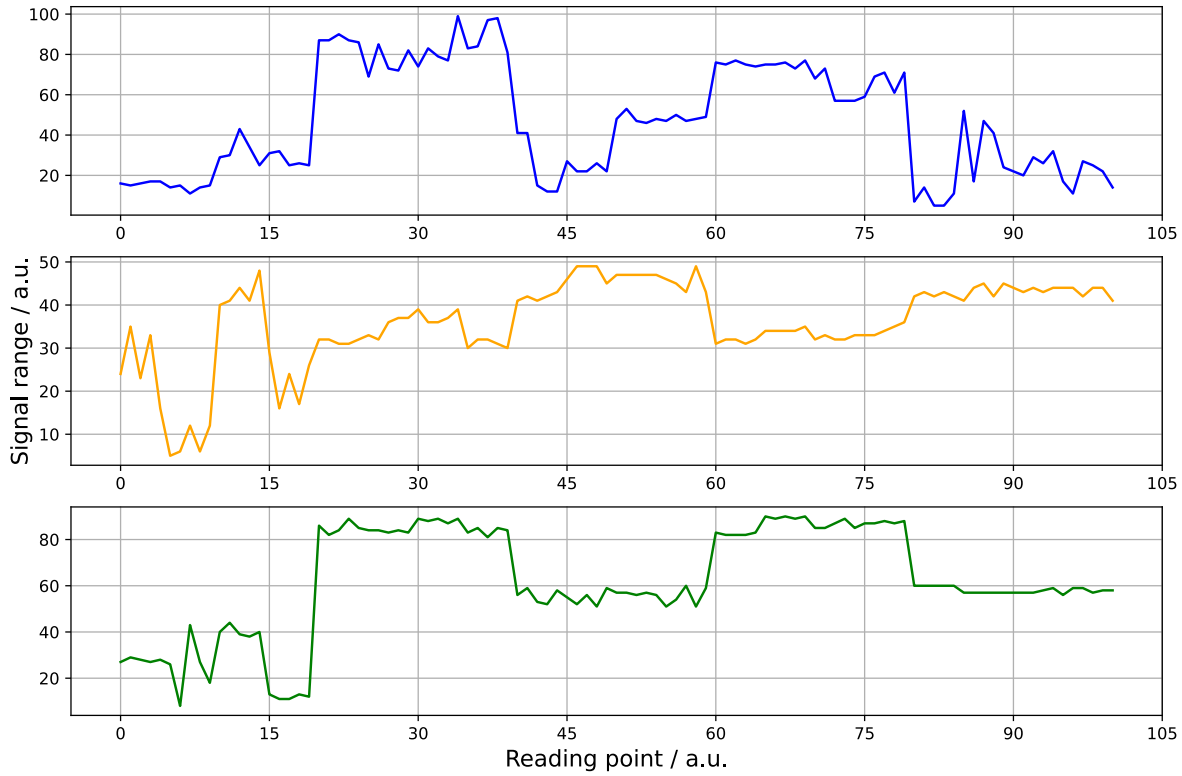

Figure 3.4: Multivariate time series used for further definitions and tests.

Independent of the metric used for measuring the homogeneity of segments found in a time series, finding the exact solution to discussed problem of determining an appropriate number of segments and precise segmentation indices requires a brute force evaluation of every combination of segmentation indices for all segments of a time series. Finding the optimal solution by trying each possible segmentation index for each segment leads to a time complexity

of $\mathcal{O}(n^{m-1})$ in the big O notation, with n being the number of columns in the data set and m being the predefined number of segments. An approximate solution with a lower time complexity is presented in chapter 5.

This work's definition of a segmentation with high homogeneity is illustrated on above multivariate test signal. Figure 3.5 shows the three channels (i.e. signals) of the test signal plotted together.

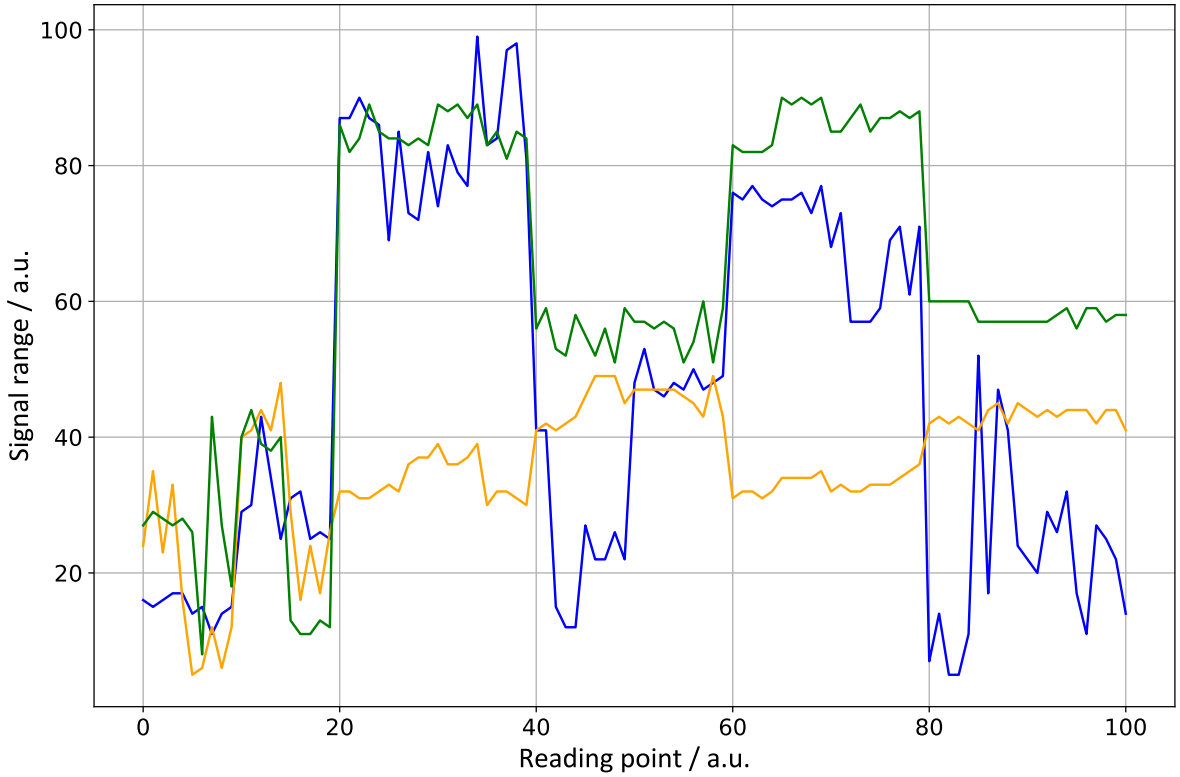

Figure 3.5: Multivariate test signal used for further definitions and tests (consolidated).

An intuitive solution to the segmentation of this multivariate test signal is a c-segmentation with 5 segments and segmentation indices at reading points 20, 40, 60, and 80, as shown in figure 3.6.

The segments one, three, and five contain parts of the signals in the signal range 0 to 60, the segments two and four contain parts of the signals in the signal range 30 to 100 with sharp jumps between the segments of each signal. In this work, it is assumed that the sharp jumps and subsequent stable signal sequences observed in signals from software systems indicate some change in the software system such as a function execution during runtime. The goal of the segmentation of such signals is to extract the singles states (i.e. changes) in the signal. As an extended example, the blue signal from the plot shows a starting signal sequence in the range of roughly 30, before jumping to approximately 80, dropping back to roughly 30, jumping to

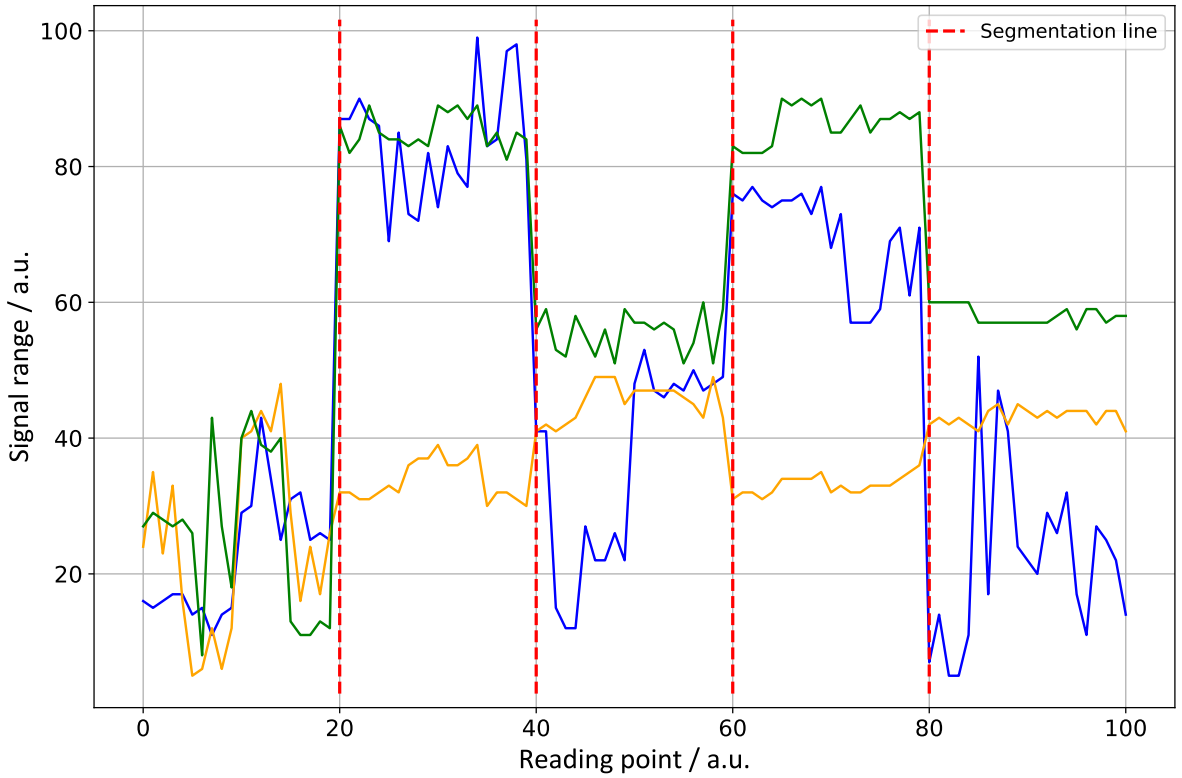

Figure 3.6: Multivariate test signal with the assumed optimal c = 5 segmentation.

70 and dropping back to around 20. Intuitively, the behavior observed in the plot justifies a c-segmentation of 5 with the segmentation indices at reading points 20, 40, 60, and 80. Therefore, the c-segmentation of 5 with the aforementioned segmentation indices is defined as a for this work informal interpretation of an optimal segmentation for the multivariate test signal. The test signal along with this work's interpretation of its optimal segmentation is used as a model for assessing various homogeneity metrics for time series segmentation in the second section of chapter 4. One of these metrics is a custom homogeneity measure specifically developed for this study, which arguably outperforms the state-of-the-art homogeneity measures used for quantifying the quality of multivariate time series segmentations.

Therefore a metric for measuring the homogeneity of segments in a multivariate time series has been developed. The metric measures the mean deviation of the maximum and minimum values in each segment of every univariate time series contained in a multivariate time series. Moreover, the ratio of segments to the number of data points in the multivariate time series is included. The metric works in a way that it rewards the minimization of variance in a segment and at the same time punishes segments that are relatively long (compared to the remaining segments). Additionally, the metric favors a lower number of segments for any given segmentation while at the same time penalizing an excessively low number with the criterion of

segment length. Therefore, the developed metric finds a balance between the subproblems of minimizing the number of segments in a segmentation and maximizing the homogeneity within each segment (i.e. maximizing the number of segments). The metric is denoted as follows:

$$Homogeneity := \frac{\frac{\sum_{p=1}^{l}\sum_{s=1}^{S}(max(s_p)-min(s_p))\cdot n_s}{\sum_{p=1}^{l}(max(p)-min(p))\cdot n} + \frac{\sqrt{n}\cdot S}{1.7n}}{1+\frac{\sqrt{n}}{1.7}} \tag{3.1}$$

The metric is normalized and outputs a measurement in the range of zero to one. A lower measurement describes a higher homogeneity for a given segmentation, therefore the goal is to minimize it. The higher the number of segments in a segmentation, the larger the term $\frac{\sqrt{n}\cdot S}{1.7n}$ becomes, hence acting as a regulation term for the number of segments from a segmentation. The term $1+\frac{\sqrt{n}}{1.7}$, which is in the denominator of the whole fraction, is for normalization of the metric. Of special interest for the criterion of homogeneity is the term $\frac{\sum_{p=1}^{l}\sum_{s=1}^{S}(max(s_p)-min(s_p))\cdot n_s}{\sum_{p=1}^{l}(max(p)-min(p))\cdot n}$. This term describes the mean deviation of the maximum and minimum values (i.e. the variance) in each segment of every univariate time series contained in a multivariate time series. n_s, which is the number of data points n in segment s acts as a regulator to make segmentations with relatively large (i.e. long) segments (compared to the remaining segments) unfavorable. Using this homogeneity metric, a high homogeneity for a segmentation can be achieved by choosing the lowest number of segments needed to retrieve a relatively low signal fluctuation (i.e. variance) within each segment while at the same time ensuring that the length of each segment (compared to all other segments) is relatively even.

In this work, the metric is mainly used to quantify the segmentation quality of conventional multivariate time series segmentation models. However, as an assessment for the suitability of the metric for quantifying the quality of multivariate time series segmentations, the metric is tested with the use of the multivariate time series introduced in figure 3.4. For the assessment, every possible combination of segmentation indices for all segments of a c = 5 segmentation of the multivariate test signal is measured with the new homogeneity metric. If the segmentation with the highest homogeneity is the same as in figure 3.6 (i.e. this work's definition of the optimal segmentation for the multivariate test signal) then the metric quantifies this work's informal definition of an optimal segmentation and is therefore suitable for benchmarking different multivariate time series segmentation models. After running the test, the segmentation that achieved the highest homogeneity is the one demonstrated in figure 3.7.

The plot shows the same segmentation as in figure 3.6 and thus matches this work's definition of an optimal segmentation. Therefore, the metric is a suitable candidate for quantifying the segmentation quality of different segmentation models as per the homogeneity of a segmentation.

It is important to note that the brute force method for finding the optimal solution to the

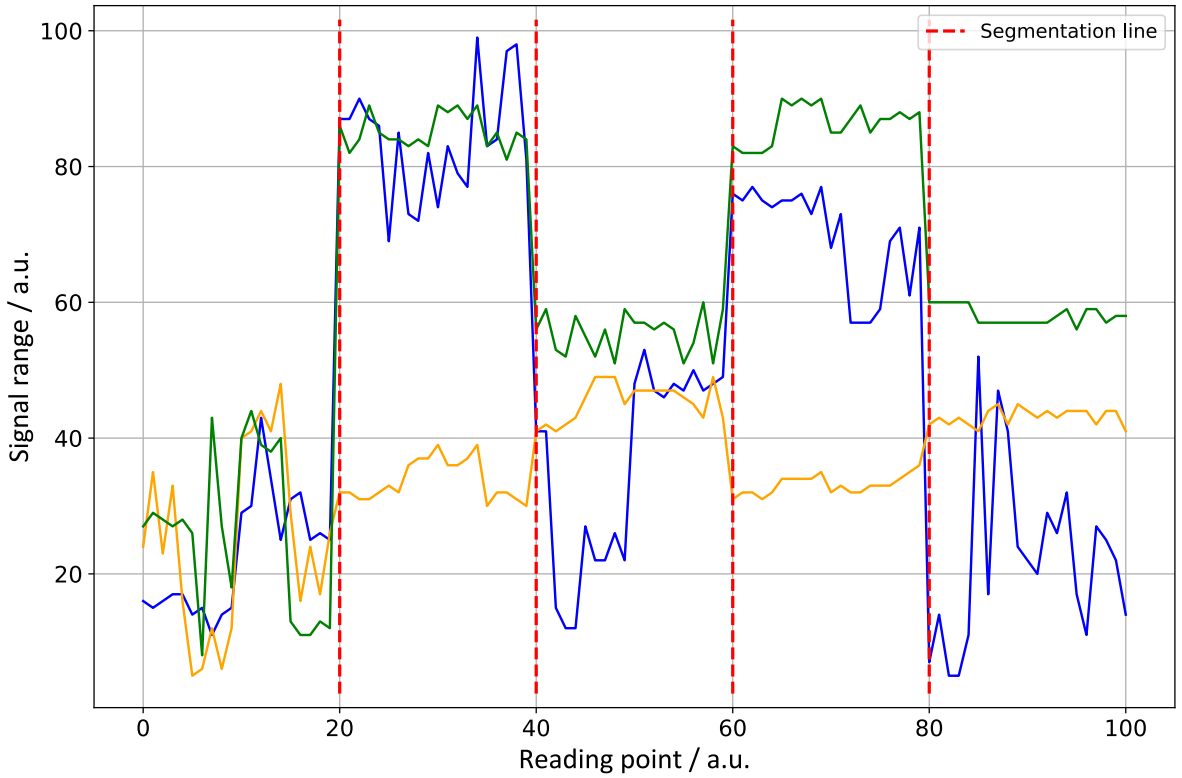

Figure 3.7: Segmentation with highest homogeneity as per the homogeneity metric.

segmentation problem for an arbitrary multivariate time series is not feasible, as the runtime of this method is relatively high. Finding the optimal segmentation with the multivariate test signal by trying every possible combination of segmentation indices took roughly three hours with the hardware used specifically for this work. To achieve comparability, all other time measurements are made with the same hardware. That given, the needed time can be measured relatively. As will be demonstrated in later chapters, the c = 5 segmentation of the multivariate test signal can be achieved in a much shorter time with the use of optimized segmentation models. Therefore, the homogeneity metric formalized above is only used for existing segmentation models to quantify their segmentation capabilities for multivariate time series. Much more efficient ways of dividing a time seris into homogeneous segments, as for instance the principal component analysis based fuzzy clustering (PCAFC), are presented in chapter 5. The possibility of using the homogeneity metric in a cost function for a segmentation model still remains open and is discussed at the end of this work.

4 Tools for Multivariate Time Series Segmentation

Multivariate time series segmentation has been researched in the domain of data mining since a relatively long time, with major advancements being made in the last two decades. Recurring methods for multivariate time series segmentation, which are discussed in recent publications, are the dimensionality reduction of multivariate data sets, (internal) homogeneity criteria for determining precise segmentations, and the sliding window or hierarchical segmentation techniques. A special segmentation technique, encompassing regression analysis, is discussed in the last section of this chapter. Some of the methods that are discussed here are used in the segmentation models presented in chapter 5. Additional methods that can be used for multivariate time series analysis and which are not applied in this work are presented for completeness.

4.1 Dimensionality Reduction

Dimensionality reduction is a way of finding fundamental features in data sets and a prominent method of reducing the complexity of multivariate time series while preserving most of the original information [18, p. xi], [2, p. 28]. The complexity of multivariate time series stems from the curse of dimensionality which, in short, is the exponential increase in data required for analysis with the increase in dimensionality of a data set [18, p. 47]. Ultimately, a more compact and smaller data set makes the use of expensive (i.e. complex) algorithms more viable [2, p. 37].

The main approaches for dimensionality reduction are feature selection, which is the selection of a subset of features from a data set, and feature extraction, which is the transformation of a data set with its features into a space with lower dimensions. Principal component analysis (PCA), the main method used for dimensionality reduction in this work, is the main focus of this section. Two additional methods for dimensionality reduction, which are prominent in the field of multivariate time series analysis, are discussed for completeness.

4.1.1 Principal Component Analysis

In his work [19] on the relation between sets of variates, Harold Hotelling describes the fundamental principle of PCA. As an example, he depicts the analysis of wind in two different

places. Instead of observing many different measures, which can be used for comparing winds, the two components of velocity, namely the magnitude and the direction, can be used for the analysis, thus reducing the dimension of variables (i.e. features) needed for an analysis down to two. The same principal can be applied to a multivariate data set by reducing the observable features of the data set with PCA to the n principal components (PCs) required to represent a certain percentage of the variance in the data set. The dimensionality reduction, however, always involves some loss of information. The loss depends on the number of PCs used for further analysis.

The basic principle behind PCA is the mapping of features of a data set on a hyperplane. In the process, the data is rotated in a separate axis system [2, p. 42]. More formally, PCA "takes a data set $X = [x_1, ..., x_N]^T$ where $x_k = [x_{1,k}, ..., x_{n,k}]^T$ is the kth sample or data point in a given orthonormal basis in R^n and finds a new orthonormal basis, $U = [u_1, ..., u_n]$, $u_i = [u_{1,i}, ..., u_{n,i}]^T$, with its axes ordered" [18, p. 48].

An example of dimensionality reduction using PCA is shown in the following. Figure 4.1 depicts a shortened version of the multivariate data set presented in figure 3.4 containing three channels (i.e. signals). Such signals are also called variables in the context of multivariate time series analysis.

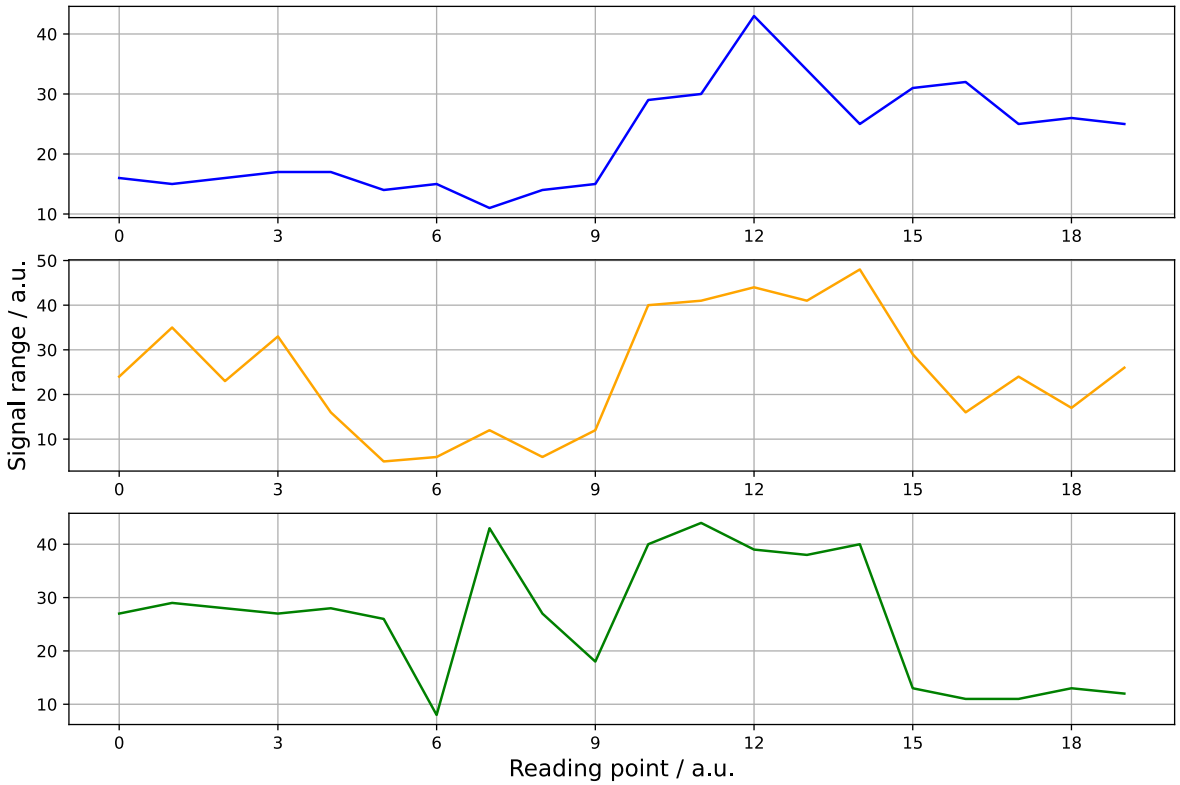

Figure 4.1: Short version of the multivariate time series.

