NOVEL CLUSTERING METHODS FOR COMPLEX CLUSTER STRUCTURES IN BEHAVIORAL SCIENCES

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de Rector Magnificus prof. dr. W.B.H.J. van de Donk, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op maandag 12 december 2022 om 13:30 uur

door

Shuai Yuan

geboren te Shanghai, China

Promotor:	prof. dr. J. Vermunt (Tilburg University)
Copromotores:	dr. K. Van Deun (Tilburg University)
	dr. K. De Roover (Tilburg University)
leden promotiecommissie:	prof. dr. M.J.P.M. van Veldhoven (Tilburg University)
	prof. dr. E. Ceulemans (KU Leuven)
	prof. dr. M.E. Timmerman (RU Groningen)
	dr. L.V.D.E. Vogelsmeier (Tilburg University)
	dr. J.F. Wilderjans (Leiden University)

This research is funded by the Netherlands Organization for Scientific Research NWO) [grant project number 406-17-526].
Printing was financially supported by Tilburg University.

ISBN

Cover Designed by Qingqing
Printed by proefschrift-aio

To my family and friends who have supported me all along

CONTENTS

1

INTRODUCTION

In this digital era marked by new technologies of data reporting and collection, more frequently than ever before, behavioral scientists and practitioners use large-scale data sets (i.e., also known as *big data*) to answer their research questions. The term "large-scale data sets" can refer to both data sets with an exceeding number of observations (e.g., data collected from cross-national surveys) and data sets with a large number of variables (e.g., text data extracted from social media). This dissertation focuses on the latter, and studies novel methods and computational tools to detect hidden clusters in data sets with a large number of variables. As these data sets cover a broader range of features than used by behavioral scientists until recently, they offer unprecedented opportunities in two important ways to identify new clusters and advance theories on understanding the heterogeneity of human nature. First, these data sets integrate information hidden in a large number of variables, thus providing unique opportunities to discover novel types of clusters. For example, several studies have recently emerged in clinical and biological psychology - albeit on a relatively small scale - that identified novel psychotic subtypes through analyzing such large-scale data sets (Chen et al., 2020). Second, as features pertaining to large-scale data sets likely come from different sources (e.g., a combination of text, survey, and genetic data), joint cluster analysis of these data sets can potentially uncover

clusters that would only emerge from the concerted efforts of the many sources, something that was not achievable in previous studies because the different data sources were only considered in isolation from each other. This assertion is substantiated, for example, in Mothi et al. (2019) where the authors identified three subtypes of psychotic patients from a collection of fMRI signals, clinical diagnoses, and laboratory measures. Clearly, the above studies shed new light on theories of subtyping psychological disorders and open new venues for researching and developing personalized treatments for these subtypes

Despite the promising applications of clustering on data sets with a large number of variables, there are at least two major challenges, namely selecting the variables most relevant to the analysis and handling the diverse groups of variables collected from different sources. This dissertation aims to address these two challenges by proposing novel clustering methods and computational tools. The novel methods proposed in the dissertation are derived from two types of clustering methods used in the previous literature. In what follows, we first describe the basic ideas of the two types and the cluster structures that these methods are able to identify. Then, the two challenges are presented and discussed separately. Following the description of each challenge, we briefly propose how the challenge is addressed in this dissertation. In the final section of the introduction, we detail the content of each chapter.

This dissertation is fully committed to open science practices. In addition to providing all of the computational code used in simulation studies and empirical analyses, we have also developed three publicly accessible R packages that implement the methods described in this dissertation and a ShinyApp that supports users in visualizing the clusters resulting from their analysis.

1.1. TWO TYPES OF CLUSTERING METHODS

This dissertation develops two types of clustering methods: methods that identify clusters based solely on their average scores across variables (which we will refer to as the Type 1 methods) and methods that identify clusters based on their average scores as well as subspaces underlying the within-cluster covariance structures (which we will refer to as the Type 2 methods). A well-known example of the Type 1 methods is K-means (KM), arguably one of the most widely

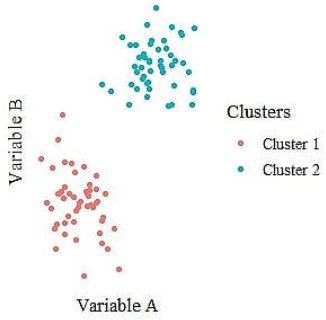


Figure 1.1: Two clusters that differ in the mean structures (but not in the covariance structures)

used clustering methods in psychological sciences (Jain, 2010; Steinley, 2006). The modeling strategy in KM and other Type 1 methods is to find clusters that maximize the total distance between cluster centroids. Here, the cluster centroids are defined as the average scores of the variables per cluster. Thus, the Type 1 methods do not model the within-cluster variances and covariances of the variables. The sole focus on the between-cluster mean structures brings computational benefits, but it limits the range of cluster structures that these methods can recover. Consider a hypothetical example with two variables A and B and with two clusters. For the data structure depicted in Figure 1.1 where the observations are centered upon their corresponding centroids with minimal overlap, the Type 1 methods are able to recover the two clusters with high accuracy. However, for the data structure displayed in Figure 1.2, where the two clusters cannot be distinguished by cluster centroids (in fact the two clusters have identical centroids), only the Type 2 methods - but not the Type 1 methods - are useful to recover the clusters with reasonable accuracy.

In this dissertation, the specific Type 2 methods we build upon belong to a genre of clustering methods that model within-cluster subspaces in addition to between-cluster mean structures (some well-known examples from this genre include D. Wang et al. (2009) and Timmerman et al. (2013)). This strategy of modeling clusters not only offers a flexible way to identify various types of cluster structures but also provides a wealth of information for interpretation: the subspace(s) discovered for each cluster summarize, in a few dimensions, the pat-

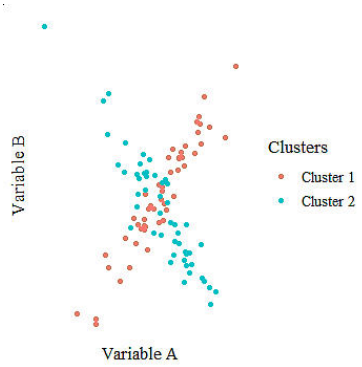


Figure 1.2: Two clusters that differ in the covariance structures (but not in the mean structures)

terns of covariances within this cluster. Since the subspaces contain significantly fewer dimensions than the full space spanned by all variables, the Type 2 methods are particularly promising when dealing with large data sets.

These two types of clustering methods described above serve as a starting point for the development of novel clustering methods for complex cluster structures which contain a substantial proportion of irrelevant variables and (or) variables from different sources. We now detail the two major challenges in identifying and recovering these complex cluster structures and discuss how the novel methods proposed in this dissertation address both challenges.

1.2. FIRST CHALLENGE: VARIABLE SELECTION

When dealing with data sets with a large number of variables, many scholars have highlighted the challenge that a large proportion of variables may be irrelevant to the data analysis under consideration (e.g., Adjerid & Kelley, 2018; Janssen et al., 2017; Qiu et al., 2018; Serang et al., 2017; Waldherr et al., 2017). The irrelevant variables defined in the Type 1 methods are those that do not separate clusters well and mask cluster structures (Brusco & Cradit, 2001; Raftery & Dean, 2006; Steinley & Brusco, 2008b). For the Type 2 methods, the irrelevant variables are those that are unrelated to within-cluster subspaces. Failing to identify these irrelevant variables is problematic for two reasons. First, the presence of irrelevant variables may largely complicate the interpretation of clusters. Sec-

ond and perhaps more importantly, since the irrelevant variables obscure true cluster structures (i.e., between-cluster mean structures and (or) within-cluster subspace structures), cluster analysis with these irrelevant variables is likely to yield poor cluster recovery (Steinley & Brusco, 2008b).

To address the challenge of variable selection, this dissertation utilizes the sparseness approach (Hastie et al., 2019) for automatic variable selection. This approach was originally proposed to address the problem of overfitting in linear regression: it imposes additional constraints on the regression coefficients and reduces the smallest coefficients to 0 (Tibshirani, 1996). As a result, only a subset of predictors have nonzero coefficients and are thus considered relevant variables (also termed signaling variables in the dissertation). Recently, the sparseness approach for variable selection has been successfully implemented in other models, including network models (Epskamp & Fried, 2018), structural equation models (Jacobucci et al., 2016), and principal component analysis (H. Shen & Huang, 2008). Following the footsteps of previous studies, this dissertation employs the sparseness approach to achieve variable selection in both Type 1 and Type 2 clustering methods. More specifically, for the Type 1 methods, sparseness is introduced in the calculation of the total distance between cluster centroids, so that the distances pertaining to the identified irrelevant variables are excluded from the calculation. For the Type 2 methods, sparseness is induced within the process of extracting subspaces, so that the irrelevant variables are not related to the identified subspaces.

1.3. SECOND CHALLENGE: DATA INTEGRATION

Aside from self-reported survey data, many behavioral studies conducted in the digital age supplemented their analyses with novel types of data (e.g., genetic, sensor, and GPS data, etc.,). Data sets consisting of measurements collected from the same group of observations but from different data sources (presented as different data blocks) are referred to as multi-block data, and the type of analyses that integrate information from these multiple data blocks are known as data integration (Van Deun et al., 2009; Van Deun et al., 2011) or data fusion (Lahat et al., 2015). As pointed out in Van Deun et al. (2009) and Lock et al. (2013), a key challenge in data integration is to discern two types of variation: variation pertaining

to the same data block or only a few data blocks (also called distinctive variation) and structural variation pertaining to all data blocks (also called common variation). To illustrate, consider a multi-block data set that includes one block of self-reported personality scores and one block of GPS data: while the common, structural variation may indicate the relationships between one's personalities and distance traveled within a month, the distinctive variation pertaining to self-reported data may reveal response styles underlying personality scores. When analyzing multi-block data sets, it is desirable - and sometimes even necessary - to separate common and distinctive variations for two critical reasons. First, oftentimes, researchers are particularly interested in the joint forces of several behavioral factors manifested in common variation (e.g., how personalities relate to the distance traveled within a month), which can only be reliably interpreted after its complete disentanglement from distinctive variation. Second, when such multi-block data are used for predictive analysis, it is not unlikely that only common variation or distinctive variation - but not a mix of the two - is predictive of outcomes; therefore, separating these two sources of variation is beneficial for accurate prediction. In this dissertation, we focus on addressing the challenge of distinguishing common and distinctive variation in data integration for the Type 2 clustering methods.

To address the challenge of data integration, a modified version of Principal Component Analysis or PCA (Guerra-Urzola et al., 2021; Van Deun et al., 2009; Van Deun et al., 2011) is applied in this dissertation. PCA, one of the most well-known methods in psychological research, summarizes the covariances of variables by a few components. Because the resulting components are correlated with all variables, it is uncertain whether they reveal common or distinctive variation, or a mix of both types. To distinguish between these two types of variation, Gu and Van Deun (2019) proposed Common and Distinctive Simultaneous Component Analysis (CD-SCA) that imposes blocks of zero loadings onto the component loading matrix: components with blocks of zero loadings indicate distinctive variation because they are unrelated to the variables associated with these zero loadings. We show that this strategy can be incorporated into the Type 2 clustering methods, allowing for accurate estimation of within-cluster subspaces that reveal both common and distinctive variation.