Due to the data set containing three variables and assuming that the variables show some correlation between each other, the data set can be plotted in a three dimensional space for better visualization, as shown in figure 4.2.

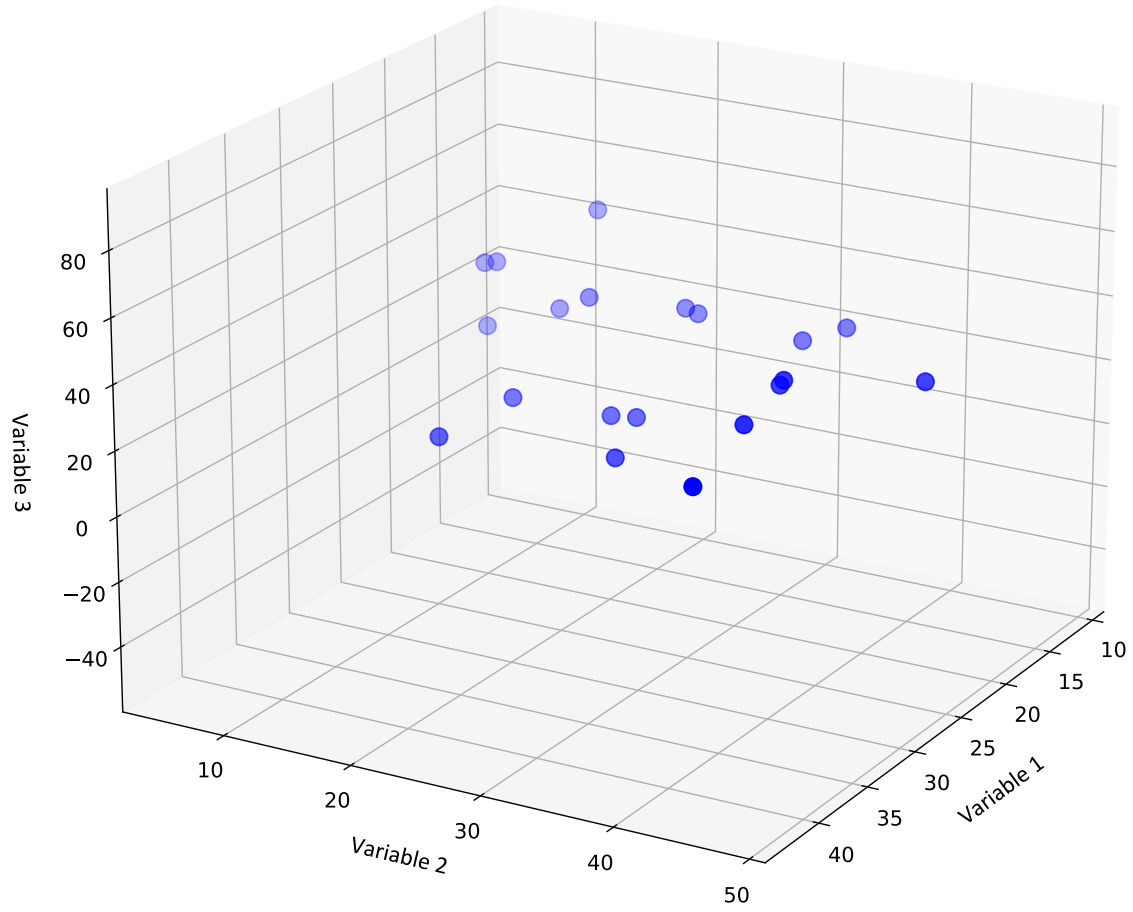

Figure 4.2: 3D plot of the multivariate time series.

Observing the three dimensional plot, it can be agreed that it is hard to see clear dependencies between the data points in the three dimensional space. Thus finding correlations between the variables could present a challenge. Using PCA however, the most prominent features of the data set can be mapped on a hyperplane in the three dimensional space, rotated, and transferred into a two dimensional coordinate system. A visualization of the hyperplane capturing most of the variance in this multivariate data set can be observed in figure 4.3.

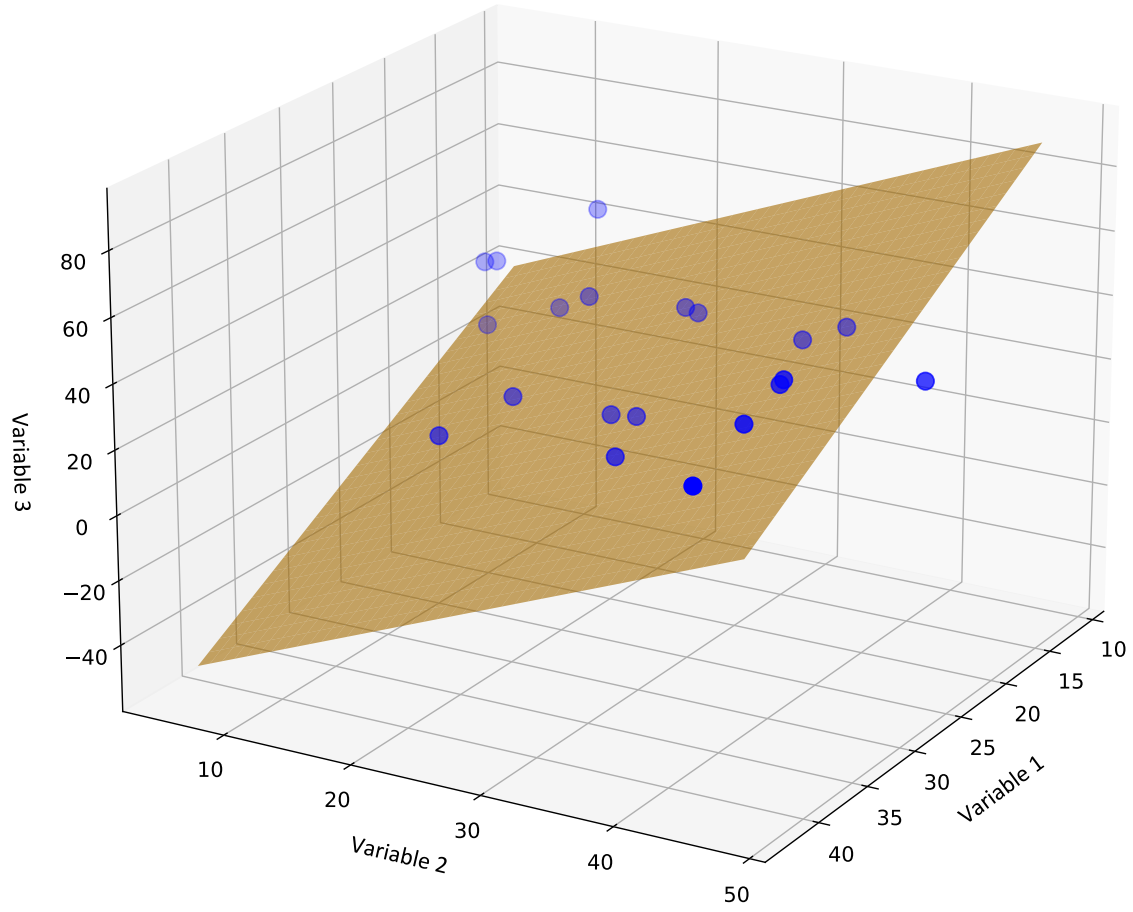

Figure 4.3: 3D plot of the multivariate time series containing the hyperplane from the principal component analysis.

When mapped on this hyperplane (i.e. orthonormal basis), the two components (i.e. features) captured in the two dimensional space hold more than 95 % of the variance from the original data set. Therefore, two components found in the PCA are enough for capturing a high percentage of the information from the original data set. The result of the PCA transformation with two PCs can be seen in figure 4.4.

Using only the two most prominent components (i.e. features) from the data set, and thus reducing the dimension from three to two, facilitates the analysis of the data (e.g. for clustering, correlation analysis, etc.). PCA can be applied to data sets with an arbitrary

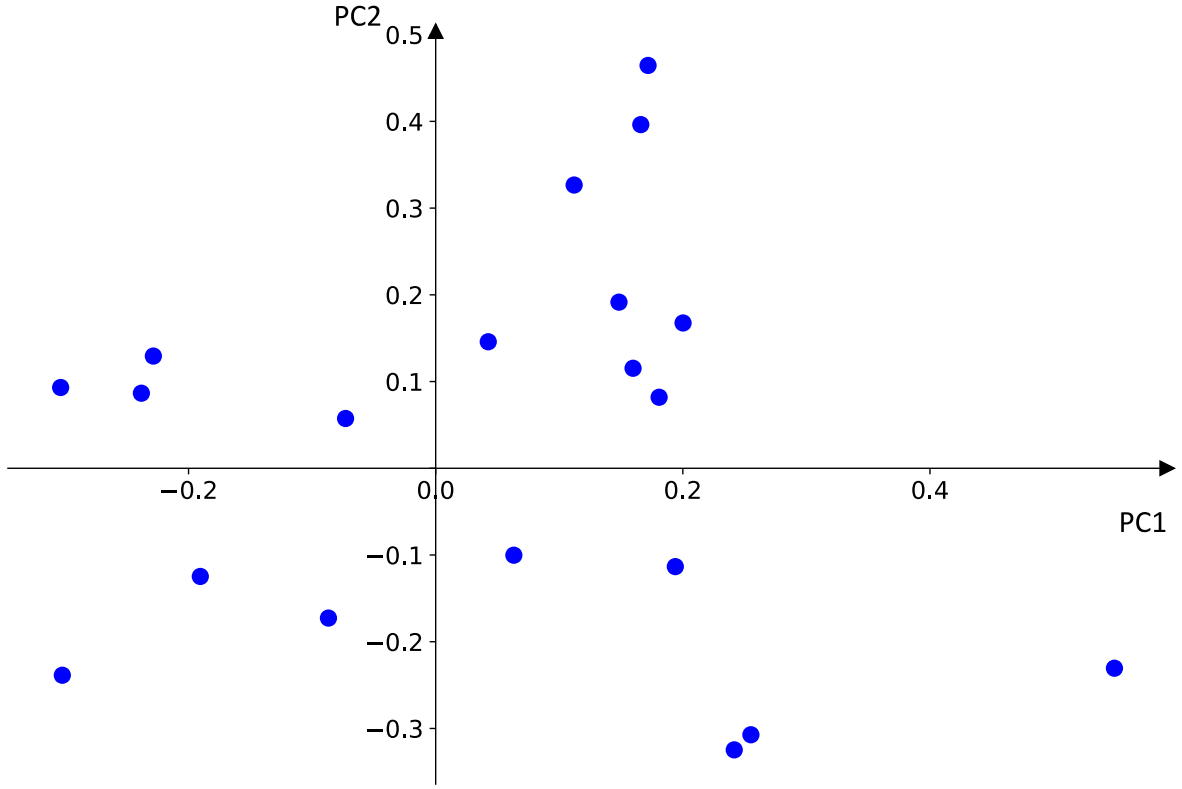

Figure 4.4: Principal component analysis of the multivariate time series containing first two principal components.

number of dimensions.

As shown in chapter 5, this characteristic makes PCA (among others) a suitable candidate for the procedure of analyzing and segmenting multivariate time series.

4.1.2 Locally-Linear Embedding

While PCA is an example of a linear dimensionality reduction technique, locally-linear embedding (LLE) is an example of a nonlinear one. Here, linearity refers to the relationship of the original high-dimensional data. Therefore, nonlinear dimensionality reduction methods are applied when the original high-dimensional data contains nonlinear relationships. LLE finds the nearest neighbors for each data point and creates a weight matrix, describing each point as a linear combination of its neighbors. Finally, LLE finds a low-dimensional embedding of the points, while preserving the linear combination of one point to its neighbors [20, p. 1], [21, pp. 1–2]. LLE has been used in some works involving clustering (i.e. segmentation), such as [9] or [22]. For this work, linear dimensionality reduction techniques, such as PCA, are sufficient to capture most of the variance from multivariate time series in a lower dimension while being computationally more efficient than LLE [23, p. 160]. Therefore, nonlinear dimensionality

reduction techniques, such as LLE are not pursued further in this work. It will, nonetheless, be of interest for future work to examine this type of dimensionality reduction technique for multivariate time series segmentation tasks.

4.1.3 Autoencoders

Autoencoders, a type of artificial neural network which has gained increasing interest in recent times, is also widely used as a method for dimensionality reduction. Autoencoders work with encoders and decoders, where the encoder transforms an input into a vector, and the decoder transforms the vector representation of the original input back to a target sequence. The model tries to recreate the original sequence from the fixed-length source sequence in an unsupervised manner. Once the decoder part of the autoencoder achieves a desired accuracy for recreating the vector representation of the original input, the decoder part is removed, with only the encoder of the autoencoder remaining. The encoder can then be used to represent the original input as a vector with reduced dimensionality [24, p. 1], [25, pp. 1724–1725]. Some examples of works including autoencoder based clustering are [26], [27], and [28]. Compared to PCA, autoencoders have a much higher computation time and require more resources [29, p. 214]. Because PCA, as shown in the next chapter, is sufficient for and more efficient in capturing features of multivariate time series in lower dimensions, autoencoders, just like LLE, are not pursued further in this work but remain open for future research in multivariate time series analysis.

4.2 Classic Segmentation Techniques

Classic segmentation techniques can be divided into three major categories: top-down and bottom-up segmentation, belonging to the hierarchical segmentation technique, and sliding window segmentation [30], [31, p. 290].

4.2.1 Top-Down Segmentation

In the top-down approach, the whole time series is first observed as one major segment. The initial segment is then recursively divided into more segments such that the difference in features between the new segments is maximized. This procedure is repeated until some stopping criterion is met (e.g. number of segments or approximation error) [3, p. 259]. An illustration of the top-down segmentation procedure can be seen in figure 4.5. One of the disadvantages of top-down segmentation is the inflexibility of the method because segmentation indices found in previous iterations of the segmentation remain throughout the segmentation process. This could become a challenge if the initial segments found with this approach do not remain optimal with an increase in segments. The probability of achieving an optimal segmentation with the first few iterations of the top-down segmentation technique is relatively small [32, p. 38]. Therefore, the top-down segmentation is not used in this work.

4.2.2 Sliding Window Segmentation

The sliding window segmentation involves a window of flexible size which scans along the time series and determines fitting segmentation indices in one pass. First, a left boundary for a segment is determined (first data point for the first segment). Then the size of the sliding window is increased along the time series, as long as the approximation error of the current segment (i.e. data inside the window) is below the user-specified threshold. If the threshold is exceeded, a segment is created and the window moves to the right border of the newly found segment. This process repeats until the end of the time series is reached. An illustration of the sliding window segmentation procedure can be seen in figure 4.6. The multivariate time series segmentation model using PCAFC, which is presented in the next chapter, requires a hierarchical segmentation technique, which is why the sliding window segmentation technique is not pursued further in this work. Another reason for not considering sliding window segmentation in this work is the fact that this type of segmentation algorithm selects segments in a greedy fashion: while scanning the data set to be segmented from left to right, the algorithm ignores possible long-term dependencies between subsequences [33, p. 1]. Despite this drawback, sliding window segmentation has shown some successful applications especially in medical research [32, p. 41] and is, due to its simplicity, an attractive candidate for further research in the field of multivariate time series segmentation.

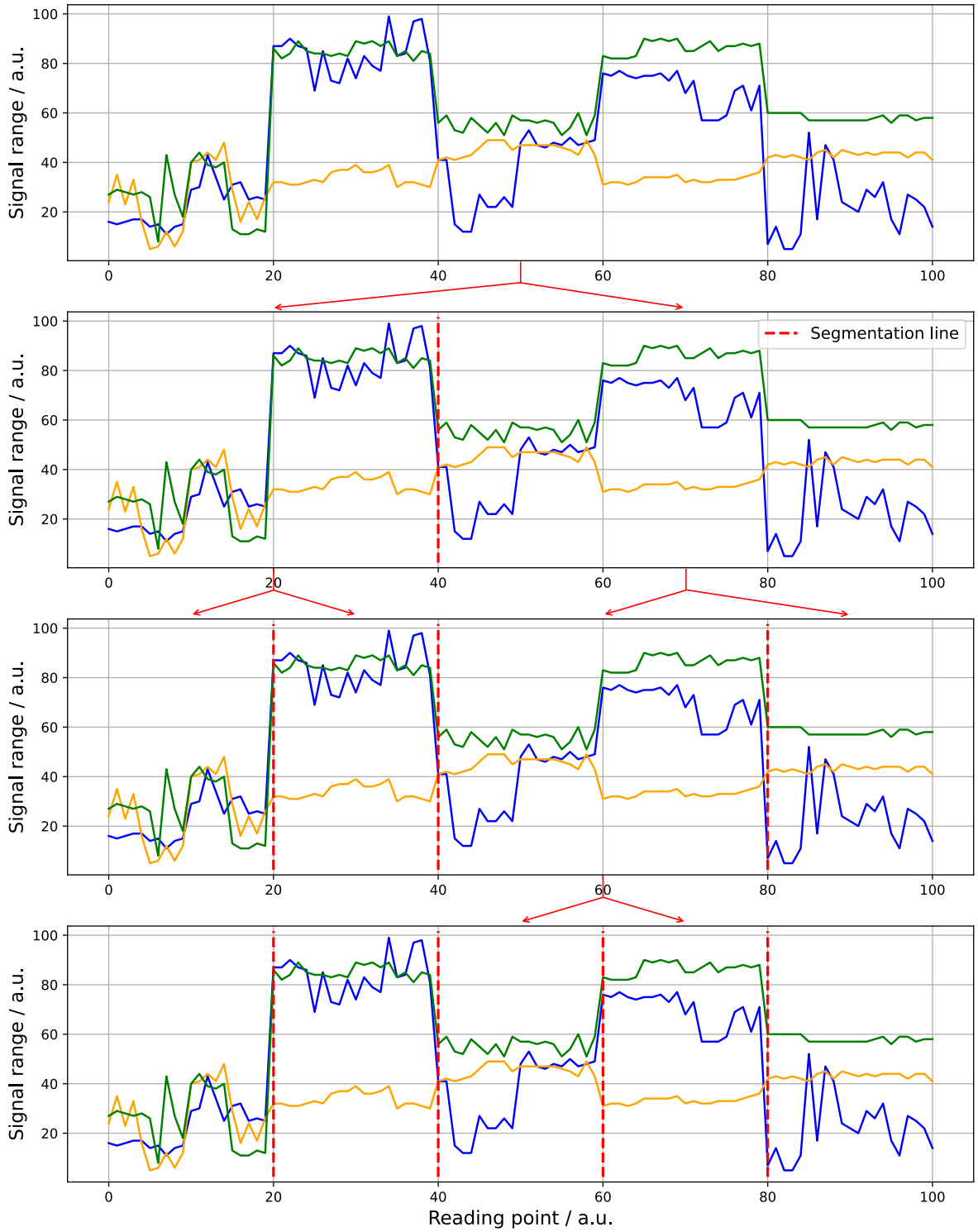

Figure 4.5: Illustration of the top-down segmentation procedure using the multivariate test signal.

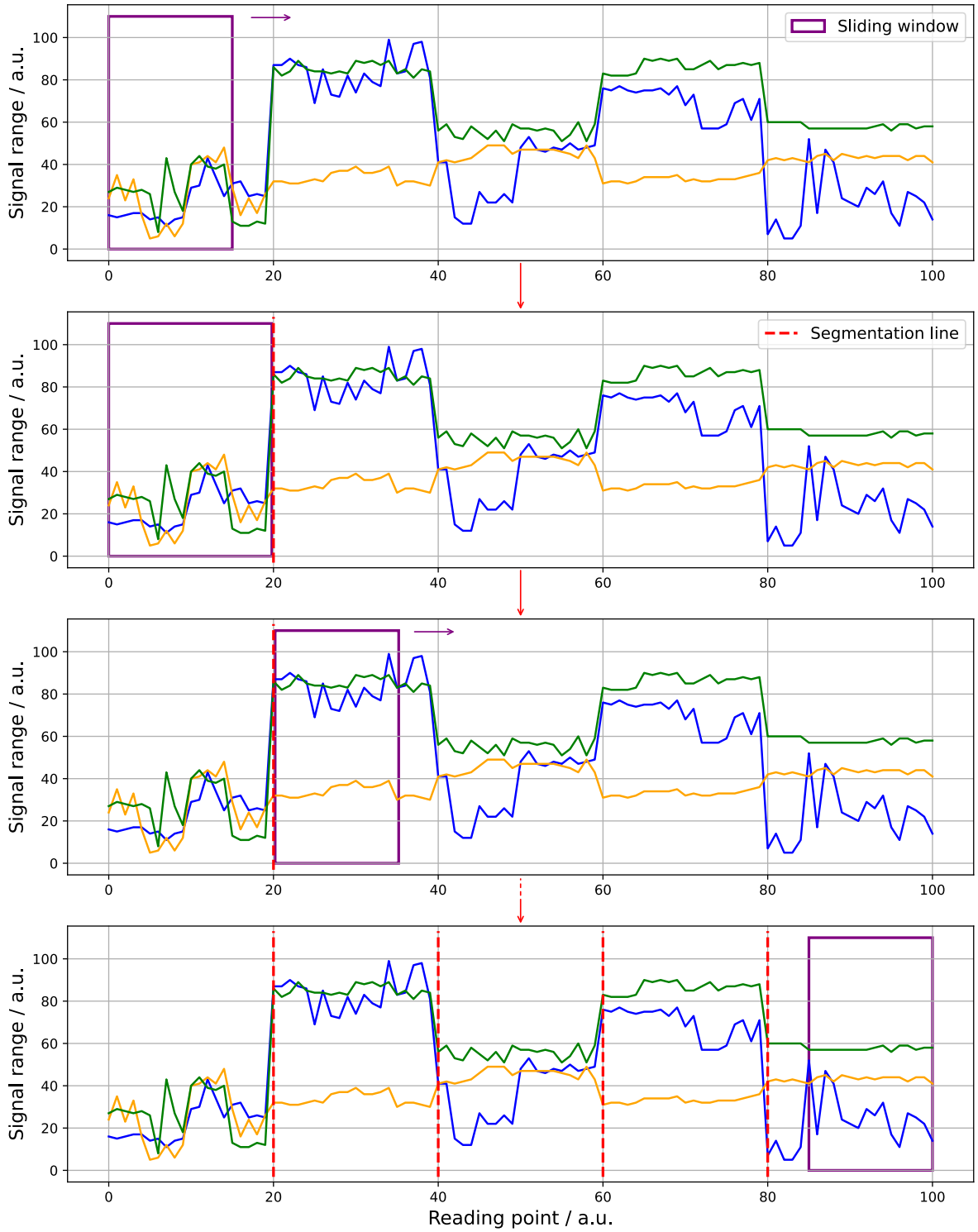

Figure 4.6: Illustration of the sliding window segmentation procedure using the multivariate test signal.