1.4. OUTLINE OF THE DISSERTATION

This dissertation aims to develop new statistical methods and computational tools for clustering complex data structures that contain a substantial proportion of irrelevant variables or variables from different sources. Chapters 2 and 3 focus on the Type 1 methods and employ the sparseness approach to address the challenge of variable selection. Chapter 2 introduces and validates a novel method, called Cardinality K-means or CKM, for simultaneous variable selection and clustering, whereas Chapter 3 provides a detailed, non-technical tutorial for applied researchers to use CKM and elaborates on various ways to validate the obtained cluster partitions. This chapter also introduces an easy-to-use ShinyApp, called `ClusterViz`, for visualizing clusters. Next, Chapters 4 and 5 present the Type 2 methods that use the modified PCA approach and the sparseness approach (only in Chapter 4) to address the challenges of data integration and variable selection, respectively. It is worth noting that special cases of the methods developed in Chapters 4 and 5 can be used to deal with single-block data (i.e., without data integration). While the clustering algorithm proposed in Chapter 4 only uses multi-block predictors to recover clusters, the method described in Chapter 5 further considers an outcome outside multi-block predictors such that the resulting within-cluster subspaces are predictive of the outcome. In other words, the method in Chapter 4 is an unsupervised learning algorithm, while the one in Chapter 5 is a supervised learning algorithm. Below, the four empirical chapters are presented in greater detail.

In Chapter 2, CKM, a novel simultaneous clustering and variable selection method, is introduced in great detail. In essence, CKM utilizes the connection between PCA and KM to transform the simultaneous clustering and variable selection problem into one that can be readily addressed by sparse principal components analysis. In two extensive simulation studies, CKM consistently outperformed its predecessor Sparse K-Means (Witten & Tibshirani, 2010) and Sparse Alternate Sum (Arias-Castro & Pu, 2017) in terms of cluster recovery and the accuracy of variable selection. In addition, inspired by Brudvig et al. (2019), Chapter 2 also proposes a novel model selection procedure that determines the number of clusters after extracting a small set of stable variables that are identified as signaling variables for all potential numbers of clusters. The R package `CKM`

is developed to implement the CKM method and the described model selection procedure.

Based on CKM and the novel model selection procedure proposed in Chapter 2, Chapter 3 serves as a tutorial to illustrate how applied researchers can apply these methods in their own research. An example of clustering subjects based on their political attitudes is provided. More importantly, Chapter 3 delves into the topic of cluster validation, which is an important yet often overlooked aspect of cluster analysis. More specifically, this chapter utilizes a recently proposed framework (Ullmann et al., 2021) to discuss and illustrate three validation approaches: visual validation, stability validation and validation by replication. Last but not least, a new ShinyApp `ClusterViz` is introduced in this chapter to aid visual validation, allowing users to visualize clusters without any programming skills.

Chapter 4 proposes a clusterwise extension to CD-SCA. This method, referred to as Clusterwise Sparse Simultaneous Component Analysis or CSSCA, employs an alternating algorithm in conjunction with a random-start procedure to achieve successful estimation. The partitions obtained by CSSCA allow users to identify subjects that are driven by the same set of (common and distinctive) components. Another desirable feature of CSSCA is that it uses the sparseness approach for automatic variable selection, thus improving the interpretability of the results. Chapter 4 also reports two simulation studies and demonstrates the conditions that are less favorable to CSSCA. CSSCA is applied to an empirical study that identifies clusters based on a block of personality scores linked with a block of nonverbal behaviors. Last, the R package `ClusterSSCA` has been developed for readers to employ CSSCA in their own research.

While CSSCA deals with unsupervised learning tasks, Chapter 5 extends the Type 2 methods for supervised learning tasks (i.e., prediction tasks). To this end, inspired by the development of Principal Covariates Regression (De Jong & Kiers, 1992; Vervloet et al., 2015), Chapter 5 proposes a novel method called Clusterwise Simultaneous Covariates Regression or CSCR, which identifies clusters of observations with different common and distinctive components and (or) different regression weights when regressing the outcome on these components. Importantly, CSCR is more advantageous than a conventional, two-step approach

linking CSSCA and ordinary least squares (OLS) regression because the components estimated from CSCR are guaranteed to be predictive of the outcome. In a large-scale simulation study, the performance of CSCR is compared with four alternative methods. The findings confirm that CSCR performs well in general, but they also identify the conditions under which it needs to be used with great care. Like other chapters, an R package `CSCR` is developed for the implementation of CSCR.

Since the chapters were written independently as potential journal articles, there may be some overlaps and inconsistencies in terminology across the chapters.

2

SIMULTANEOUS CLUSTERING AND VARIABLE SELECTION: A NOVEL ALGORITHM AND MODEL SELECTION PROCEDURE

The growing availability of high-dimensional data sets offers behavioral scientists an unprecedented opportunity to integrate the information hidden in novel types of data (e.g., genetic data, social media data, and GPS tracks, etc.,) and thereby obtain a more detailed and comprehensive view towards their research questions. In the context of clustering, analyzing the large volume of variables could potentially result in an accurate estimation or a novel discovery of underlying clusters. However, a unique challenge is that the high-dimensional data sets likely involve a significant amount of irrelevant variables. These irrelevant variables do not contribute to the separation of clusters and may mask cluster partitions. The current paper addresses this challenge by introducing a new clustering algorithm, called

This chapter is published as **Yuan, S.**, De Roover, K., & Van Deun, K. (in press). Simultaneous Clustering and Variable Selection: a Novel Algorithm and Model Selection Procedure. *Behavior Research Methods*

Cardinality K-means or CKM, and by proposing a novel model selection strategy. CKM is able to perform simultaneous clustering and variable selection with high stability. In two simulation studies and an empirical demonstration with genetic data, CKM consistently outperformed competing methods, in terms of recovering cluster partitions and identifying signaling variables. Our proposed model selection strategy determines the number of clusters based on a subset of variables that are most likely to be signaling variables. Through a simulation study, this strategy was found to result in a more accurate estimation of the number of clusters, compared to the conventional strategy that utilizes the full set of variables. Our proposed CKM method and the novel model selection strategy have been implemented in a freely accessible R package.

2.1. INTRODUCTION

Recent technological developments have made it fairly easy to collect a large number of variables within a single study in social and behavioral sciences. Examples include examinations of genetic influences in organizational psychology (e.g., Arvey et al., 2016; Chi et al., 2016), personality psychology (e.g., Davis et al., 2019) and social psychology (e.g., Feldman et al., 2016); studies on neuroscientific foundations of behaviors in management (e.g., Waldman et al., 2019) and psychiatry research (e.g., Sun et al., 2009); research aiming to predict personality from social media footprints (e.g., G. Park et al., 2015); questionnaire-based studies that simply collected a comprehensive set of variables (e.g., Joel et al., 2017); as well as a combination of all these types of data (e.g., Bzdok & Meyer-Lindenberg, 2018).

A noteworthy advantage of data sets including many variables is that they provide a detailed and comprehensive view. Here, the definition of "many variables" is rather subjective and depends largely on the field of research. In behavioral sciences, one can think of data sets with more than 100 variables (Groeneveld & Rumsfeld, 2016). These types of data sets become increasingly common due to the fact that novel types of data sources are more and more often collected. Some special examples are so-called 'high-dimensional' data sets where the number of variables exceeds the number of observations. In the context of cluster analysis - where the intent is to group observations in such a way that those in the same cluster are similar to each other - using data with many variables will likely result in a more accurate estimation of clusters and (or) a discovery of novel clusters. In one of the very few reported attempts to cluster data sets with many variables, Mothi et al. (2019) combined clinical measures, laboratory measures, and measures derived from MRI scans of psychotic patients to form a combined data set, on which they conducted a cluster analysis and identified three sub-types of psychoses. Evidently, clustering high-dimensional data sets grants researchers an unprecedented opportunity to clarify and deepen our understanding of the heterogeneity in various social phenomena.

Although research that exploits data sets with many variables to identify clusters is promising, it also comes with challenges. One of the most compelling challenges, as stressed by a number of scholars (e.g., Bzdok & Meyer-Lindenberg,

2018; Waldherr et al., 2017; Yarkoni & Westfall, 2017), is that these data sets may comprise a large amount of "irrelevant variables" (Fowlkes & Mallows, 1983). They are variables that do not separate clusters well and therefore do not define cluster structures. These irrelevant variables may hinder cluster discovery by masking the cluster structure under investigation (Steinley & Brusco, 2008b). Therefore, a cluster analysis should effectively recover the cluster structure while simultaneously filtering out irrelevant variables.

The variable selection problem in cluster analysis is not a new topic and has been extensively studied since the 1980s. For example, Steinley and Brusco (2008b) have compared the performance of eight different procedures to address this problem. These approaches - most notably the Variable Selection in K-Means (i.e., VS-KM; Brusco & Cradit, 2001), model-based variable selection (Raftery & Dean, 2006), the Clustering Objects on Subsets of Attributes (i.e., COSA; Friedman & Meulman, 2004) and the relative clusterability weighting method (Steinley & Brusco, 2008a) - are well designed and have been extensively validated. However, these methods are computationally prohibitive in the presence of many variables, as the computational demand grows exponentially with the number of variables. For example, Steinley and Brusco (2008a) proposed to test all subsets of variables that pass the initial screening, where the theoretical maximum number of tests can be as high as $2^J - 1$ (with J indicating the number of variables in the data set). Raftery and Dean (2006) and Brusco and Cradit (2001) have both proposed a forward-searching strategy that starts with an initial pair of two signaling variables and, after searching all remaining variables, adds other signaling variables one by one. This strategy, too, becomes very inefficient when there are more than 100 variables.

Other methods are available, however, that are able to simultaneously perform variable selection and clustering, with reasonable computational time for large data sets with many variables. They are, for example, Sparse k-means (SKM; Witten & Tibshirani, 2010) and Sparse Alternate Sum (SAS; Arias-Castro & Pu, 2017). Importantly, these methods have been verified in several simulation studies to entail a better performance than competing approaches, such as the aforementioned COSA (Witten & Tibshirani, 2010).

One of the important contributions of the current study is to present a novel method, which we named Cardinality k-means or CKM, for simultaneous variable selection and clustering (see Yamashita and Adachi (2020) for another application of the cardinality constraint on clustering). CKM essentially exploits the fact that principal component analysis (PCA) offers reasonable starting partitions to the k-means algorithm (hereafter called KM; Ding and He, 2004; Xu et al., 2015), especially in high-dimensional data sets. Based on this connection, CKM approximates clustering solutions through sparse principal component analysis (SPCA; H. Shen & Huang, 2008) and, based on the initial results of SPCA, continuously updates partitions until convergence is reached. Here, the algorithm is considered to converge when all observations remain in the same cluster after another iteration of cluster updates. Section 2.2 illustrates how CKM theoretically relates to SKM and SAS, while section 2.3 reports how their performance compared.