4.2.3 Bottom-Up Segmentation

The bottom-up segmentation method is the complement to the top-down method and ideally starts the segmentation with a maximally fine segmentation (i.e. one segment per data point). In each iteration, two adjacent segments are merged, creating one new segment. The two segments for merging are chosen by calculating the similarity (e.g. closeness or parallelism) between all adjacent segments and selecting the two segments with the highest similarity (or the lowest merging error). This process is repeated until some stopping criterion is reached, as for instance a desired number of segments or the excess of some threshold [32, p. 40], [3, pp. 259–260]. An illustration of the bottom-up segmentation procedure can be seen in figure 4.7.

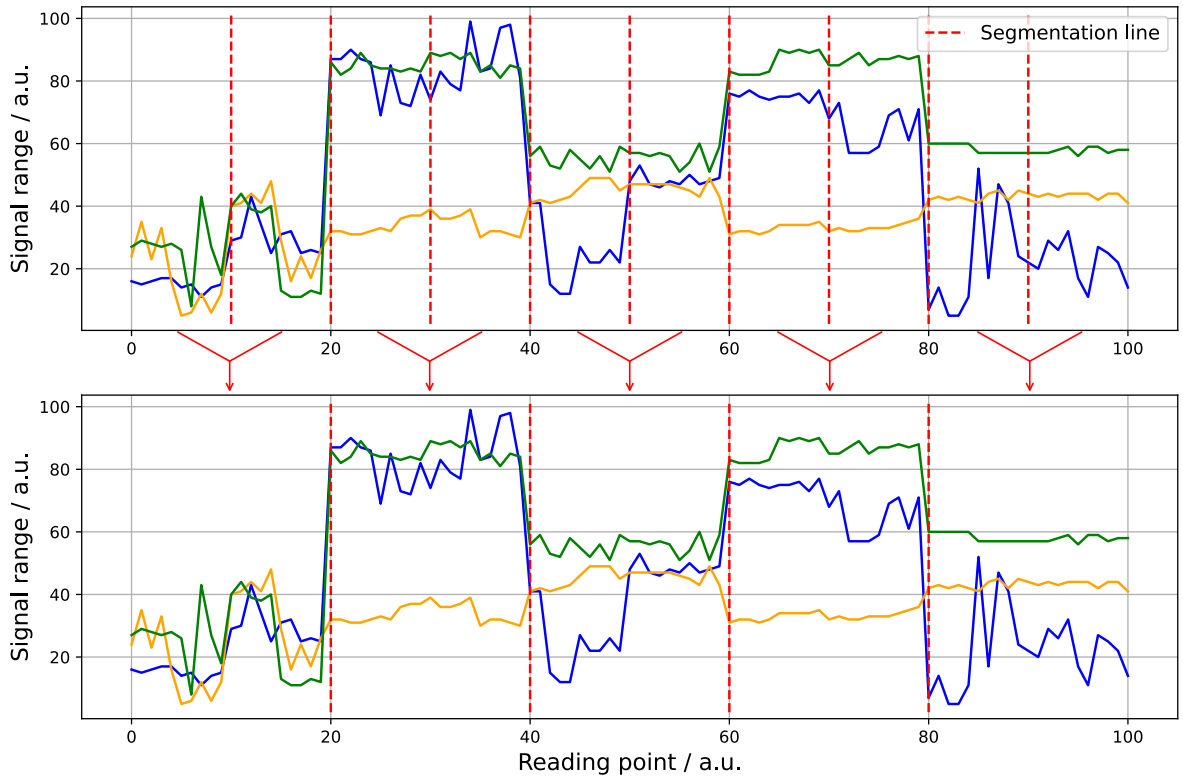

Figure 4.7: Illustration of the bottom-up segmentation procedure using the multivariate test signal.

Due to the bottom-up segmentation starting with a fine segmentation, it overcomes the inflexibility of the top-down segmentation technique. Moreover, by merging adjacent segments, the bottom-up approach considers long-term dependencies in a data set and therefore overcomes the drawbacks of the sliding window approach. For one of the two models used in this work (PCAFC), a combination of a bottom-up segmentation with PCA is used for segmenting multivariate time series. First, PCA is performed on the whole multivariate time series and an adequate number of PCs is selected. The time series is then divided into a relatively high

number of equispaced segments. Finally, adjacent segments are merged based on their similarity: the similarity is calculated by performing PCA on each segment and comparing the angle of the segments' PCs as well as the distance of their centers (i.e. segment centers) to one another. The procedure stops once a desired number of segments is found or similarities between segments fall below a certain threshold. The complete segmentation models are described in more detail in the upcoming chapter.

4.3 Homogeneity Criteria

A core issue in time series segmentation is the quantification of segmentation quality for a time series. As mentioned in the beginning of this work, the homogeneity of segments from time series can be defined as the internal consistency of the single segment, i.e. to which degree the data contained in the segments carry the same information. As shortly mentioned in the previous subsection on segmentation techniques, the PCAFC model for segmenting multivariate time series used in this work already incorporates a type of homogeneity measure for merging adjacent segments. This measure calculates the homogeneity (i.e. similarity) of segments by comparing their PCs and distances to one another. A topic, which has gained interest in the field of time series segmentation in recent times and which plays a fundamental role in this work, is the evaluation of time series segmentations and the quantification of the quality of different time series segmentation models. These topics are discussed for instance in [16] and [34]. Thus far, there has been no substantial progress in the domain of measuring the quality of time series segmentations. In contrast, there has been significant advances in the research of evaluating the segmentation quality in the domain of data clustering. The difference between data sets used for clustering and time series is that time series involve an additional dimension of time. More precisely, the data points in a time series are ordered by time, whereas data sets used for clustering are usually not [3, pp. 259–260]. This means that metrics used for evaluating the quality of clusters in a clustering problem disregard the dimension of time, which could lead to unsatisfactory results when applying these metrics to time series. Nevertheless, in the subsection 4.3.3, different types of evaluation metrics used for clustering are observed and assessed for suitability in the domain of multivariate time series analysis, as the relative and not the absolute segmentation quality for different segmentations is of interest. In the following subsections, different evaluation methods, which are widely used in the domains of time series segmentation as well as clustering, are presented.

4.3.1 Metrics for Time Series Segmentation

The current state-of-the-art metrics used for quantifying the quality of time series segmentations are accuracy, F_1 score, Matthew correlation coefficient, precision, and recall [35, p. 173]. As all of these metrics use a ground truth (i.e. an expert segmentation) for measuring the quality of a segmentation, these metrics are not suitable for this work, which focuses on the unsupervised segmentation of time series without any prior knowledge about optimal segmentations.

In the following two subsections, the external and internal measures for quantifying segmentation quality in clustering tasks are presented.

4.3.2 External Cluster Metrics

The external evaluation criteria for clusters in a data set are used when the underlying data points are labeled, i.e. each data point of the data set belongs to a class. Examples for external cluster metrics are the Rand index [36], adjusted mutual information [37], Fowlkes Mallows index [38], homogeneity [17], and completeness [17]. As this work focuses on data sets containing unlabeled data points, these measures are not further considered.

4.3.3 Internal Cluster Metrics

The internal cluster metrics are used to quantify the goodness of clusters and they do not require the data points of data sets being labeled. Examples for internal cluster metrics are the silhouette coefficient [39], Davies-Bouldin index [40], Dunn index [41], [42], and the I-index [43]. The silhouette coefficient measures the cohesion of a cluster compared to adjacent clusters by taking the average intra-cluster distance, i.e. the average distance between each point within a cluster, and the average inter-cluster distance, i.e. the average distance between all clusters, into account. The Davies-Bouldin index uses the separation between clusters as well as the scattering within each cluster to measure the goodness of clusters. The Dunn index is measured by the compactness of clusters, calculating the variance, separation, and the means between clusters. The index I is calculated with the number of clusters and the separation between clusters over all pairs of clusters. For this work, a slightly modified version of the index I is used, which incorporates the minimum separation between any two clusters in a data set. These evaluation metrics are discussed in detail in [39], [40], [41], [42], and [43].

As mentioned above, the time series segmentations in this work are performed in an unsupervised fashion, more precisely, there is no a priori knowledge about clusters and class labels for segments. Therefore, internal evaluation metrics for clustering are used and assessed for suitability in the domain of time series segmentation. As mentioned in chapter 3, the multivariate test signal from figure 3.4 is used for the evaluation of segment homogeneity, using the above discussed internal measures as well as the newly developed homogeneity metric. For this, the optimal segmentation of the test signal is compared with two disadvantageous segmentations of the same signal. The three different segmentations are shown in figure 4.8.

It is assumed that the optimal segmentation correlates with high homogeneity scores (relative to the disadvantageous segmentations) when using the internal cluster measures and therefore proves the suitability of these measures for quantifying the segmentation quality of multivariate times series segmentations. After applying all metrics (i.e. internal cluster metrics and homogeneity metric) to the three segmentations, the result from table 4.1 can be obtained. Lower values for the Davies-Bouldin index and the homogeneity metric, and higher values for the Dunn index, the silhouette coefficient, and the custom I-index are preferred. As shortly

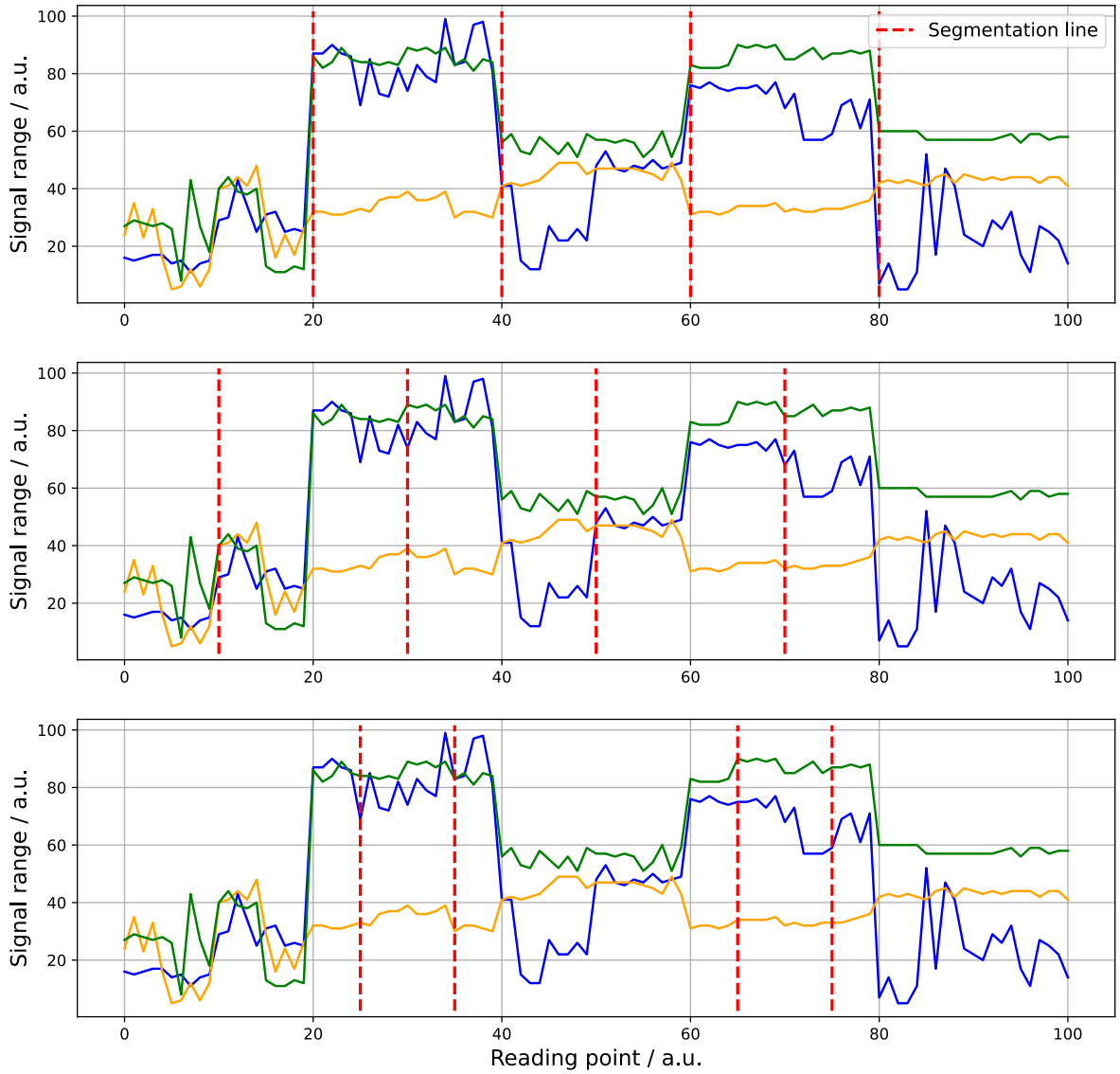

Figure 4.8: Optimal segmentation and two disadvantageous segmentations of the multivariate time series.

mentioned in the previous chapter, the value of the homogeneity metric can be between zero and one.

The results show that the best scores for all metrics correlate with the optimal segmentation of the multivariate test signal. In contrast, the scores for the two disadvantageous segmentations are relatively bad for all metrics. Especially the silhouette coefficient, the custom I-index, as well as the homogeneity metric show a good suitability for measuring the quality of time series segmentations, as the values for these metrics show a high distinction between the optimal segmentation and the two disadvantageous ones. As expected, the segmentation which,

Segmentation method	Evaluation methods for time series segmentation				
	Davies-Bouldin	Dunn index	Silhouette	Custom index I	Homogeneity
Optimal segmentation	**1.26**	**0.23**	**0.24**	**0.11**	**0.08**
Disadvantageous segmentation I	4.96	0.19	-0.01	0.05	0.12
Disadvantageous segmentation II	1.81	0.09	-0.05	0.05	0.13

Table 4.1: Evaluation of segmentation quality for one optimal and two disadvantageous segmentations of the multivariate test signal.

by this work's definition, is optimal correlates with high homogeneity scores, as seen in the results above. However, even if the first four metrics show relatively promising results, these metrics still need to be validated for their capability of quantifying the quality of time series segmentations. Similarly to the homogeneity metric developed in this work, all possible combinations of segmentation indices for a c = 5 segmentation of the multivariate test signal are measured with the four metrics Davies-Bouldin index, Dunn index, silhouette coefficient, and index I. For each metric, the segmentation with the best score is plotted. Figures 4.9 and 4.10 show the segmentations which have the highest scores as per the four metrics.

As the results show, the standard metrics used for measuring cluster purity do not reflect the informal definition of an optimal segmentation for the multivariate test signal. Therefore these metrics are not suitable for quantifying the quality of multivariate time series segmentations. For the quantification of the segmentations presented in the next chapter, only the newly developed homogeneity metric is used.

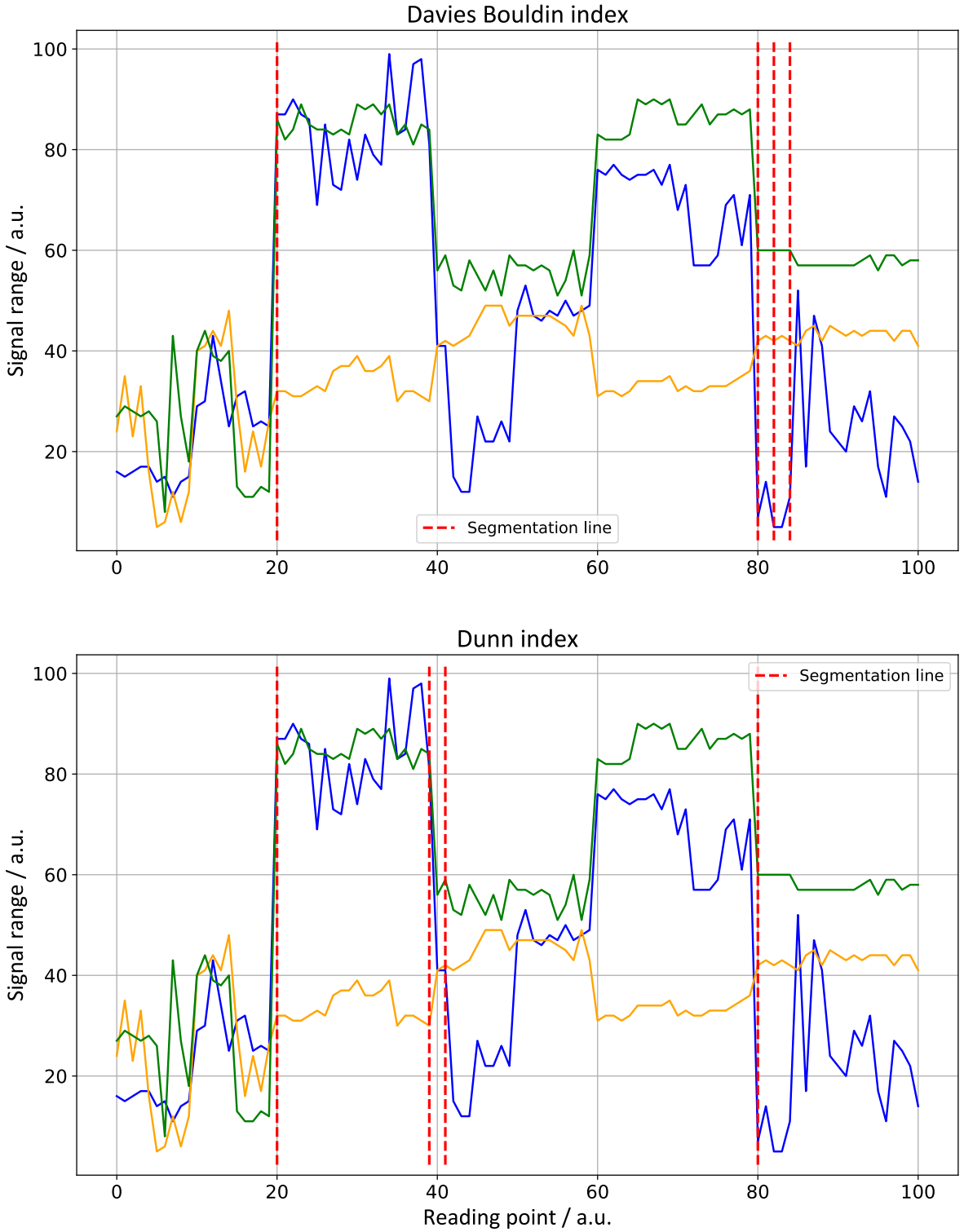

Figure 4.9: Segmentations with highest scores as per the metrics Davies Bouldin index and Dunn index.

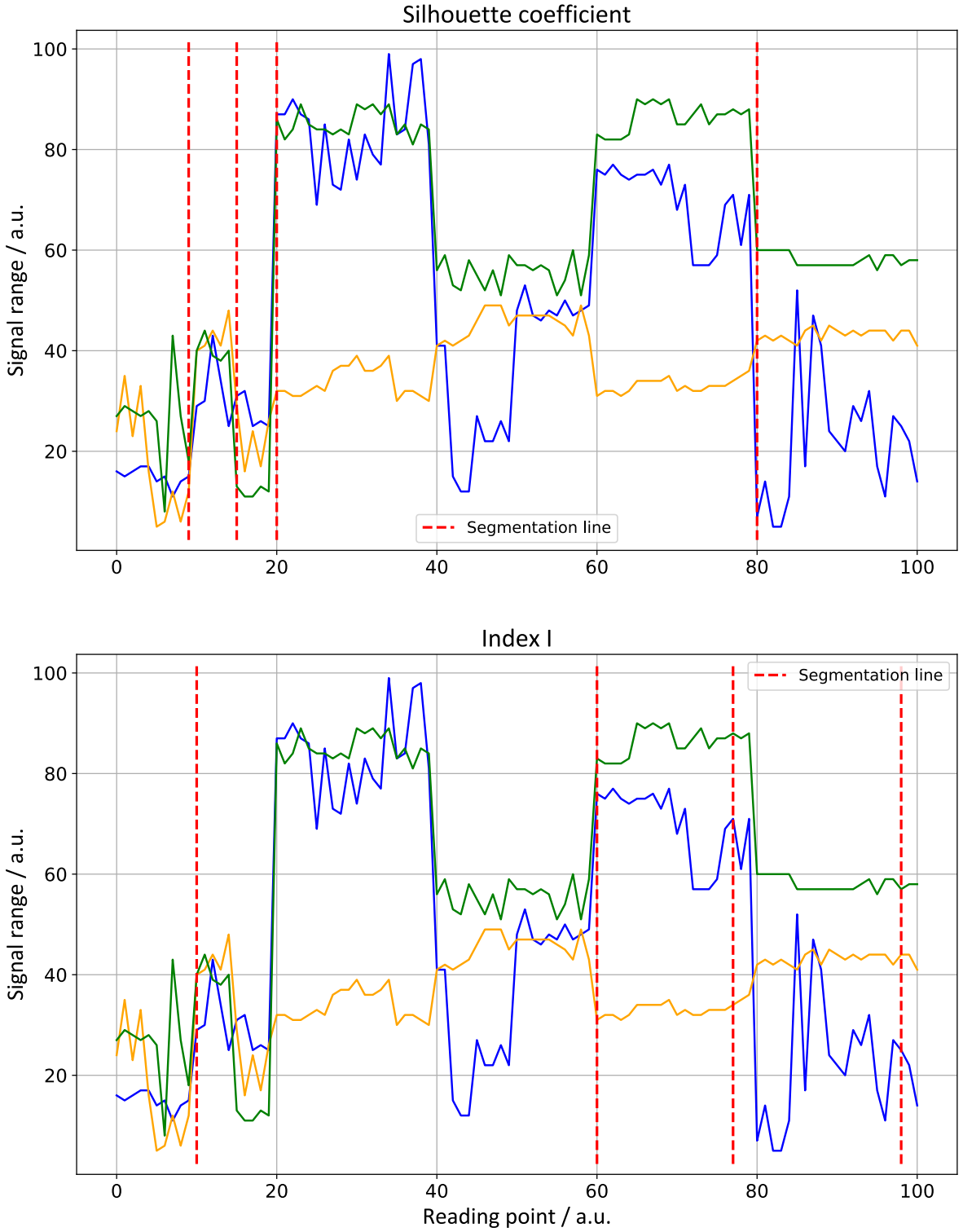

Figure 4.10: Segmentations with highest scores as per the metrics silhouette coefficient and index I.

4.4 Regression Analysis for Time Series Segmentation

Regression analysis, a set of statistical tools used for estimating relationships between variables in data sets, can be used for segmenting complex multivariate time series. In [44] and [45] the clustering and segmentation of multivariate time series with hidden process regression is demonstrated. In their work, the authors present a statistical model, which assumes that certain sequences in a multivariate data set are governed by sequences of hidden activities (i.e. hidden Markov chains). Moreover, an expectation-maximization (EM) algorithm with a log-likelihood function is used to iteratively fit a regression model to the data set and in turn find precise segments for it.

The most common form of regression analysis is linear regression. In this form of regression, a line is sought-after for fitting data according to some criterion. Figure 4.11 shows the linear regression of some test data with two variables.

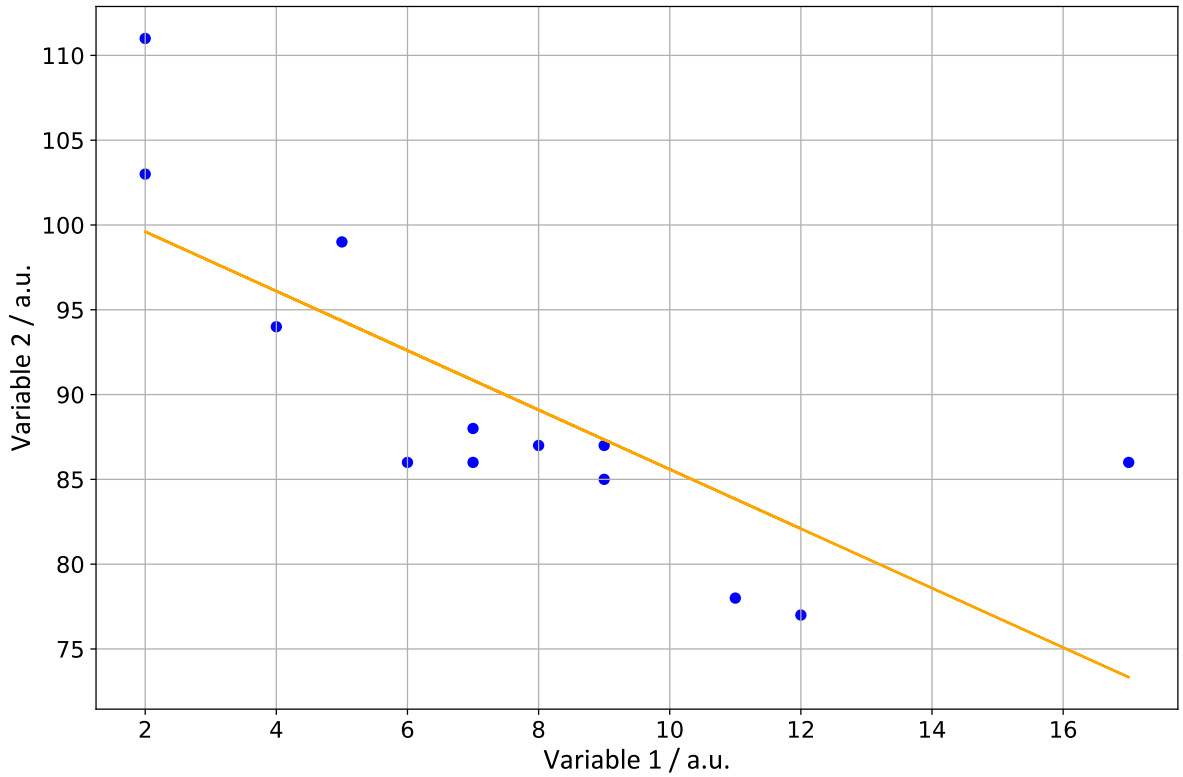

Figure 4.11: Linear regression of a test data set containing two variables.

The function (i.e. line) that in this case best captures the cluster of data (blue dots) is expressed with the orange line. Depending on the use case, polynomial regression can also be applied, which finds the best fitting polynomial function for a data set. More complex regression models

can fit multidimensional data sets with hyperplanes, however, for this work, an EM algorithm with a simple linear regression model is applied.

The detailed steps for multiple hidden markov model regression (MHMMR) for segmentation are demonstrated on a multivariate test signal in the next chapter.

5 Implementation of Multivariate Time Series Segmentation

In this chapter, selected tools for multivariate time series segmentation, shortly presented in the previous chapter, are applied. First, the multivariate test signal shown in figure 3.4 is segmented, before presenting, preprocessing, and segmenting the multivariate time series provided by the automotive partner. The two models used for multivariate time series segmentation, namely MHMMR and PCAFC, are chosen due to their advanced use of statistical methods for multivariate time series analysis and their high performance when it comes to multivariate time series segmentation. These models are demonstrated more detailed in the following sections.

5.1 Segmentation of a Test Signal

The first segmentation model used for segmenting the multivariate test signal as well as the multivariate time series provided by the automotive partner is a modification of the Gath-Geva clustering algorithm[1], called PCAFC segmenation. The PCAFC segmentation was first introduced by János Abonyi and Balázs Feil in their work [18] on cluster analysis for data mining. Their state-of-the-art segmentation model involves statistical methods for a bottom-up multivariate time series segmentation. The second segmentation model is the MHMMR, which is a Markov model based regression model for segmenting multivariate time series. Some of the tools used in these models were introduced in the previous chapter and are demonstrated in detail in the next subsections. The PCAFC as well as the MHMMR segmentation model are demonstrated and validated by first applying them to the multivariate test signal.

5.1.1 Principal Component Analysis Based Fuzzy Segmentation

First, the PCAFC segmentation model is used to segment the multivariate test signal. The process of segmenting a multivariate time series with the PCAFC can be explained in three parts: the feature extraction with PCA, soft segmentation with fuzzy clustering, and the bottom-up segmentation with the use of a cost function.

[1] See I. Gath and A.B. Geva. "Unsupervised optimal fuzzy clustering". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence 11.7.*

Feature Extraction with Principal Component Analysis

An essential part of the PCAFC segmentation algorithm is the selection of the right number of PCs for the segmentation of some multivariate time series. As mentioned in the previous chapter, PCA is used to extract the main features from data sets. Additionally, in this work, it is used to calculate the similarity between adjacent segments for the bottom-up segmentation, described in the upcoming subsection 5.1.1. The analysis of the eigenvalues of the time series shown in chapter 3 facilitates the selection of the right number of PCs for further analysis. With a scree plot (a line plot of the eigenvalues of principal components in a statistical analysis), as shown in 5.1, the eigenvalues of a data set ordered by their contribution to the variance of data can be plotted.

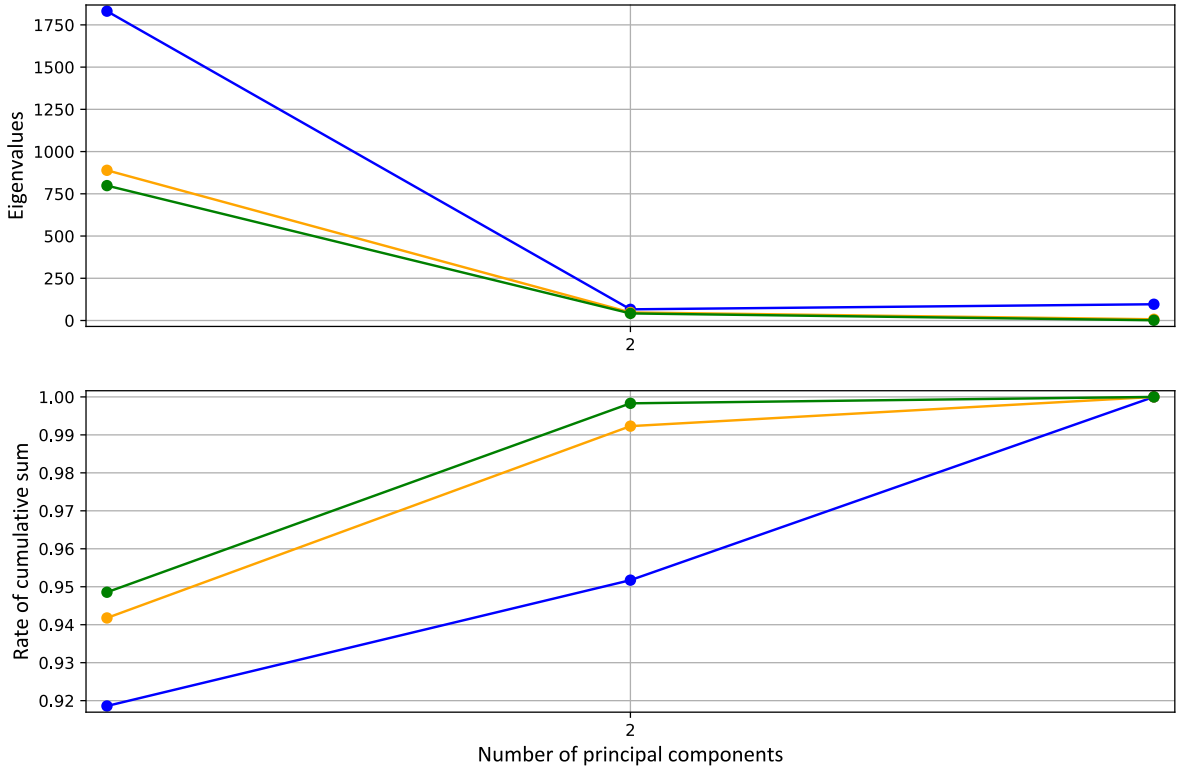

Figure 5.1: Scree plot for the principal component analysis of the multivariate test signal.

The higher the variance from the eigenvalues, the higher the captured information (i.e. distribution of the data) from the original data set is. Using all eigenvalues from one data set for further analysis (three in above case), the distribution of data in the data set can be explained with 100 % accuracy. As shown in this scree plot, two PCs are sufficient to capture the distribution of the data with roughly 95 % accuracy. Depending on the required accuracy, the number of PCs can be selected arbitrarily.

For the cost function used in the upcoming PCAFC segmentation of the multivariate test signal, two PCs are used.

Soft Segmentation with Fuzzy Clustering

Conventional segmentation algorithms assume crisp memberships of data points to a single segment, meaning that one data point in a data set can only belong to a single segment, as shown with $\beta_i(t_k) \in 0, 1$ in equation 5.1.

$$\beta_i(t_k) \in 0, 1 = \begin{cases} 1, & \text{if } s_{i-1} < k \leq s_i \\ 0, & \text{otherwise,} \end{cases} \tag{5.1}$$

with $\beta_i(t_k) \in 0, 1$ standing for the crisp membership of the kth data point to the ith segment.

As changes in the curve progression of time series are usually not extreme and rather continuous, it is not practical to define crisp bounds for segments of a time series. Therefore, in the PCAFC segmentation model the $\beta_i(t_k) \in 0, 1$ membership function is replaced with a Gaussian membership function. The algorithm assumes a normal (i.e. Gaussian) distribution of the data and minimizes the sum of weighted squared distances d between the z_k data points and η_i segments:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{i,k})^m d^2(z_k, \eta_i), \tag{5.2}$$

where $\mu_{i,k}$ is the degree of membership of z_k to the ith cluster and $m \in [1, \infty)$ is a weight for the fuzziness of the segments [18].

Cost Function for Bottom-up Segmentation

In order to cluster data points from a time series during the bottom-up segmentation, the internal homogeneity of segments needs to be measured for merging adjacent segments. In order to achieve this goal, a cost function can be defined, which is used during the bottom-up merging process of a segmentation procedure as shown in equation 5.3.

$$cost(S_T^c) = \sum_{i=1}^{c} cost(S_i). \tag{5.3}$$

This cost function, which can be any arbitrary function, measures the internal homogeneity of individual data points (or segments).

The cost function used for the PCAFC segmentation uses the combination of the similarity of PCA models and the distance of segment (i.e. cluster) centers. The similarity of two PCA models can be calculated with the sum of the squares of cosines of the angles between each

principal component q of the segments S:

$$S_{PCA}^{i,j} = \frac{1}{q} \sum_{i=1}^{q} \sum_{j=1}^{q} cos^2 \theta_{i,j}. \tag{5.4}$$

The distance d among segment centers v_i^x can be calculated as follows:

$$d(v_i^x, v_j^x) = \left\| v_i^x - v_j^x \right\|. \tag{5.5}$$

The compatibility criteria from equations 5.4 and 5.5 are calculated for each pair of clusters. This results in a compatibility matrix for the whole multivariate time series. Additionally, membership functions are defined for the compatibility criteria, as shown in figure 5.2.

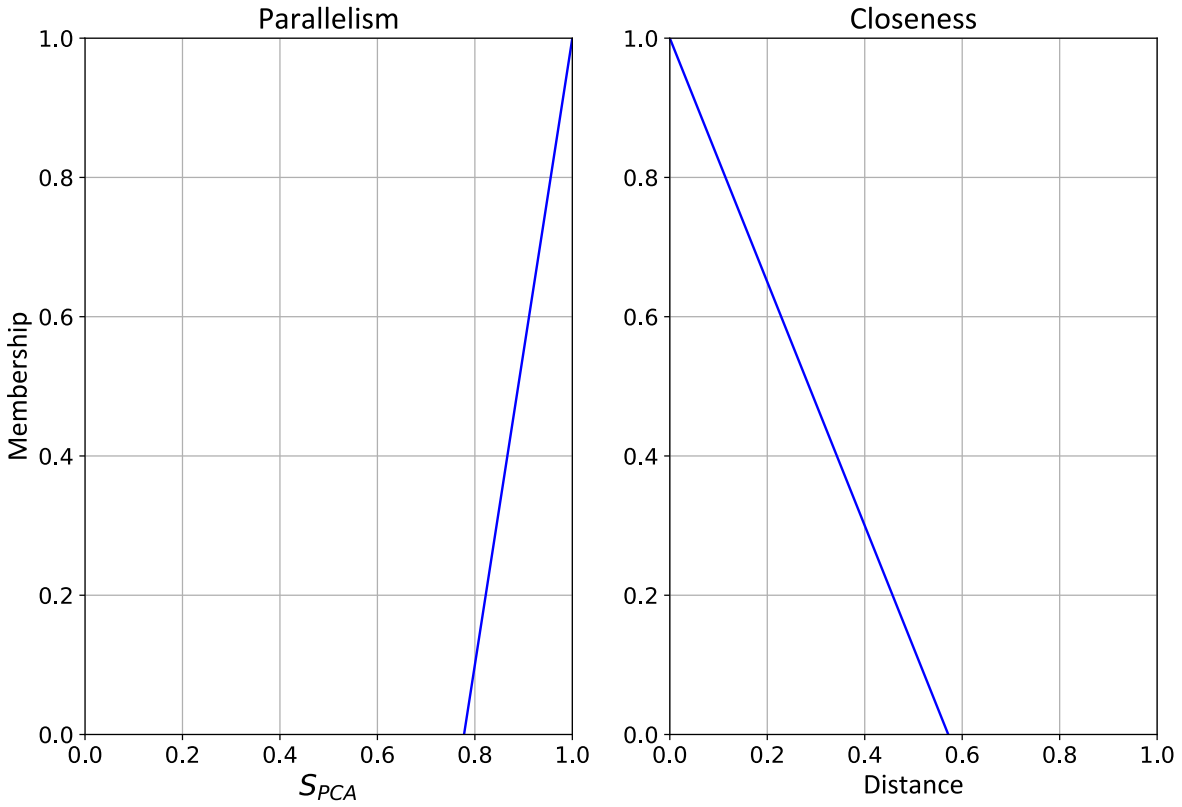

Figure 5.2: Membership functions for parallelism and closeness of segments.

The greater the similarity of PCA models between any two segments, the higher their membership is. The greater the distance between any two segment centers, the lower the membership of these segments. Applying the membership functions to the similarity described in equation 5.4 and distance in equation 5.5 results in the $\mu_{i,j}^1$ degree of parallelism and $\mu i, j^2$ degree of closeness. Finally, the overall compatibility matrix O for the whole multivariate time series is

calculated with both $\mu_{i,j}^1$ and $\mu_{i,j}^2$, as shown in the following equation:

$$cost(S_T^c) = O_{i,j} = \left[\frac{(\mu_{i,j}^1)^2 + (\mu_{i,j}^2)^2}{2}\right]^{1/2}. \tag{5.6}$$

In this model, only adjacent pairs of segments that are most similar to each other are merged, and only if the merged segment $O_{i,i+1}$ surpasses a certain threshold γ. This threshold can be chosen arbitrarily [18].[2]

Evaluation of the Principal Component Analysis Based Fuzzy Segmentation

After applying the PCAFC segmentation procedure to the multivariate test signal, the segmentation shown in figure 5.3 can be obtained.

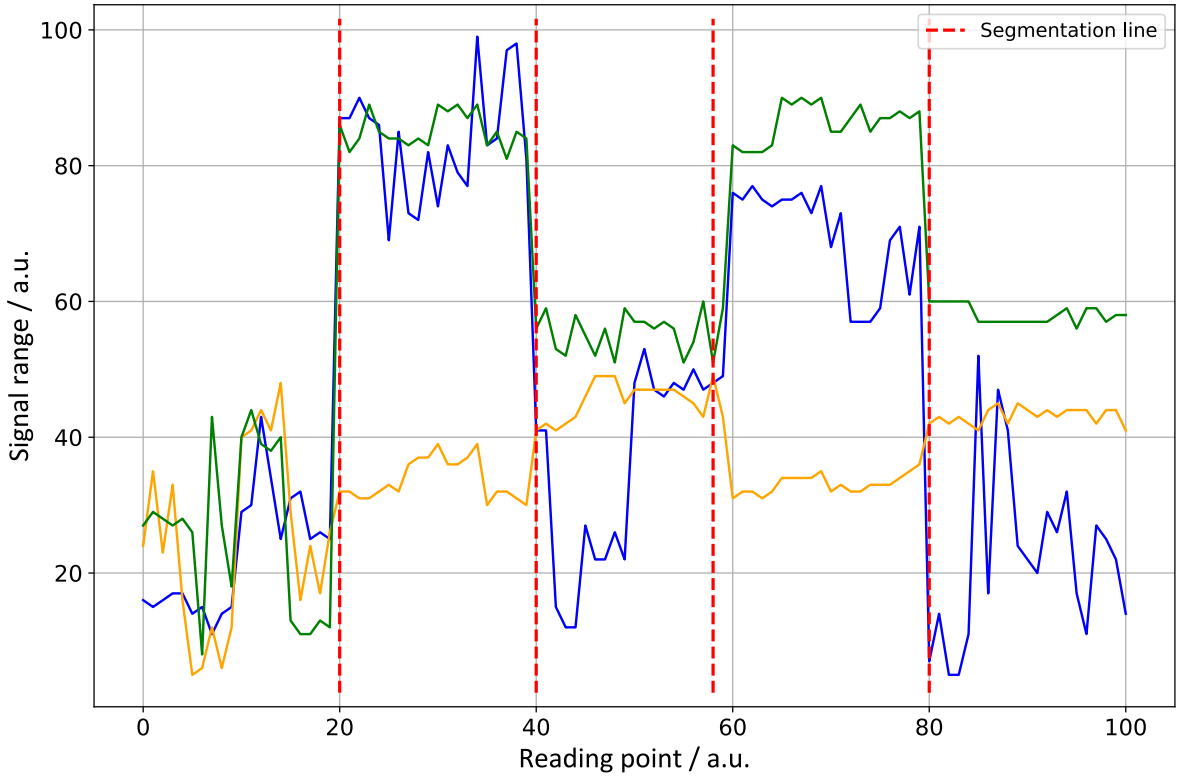

Figure 5.3: Principal component analysis based fuzzy segmentation of the multivariate test signal.

The segmentation bears a close resemblance to the optimal segmentation for the test signal defined in the previous chapter. In total, the program ran for approximately one minute before

[2]See "Segmentation of Multivariate Time-series". In: *Cluster Analysis for Data Mining and System Identification*, pp. 253–273, ISBN: 978-3-7643-7988-9 for supplementary material and a detailed explanation of the PCAFC segmentation algorithm.

outputting a result. Applying the benchmark homogeneity metric results in a homogeneity score of 0.30 for the PCAFC segmentation of the multivariate test signal. The relatively good homogeneity score as well as the short runtime for this segmentation method justifies the use of the model for (further) multivariate time series segmentations.

5.1.2 Multiple Hidden Markov Model Regression for Segmentation

As discussed in chapter 4, regression analysis is another statistical technique which can be used for multivariate time series segmentation. In [44] Chamroukhi et al. present a statistical model for segmenting multivariate data sets with an EM algorithm and regression analysis. The authors assume that certain sequences in a multivariate data set are governed by sequences of hidden activities (i.e. hidden Markov chains). In the following subsections, the steps for performing a multivariate time series segmentation with MHMMR are demonstrated.

Multiple Hidden Markov Regression Model Definition

The MHMMR model assumes that time series can be issued from a predefined number of segments. Moreover, within each segment, the time series that makes up the segment is generated by some polynomial regime. The transition from one polynomial regime (i.e. segment) to another is controlled by a homogeneous Markov chain. Each segment in the time series can be described by a polynomial regression model, in this case a linear regression model [44, p. 2].

Maximization of Log-likelihood with Expectation-Maximization Algorithm

The parameter estimation of the MHMMR model is performed by maximizing the log-likelihood of the model parameter. The maximum likelihood estimation is a statistical method for estimating the parameters of a model describing some observed data. As the maximization of the log-likelihood with the MHMMR model cannot be performed in a closed form, an EM algorithm is used.

The simplified pseudo code of the complete MHMMR algorithm is denoted as follows:

Inputs: time series Y, sampling time T, number of polynomial components K (i.e. segments), polynomial degree p.

```
Initialize MHMMR model parameter vector ψ,
define threshold ϵ > 0, set q ← 0 (EM iterations)
while increment in log-likelihood > ϵ do
  # E-step:
  for k = 1, ..., K do
    compute τ_{i,k} for i = 1, ..., n
  end for
  # M-step:
  for k = 1, ..., K do
    compute α_k
  end for
  q ← q + 1
end while
```

Output: MHMMR model parameter vector ψ

Table 5.1: High-level description of the MHMMR algorithm

with $\tau_{i,k}$ being the posterior probability that y_i originates from the polynomial regression model that describes the kth activity and α_k being the mixing proportions of the cluster (i.e. the cluster sizes) summing up to 1.[3]

In the next subsection, the algorithm is applied to the multivariate test signal and the segmentation is evaluated using the predefined homogeneity metric for segmentation.

Evaluation of the Multiple Hidden Markov Model Regression Segmentation

After applying above algorithm to the multivariate test signal, the segmentation shown in figure 5.4 can be obtained.

The segmentation generated by the MHMMR model is identical to the optimal segmentation of the multivariate test signal. Applying the benchmark homogeneity metric to this segmentation results in a score of 0.08. The exceptionally good homogeneity score, as well as the output being the optimal segmentation for the multivariate time series justifies the use of the MHMMR segmentation model for (further) multivariate time series segmentations.

[3]See "Model-based clustering with Hidden Markov Model regression for time series with regime changes". In: *The 2011 International Joint Conference on Neural Networks (2011)*, pp. 2814-2821 for supplementary material and a detailed explanation of the MHMMR segmentation algorithm.