As another important contribution, this study tackles the problem of selecting the correct number of clusters in the presence of (many) irrelevant variables. To date, despite calls to research this problem (e.g., Steinley & Brusco, 2008b, 2011b), to the best of our knowledge, only Brudvig et al. (2019) has empirically addressed this issue. Brudvig et al. (2019) argued convincingly that the selection of the number of clusters is a central issue, and, perhaps more importantly, pointing out that the common practice of selecting the number of clusters using all variables may be misleading, as the irrelevant variables could mask the cluster separation, resulting in an erroneous estimation of the number of clusters. Building on Steinley and Brusco (2008a), the authors have proposed a new index to simultaneously select the number of signaling variables and the number of clusters. Unfortunately, the calculation of this index is prone to computational difficulties when dealing with data sets with a large number of variables. In the current study, we aim to expand this line of research in two ways: 1) we propose a novel strategy to select the number of clusters that might be more suitable in the presence of a large proportion of irrelevant variables and 2) within the framework of our novel strategy, we compare several methods to select the number of clusters in a simulation study. The novel strategy is based on the idea of extracting a "stable" set of variables that are deemed to be signaling variables given any num-

ber of clusters. To evaluate the novel model selection strategy, we obtained the accuracy of the novel and competing model selection strategies when applied in conjunction with various clustering methods and with various test statistics.

The paper is organized as follows. We present the CKM model and the accompanying algorithm in Section 2.2, where we also discuss the novel strategy to determine the number of clusters and several methods related to CKM. Three simulation studies are presented in Section 2.3. In the first two simulation studies, CKM is validated and compared with SKM and SAS across various conditions; while both the number of irrelevant variables and the number of clusters are treated as known information in the first simulation, only the latter is treated as known in the second. In the third simulation study, we illustrate the relative performance of the novel model selection strategy that utilizes the stable set of variables as opposed to the strategy that utilizes the full set of variables. We then proceed to illustrate the usage of CKM on a large data set that consists of over forty thousand variables in Section 2.4. Finally, in Section 2.5, we discuss the practical implication of CKM and the novel model selection strategy, address their limitations, and propose future research directions. To promote our methods, we implemented CKM and the model selection procedure in a user-friendly *R* package `CKM` (available at `https://github.com/syuanuvt/CKM`).

2.2. METHODS

To develop CKM, we rely on results proven in Ding and He (2004) and Xu et al. (2015). They have shown how principal component analysis (PCA) can be used to obtain the subspace in which the clusters reside. A key advantage of this proposal, as discussed and illustrated in Xu et al. (2015), is the stability of the clusters obtained and improved accuracy in recovering the clusters, given that the clustering process mainly operates on the reduced (i.e., low-dimensional) space. In the current paper, we develop CKM that builds upon these results in the context of sparse PCA (i.e., Adachi & Trendafilov, 2016; H. Shen & Huang, 2008) for effective variable selection. First, we discuss the assumed clustering model (i.e. the KM model) and how it links up to PCA. Then, we illustrate our novel idea of incorporating sparseness in a PCA-like framework to filter out irrelevant variables in the KM model. After that, we introduce an efficient algorithm designed for

CKM, followed by an overview and comparison with related methods. Last, We formally introduce our novel strategy to determine the number of clusters in the presence of many irrelevant variables.

2.2.1. MODEL SPECIFICATION

A PCA APPROACH TO SOLVE THE KM PROBLEM

Prior to our discussion of CKM, we briefly show the connection between KM and PCA. That PCA can be effectively used to find the subspace in which the clusters reside was first shown in Ding and He (2004) and later in Xu et al. (2015). Interested readers are referred to those articles for detailed derivations and proofs of the main results reported here.

For a variable-wise standardized data matrix $\mathbf{X}$ (i.e. each variable is mean-centered and re-scaled to unit variance) with N subjects and J variables (and $\mathbf{x}_i$ denotes the response vector of subject i where $i \in 1,2,...,N$), we assume a total number of K clusters to be present in the data. We define an indicator vector $\mathbf{c}$ in such a way that $\mathbf{c}(i)$ represents the cluster assignment of observation i and $\mathbf{c}^{-1}(k)$ comprises the indices of all N_k subjects in cluster k. The objective of KM is given in

$$\begin{aligned} &\mathbf{argmin}_{\mathbf{c}} \sum_{k=1}^{K} \sum_{i \in \mathbf{c}^{-1}(k)} ||\mathbf{x}_i - \mathbf{m}_k||_2^2 \\ &with\, \mathbf{m}_k = \frac{1}{N_k} \sum_{i \in \mathbf{c}^{-1}(k)} \mathbf{x}_i, \end{aligned} \tag{2.1}$$

where $||.||_2^2$ refers to the squared Euclidean norm (i.e., for $\mathbf{x} = (x_1, x_2, ..., x_J)$, $||\mathbf{x}||_2^2 = x_1^2 + x_2^2 + ... + x_J^2$).

Because the optimization problem in Equation (2.1) is a discrete one, typically an alternating algorithm with multiple starts is employed where each indicator vector is generated randomly and updated until convergence. From the multiple converged solutions, the best one is retained as the final solution; however, there is no guarantee this solution is optimal.

The major contribution of Ding and He (2004) and later Xu et al. (2015) is the proof of the equivalence between PCA and a continuous relaxation of KM and henceforth the proposal of solving KM with the help of PCA. To see this, they first introduced a partition matrix $\mathbf{H}$ $(N \times K)$ to specify the correspondence between

subjects and clusters. More specifically, the element h_{ik}, located at the i^{th} row and the k^{th} column of $\mathbf{H}$, is constructed as follows,

$$h_{ik} = \begin{cases} 1 & i \in \mathbf{c}^{-1}(k) \\ 0 & i \notin \mathbf{c}^{-1}(k) \end{cases} \tag{2.2}$$

This specification results in $\mathbf{H}$ having orthogonal columns. Moreover, $\mathbf{H}$ is directly linked with $\mathbf{m}_k$, according to

$$\mathbf{m}_k = \frac{1}{\sqrt{N_k}} \mathbf{h}_k' \mathbf{X} \tag{2.3}$$

where $\mathbf{h}_k$ denotes the k^{th} column of $\mathbf{H}$.

Combine Equations (2.3) and (2.1), and perform some algebraic operations (detailed in Appendix 2.A), we arrive at

$$\begin{aligned} &\mathbf{argmax}_{\mathbf{H}}\ Tr\mathbf{H}'\mathbf{X}\mathbf{X}'\mathbf{H} \\ &s.t. \mathbf{H}'\mathbf{H} = \mathbf{I}_K, h_{ik} \in \{\frac{0}{\sqrt{N_k}}, \frac{1}{\sqrt{N_k}}\}. \end{aligned} \tag{2.4}$$

Equation (2.4) can be viewed as another way to formulate the objective of KM.

Instead of directly solving Equation (2.4), Ding and He (2004) proposed to first address a more convenient problem by releasing the constraint that h_{ik} should be either 0 or $\frac{1}{\sqrt{N_k}}$. To do so, they introduced $\hat{\mathbf{H}}$ as the continuous relaxation of $\mathbf{H}$ that satisfies $\hat{\mathbf{H}} = \mathbf{H}\mathbf{R}$ where $\mathbf{R}$ is a rotation matrix subject to $\mathbf{R}\mathbf{R}' = \mathbf{I}_K$. Also, to illustrate more explicitly the connection of Equation (2.4) and PCA, $\mathbf{Z} = \mathbf{X}'$ is brought in. Then, Equation (2.4) could be rephrased in

$$\begin{aligned} &\mathbf{argmax}_{\hat{\mathbf{H}}}\ Tr\hat{\mathbf{H}}'\mathbf{Z}'\mathbf{Z}\hat{\mathbf{H}} \\ &s.t. \hat{\mathbf{H}}'\hat{\mathbf{H}} = \mathbf{I}_K, \end{aligned} \tag{2.5}$$

which is the PCA formulation yet formulated on the transposed data. A solution is attained when $\hat{\mathbf{H}}$ equals the first K left eigenvectors of $\mathbf{Z}'\mathbf{Z}$ that correspond to the K largest eigenvalues. Xu et al. (2015) proposed to estimate the partition matrix $\mathbf{H}$ from this K-dimensional representation of the data with a two-step approach: (1) obtain an initial partition by employing a multi-start KM algorithm

on $\hat{\mathbf{H}}$; (2) use the partition resulting from the first step as a rational start for a KM analysis of the original data $\mathbf{X}$.

We note that the objective in Equation (2.5) can also be written as

$$\begin{aligned} &\mathbf{argmin}_{\hat{\mathbf{H}},\mathbf{P}} ||\mathbf{X} - \hat{\mathbf{H}}\mathbf{P}'||_2^2 \\ &s.t. \hat{\mathbf{H}}'\hat{\mathbf{H}} = \mathbf{I}_K, \end{aligned} \tag{2.6}$$

where $\mathbf{P}$ serves as the loading matrix and the expression can be seen as the least-squares formulation of PCA (for more details, the reader is referred to Guerra-Urzola et al., 2021). In Equation (2.6), if the t^{th} row in $\mathbf{P}$ contains all zero elements, the t^{th} variable does not contribute to cluster separation and is therefore viewed as an irrelevant variable. Therefore, the contribution of the variables can be obtained by controlling $\mathbf{P}$, e.g., by regularizing the variable contributions such that variables that are not associated with cluster separation are regularized to have only zero loadings. This forms the basis for the development of CKM, as described below.

A SPARSE PCA APPROACH TO SOLVE KM IN THE PRESENCE OF IRRELEVANT VARIABLES

Let us reconsider the cluster analysis of $\mathbf{X}$ and assume that, out of all J variables, a total of V variables are irrelevant variables that do not separate clusters. The remaining $(J-V)$ variables are therefore signaling variables. The vector $\mathbf{g}$ contains the indices of all V irrelevant variables, while $\mathbf{X}_\mathbf{g}$ and $\mathbf{X}_{-\mathbf{g}}$ denote the subsets of the original data set that involves only the irrelevant and signaling variables, respectively. In light of Equation (2.1) and Equation (2.3), we define the objective of KM in the presence of V irrelevant variables:

$$\begin{aligned} &\mathbf{argmin}_{\mathbf{c},\mathbf{g}} (||\mathbf{X}_\mathbf{g}||_2^2 + \sum_{k=1}^{K} \sum_{i \in \mathbf{c}^{-1}(k)} \sum_{j \notin \mathbf{g}} (x_{ij} - m_{kj})^2 \\ &with\ m_{kj} = \frac{1}{N_k} \sum_{i \in \mathbf{c}^{-1}(k)} x_{ij}, \end{aligned} \tag{2.7}$$

where x_{ij} and m_{kj} are the individual score of subject i and the mean score of cluster k on variable j, respectively. The objective represented by Equation (2.7) is to minimize the total within-cluster sum of squares (also called within-SS)

across all observations and variables. The first term, $||\mathbf{X_g}||$, summarizes the within-SS over all irrelevant variables. To see this, note that a variable is considered irrelevant if its cluster-specific centroids are assumed equal; hence, these centroids are further equal to the grand mean (i.e., 0, since all variables are column-wise centered). The second term of Equation (2.7) calculates the within-SS over all signaling variables. Note that $\mathbf{g}$ is added as a parameter over which Equation (2.7) is optimized.

For the second part of Equation (2.7), with a set of operations similar to those listed in Appendix 2.A and 2.B, we obtain an equivalent formulation

$$\begin{aligned} &\mathbf{argmax}_{\mathbf{H},\mathbf{g}} Tr\mathbf{H}'\mathbf{X_{-g}}\mathbf{X_{-g}}'\mathbf{H} \\ &s.t. \mathbf{H}'\mathbf{H} = \mathbf{I}_K, h_{ik} \in \{\frac{0}{\sqrt{N_k}} \frac{1}{\sqrt{N_k}}\}, \end{aligned} \tag{2.8}$$

where $\mathbf{g}$ contains V irrelevant variables and $\mathbf{X_{-g}}$ denotes the subset of the original data set that only contains signaling variables. In the next section, We propose a set of procedures to determine V. Again, $\hat{\mathbf{H}}$, the continuous relaxation of $\mathbf{H}$, can be used to replace $\mathbf{H}$ in Equation (2.8), resulting in

$$\begin{aligned} &\mathbf{argmax}_{\hat{\mathbf{H}},\mathbf{g}} Tr\hat{\mathbf{H}}'\mathbf{X_{-g}}\mathbf{X_{-g}}'\hat{\mathbf{H}} \\ &s.t. \hat{\mathbf{H}}'\hat{\mathbf{H}} = \mathbf{I}_K. \end{aligned} \tag{2.9}$$

Furthermore, in the same vein as Equation (2.6), Equation (2.9) can be reframed as a minimization problem. Adding the first part of Equation (2.7), we obtain an optimization problem

$$\begin{aligned} &\mathbf{argmin}_{\hat{\mathbf{H}},\mathbf{P}} ||\mathbf{X} - \hat{\mathbf{H}}\mathbf{P}'||_2^2 \\ &s.t. \hat{\mathbf{H}}'\hat{\mathbf{H}} = \mathbf{I_k}, \sum_{j=1}^{J} [\mathbf{row}(\mathbf{P})_j = 0] = V, \end{aligned} \tag{2.10}$$

where $\mathbf{row}(\mathbf{P})_j$ indicates the j^{th} row of the loading matrix $\mathbf{P}$ and $[.]$ refers to the Iverson bracket: $[Q] = 1$ if Q is true and $[Q] = 0$ if Q is false. Equation (2.10) can be solved with a modification of the SPCA algorithm introduced by Adachi and Trendafilov (2016). Similar to the proposal in Xu et al. (2015), a KM analysis is then performed on $\hat{\mathbf{H}}$, resulting in an initial partition, $\mathbf{c_0}$, that is used for com-

puting the final solution of the CKM analysis. Furthermore, the SPCA analysis produces an initial set of irrelevant variables **g** by selecting variables whose loadings on K components all equal zero. Subsequently, following a similar strategy as SKM and SAS, and as detailed in the next section, **c** and **g** are updated iteratively to solve Equation (2.7). [1]

2.2.2. ALGORITHM

In this section, we present the details of the algorithm for CKM with the number of clusters K and irrelevant variables V assumed to be known. The discussion on how to select K and V is deferred to section 2.4. In essence, the algorithm consists of two parts. First, the sparse PCA problem defined by Equation (2.10) is solved with a modified version of Unpenalized Sparse Loading PCA (USLPCA; Adachi and Trendafilov, 2016). The modified version revises the structure of the imposed cardinality constraint so that the algorithm returns a selection of variables across all components (instead of per component). This optimization procedure is used because it has proven to be one of the most efficient algorithms to solve the SPCA problem with loading matrices subject to a cardinality constraint. Therefore, the result of this modified procedure is an accurate and efficient solution to the optimization problem presented in Equation (2.10). From this procedure, the initial set of irrelevant variables $\mathbf{g}_0$ is obtained. Furthermore, the initial indicator vector $\mathbf{c}_0$ is obtained by performing a multi-start KM analysis on the component scores estimated from SPCA. In the second part, we solve the sparse KM problem defined in Equation (2.7) by updating **c** and **g** iteratively. Both USLPCA and the sparse KM procedure are of an alternating least squares type and, in practice, they both converge to a local optimum. The full algorithm is presented in the form of pseudocode in Algorithm 1. In Appendix (c), we show the derivation behind the optimization of $\hat{\mathbf{H}}$.

Here are four remarks on Algorithm 1. First, we solve the sparse PCA problem formulated in Equation (2.10) with one rational start based on the singular value decomposition of **X**. This choice was made because this step is computationally demanding, and, in our experiments, increasing the number of starts

[1] We have also tested the direct use of **c** and **g** as the partition of the samples and set of irrelevant variables, respectively. This procedure gave unsatisfactory results.

Algorithm 1: The CKM algorithm

Input : the data matrix $\mathbf{X}(N \times J)$, the number of clusters K, the number of irrelevant variables V, the convergence criteria ϵ and the maximal number of iterations $iter_{max}$

Output: the set of irrelevant variables $\mathbf{g}$ and the indicator vector $\mathbf{c}$

Initialize $\hat{\mathbf{H}} = \mathbf{U}_K$ and $\mathbf{P} = \mathbf{V}_K \Sigma_K$ where $\mathbf{U}_K \Sigma_K \mathbf{V}_K'$ is the rank-k truncated SVD solution of $\mathbf{X}$

Initialize the current number of iterations $iter = 0$

Initialize $L = \Delta L = ||\mathbf{X} - \hat{\mathbf{H}}\mathbf{P}'||_2^2$

while $\Delta L > \epsilon$ *and* $iter < iter_{max}$ **do**

  Update $\hat{\mathbf{H}} = \mathbf{V}\mathbf{U}'$ where $\mathbf{U}$ and $\mathbf{V}$ are obtained from the SVD solution of $\mathbf{P}'\mathbf{X}'$

  Update $\mathbf{P}$ with two steps: (1) $\mathbf{P} = \mathbf{X}'\hat{\mathbf{H}}$ and (2) set the V rows of $\mathbf{P}$ with the smallest sum-of-squares to zero

  Update $\Delta L = L - ||\mathbf{X} - \hat{\mathbf{H}}\mathbf{P}'||_2^2$

end

Initialize $\mathbf{g}$ with the indices of the rows having only zero loadings in $\mathbf{X}_{-\mathbf{g}}$

Initialize $\mathbf{c}$ as the result of a KM analysis with multiple starts on $\hat{\mathbf{H}}$

Initialize $L = \Delta L = \mathbf{argmin}_{\mathbf{c},\mathbf{g}}(||\mathbf{X}_{\mathbf{g}}||_2^2 + \sum_{k=1}^{K} \sum_{i \in \mathbf{c}^{-1}(k)} \sum_{j \notin \mathbf{g}} (x_{ij} - m_{kj})^2)$

while $\Delta L > \epsilon$ **do**

  Update $\mathbf{g}$, conditional on $\mathbf{c}$, by maximizing $\sum_{k=1}^{K} \sum_{i \in \mathbf{c}^{-1}(k)} \sum_{j \notin \mathbf{g}} (x_{ij}^2 - (x_{ij} - m_{kj})^2)$

  Update $\mathbf{c}$, conditional on $\mathbf{g}$, by a KM analysis on $\mathbf{X}_{-\mathbf{g}}$ with the current $\mathbf{c}$ as the (single) informative start

  Update $\Delta L = L - (||\mathbf{X}_{\mathbf{g}}||_2^2 + \sum_{k=1}^{K} \sum_{i \in \mathbf{c}^{-1}(k)} \sum_{j \notin \mathbf{g}} (x_{ij} - m_{kj})^2)$

end

only marginally improved the performance of the algorithm. Second, a (standard) KM analysis with 10 random starts is proposed to obtain the initial cluster partition from the matrix $\mathbf{X_{-g}}$ with the initial set of signaling variables (i.e., $\mathbf{g}$) obtained from SPCA. Third, if cluster recovery - but not computational efficiency - is of concern, then an additional KM analysis with 10 starts can be conducted on the subset of the data set with only the selected signaling variables. The loss value from this additional analysis can then be compared to the original loss value and a final solution can be determined that minimizes this loss value. Fourth, to update the index vector of the irrelevant variables $\mathbf{g}$, we propose to maximize $(\sum_{k=1}^{K}\sum_{i\in\mathbf{c}^{-1}(k)}\sum_{j\notin\mathbf{g}}(x_{ij}^2-(x_{ij}-m_{kj})^2)$, conditional on $\mathbf{c}$. This can be conveniently solved by selecting the V variables corresponding to the V largest values computed from $\sum_{k=1}^{K}\sum_{i\in\mathbf{c}^{-1}(k)}(x_{ij}^2-(x_{ij}-m_{kj})^2)$.

2.2.3. RELATED METHODS

As discussed in the introduction, other algorithms that are developed from KM have been proposed to perform cluster analysis in the presence of a large number of variables. These methods could be generally classified into three types: dimension reduction, subspace clustering and variable selection. Our proposed CKM falls into the category of variable selection methods. Therefore, in the current paper, we only consider other methods from this category. Readers who might be interested in a broad review of all existing methods are referred to review articles and textbooks, for example, Bouveyron and Brunet-Saumard (2014) and Bouveyron et al. (2019).

Sparse K-means (Witten & Tibshirani, 2010) was built upon the weighted k-means framework (Tseng, 2007) where a weight is assigned to each variable to quantify the relative importance of the variable. The objective function of SKM can be formulated in

$$\begin{aligned}&\mathbf{argmax}_{\mathbf{c},w_1,\ldots,w_J}\sum_{j=1}^{J}w_j\sum_{k=1}^{K}\sum_{i\in\mathbf{c}^{-1}(k)}(x_{ij}^2-(x_{ij}-m_{kj})^2)\\&s.t.\ w_j\geq 0,||\mathbf{w}||_2^2\leq 1,||\mathbf{w}||_1\leq s\end{aligned}\tag{2.11}$$

where w_j denotes the weight associated with the variable j, $||\mathbf{w}||_1 = \sum_{j=1}^{J} |w_j|$ refers to the l_1 norm, and s is the hyper-parameter that is determined during model tuning.