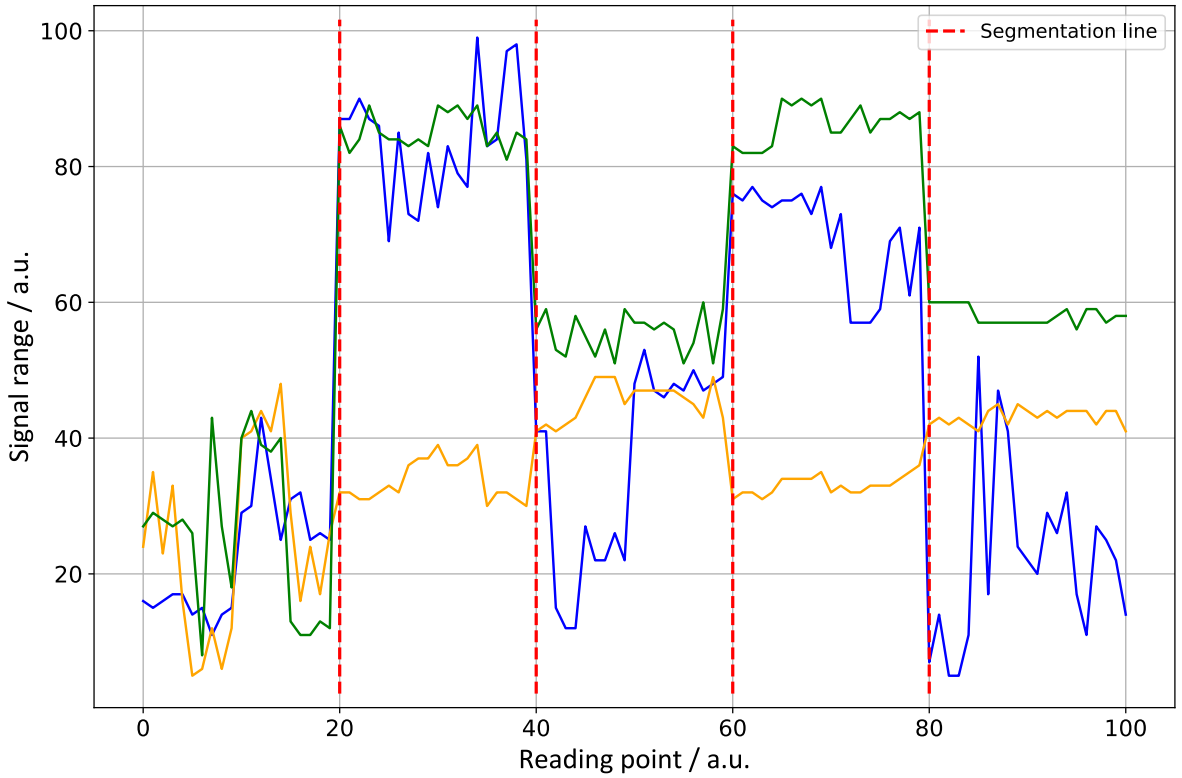

Figure 5.4: Multiple hidden Markov model regression segmentation of the multivariate test signal.

In the next section, the multivariate time series provided by the automotive partner is first presented, before applying the two segmentation models to it.

5.2 Preparation and Segmentation of Multivariate Time Series from a Software System

The multivariate time series provided by the automotive partner consists of 939 variables and 402 data points per variable. Each variable shows a signal (i.e. process running in the software system) consisting of runtime values. Figure 5.5 shows all processes in one plot.

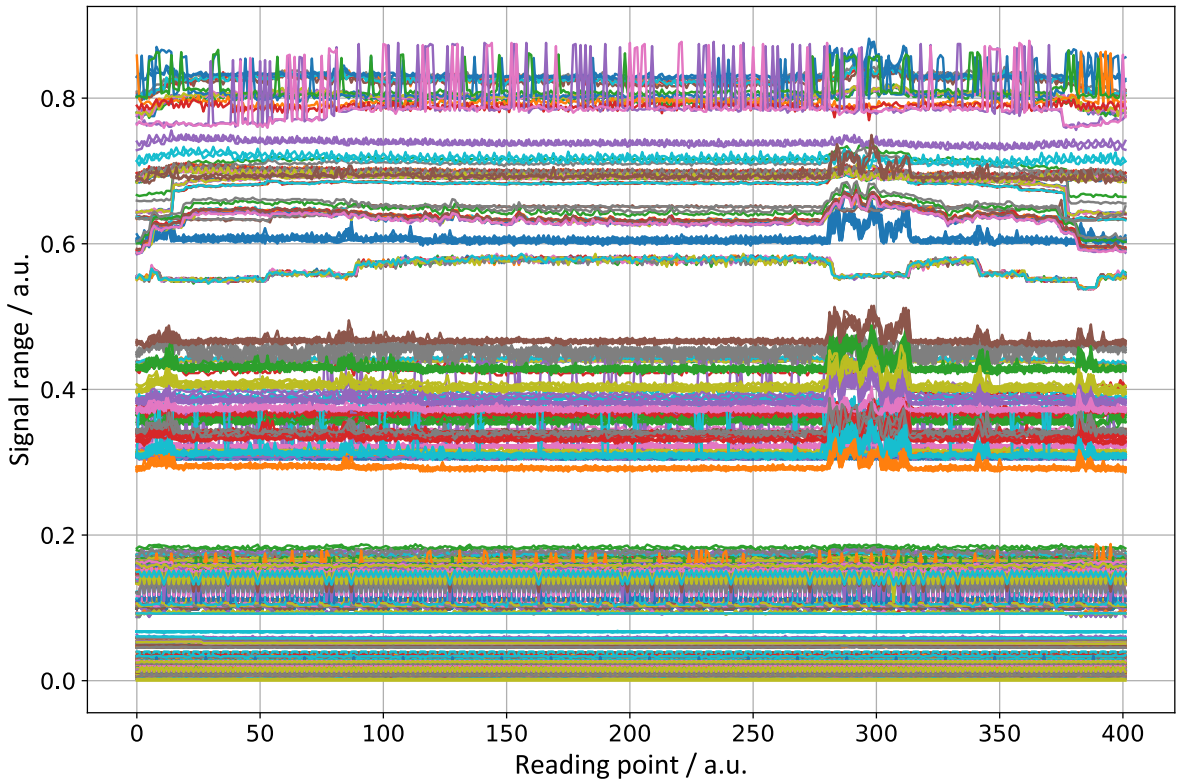

Figure 5.5: Plot of all processes contained in the multivariate time series.

An illustration of one process can be seen in figure 5.6. The data set contains a high number of processes, which could be difficult to analyze collectively. Additionally, many of these processes are not showing much information or contain an excessive amount of noise. Therefore, the data set needs to be preprocessed before applying discussed models for multivariate time series segmentation. In the next subsections, the methods for filtering the most significant processes from the complete data set are presented.

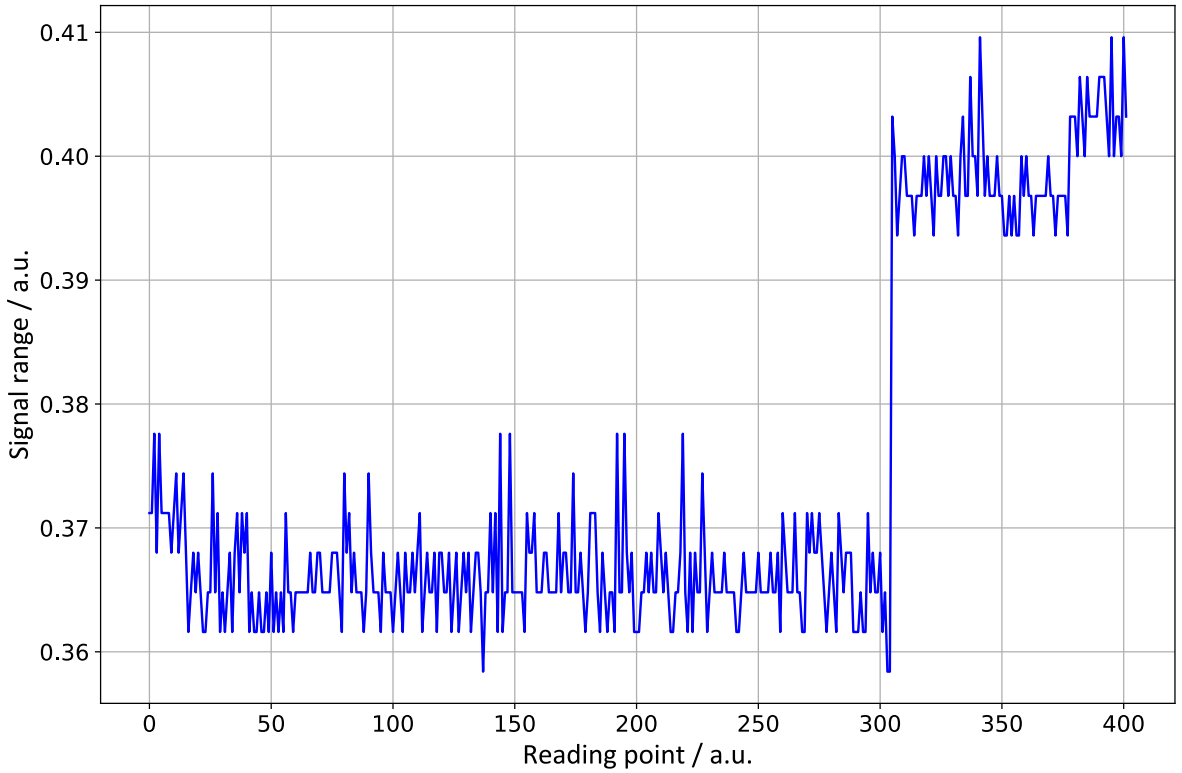

Figure 5.6: Sample from the set of processes contained in the multivariate time series.

5.2.1 Data Preparation

For better visualization and easier application of methods for multivariate time series segmentation, the data set is filtered by means of transformation and selection procedures. The transformation works by subtracting the minimum value of each process from each data point of respective process (offset reduction) and subsequently selecting the n processes with the highest peak-to-valley ratio. Figure 5.7 shows the remaining process after applying said offset reduction.

More formally, the offset reduction subtracts a constant c from each data point of a process p_n. The constant c_{p_n} is equal to the minimum value of process p_n and is subtracted from each data point of p_n. This can be expressed as follows:

$$\begin{aligned} c_{p_n} &= min(p_n) \\ {p_n}^{new} &= p_n - c_{p_n} \end{aligned} \tag{5.7}$$

After applying the offset reduction to the data set, the minimum values of all processes are at y = 0. Figure 5.7 shows the result of the transformation.

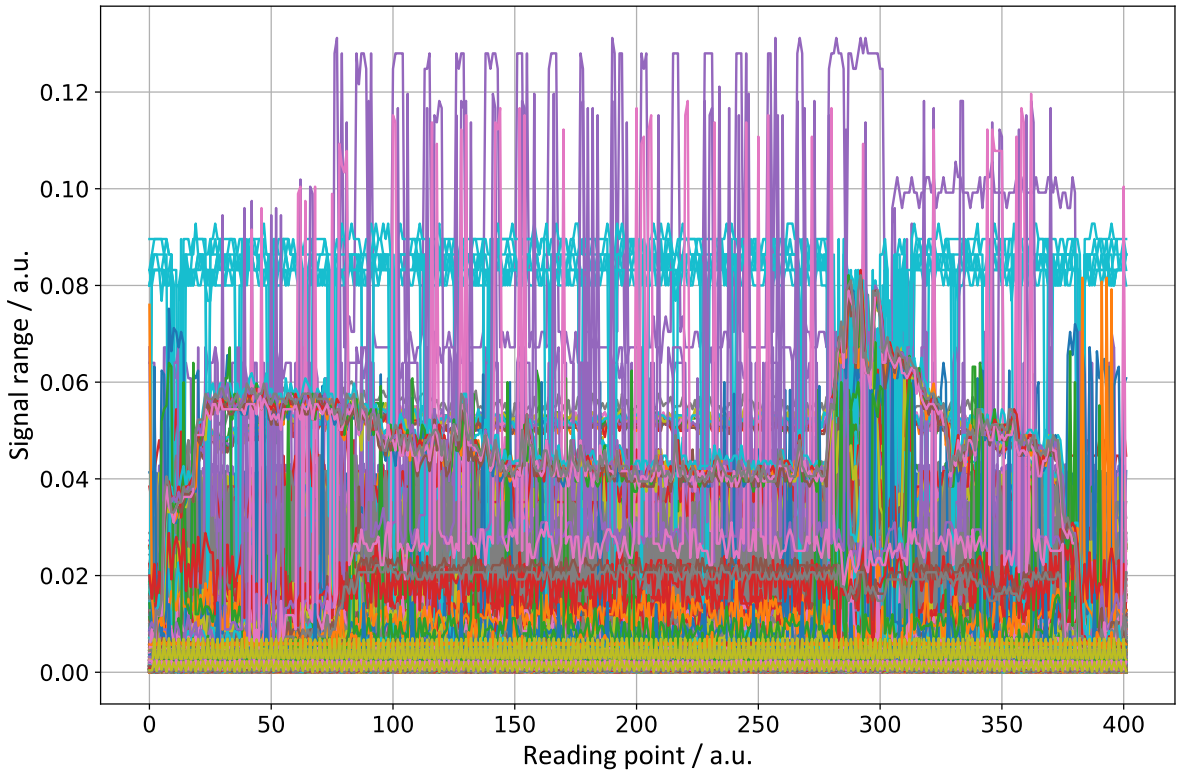

Figure 5.7: Plot of all processes transformed through offset reduction.

Finally, the n processes with the highest peak-to-valley ratio can be selected. The peak-to-valley ratio is the ratio of the maximum value to the minimum value in a data set (i.e. one signal). The offset reduction described above is necessary before selecting the n processes with the highest peak-to-valley ratio in order to get a relative ratio for all processes. In this case, from the 939 processes that were transformed by offset reduction, the processes with a ratio higher than 60 % of the highest peak-to-valley ratio are selected and kept for further analysis. After applying this filtering method, the previously 939 processes are reduced to 23 processes. These 23 processes are then returned to their original signal range to preserve their original features. The final result after filtering the data set can be observed in figure 5.8.

The final data set shows the most significant processes contained in the original data set. The intuition is that the maximum and minimum value of these processes are further apart than those of the signals that were filtered. Therefore, the selected processes show a more distinguished behavior.

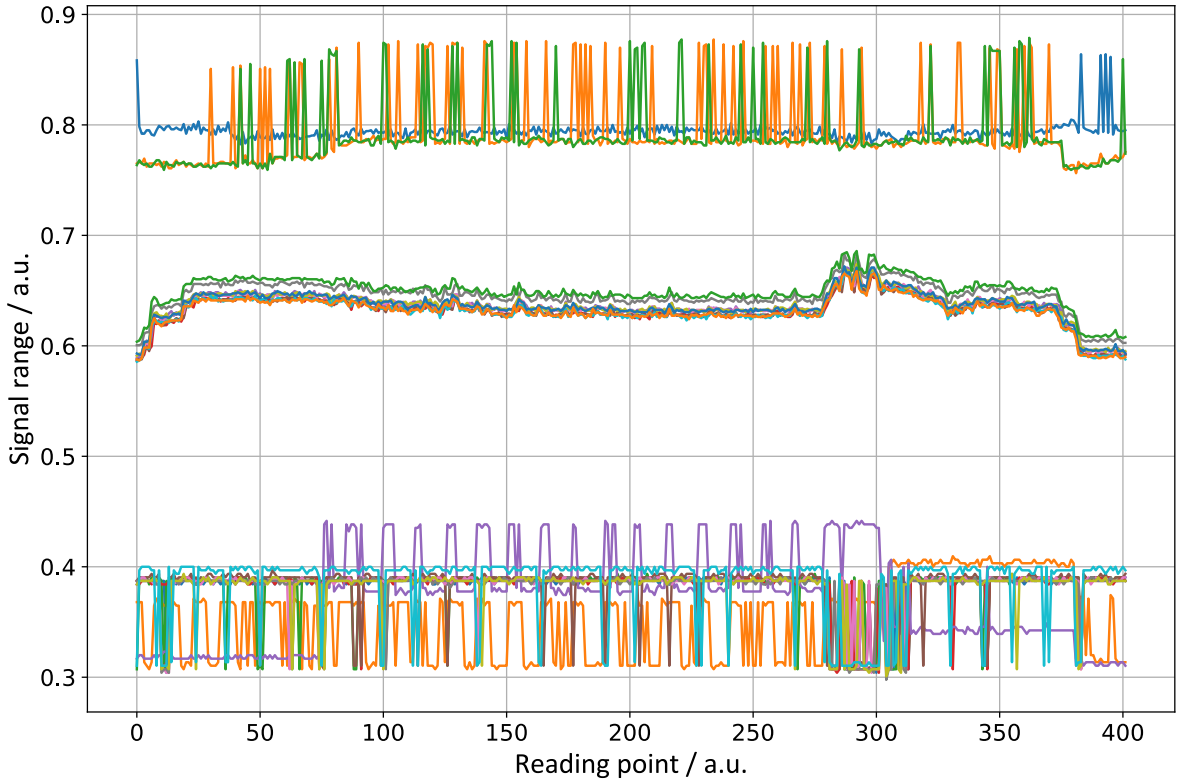

Figure 5.8: Top 23 processes by peak-to-valley-ratio.

In the next sections, the filtered multivariate time series provided by the automotive partner is segmented using the two segmentation models, before evaluating the segmentations with the homogeneity metric.

5.2.2 Principal Component Analysis Based Fuzzy Segmentation

The same steps for the PCAFC segmentation that were performed on the multivariate test signal are applied to the filtered multivariate time series provided by the automotive partner. The PCA of the data set shows that three PCs are enough to capture roughly 85 % of variance in the data set. Therefore, three PCs are used for further analysis. Furthermore, due to unavailable a priori knowledge about the optimal amount of segments for the multivariate time series, segmentations with four to fourteen segments are tested and evaluated with the homogeneity metric. The figures A1, A2, A3, A4, A5, and A6 in the appendix show the final segmentations for the filtered processes using the PCAFC segmentation. An exemplary PCAFC segmentation of the multivariate time series can be seen in figure 5.9.

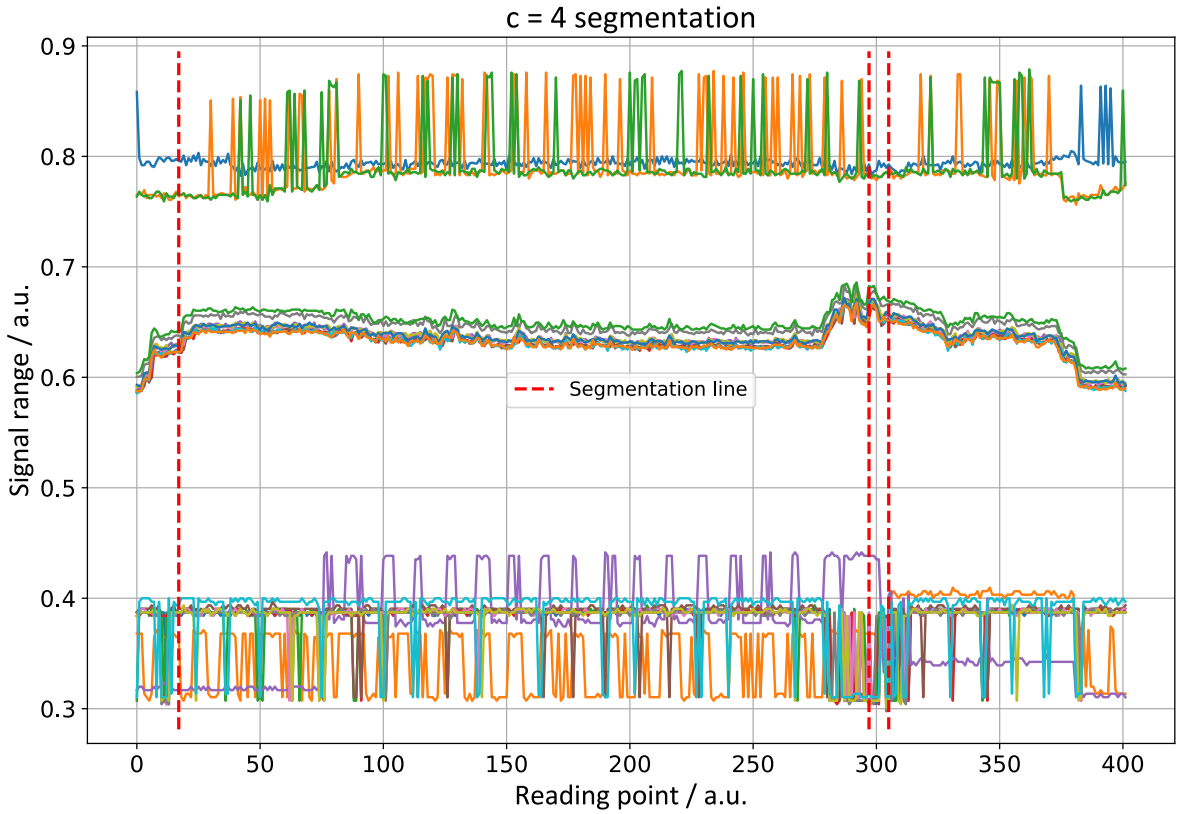

Figure 5.9: c = 4 PCAFC segmentation of the filtered multivariate time series.

5.2.3 Multiple Hidden Markov Model Regression for Segmentation

After applying the MHMMR algorithm to the filtered multivariate multivariate time series, the segmentations shown in figures A7, A7, A9, A10, A11, and A12 in the appendix can be obtained. An exemplary MHMMR segmentation of the multivariate time series can be seen in figure 5.10.

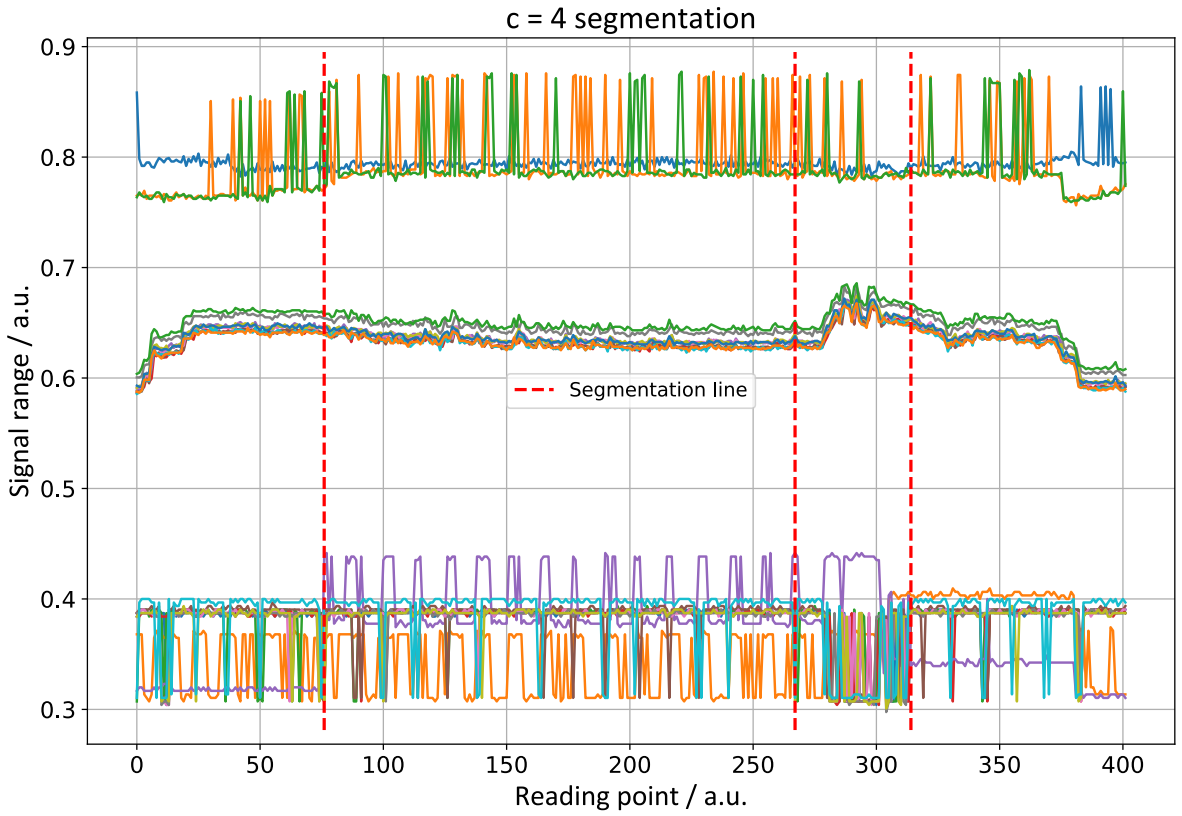

Figure 5.10: c = 4 MHMMR segmentation of the filtered multivariate time series.

In the next chapter, the results for the segmentations of the filtered multivariate time series performed with the PCAFC and the MHMMR model are presented and evaluated.

6 Results

After applying the PCAFC and MHMMR segmentation to the filtered multivariate time series, the resulting segmentations can be evaluated with the predefined homogeneity metric. Table 6.1 shows the benchmark results of the segmentations with four to fourteen segments for each segmentation model. The best scores for each metric are marked in green. As mentioned previously, the lower the homogeneity score the better the underlying segmentation is.

# of segments	Segmentation algorithm	
	PCAFC	MHMMR
Four	0.070	0.063
Five	0.071	**0.060**
Six	0.072	0.062
Seven	0.067	0.060
Eight	0.066	0.062
Nine	**0.064**	0.061
Ten	0.065	0.062
Eleven	0.065	0.061
Twelve	0.066	0.063
Thirteen	0.066	0.065
Fourteen	0.067	0.067

Table 6.1: Evaluation of the PCAFC and MHMMR segmentation quality for the different c-segmentations of the filtered multivariate time series.

The metric values are plotted and presented in figure 6.1. For the PCAFC segmentation model, the segmentation with the best homogeneity score is the one with nine segments, the segmentation with the worst score is the one with six segments. The scores are 0.064 and 0.072 respectively. The segmentation performed with the MHMMR model shows the best homogeneity score at five segments and the worst at fourteen segments with values of 0.060 and 0.067. For all numbers of segments except fourteen, the MHMMR segmentation model performs better than the PCAFC with the segmentations showing a higher homogeneity. The MHMMR model shows a trend towards a worse homogeneity metric with a higher number of segments.

The PCAFC model has relatively large homogeneity values for lower number of segments but also shows a slight trend towards larger homogeneity scores with an increase in the number of segments.

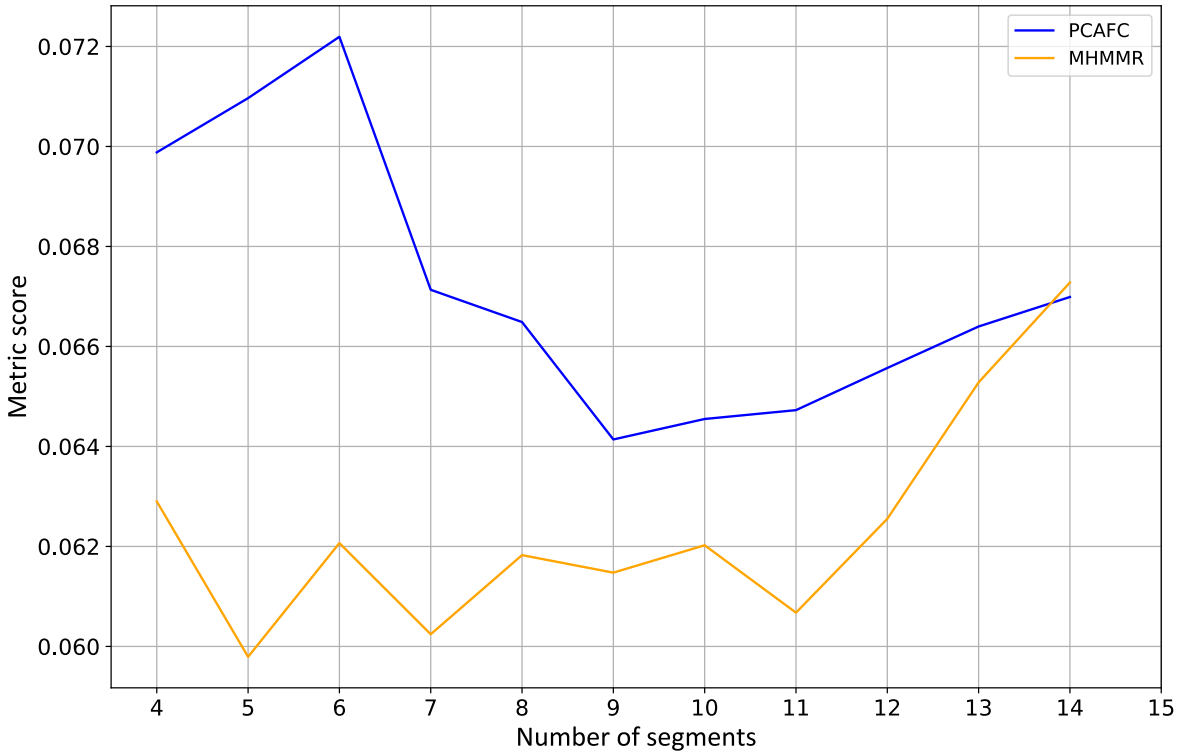

Figure 6.1: Normalized metric values for the different c-segmentations.

In the next chapter, the implications of the results in the context of multivariate time series segmentation are discussed.

7 Discussion

In general, the homogeneity scores for the two segmentation models shown in chapter 6 are relatively small throughout all numbers of segments. This indicates a high capability of the segmentation models for dividing multivariate time series into homogeneous segments. Inspecting the different c-segmentations in detail, the segments found by the models are, from an intuitive standpoint, relatively precise. Especially the MHMMR segmentation with four segments depicted in figure A12 shows segments that hold parts of the multivariate time series with unique and separate behavior. The third segment, for instance, holds a part of the signals which are in the signal range between 0.3 and 0.7. These signals show a relatively high fluctuation and some behavior, which is not present in the adjacent segments two and four.

The results show that the PCAFC and MHMMR segmentation models have a high capability for dividing a multivariate time series into relevant and homogeneous segments. With the newly developed homogeneity metric, the results obtained from a segmentation model can be measured and the segmentation with the highest homogeneity from a set of segmentations can be selected. Previously, it was difficult – if not impossible – to quantify the quality of a multivariate time series segmentation. However, as shown by the results of this work, with the homogeneity metric it is now possible to measure the quality of a multivariate time series segmentation efficiently and effectively.

The limitation taken from the results is the interpretability of the homogeneity scores. Without knowledge about the homogeneity scores for each number of segments for the segmentation of a time series, it is not apparent if the measured homogeneity score for a single segmentation is high or low. Hence, results can only be interpreted in relative terms. Therefore, the ideal number of segments for the segmentation of a multivariate time series can only be determined through an exhaustive search using the homogeneity metric. Due to the aim of this work not being the identification of the ideal number of segments in a multivariate time series, this topic should be part of future research. A limitation of the segmentation models is that they generally give approximate solutions for a multivariate time series segmentation. Because not all combinations of segmentation indices for a predefined number of segments are tested for an arbitrary multivariate time series, optimums that are found with the models and the homogeneity metric can only be seen as local (i.e. relative) optimums, not global (i.e. absolute) ones.

The questions of how homogeneity in segments of a multivariate time series can be defined and measured, and how precise segmentation indices for maximizing homogeneity in segments can be determined have been answered. Finding precise segmentation indices for the segmenta-

tion of a multivariate time series requires adequate statistical methods and techniques, such as the PCA based bottom-up segmentation. As shown in this work, the PCAFC and MHMMR segmentation models are capable for segmenting multivariate time series and have, due to internal homogeneity metrics used during segmentation, the ability to define precise segmentation indices for maximizing the homogeneity of a multivariate time series segmentation. The external homogeneity metric defined in this work quantifies the quality of a multivariate time series segmentation and presents a solution to the problem of measuring the homogeneity in segments of a multivariate time series.

In the next chapter, this work is concluded and a future outlook for the research area of multivariate time series segmentation is presented.

8 Conclusion and Outlook

This work was motivated by the problem of state extraction from multivariate time series by dividing the main problem into the subproblems of defining and measuring the homogeneity in segments, determining the ideal number of segments, and finding the precise segmentation indices for the segmentation of a multivariate time series.

Two segmentation models, namely PCAFC and MHMMR, that give an approximate solution to the subproblem of finding suitable segmentation indices with a given amount of segments for a multivariate time series, were presented. The capabilities of the segmentation models were shown by applying them first to a multivariate test signal and then to a multivariate time series taken out of industry data from an automotive partner. The chosen algorithms are capable of finding relatively good segmentation indices even for complex multivariate time series. The results obtained from the segmentation models were then measured with a newly developed homogeneity metric, scoring segmentations with different numbers of segments, and ultimately giving a solution to the subproblem of defining and measuring the homogeneity in segments after a multivariate time series segmentation. The subproblem of finding an ideal amount of segments for a multivariate time series segmentation remains open and will be investigated in the future. Also, the overall complexity of the problem has not been analyzed yet and is a topic, which should be considered in future research.

The tools for multivariate time series analysis, that were shortly presented in chapter 4 and not pursued further in this work, show some potential for use in multivariate time series segmentation algorithms. These tools should be investigated for their practicability and efficiency by comparing them to tools used in state-of-the-art multivariate time series segmentation algorithms. A big potential exists for the modification of state-oft-the-art time series segmentation models with the addition of the newly developed homogeneity metric used in this work. Using this external homogeneity metric as an internal homogeneity metric in algorithms performing segmentation could show fruitful results in finding segmentations with relatively high homogeneity. The prerequisite, however, is an efficient and effective implementation, as an exhaustive search with this metric is time consuming even with relatively small and simple multivariate time series.

A complementing field of research, which could be part of future work in the domain of multivariate time series segmentation is the classification of segments. As the aim of a time series segmentation is to divide a time series into unique parts containing single states, the natural next step is to classify these states (i.e. segments) and group them into buckets (i.e. classes). The measures used in the PCAFC segmentation model for quantifying the similarities

between segments during the bottom-up merge of adjacent segments could be considered for the classification of segments. Other models, such as deep neural networks, are already in use for time series classification and should be considered in future works.

References

[1] Andrew T. Jebb and Louis Tay. "Introduction to Time Series Analysis for Organizational Research: Methods for Longitudinal Analyses". In: *Organizational Research Methods* 20.1 (2017), pp. 61–94. DOI: 10.1177/1094428116668035. eprint: https://doi.org/10.1177/1094428116668035. URL: https://doi.org/10.1177/1094428116668035.

[2] Charu C. Aggarwal. *Data Mining: The textbook*. Springer, 2015.

[3] "Segmentation of Multivariate Time-series". In: *Cluster Analysis for Data Mining and System Identification*. Basel: Birkhäuser Basel, 2007, pp. 253–273. ISBN: 978-3-7643-7988-9. DOI: 10.1007/978-3-7643-7988-9_6. URL: https://doi.org/10.1007/978-3-7643-7988-9_6.

[4] Janos Abonyi et al. "Fuzzy Clustering Based Segmentation of Time-Series". In: *Advances in Intelligent Data Analysis V*. Ed. by Michael R. Berthold et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 275–285. ISBN: 978-3-540-45231-7.

[5] Janos Abonyi et al. "Modified Gath–Geva clustering for fuzzy segmentation of multivariate time-series". In: *Fuzzy Sets and Systems* 149.1 (2005). Fuzzy Sets in Knowledge Discovery, pp. 39–56. ISSN: 0165-0114. DOI: https://doi.org/10.1016/j.fss.2004.07.008. URL: https://www.sciencedirect.com/science/article/pii/S0165011404003069.

[6] Ashish Singhal and Dale E. Seborg. "Clustering multivariate time-series data". In: *Journal of Chemometrics* 19.8 (2005), pp. 427–438. DOI: https://doi.org/10.1002/cem.945. eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cem.945. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.945.

[7] Shaowen Lu and Shuyu Huang. "Segmentation of Multivariate Industrial Time Series Data Based on Dynamic Latent Variable Predictability". In: *IEEE Access* 8 (2020), pp. 112092–112103. DOI: 10.1109/ACCESS.2020.3002257.

[8] Laszlo Dobos and János Abonyi. “On-line detection of homogeneous operation ranges by dynamic principal component analysis based time-series segmentation”. In: *Chemical Engineering Science* 75 (June 2012), pp. 96–105. DOI: 10.1016/j.ces.2012.02.022.

[9] Dazhuo Zhou, JinXia Li and WenXiu Ma. “Clustering Based on LLE For Financial Multivariate Time Series”. In: *2009 International Conference on Management and Service Science*. 2009, pp. 1–4. DOI: 10.1109/ICMSS.2009.5305089.

[10] Röpke Karsten et al. “Data Analysis I”. In: *International conference on calibration methods and Automotive Data Analytics*. expert, 2019, pp. 1–21.

[11] Y.L. Murphey et al. “Automotive fault diagnosis - part II: a distributed agent diagnostic system”. In: *IEEE Transactions on Vehicular Technology* 52.4 (2003), pp. 1076–1098. DOI: 10.1109/TVT.2003.814236.

[12] David Guijo-Rubio et al. “Time-Series Clustering Based on the Characterization of Segment Typologies”. In: *IEEE Transactions on Cybernetics* PP (Jan. 2020), pp. 1–14. DOI: 10.1109/TCYB.2019.2962584.

[13] Theodoros Giannakopoulos and Aggelos Pikrakis. “Audio Segmentation”. In: *Introduction to audio analysis a MATLAB approach*. 1st ed. Academic Pr, 2014, pp. 154–177.

[14] Hui Zhang et al. “Unsupervised Feature Extraction for Time Series Clustering Using Orthogonal Wavelet Transform.” In: *Informatica (Slovenia)* 30 (Oct. 2006), pp. 305–319.

[15] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi and Teh Ying Wah. “Time-series clustering – A decade review”. In: *Information Systems* 53 (2015), pp. 16–38. ISSN: 0306-4379. DOI: https://doi.org/10.1016/j.is.2015.04.007. URL: https://www.sciencedirect.com/science/article/pii/S0306437915000733.

[16] Ali Javed, Byung Suk Lee and Donna M. Rizzo. “A benchmark study on time series clustering”. In: *Machine Learning with Applications* 1 (2020), p. 100001. ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2020.100001. URL: https://www.sciencedirect.com/science/article/pii/S2666827020300013.

[17] Andrew Rosenberg and Julia Hirschberg. “V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure”. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic:

Association for Computational Linguistics, June 2007, pp. 410–420. URL: https://aclanthology.org/D07-1043.

[18] János Abonyi and Balázs Feil. *Cluster Analysis for Data Mining and System Identification*. Jan. 2007. ISBN: 376437988X, 9783764379889. DOI: 10.1007/978-3-7643-7988-9.

[19] HAROLD HOTELLING. "RELATIONS BETWEEN TWO SETS OF VARIATES*". In: *Biometrika* 28.3-4 (Dec. 1936), pp. 321–377. ISSN: 0006-3444. DOI: 10.1093/biomet/28.3-4.321. eprint: https://academic.oup.com/biomet/article-pdf/28/3-4/321/586830/28-3-4-321.pdf. URL: https://doi.org/10.1093/biomet/28.3-4.321.

[20] Kanghua Hui and Chunheng Wang. "Clustering-based locally linear embedding". In: *2008 19th International Conference on Pattern Recognition*. 2008, pp. 1–4. DOI: 10.1109/ICPR.2008.4761293.

[21] Lawrence Saul and Sam Roweis. "An introduction to locally linear embedding". In: *Journal of Machine Learning Research* 7 (Jan. 2001).

[22] Lori Ziegelmeier, Michael Kirby and Chris Peterson. "Locally Linear Embedding Clustering Algorithm for Natural Imagery". In: (Feb. 2012).

[23] Jelena Zubova, Olga Kurasova and Marius Liutvinavičius. "Dimensionality reduction methods: The comparison of speed and accuracy". In: *Information Technology And Control* 47 (Mar. 2018). DOI: 10.5755/j01.itc.47.1.18813.

[24] Nitish Srivastava, Elman Mansimov and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs". In: *CoRR* abs/1502.04681 (2015). arXiv: 1502.04681. URL: http://arxiv.org/abs/1502.04681.

[25] Kyunghyun Cho et al. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: https://aclanthology.org/D14-1179.

[26] Si Lu. *DAC: Deep Autoencoder-based Clustering, a General Deep Learning Framework of Representation Learning*. Feb. 2021.

[27] Chunfeng Song et al. “Auto-encoder Based Data Clustering”. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Ed. by José Ruiz-Shulcloper and Gabriella Sanniti di Baja. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 117–124. ISBN: 978-3-642-41822-8.

[28] Siquan Yu et al. “Representation learning based on Autoencoder and deep adaptive clustering for image clustering”. In: *Mathematical Problems in Engineering* 2021 (2021), pp. 1–11. DOI: 10.1155/2021/3742536.

[29] Quentin Fournier and Daniel Aloise. “Empirical Comparison between Autoencoders and Traditional Dimensionality Reduction Methods”. In: *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. 2019, pp. 211–214. DOI: 10.1109/AIKE.2019.00044.

[30] G. Stephanopoulos and C. Han. “Intelligent systems in process engineering: A review”. English (US). In: *Computers and Chemical Engineering* 20.6-7 (1996), pp. 743–791. ISSN: 0098-1354. DOI: 10.1016/0098-1354(95)00194-8.

[31] E. Keogh et al. “An online algorithm for segmenting time series”. In: *Proceedings 2001 IEEE International Conference on Data Mining*. 2001, pp. 289–296. DOI: 10.1109/ICDM.2001.989531.

[32] Miodrag Lovric, Marina Milanovic and Milan Stamenkovic. “Algorithmic methods for segmentation of time series: An overview”. In: *Journal of Contemporary Economic and Business Issues (JCEBI)* 1 (Jan. 2014), pp. 31–53.

[33] Li Zeng et al. “SegTime: Precise Time Series Segmentation without Sliding Window”. In: *International Conference on Learning Representations* (2021).

[34] T. Warren Liao. “Clustering of time series data—a survey”. In: *Pattern Recognition* 38.11 (2005), pp. 1857–1874. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2005.01.025. URL: https://www.sciencedirect.com/science/article/pii/S0031320305001305.

[35] Huynh Thi Thu Thuy, Duong Tuan Anh and Vo Thi Ngoc Chau. “Comparing three time series segmentation methods via novel evaluation criteria”. In: *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. 2017, pp. 171–176. DOI: 10.1109/ICITISEE.2017.8285489.

[36] Lawrence J. Hubert and Phipps Arabie. “Comparing partitions”. In: *Journal of Classification* 2 (1985), pp. 193–218.

[37] Simone Romano et al. “Adjusting for Chance Clustering Comparison Measures”. In: *J. Mach. Learn. Res.* 17 (2016), 134:1–134:32.

[38] E. B. Fowlkes and C. L. Mallows. “A Method for Comparing Two Hierarchical Clusterings”. In: *Journal of the American Statistical Association* 78.383 (1983), pp. 553–569. DOI: 10.1080/01621459.1983.10478008. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1983.10478008. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10478008.

[39] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257.

[40] David L. Davies and Donald W. Bouldin. “A Cluster Separation Measure”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.

[41] J. C. Dunn. “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters”. In: *Journal of Cybernetics* 3.3 (1973), pp. 32–57. DOI: 10.1080/01969727308546046. eprint: https://doi.org/10.1080/01969727308546046. URL: https://doi.org/10.1080/01969727308546046.

[42] J. C. Dunn†. “Well-Separated Clusters and Optimal Fuzzy Partitions”. In: *Journal of Cybernetics* 4.1 (1974), pp. 95–104. DOI: 10.1080/01969727408546059. eprint: https://doi.org/10.1080/01969727408546059. URL: https://doi.org/10.1080/01969727408546059.

[43] U. Maulik and S. Bandyopadhyay. “Performance evaluation of some clustering algorithms and validity indices”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.12 (2002), pp. 1650–1654. DOI: 10.1109/TPAMI.2002.1114856.

[44] Faicel Chamroukhi et al. “Model-based clustering with Hidden Markov Model regression for time series with regime changes”. In: *The 2011 International Joint Conference on Neural Networks* (2011), pp. 2814–2821.

[45] Faicel Chamroukhi et al. “Joint segmentation of multivariate time series with hidden process regression for human activity recognition”. In: *Neurocomputing* 120 (2013), pp. 633–644.

Appendix

M = [[16, 15, 16, 17, 17, 14, 15, 11, 14, 15, 29, 30, 43, 34, 25, 31, 32, 25, 26, 25, 87, 87, 90, 87, 86, 69, 85, 73, 72, 82, 74, 83, 79, 77, 99, 83, 84, 97, 98, 81, 41, 41, 15, 12, 12, 27, 22, 22, 26, 22, 48, 53, 47, 46, 48, 47, 50, 47, 48, 49, 76, 75, 77, 75, 74, 75, 75, 76, 73, 77, 68, 73, 57, 57, 57, 59, 69, 71, 61, 71, 7, 14, 5, 5, 11, 52, 17, 47, 41, 24, 22, 20, 29, 26, 32, 17, 11, 27, 25, 22, 14],
[24, 35, 23, 33, 16, 5, 6, 12, 6, 12, 40, 41, 44, 41, 48, 29, 16, 24, 17, 26, 32, 32, 31, 31, 32, 33, 32, 36, 37, 37, 39, 36, 36, 37, 39, 30, 32, 32, 31, 30, 41, 42, 41, 42, 43, 46, 49, 49, 49, 45, 47, 47, 47, 47, 47, 46, 45, 43, 49, 43, 31, 32, 32, 31, 32, 34, 34, 34, 34, 35, 32, 33, 32, 32, 33, 33, 33, 34, 35, 36, 42, 43, 42, 43, 42, 41, 44, 45, 42, 45, 44, 43, 44, 43, 44, 44, 44, 42, 44, 44, 41],
[27, 29, 28, 27, 28, 26, 8, 43, 27, 18, 40, 44, 39, 38, 40, 13, 11, 11, 13, 12, 86, 82, 84, 89, 85, 84, 84, 83, 84, 83, 89, 88, 89, 87, 89, 83, 85, 81, 85, 84, 56, 59, 53, 52, 58, 55, 52, 56, 51, 59, 57, 57, 56, 57, 56, 51, 54, 60, 51, 59, 83, 82, 82, 82, 83, 90, 89, 90, 89, 90, 85, 85, 87, 89, 85, 87, 87, 88, 87, 88, 60, 60, 60, 60, 60, 57, 57, 57, 57, 57, 57, 57, 57, 58, 59, 56, 59, 59, 57, 58, 58]]

Table A1: Multivariate test signal used for definitions and tests.