As illustrated in Equation (2.11), to achieve variable selection, SKM includes a constraint with an l_1 norm and a constraint with an l_2 norm on the weights. The former enforces some of the weights to become exactly zero, indicating that the corresponding variables of these weights do not contribute to cluster separation. The latter prevents putting all the weights on only one or a small set of variables for which the separation of the clusters is the largest. To solve Equation (2.11), an alternating algorithm is developed that updates the weights and the cluster assignments iteratively. Typically, a set of equal weights is used to initialize the algorithm.

When tested on simulated data, SKM enjoyed a clear advantage over KM in terms of the accuracy of cluster recoveries, for data sets with a large proportion of irrelevant variables. However, it performed slightly worse than KM when the vast majority of variables were signaling variables.

Inspired by SKM, Arias-Castro and Pu (2017) proposed SAS, which applies a similar model as SKM, except for the fact that SAS uses binary weights w_j: $w_j = 1$ indicates that the j^{th} variable is included in determining the cluster structure (i.e., the variable is regarded as a signaling variable) while $w_j = 0$ indicates that it is an irrelevant variable. Similar to SKM, an alternating estimation procedure has been proposed that updates the weights and the cluster assignments iteratively; the authors suggested to initialize the procedure with multiple sets of randomly selected variables. [2] In simulation studies, compared to SKM, SAS took considerably less time to achieve better performance in terms of cluster recovery in most scenarios. However, its edge over SKM in cluster recovery vanished when a vast majority of variables were irrelevant variables. We argue this is probably because the initial set of signaling variables used in SAS analysis is often far from the underlying model. CKM, on the other hand, uses initial values that stem from a sparse SPCA analysis of the original data; as a result, the starting set of signaling variables should be much closer to the underlying model. Therefore, we expect

[2] They also reported other suggestions for initialization; yet these ways of initialization all lead to similar results.

CKM to outperform SAS especially when the data set under consideration involves a large proportion of irrelevant variables.

2.2.4. MODEL SELECTION

One of our contributions is to propose a novel procedure to select K while taking the presence of irrelevant variables into account; in the current section, we introduce this procedure in detail. Despite the fact that numerous criteria and procedures have been proposed to select K in deterministic clustering algorithms in general (some of the best-performing algorithms include Tibshirani et al. (2001) and J. Wang (2010); see Steinley (2006) for a comprehensive review), it is still largely unclear how the selection of the number of clusters should be done for these methods in the presence of irrelevant variables. In previous studies, the common practice was to apply a specific criterion on the full data set, as if irrelevant variables did not influence the selection of the optimal number of clusters. We argue, however, this procedure will likely result in selecting a wrong number of clusters when a majority of variables are irrelevant and may therefore hamper an accurate recovery of the clusters. Therefore, we propose a novel strategy that filters out irrelevant variables as much as possible before selecting K. The procedure applies a three-step procedure, as follows. In the first step, for each possible number of clusters K (K = 1,2,...,K_{max}), the optimal number of irrelevant variables V_K as well as the subset of signaling variables $\mathbf{s}_K$ are determined. Second, a set of variables - called the stable set or $\mathbf{s}_{stable}$ - are obtained that are considered signaling variables over different values of K. In the third step, the optimal value of K (denoted by K_{opt}) is determined while the associated V_K and $\mathbf{s}_K$ - computed during the first step - are retrieved as the optimal value of V and the optimal set of signaling variables, respectively.

We now first introduce the procedure to select V_K and $\mathbf{s}_K$ with a pre-determined value of K. The procedure is based on the Gap Statistic (Tibshirani et al., 2001), which has demonstrated good performance in selecting the number of clusters in previous studies (e.g., Arias-Castro and Pu, 2017). More specifically, for each possible value of the number of irrelevant variables V ($V = 0,1,2,\ldots,J-2$), a CKM analysis is conducted on $\mathbf{X}$. Note, we recommend including at least two signaling variables to avoid identification problems. From the analysis, the set of sig-

naling variables $\mathbf{s}_K(V)$ is selected and its corresponding between-cluster sum of squares is calculated as $O(V)$. Then, B random data sets are generated based on the subset $\mathbf{s}_K(V)$ by independently permuting the observations within each variable. For each of the permuted data sets, a KM analysis is conducted, from which the between-cluster sum of squares is recorded as $O_b(V)$. Consequently, the Gap statistic is defined in

$$Gap(V) = log\, O(V) - \frac{\sum_{b=1}^{B} log\, O_b(V)}{B}. \tag{2.12}$$

The intuition is that, as the permuted data contain no clusters, a larger value of $Gap(V)$ indicates a more salient cluster structure. Therefore, the value of V that maximizes $Gap(V)$ is selected. The corresponding set of signaling variables is consequently picked up as $\mathbf{s}_K$.

As the set of estimated irrelevant variables at each value of K likely differs, we identify a set of variables - the stable set of variables $\mathbf{s}_{stable}$ - that are consistently selected as signaling variables regardless of the value of K. More formally, $\mathbf{s}_{stable}$ is calculated as follows: $\mathbf{s}_{stable} = \cap_{K=2}^{K_{max}} \mathbf{s}_K$, where $\cap$ denotes the operation of extracting the intersection over all vectors. The resulting subset of variables $\mathbf{s}_{stable}$ hence consists of signaling variables that were consistently identified as relevant to cluster separation for each and every value of K.

Once the stable set of signaling variables is determined, existing criteria to determine K can be used. Given the promising performance of the Gap statistic in recovering the true number of clusters in previous research, the Gap statistic is set as the default criterion in the implementation of our model selection procedure. However, other popular indices such as the KL index (Krzanowski & Lai, 1988) and the Dindex (Lebart et al., 1995) are interesting alternatives. In simulation study 3 described below, we assessed the performance of these criteria in terms of the accuracy in recovering the true number of clusters K across various conditions.

Last, to make the selection of V more precise, an additional step is recommended. This additional step determines the value of V from a set of candidates that are located around the selected V resulting from the previous step based on the Gap statistic. With respect to the size of the set of candidates, according

to our experience, a set of 10 alternative values is generally sufficient for the task. Specifically, the between-cluster sum-of-squares is calculated for each candidate value and an elbow point is determined to be the optimal value of V. [3]

A potential risk of deriving a stable set of variables in this way is that too many variables have been left out. Nevertheless, our experience in analyzing simulated and empirical data sets is that as long as K_{max} is set at a reasonable value, the identified $\mathbf{s}_{stable}$ always contains an adequate set of variables for selecting K.

Algorithm 2 summarizes the proposed model selection procedure that consists of the selection of the number of clusters K, and the set of signaling variables.

When the number of variables J is small, it is feasible to search the full grid (i.e. from 1 to $J-2$) in selecting V_{opt}. However, this approach is computationally prohibitive with a large J (e.g., $J > 100$). Thus, in these cases, an adaptive grid search algorithm that progressively zooms in on smaller areas in the solution space is employed that effectively reduces the computational demand while maintaining reasonable accuracy. More specifically, this "zoom-in" strategy is an iterative procedure that gradually narrows the search space for the number of signaling variables until it converges to a single number. The algorithm starts with 10 evenly-spaced numbers ($a_1 < a_2 < \ldots < a_{10}$), where a_1 takes the smallest possible value and a_{10} takes the largest possible value. For each of these 10 candidate numbers of signaling variables, a CKM solution is obtained and the optimal number is selected with the Gap statistic. The algorithm then zooms in to $[a_{i-1}+1,\ a_{i+1}-1]$ (both sides included) and creates 10 new evenly spaced numbers. This step is repeated until convergence.

2.3. SIMULATION STUDIES

To evaluate the performance of CKM and of the proposed model selection strategy, three simulation studies were carried out. In the first two simulation studies, we compared the performance of CKM in recovering the clusters and the status of the variables (signaling versus irrelevant) with that of SAS and of SKM. The

[3] Alternatively, this optimal value can be found automatically by identifying the global or local maximum of scree ratios. See an illustration in De Roover, Ceulemans, Timmerman, Vansteelandt, et al. (2012).

Algorithm 2: Proposed procedure to determine V and K

Input : the data matrix $\mathbf{X}$, the maximal number of clusters K_{max}, the number of permutation samples B

Output: the optimal number of clusters K_{opt}, the optimal number of irrelevant variables V_{opt}, and the selected set of signaling variables $\mathbf{s}_{opt}$

for $K = 2$ **to** K_{max} **do**
  for $V = 1$ **to** $J - 2$ **do**
    Run **Algorithm** 1 with K and V. Denote the resulting between-cluster sum of squares by $O(V)$ and the set of signaling variables by $\mathbf{s}_K(V)$
    Obtain the subset of $\mathbf{X}$ that contains only the signaling variables
    for $b = 1$ **to** B **do**
      Randomly permute the values of each variable in the above subset
      Run KM on the permuted data set, resulting in $O_b(V)$
    end
    Compute $Gap(V)$: $Gap(V) = log\, O(V) - \frac{\sum_{b=1}^{B} log\, O_b(V)}{B}$
  end
  Set V_K equal to the V that maximizes $Gap(V)$, while $\mathbf{s}_K$ denotes the corresponding set of signaling variables
end
Obtain $\mathbf{s}_{stable}$: $\mathbf{s}_{stable} = \cap_{K=2}^{K_{max}} \mathbf{s}_K$
Determine K_{opt}: use a criterion (e.g., the Gap statistic) to determine the number of clusters based on the subset of $\mathbf{X}$ (i.e., only those variables whose indices are in $\mathbf{s}_{stable}$)
Update $V_{opt} = V_{K_{opt}}$. Update $\mathbf{s}_{opt} = \mathbf{s}_{K_{opt}}$
NOTE: The following step is an optional step, and it is only recommended when K_{opt} is large (e.g., > 20).
for $V = V_{opt} - 5$ **to** $V = V_{opt} + 5$ **do**
  Run Algorithm 1 with K_{opt} and V, and obtain $O(V)$ and $\mathbf{s}_{K_{opt}}(V)$.
end
Determine the elbow point on the resulting sets of $O(V)$, and update V_{opt}. Update $\mathbf{s}_{opt} = \mathbf{s}_{K_{opt}}(V_{opt})$.

two simulation studies differed in the amount of prior information: while both K (i.e., the number of clusters) and V (i.e., the number of irrelevant variables) were assumed to be known in simulation study 1, only the true value of K was provided in simulation study 2. In addition to SAS and SKM, in simulation study 2, CKM was also compared to KM. In simulation study 3, our proposed strategy that relies on the stable set of signaling variables for selecting the number of clusters and identifying the set of signaling variables was compared to the alternative - and widely applied - selection strategy that selects K based on the full set of variables.

All of the analyses were carried out in the statistical software R. We used our self-developed package `CKM` for the CKM algorithm, the package `stats` for the KM algorithm, and the package `sparcl` for the SKM algorithm. The SAS algorithm was available from standalone functions that were extracted from the Github page (see Arias-Castro and Pu, 2017). When running CKM, SAS, and SKM in simulation studies 2 and 3, one hyper-parameter must be tuned for each method to select the optimal number of signaling variables. For CKM, we have elaborated the procedure to tune the cardinality constraint in section 2.2.4. The procedure to tune the hyper-parameter for SAS is similar to that for CKM: according to Arias-Castro and Pu (2017), here too the optimal number of signaling variables is determined by maximizing the Gap statistic calculated from Equation (2.12). For SKM, the tuning parameter s, associated with the l_1 norm, should be decided for each of the simulations. s is tuned from a grid consisting of 200 evenly spaced values ranging from 1.001 to 10. For simulation study 1 where the number of irrelevant variables V is known prior to data analysis, we first determine the number of irrelevant variables V_0 for each value s_0 on the grid. Then, the tuning parameter s is selected such that its corresponding V_0 equals V. In case multiple V_0 equal V, the average value of their associated s_0 is used. For simulation studies 2 and 3 where V is determined during data analysis, the optimal value is selected that results in the simplest model (i.e., the model with the fewest number of signaling variables) with a Gap statistic less than $1SE$ away from the maximum. In other words, the tuning procedure for SKM follows the well-known

$1SE$ rule, as proposed in Witten and Tibshirani (2010). [4] In the above tuning process, the Gap statistic must be computed for each candidate value; here, we set the number of permutation samples to 20 for all analyses.

2.3.1. SIMULATION STUDY 1

In this simulation study, we compared the accuracy of CKM in recovering the clusters and signaling variables with SAS and SKM; where the values of K and V were set at pre-defined values. To facilitate a systematic comparison with other studies, we adopted, as closely as possible, the data generation procedure from Witten and Tibshirani (2010) and Arias-Castro and Pu (2017). More specifically, the simulation was designed as follows: (1) the number of clusters K was either 3, 5, or 30; (2) the number of observations per cluster was 50; (3) the number of irrelevant variables V took one of the following four values: 5, 50, 250, and 1000; (4) the number of signaling variables (i.e., $J - V$) was 50 and (5) the distance of centroids for each variable between neighboring clusters $\Delta\mu$ equaled one of the following four values: 0.6, 0.7, 0.8, 1. A fully crossed design was used, resulting in $3 \times 1 \times 4 \times 1 \times 4 = 48$ conditions.

To generate the data, each observation was assigned to one of the K clusters such that all clusters were of equal size. Then, irrelevant variables were generated by drawing from the standard normal distribution. The responses on the signaling variables were sampled independently for each cluster from a normal distribution with a cluster-specific mean and a standard deviation of 1. The cluster-specific mean values were determined such that the grand mean calculated over all clusters was 0 while differences in neighboring clusters were fixed at $\Delta\mu$. For example, when $\Delta\mu$ equaled 0.6, the cluster-specific mean values of the three clusters for each variable were respectively -0.6, 0, and 0.6. Obviously, a smaller $\Delta\mu$ corresponds to closer cluster centroids, and thus results in a more difficult task to recover the clusters.

For each condition, 40 data sets were generated. Therefore, a total of 1920 data sets were generated and analyzed by CKM, SAS, and SKM. Note that, SKM

[4] Note that, for SKM, we also tried in a small-scale simulation to determine the hyperparameter by maximizing the Gap statistic; however, the results of the simulation were more in favor of the selection with the 1SE rule.

was eventually dropped for the data sets generated in the conditions with 30 clusters because of its slow computation.

Following Chipman and Tibshirani (2006), Witten and Tibshirani (2010), and Arias-Castro and Pu (2017), we used classification error (CE) as the evaluation criterion of cluster recovery. By reporting CE, we hope to provide future research with a consistent point of comparison, which is particularly beneficial for studies where different methods are synchronized and (or) compared. CE indicates the similarity between the true cluster assignment $\mathbf{c_{true}}$ and the assignment $\mathbf{c_{est}}$ resulting from a particular clustering algorithm. To illustrate, we introduce the following notation: $1_{\mathbf{c}(i,i')}$ equals 1 when observations i and i' belong to the same cluster and 0 when they do not. Then, CE is defined as follows,

$$CE = \frac{\sum_{i>i'} |1_{\mathbf{c_{true}}(i,i')} - 1_{\mathbf{c_{est}}(i,i')}|}{N(N-1)/2}, \tag{2.13}$$

where N is the total number of observations.

CE in Equation (2.13) takes values between 0 and 1; CE=0 indicates a perfect agreement between $\mathbf{c_{true}}$ and $\mathbf{c_{est}}$ while a higher value indicates a larger classification error and thus less agreement between these two partitions.

Furthermore, to quantify how well an algorithm retrieved the signaling variables, we computed the proportion of true signaling variables that were successfully identified by the algorithm relative to the total number of signaling variables (e.g., if 40 of the 50 signaling variables have been identified, the success rate will be 80%). Hence, a larger proportion suggests a better performance of the algorithm in detecting the signaling variables.

The relative performance of CKM, SAS, and SKM in recovering the clusters are visualized in Figure 2.1. Figures 2.1A and 2.1B shows that, when K equaled 3 or 5, CKM and SAS recovered the clusters equally well (for both methods, average CE = .012 when $K = 3$; average CE = .014 when $K = 5$) and both better than SKM (average CE = .025 when $K = 3$; average CE = .021 when $K = 5$). Furthermore, CKM (average CE = .092) outperformed SAS (average CE = .109) when $K = 30$ (see Figure 2.1C; note that, as discussed earlier, SKM was dropped in these conditions), i.e., in the presence of a more complex cluster structure.

Next, we examined how well the three methods were able to identify the set of signaling variables. We found that the task of identifying the set of signaling

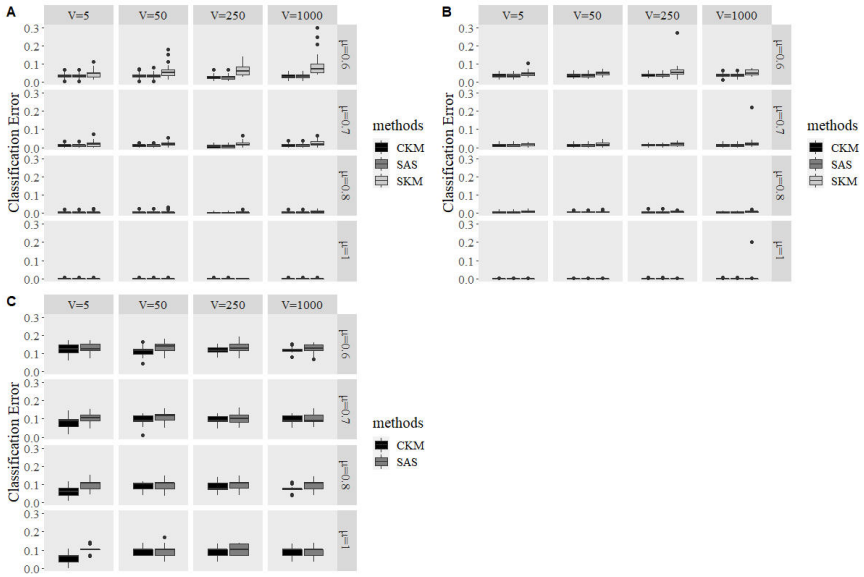


Figure 2.1: A comparison of different clustering methods for cluster recovery when both the number of clusters K and signaling variables V are given. Panel A: K = 3; Panel B: K = 5; Panel C: K = 30.

variables proved to be relatively easy given the true values of both S and K: all three methods were able to identify the set of signaling variables with a success rate of at least 99%.

2.3.2. SIMULATION STUDY 2

Our objective in simulation study 2 was to further examine the relative performance of CKM, compared to SAS and SKM, in recovering clusters and the status of variables when only K was given; hence, V as well as the subset of signaling variables had to be determined by the algorithm. Furthermore, we have also added (standard) KM - the most commonly used algorithm that does not allow for variable selection - to the comparison and evaluated the relative performance of all four methods in terms of cluster recovery. The settings and the data generation procedure were identical to those used in simulation study 1.

In simulation study 2, again a total of 48 conditions were manipulated with 40 data sets each. This resulted in a total of 1920 data sets. We assessed the performances of the four clustering algorithms primarily based on the recovery

of clusters (indicated by CE) and the number of variables identified as signaling variables. In addition, we also recorded and compared the average running time for each of the methods.

Figures 2.2A and 2.2B visualize the extent of cluster recovery by the different methods, when K equaled 3 and 5, respectively. Because the two subplots present a similar pattern of the relative performance of the four methods (CKM, SAS, SKM, and KM), we discuss the combined results here. Averaged over all conditions, CKM was the winner with an average CE of .013, followed by SAS (average CE = .016) and SKM (average CE = .023). KM, on average, produced cluster partitions with a CE equaling .10. With regard to the effect of $\Delta\mu$, the largest advantage of CKM (average CE = .035) over the other four algorithms (for SAS, average CE = .045; for SKM, average CE = .057; for KM, average CE = .16) was found when $\Delta\mu = .6$ (i.e., the smallest distance of centroids between neighboring clusters). We also examined how well these methods recovered clusters with respect to the different numbers of irrelevant variables (i.e., V). In accordance with our expectation, the three methods performing simultaneous variable selection and clustering (i.e., CKM, SAS, and SKM; for CKM, average CE = .014; for SAS, average CE = .021; for SKM, average CE = .024) recovered the clusters considerably better than KM (average CE = .26) in the presence of an exceedingly large proportion of irrelevant variables (i.e. $V = 1000$). Last, in accordance to our expectation, the performance advantage of CKM over SAS and KM in terms of cluster recovery was greatest when $K = 30$ (see Figure 2.2C; for CKM, average CE = .08, for SAS, average CE = .30, for KM, average CE = .70). This again illustrates that CKM is particularly powerful to deal with a complex cluster structure. When $K = 30$ and $V = 1000$, the difference in cluster recovery from the three methods is striking: the average CEs for CKM, SAS, and KM were .09, .28, and .77, respectively.