M = [[16 15 16 17 17 14 15 11 14 15 29 30 43 34 25 31 32 25 26 25],
[24 35 23 33 16 5 6 12 6 12 40 41 44 41 48 29 16 24 17 26],
[27 29 28 27 28 26 8 43 27 18 40 44 39 38 40 13 11 11 13 12]]

Table A2: Part of the multivariate test signal used for PCA visualization.

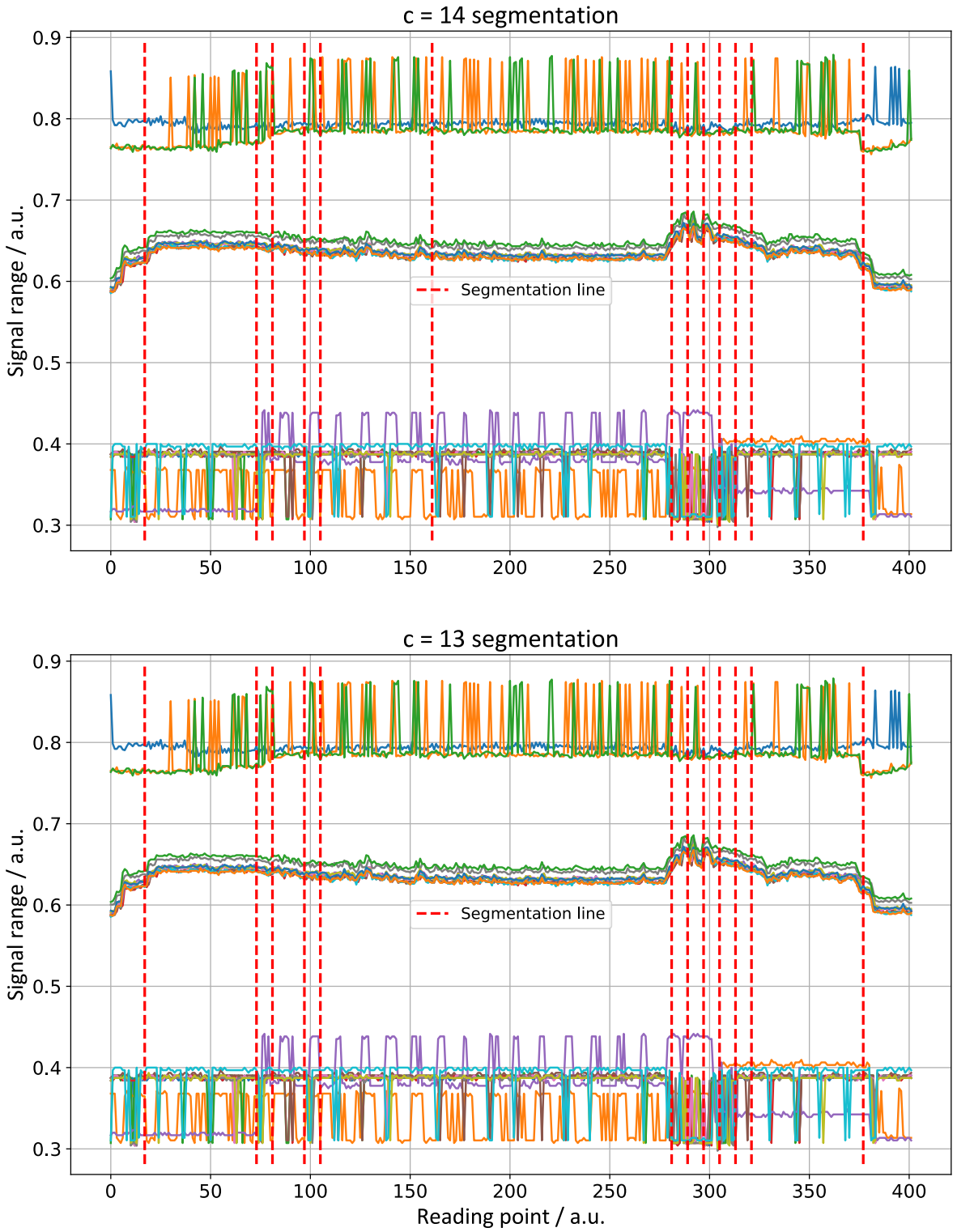

Figure A1: c = 14 and 13 PCAFC segmentation of the filtered multivariate time series.

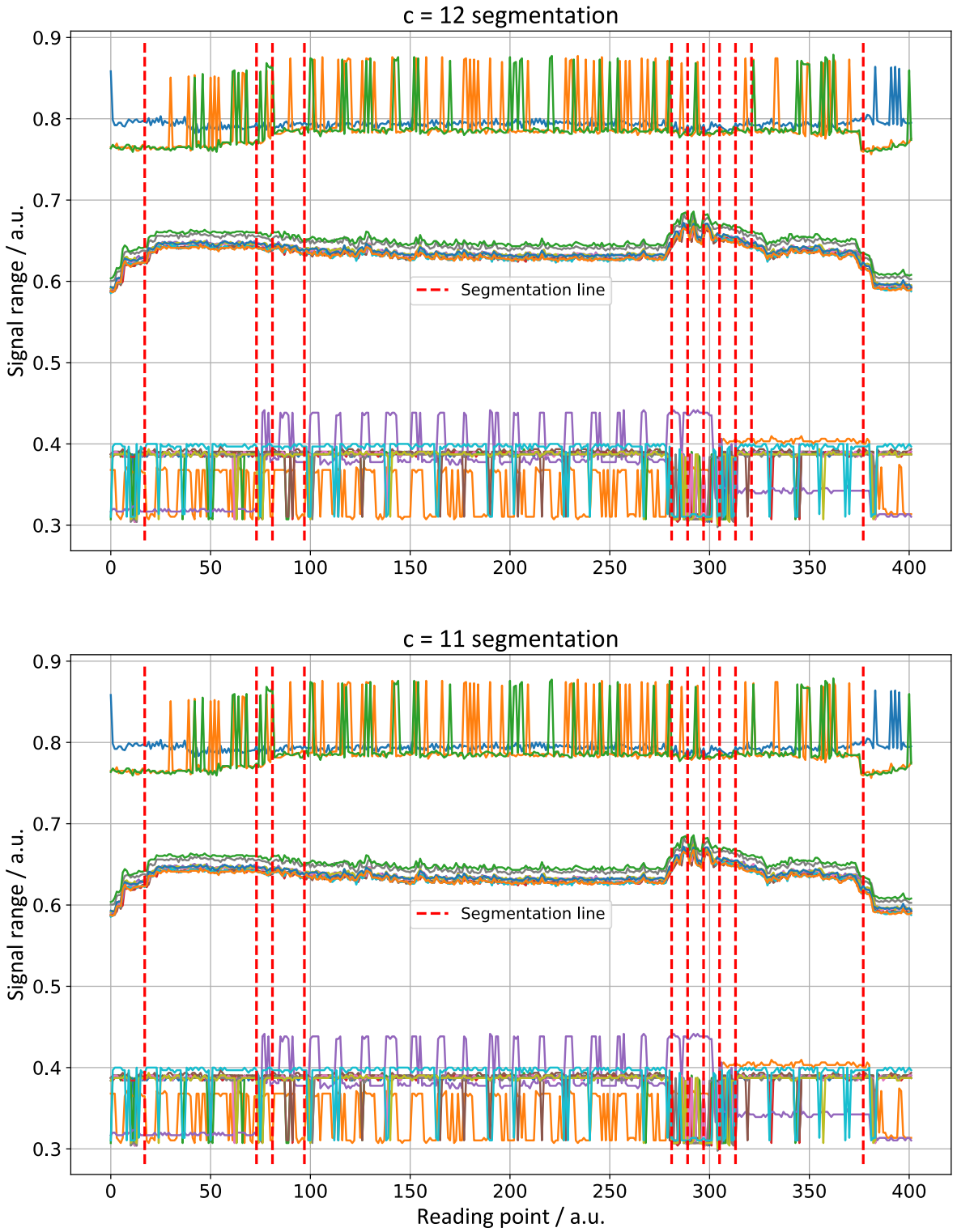

Figure A2: c = 12 and 11 PCAFC segmentation of the filtered multivariate time series.

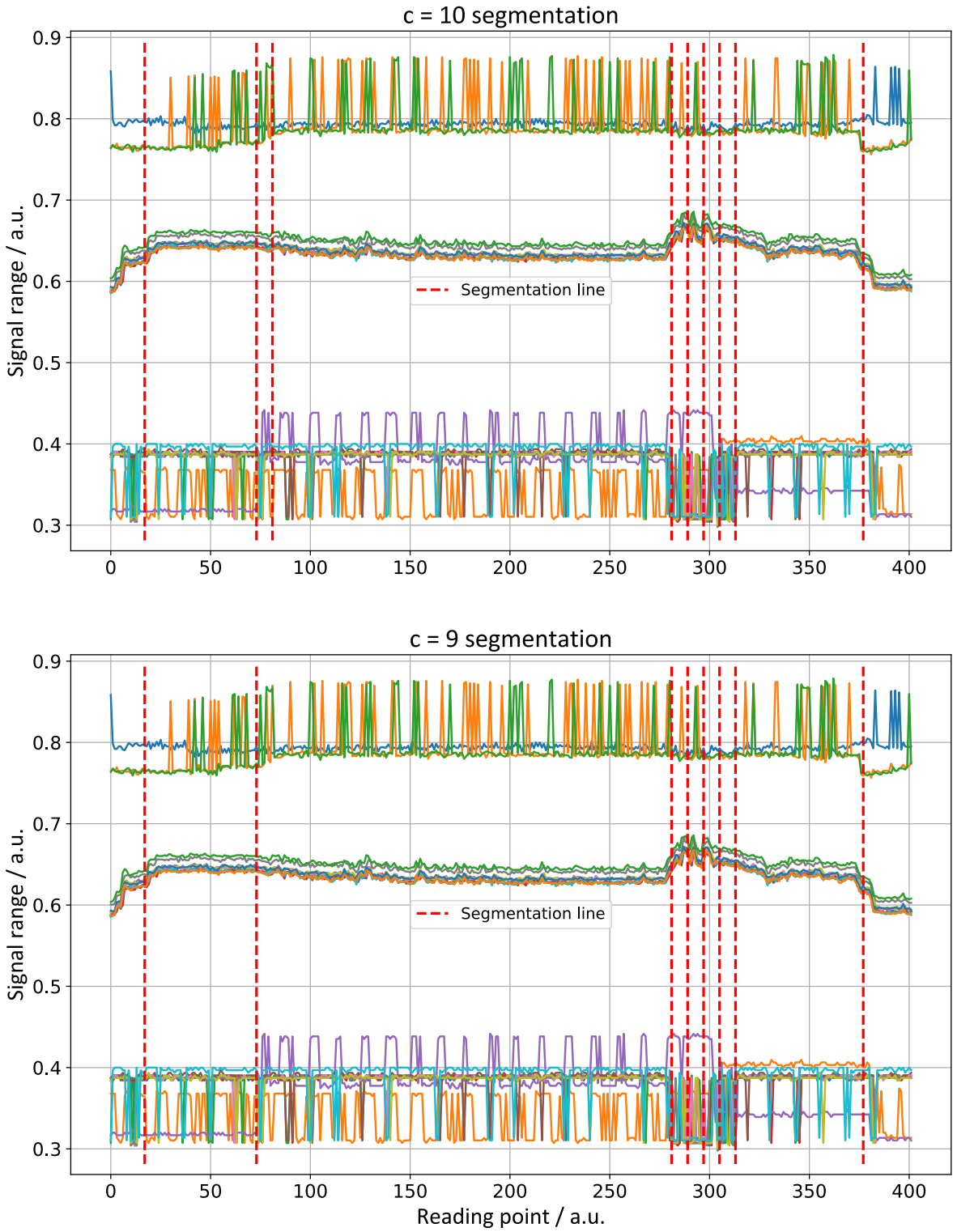

Figure A3: c = 10 and 9 PCAFC segmentation of the filtered multivariate time series.

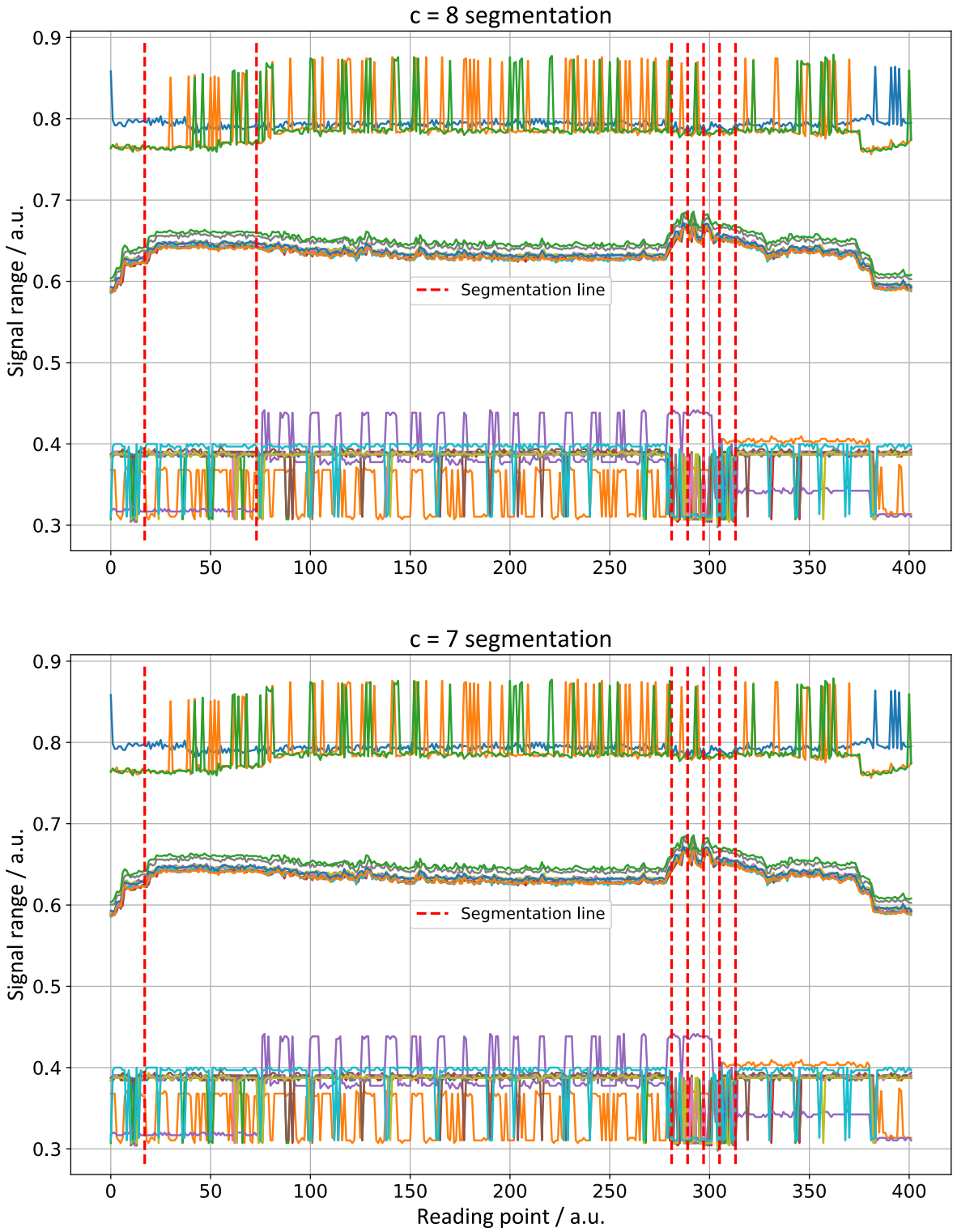

Figure A4: c = 8 and 7 PCAFC segmentation of the filtered multivariate time series.

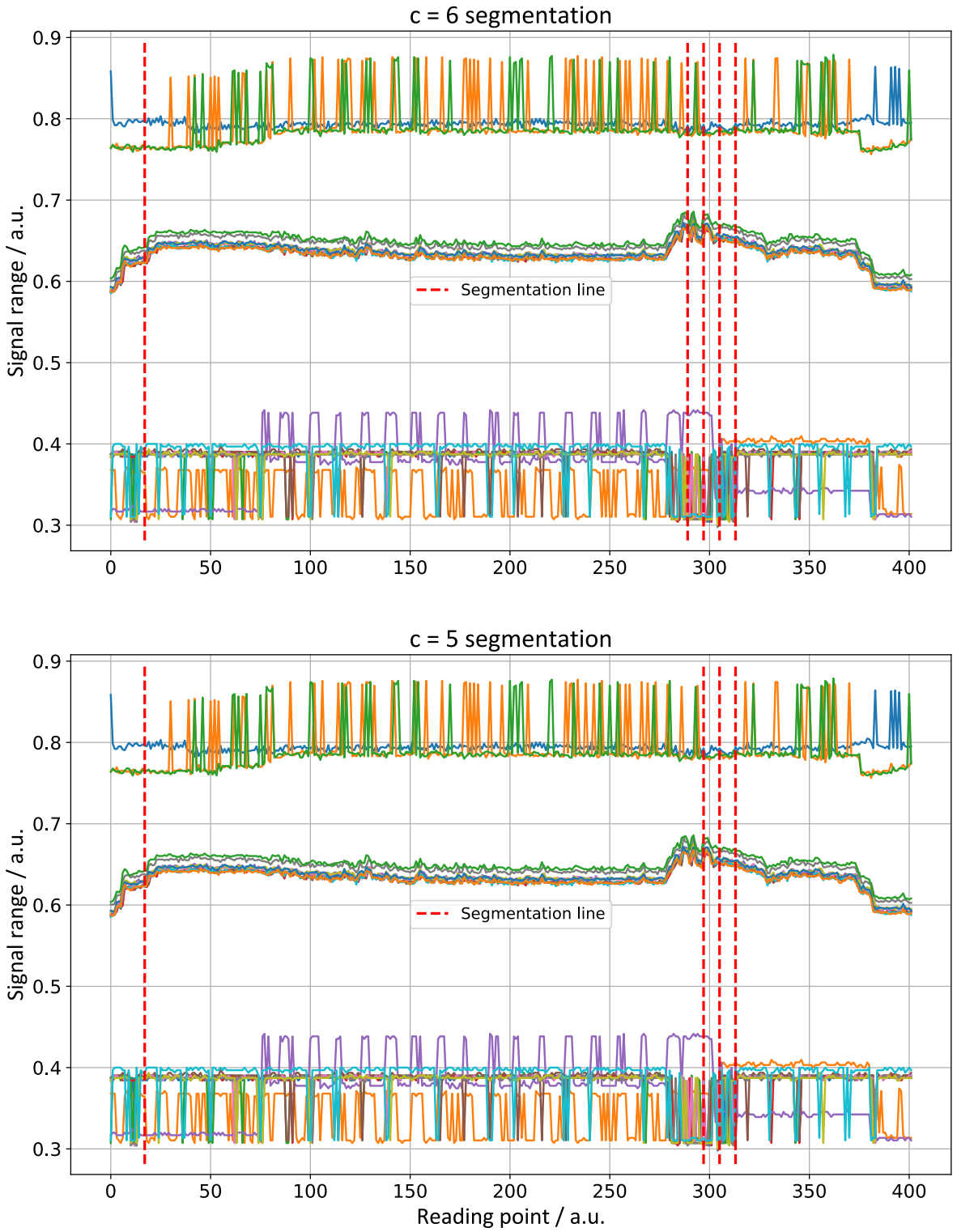

Figure A5: c = 6 and 5 PCAFC segmentation of the filtered multivariate time series.

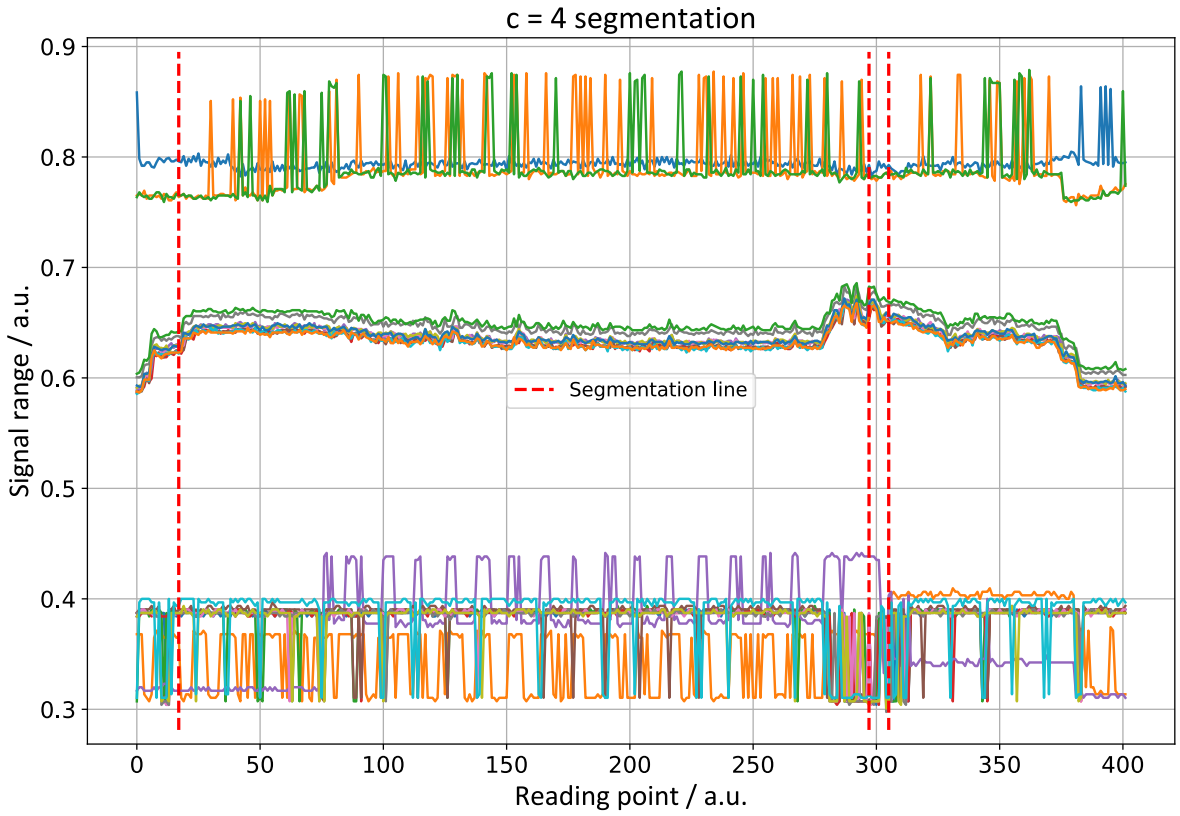

Figure A6: c = 4 PCAFC segmentation of the filtered multivariate time series.

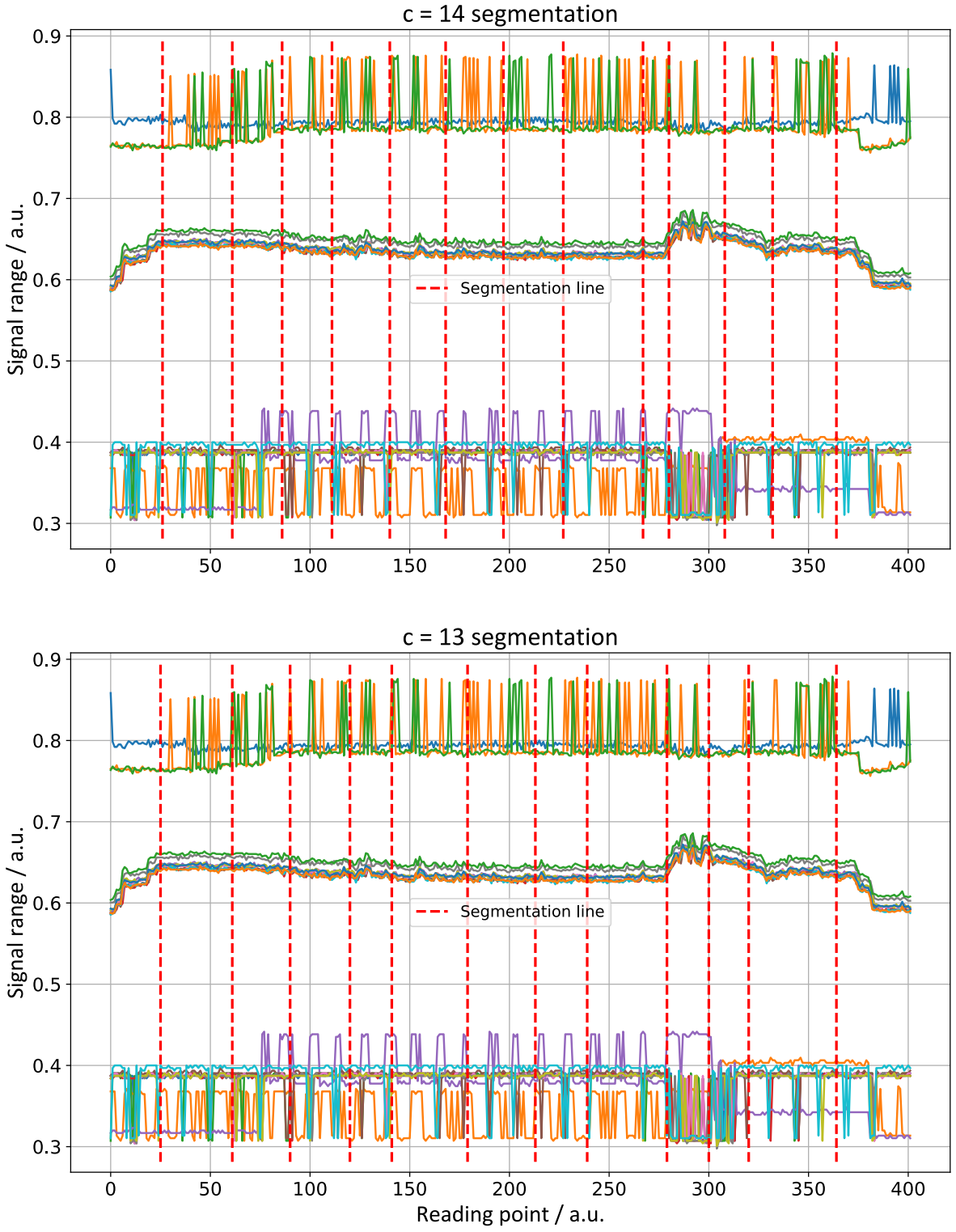

Figure A7: c = 14 and 13 MHMMR segmentation of the filtered multivariate time series.

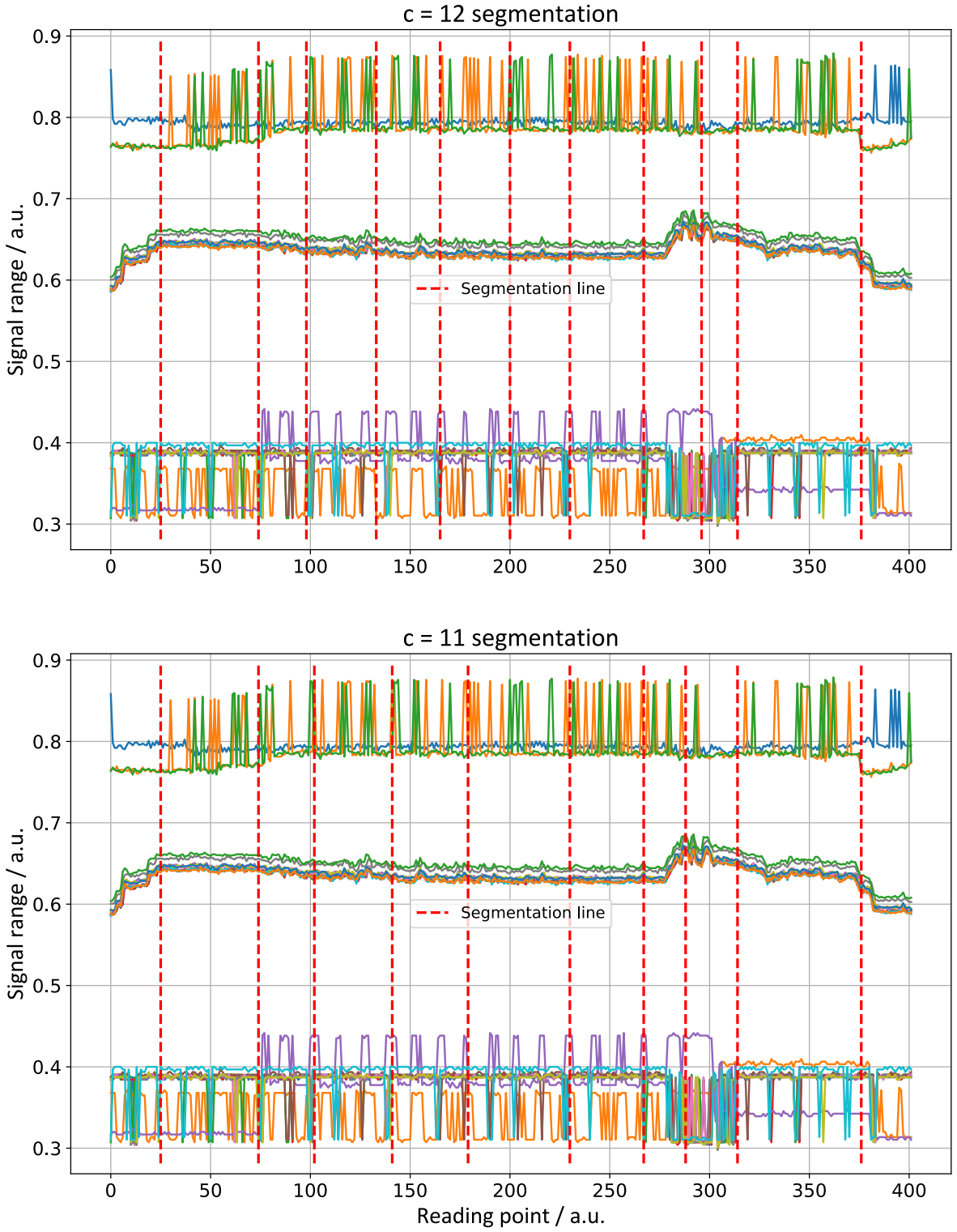

Figure A8: c = 12 and 11 MHMMR segmentation of the filtered multivariate time series.

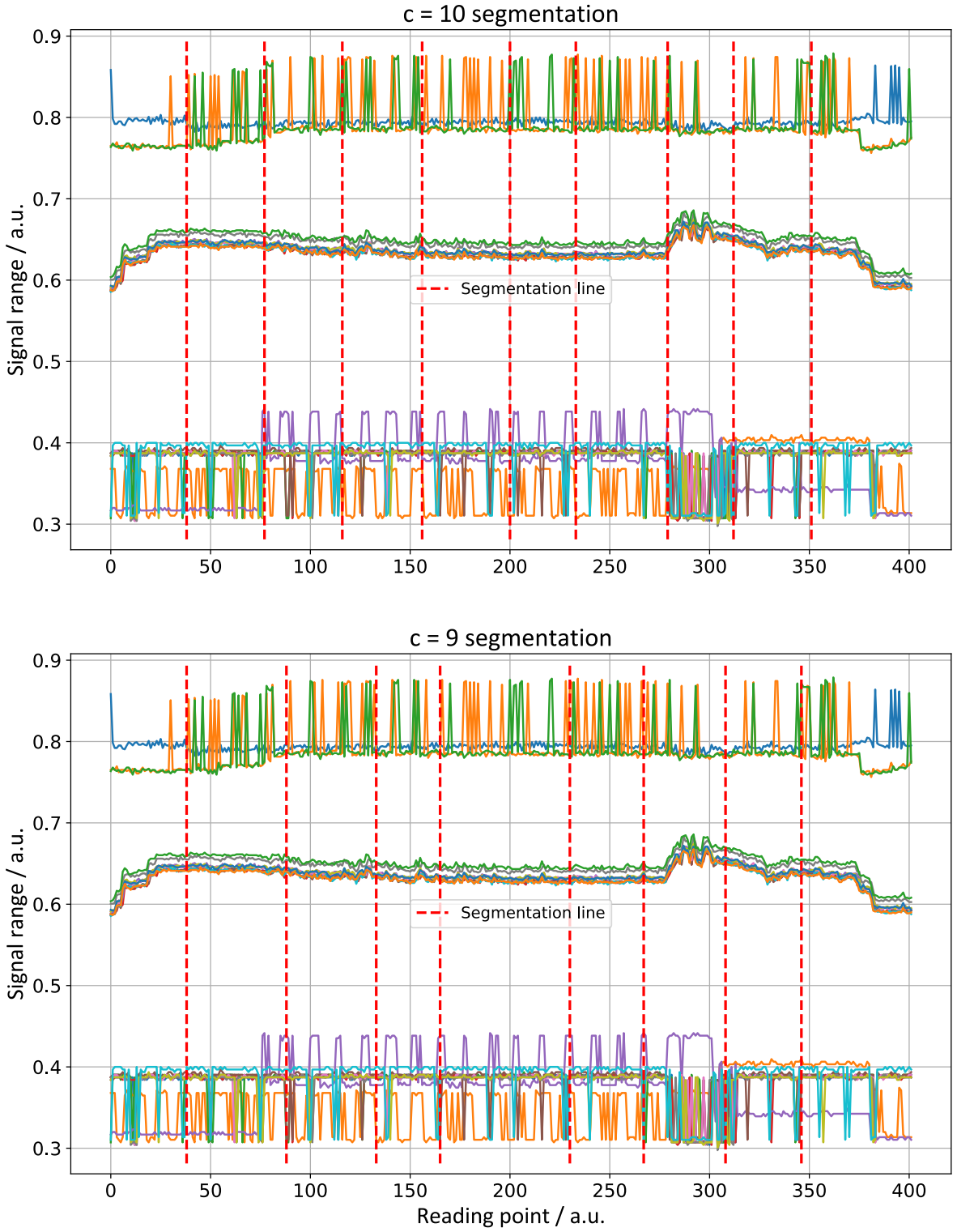

Figure A9: c = 10 and 9 MHMMR segmentation of the filtered multivariate time series.

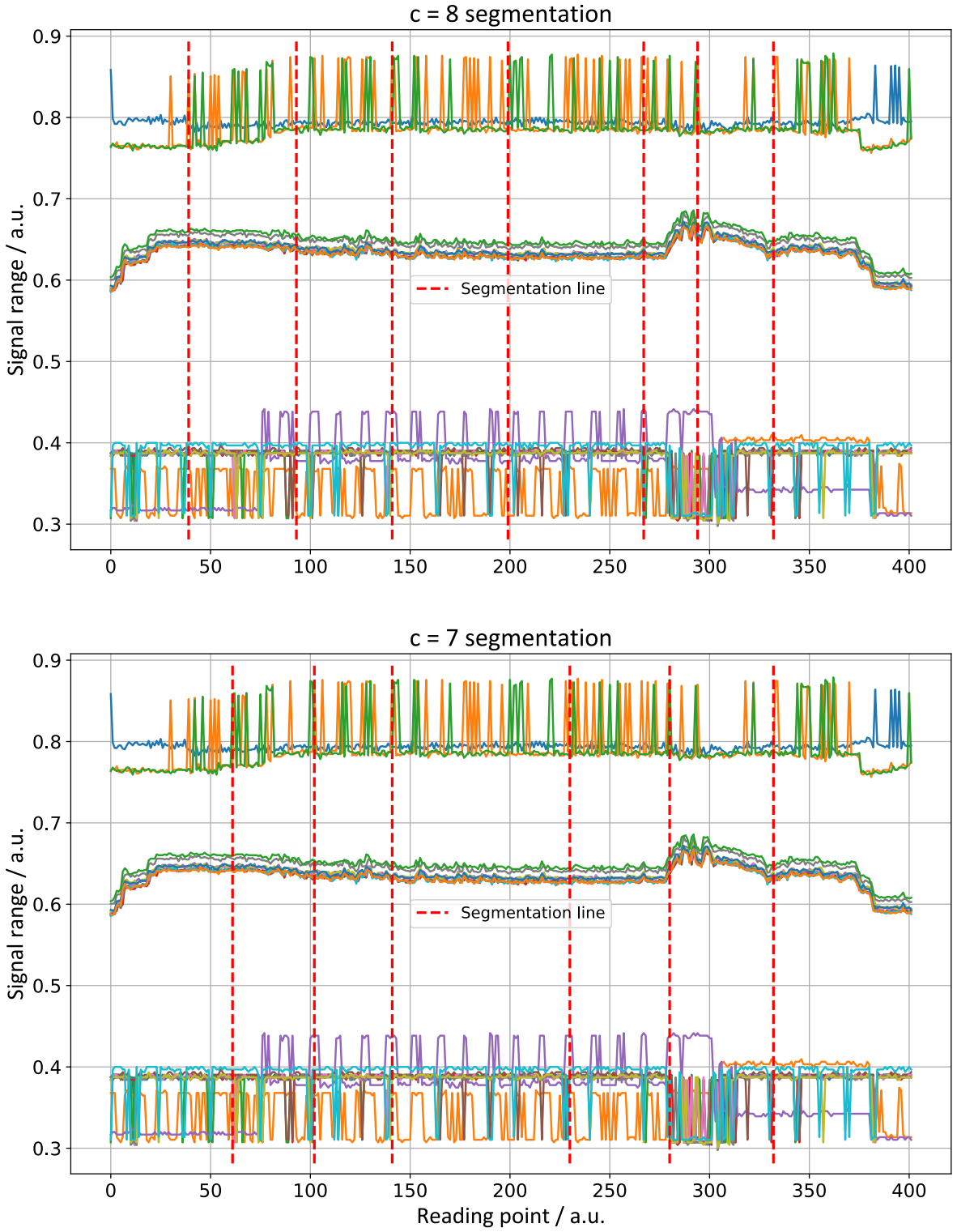

Figure A10: c = 8 and 7 MHMMR segmentation of the filtered multivariate time series.

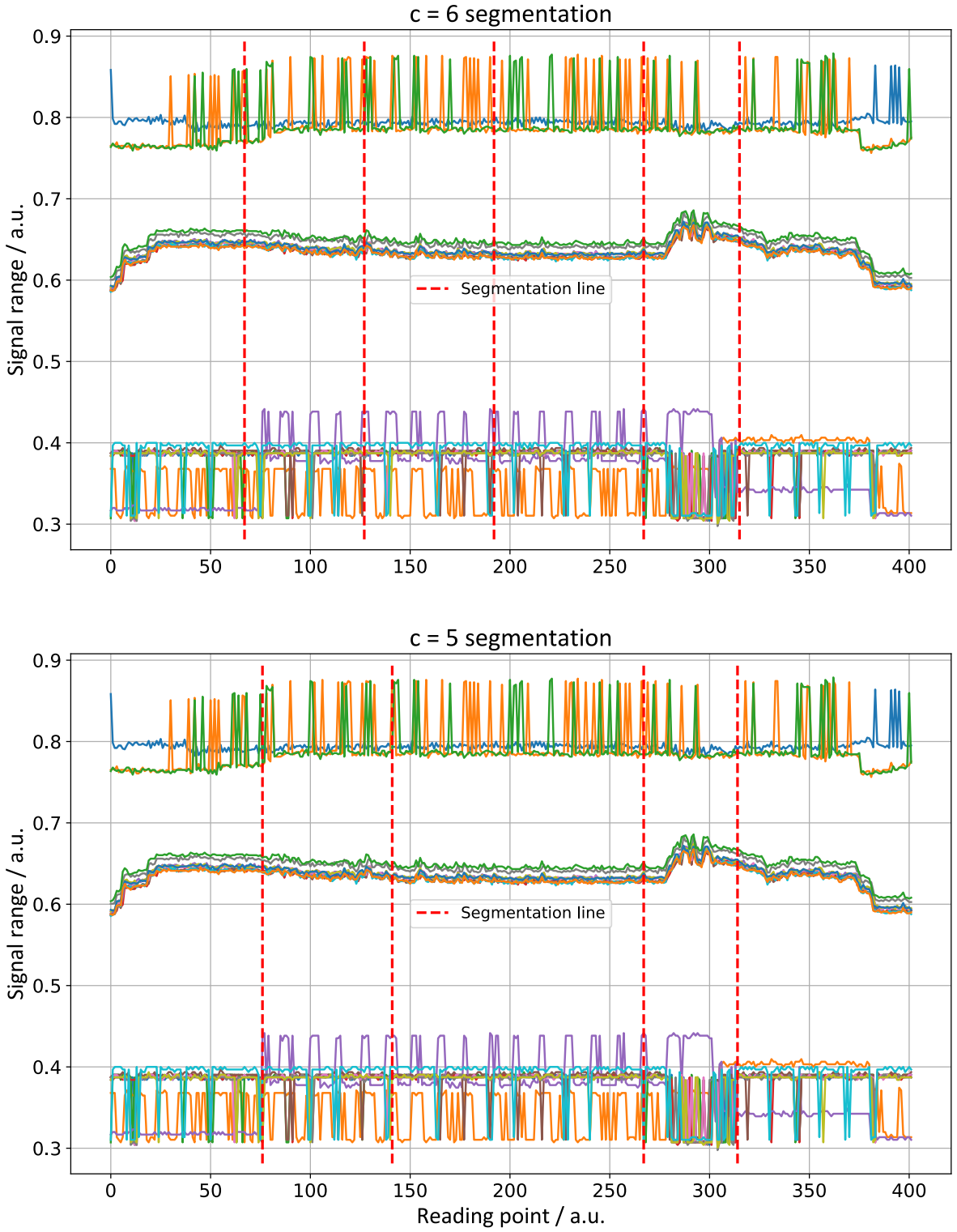

Figure A11: c = 6 and 5 MHMMR segmentation of the filtered multivariate time series.

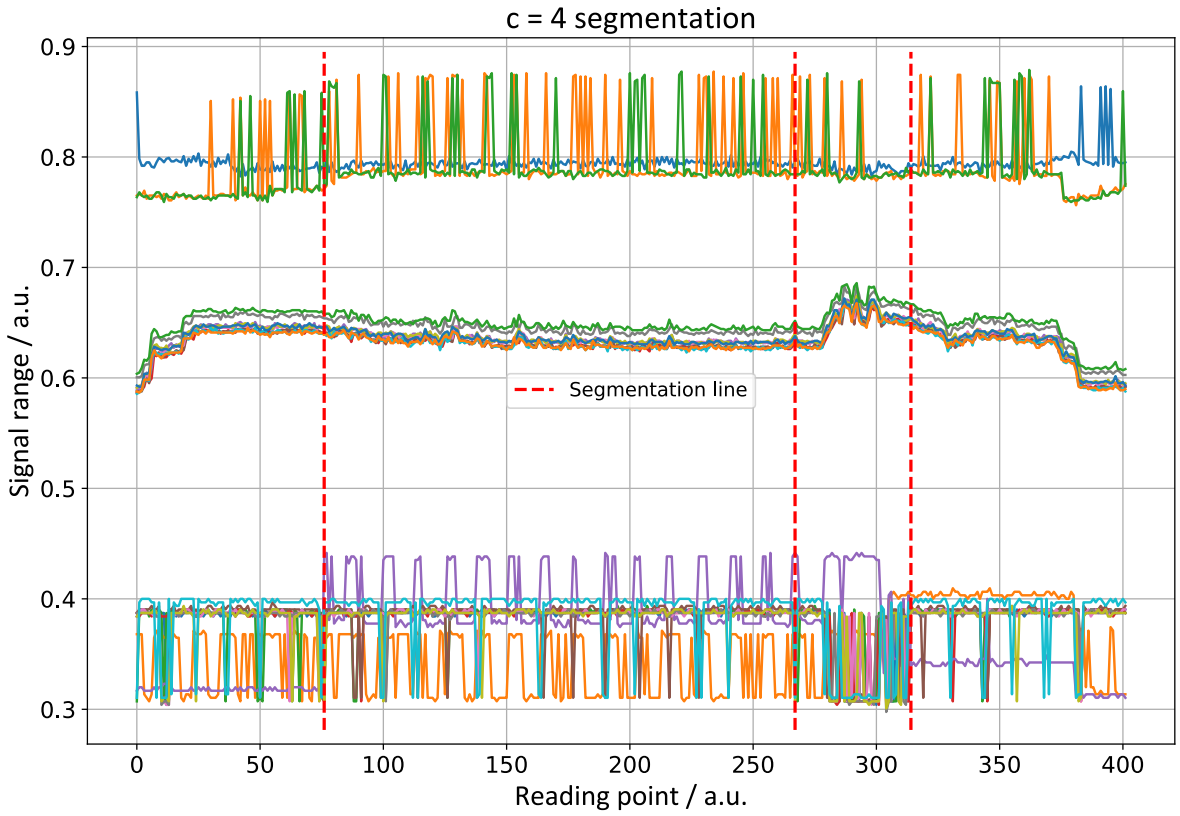

Figure A12: c = 4 MHMMR segmentation of the filtered multivariate time series.