We further evaluated how well the algorithms identified the set of 50 signaling variables when the correct number of irrelevant variables (i.e., V) was not given. Since KM is not able to explicitly single out signaling variables, the comparison only concerns CKM, SAS, and SKM - note that the true value was always 50. The results, plotted in Figure 2.3, show that CKM was the best-performing method in terms of successful variable selection, since the number of variables selected by CKM was consistently close to 50, even with $V = 1000$. In contrast,

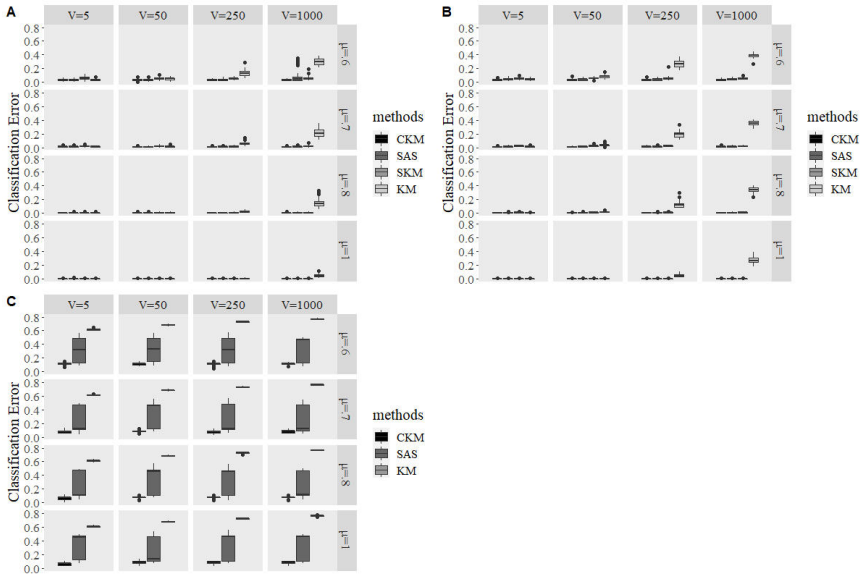


Figure 2.2: A comparison of different clustering methods for cluster recovery when only the number of clusters K is given. Panel A: K = 3; Panel B: K = 5; Panel C: K = 30.

with a larger number of irrelevant variables (i.e., V = 250 or 1000), both SAS and SKM experienced difficulty to recover the exact 50 signaling variables. Expressed in numbers, CKM recovered the exact 50 variables in 92.7 % of the cases; for SAS and SKM, this percentage of successful recoveries was only 62.9 % and 30 %, respectively.

Last, we examined the average execution time for each of the clustering methods (here, we only consider K = 3 and K = 5, because these are the typical scenarios behavioral researchers commonly encounter). With an average execution time of .16 seconds and 4.28 seconds, respectively, KM and SAS were the two fastest algorithms. CKM ranked third among all four methods, taking an average of 43.5 seconds to analyze a data set. In our opinion, its speed is acceptable for most empirical studies. SKM, with an average of 293.6 seconds, was a lot slower than the other three algorithms.

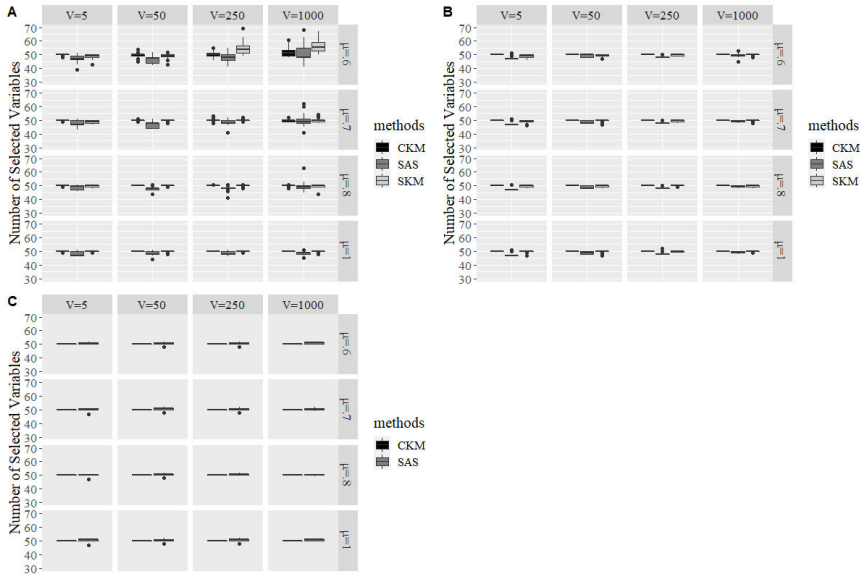


Figure 2.3: A comparison of different clustering methods for variable selection (true number of signaling variables = 50). Panel A: K = 3; Panel B: K = 5; Panel C: K = 30.

2.3.3. SIMULATION STUDY 3

Our major objective in simulation study 3 was to evaluate and compare different model selection procedures for deterministic clustering algorithms that perform simultaneous clustering and variable selection (e.g., CKM, SAS, and SKM). To achieve this, we examined the relative accuracy of selecting K with regard to (1) the set of variables used (i.e., either relying on a stable set of variables that were selected consistently across all possible numbers of clusters or the full set of variables), and (2) the selection criteria for determining the number of clusters.

A key interest in the current comparison was to compare our novel strategy that pre-selected a stable set of variables (see the previous section) with the traditional strategy that involved all variables. Our expectation was that, with a relatively large proportion of irrelevant variables, the traditional strategy considered too much noisy information and therefore resulted in a less accurate selection compared to our novel strategy. Besides, we have also implemented and tested another strategy - called the local selection strategy. This strategy first selects V

conditional upon each possible value of K with the $Gap(V)$ statistic and then selects K that maximizes the associated $Gap(V)$ statistic. However, in all conditions, this strategy consistently selected the smallest value of K (i.e., 2). Because of the poor performance of this strategy, we do not report its results any further in the current study.

In the current study, we considered some of the most popular model selection criteria, namely the "KL Index" (Krzanowski & Lai, 1988), the "DIndex" (Lebart et al., 1995), and two versions of the Gap statistics (Tibshirani et al., 2001), and examined which selection criteria determined K with the highest accuracy. Specifically, in the current study, the following two Gap-based criteria were investigated: 1) selecting K that corresponded to the global maximum of the Gap statistic, called "globalGap"; and 2) choosing K that was associated to the first local maximal value of the Gap statistic, called "firstGap". While the first one was proposed in Tibshirani et al. (2001), the second one was introduced in Maechler et al. (2012) in developing the well-known R package `Cluster`.

Furthermore, in the current study, to evaluate the generalizability with respect to the preferred selection strategy and selection criterion, we replicated our findings with both CKM and SAS (SKM was not involved because, as illustrated above, it was relatively slow compared to CKM and SAS).

To summarize, in simulation study 3, we tested the accuracy of selecting K with respect to three factors: (1) the selection strategy (i.e., the proposed strategy that utilizes a stable set of variables versus and a strategy that utilizes the full set of variables), (2) the selection criterion (i.e., "globalGap" v.s. "firstGap" v.s. "KL Index" v.s. "DIndex"), and (3) the clustering algorithm (i.e., CKM v.s. SAS).

A number of factors in the data generation process were systematically manipulated. These were largely identical to those of the first two simulation studies, yet, with the following exception. Namely, the varying number of clusters K was one of three values: 3, 5, or 15. Again, in total $3 \times 4 \times 4 = 48$ conditions were manipulated. For each of the conditions, again 40 replicate data sets were generated, leading to a total of 1920 data sets. For each data set, K was selected among models with 2 up to 10 clusters when $K = 3$ or $K = 5$ and among models with 11

up to 19 clusters when $K = 15$. [5] Specifically, three model selection strategies (i.e., utilizing the stable set of variables obtained from (1) CKM, or (2) SAS, and (3) utilizing the full set of variables) combined with four model selection criteria (i.e., (1)"globalGap", (2) "firstGap", (3)"KL Index", and (4) "DIndex") were employed to analyze each of the data sets. That is, for each data set, we applied a total of 12 different ways for selecting the number of clusters K.

Table 2.1 presents the results of simulation study 3. Most importantly, the novel selection strategy for selecting the number of clusters that relies on the stable set of variables led to an equal or higher success rate in selecting the true number of clusters, across all criteria and conditions, and both for CKM and for SAS, in comparison with using the full set of variables. This advantage was especially pertinent in the presence of a large proportion of irrelevant variables (i.e. when $V = 250$ or $V = 1000$) where these irrelevant variables likely hampered the recovery of clusters and (or) in the presence of a large number of clusters (i.e. when $K = 15$). By first filtering out the irrelevant variables and only retaining the signaling variables that clearly separate the clusters, the stable set of variables offered a more defined structure for model selection, even in the presence of a large number of clusters. In fact, the proposed model selection strategy, when coupled with the selection criteria "globalGap" or "firstGap" and the CKM or SAS algorithm, achieved a remarkable 100% recovery in all conditions examined.

[5] While the range for selecting K was limited by the scope of the simulations, we encourage applied researchers to consider a wide range of candidate values.

Table 2.1: Percentage of correct recovery of the number of clusters for 12 different strategies to determine the number of clusters

K	V	Full set of Variables				Stable set obtained with SAS				Stable set obtained with CKM			
		gp	fp	KL	Dindex	gp	fp	KL	Dindex	gp	fp	KL	Dindex
3	5	100%	100%	87.5%	94.4%	100%	100%	91.3%	100%	100%	100%	87.5%	100%
	50	66.3 %	96.9%	67.5%	100%	100%	100%	90.6%	100%	100%	100%	88.8%	100%
3	250	32.5%	75%	6.3%	61.3%	100%	100%	99.4%	82.5%	100%	100%	83.8%	100%
	1000	6.3%	70.6%	0%	31.3%	100%	100%	99.4%	82.5%	100%	100%	83.8%	100%
5	5	58.1%	60.6%	81.9%	96.9%	100%	100%	88.1%	97.5%	100%	100%	75.6%	97.5%
	50	73.1%	95.6%	0%	19.4%	100%	100%	81.3%	93.8%	100%	100%	81.3%	93.8%
5	250	27.5%	51.9%	0%	0%	100%	100%	56.9%	80.6%	100%	100%	79.4%	91.3%
	1000	0%	0%	0%	0%	100%	100%	85.6%	89.4%	100%	100%	85.6%	90%
15	5	0%	0%	19.4%	0%	100%	100%	13.1%	48.1%	100%	100%	0.6%	7.5%
	50	0%	0%	40.6%	40.6%	100%	100%	12.5%	26.3%	100%	100%	0%	6.3%
15	250	0%	0%	0%	0%	100%	100%	12.5%	26.3%	100%	100%	0%	6.3%
	1000	0%	0%	0%	0%	100%	100%	16.9%	10.6%	100%	100%	0%	5%

Note. Stable set refers to the proposed approach where only the stable set of signaling variables are used for selecting the number of clusters; full set refers to the conventional approach where all variables are used. gp = "globalGap", fp = "firstGap" (see the text for a detailed explanation of the two statistics).

2.3.4. SUMMARY OF THE SIMULATION STUDIES

In three simulation studies, we evaluated (1) the relative performance of CKM with respect to SAS, SKM, and KM in cluster recovery and the selection of signaling variables with (simulation study 1) and without (simulation study 2) a pre-determined number of irrelevant variables, and (2) the accuracy of selecting the number of clusters for all possible combinations of three variable selection strategies and four indices for determining the number of clusters. Our main findings were as follows: first, compared to the three competing methods - namely SAS, SKM, and KM, CKM was the winner in terms of cluster recovery across various conditions, with or without model selection. Second, in comparison to the other methods that are also capable of identifying signaling variables (i.e., SAS and SKM), CKM was the most accurate one when the number of irrelevant variables was unknown and the cluster structure was complex. Third, SAS enjoyed the shortest execution time in comparison to CKM and SKM. Fourth, we found that, across all conditions, the proposed model selection strategy that utilizes the stable set of variables resulted in better accuracy in selecting the number of clusters compared to the traditional strategy that utilizes the full set of variables. Finally, the best model selection procedure consisted of the combination of the proposed model selection strategy that relies on the stable set of signaling variables and the index "globalGap" or "firstGap". In our simulation setup, this procedure led to the perfect performance of CKM and SAS.

2.4. APPLICATION

Here, we demonstrate the usefulness of CKM in analyzing an empirical data set. We consider gene expression data of 13 autistic subjects and 14 healthy subjects that are publicly available from the gene expression omnibus (GEO) with accession number GSE7329[6]. For each subject, the transcription rates of 43893 probes were analyzed. Therefore, the data used in our analyses includes a total of 27 rows (subjects) and 43893 columns (variables). According to Nishimura et al.

[6]The full data set as well as the associated material could be extracted from the following address: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7329. While the original data set contained a total of 30 subjects, we were informed that three of the subjects (with series number GSM176615, GSM176589, and GSM176586) were not correctly stored in the data set and were therefore excluded from the current analysis

(2007), only a small number of probes are associated with autism - in their research, the authors selected a total of 293 probes for which the Analysis of Variance (ANOVA) tests resulted in a false discovery rate below a threshold of 5%.

Before the analysis, we pre-processed the data set such that each of the variables was mean-centered and scaled to unit sum-of-squares. Our first set of analyses was based on the full set of 43893 variables. More specifically, CKM, SAS, and KM were applied to the entire data set with K specified at 2 - to represent the autistic group and the control group. We did not try out a larger number of clusters considering the very small sample size. The three methods (i.e., CKM, KM, and SAS) all resulted in the same cluster partition: the first cluster contained the subjects with the indices 5, 6, 9, 15, 16, and 27 while the second cluster contained the remaining 21 subjects. Note that this partition was different from the assumed partition separating the patients (with the indices 1-14) and the control group (with the indices 15-27). The disparity between the known partition and the obtained partition is probably due to the presence of other biological mechanisms. To support this hypothesis, we further inspected the probes selected by the algorithms. While CKM selected a total of 958 probes, SAS selected 1238 probes. We used the free functional annotation tool DAVID (Bioinformatics Resources Version 6.8; Huang et al., 2007) to explore if the set of signaling variables identified by CKM indeed corresponds to any meaningful biological processes. The annotation picked up three groups of genes that were related to pathways that play an important role in three different types of disease: 20 genes were involved in the pathway of Parkinson's Disease; 22 in the pathway of Alzheimer's Disease; 22 in the pathway of Huntington's Disease. Given that the autistic subjects had a single gene Mendelian disorder (either a 15q11-q13 duplication or a fragile X mutation) and that the control subjects were composed of non-autistic siblings, it is not unlikely that a grouping structure is present in which autistic and control subjects are mixed. Figure 2.4 offers a visualization of cluster-specific centroids (after pre-processing) of all 958 signaling probes, with the line linking the two centroids of the same variable for the two clusters. Clearly, the two clusters showed distinctive response patterns: while a group of variables were associated with positive values in Cluster 1 and negative values in Cluster 2, the other group of variables showed a directly opposite pattern. We stress that the current

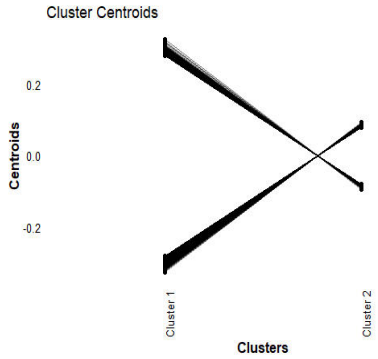


Figure 2.4: The cluster-specific centroids of the probes that were involved in key disease-related pathways

analysis should only be regarded as an exploratory analysis and further studies are needed to confirm the relevance of the two obtained clusters and their distinct genetic profiles.

We then conducted a second set of analyses where we used a subset of variables from the original data set. The subset consisted of two types of variables: the 293 signaling variables that have a significant difference in means between the autistic group and the control group and 1707 variables that were chosen randomly from the remaining variables (the new data set thereby involved a total of 2000 variables). To determine the signaling variables, we conducted a total of 43893 sets of linear regressions that regressed the transcription rate of each probe on the known partition of subjects with or without autism, and chose the 293 variables with the highest values of regression coefficients. Compared to the previous cluster analysis, we were more certain that the primary factor that divided all subjects was whether they were autistic or not. Consequently, we were able to evaluate the empirical performance of the clustering methods by examining to what extent a method successfully recovered the cluster partition and the set of signaling variables. CKM completed the task perfectly as it identified the exact 293 variables that were pre-defined as signaling variables. SAS also identified all of the 293 pre-defined signaling variables; however, in addition to this, it also erroneously picked 23 of the pre-defined irrelevant variables as if they were signaling variables.

To summarize, although the pre-existing groups were not recovered when the full data set was used, probably because of the existence of other biological processes that divided the subjects, the cluster structure was successfully recovered by CKM in a chosen subset of the data (with a total of 2000 probes). In terms of the accuracy of variable selection, in accordance with our findings in the simulation studies, CKM clearly outperformed SAS as it recovered the subset of signaling variables perfectly.

2.5. GENERAL DISCUSSION

Although behavioral sciences have a long tradition of operating in a "theory-driven way" and hence typically work with a small number of carefully selected and designed variables, they are now opening up their door to an interdisciplinary, data-rich approach where data sets involving many variables are increasingly common (Gil de Zuniga & Diehl, 2017). The growing availability of these data sets and the adoption of a data-driven approach could largely contribute to exploratory research (Fan et al., 2014; Yuan, Kroon, et al., 2021). In the context of cluster analysis, for example, the application of data-driven approaches to high-dimensional data could potentially lead to the discovery of novel clusters that are not detectable from a traditional examination (Yuan, De Roover, Dufner, et al., 2021). Yet, a unique challenge of this approach pertains to retaining only crucial variables that truly separate the clusters and filter out irrelevant variables. Successfully identifying these signaling variables is beneficial to the recovery as well as the interpretation of the underlying clusters.

To address this challenge and facilitate data exploration with high-dimensional data sets, several methods - for example, Sparse k-means (SKM) and Sparse Alternate Sum (SAS) - have been proposed that perform simultaneous clustering and variable selection. In the current study, we contributed to this line of research in two important ways. First, we presented a novel method, called Cardinality K-means, or CKM, that exploits the connection between PCA and KM to obtain, in a computationally efficient way, good starting values for a K-means (KM) procedure with variable selection. Our specific contribution is to introduce a special variant of the sparse principal component analysis (SPCA) with a cardinality constraint on the number of variables. As a result, CKM is a method

that is similar to SAS, but with a much better initiation of the parameter values. Through extensive simulations that included a number of important factors (e.g., the number of clusters, the proportion of irrelevant variables, and the distance between the centroids of adjacent clusters), we confirmed that CKM outperformed the other clustering methods (i.e., SAS, SKM, and KM) in terms of cluster recovery, especially in the presence of a large number of irrelevant variables. Furthermore, among the three methods with simultaneous variable selection (i.e., SAS, SKM, and CKM), CKM enjoyed the highest success rate in the identification of signaling variables. Compared to its predecessors SKM and SAS, CKM not only recovers clusters better but also offers a more structured and flexible approach to simultaneous clustering and variable selection. CKM uses the cardinality constraint, which offers at least the following two advantages over the l_1 penalty used in SKM. First, the application of the cardinality constraint (but not the l_1 penalty) allows users to have exact control over the number of signaling variables (Guerra-Urzola et al., 2021). This option is particularly helpful when a pre-specified number of signaling variables is desired in certain applications. Second, the l_1 penalty has long been criticized as suboptimal when the primary task is variable selection, and in such tasks, regression analysis with an l_1 penalty under-performed that with a cardinality constraint (e.g., Bertsimas et al., 2016). Moreover, thanks to the structured SPCA step, CKM can be easily extended to account for different types of analyses, which is not possible with SAS. For example, a researcher may want to find a specific structure of 4 clusters in which irrelevant variables only pertain to 2 clusters, while for the other 2 clusters, all variables are considered signaling variables. To accommodate this structure, in the first step where SPCA is performed, the cardinality constraint can be imposed for only 2 columns of the loading matrix. Furthermore, in the second step where the model parameters of CKM are iteratively updated, the loss function can be adjusted to reflect this assumption.

Another important contribution to the literature is that we proposed a novel model selection strategy to determine the number of clusters K. The proposed strategy adopts a three-step procedure that first applies a simultaneous clustering and variable selection algorithm (e.g., CKM, SAS, or SKM) to identify the most stable set of variables, i.e., those consistently identified as signaling variables

given any of the considered values of K, and then rely on this subset of variables to select the optimal value of K. Through simulation study 3, the proposed strategy - using either SAS or CKM to extract the stable set of variables - recovered K more accurately than the traditional strategy that selects K based on the full set of variables. Furthermore, we also found that, among the four evaluated model selection criteria (i.e., "globalGap", "firstGap", "KL Index", and "DIndex"), the two criteria developed from the Gap statistic (Tibshirani et al., 2001) recovered K with the highest accuracy. Overall, our study indicated that the preferred procedure of selecting K consists of two steps: (1) apply either CKM or SAS for each possible value of K and identify a stable set of variables that are consistently estimated as signaling variables; (2) determine K based on the stable set of variables with either "globalGap" or "firstGap".

To conclude, We strongly advocate the use of a simultaneous variable selection and clustering approach (e.g., CKM, SAS, and SKM) when the data contains a large number of variables and (or) it is desirable to pick up a subset of the most important variables - e.g., for the purpose of data exploration. When choosing between CKM, SAS, and SKM, according to the aforementioned results, we recommend the application of CKM when the primary objective is to recover the clusters and signaling variables as much as possible. When speed is important (e.g., in dealing with streaming data), however, SAS is the most desirable method. Last, the selection of the number of clusters is preferably based on a stable set of signaling variables that partial out irrelevant variables as much as possible.

We see several interesting future directions for CKM. First, in applications, the underlying cluster structure may be more complex than those generated in the simulations. Here, we discuss two scenarios that researchers may encounter and briefly elaborate on how CKM can be used in both scenarios. Consider a hypothetical data set with 200 variables and 6 clusters. In the first scenario, there is only one way of partitioning subjects, and different subsets of clusters are separated by different subsets of variables (e.g., the first 50 variables are relevant to Cluster 1-3 but not to Cluster 4-6, the last 50 are relevant to Cluster 4-6 but not to Cluster 1-3, and the other 100 variables are completely irrelevant to all clusters). When dealing with this data set, we expect CKM to successfully recover the 6 clusters and select variables 1-50 and 151-200 as signaling variables. Af-

ter retrieving the full set of signaling variables, users can then inspect the centroids of these variables for the 6 clusters to discover which subsets of variables are relevant to which subsets of clusters. In the second scenario, completely different partitions (i.e., with hardly any agreement between the two partitions) of the subjects pertain to different subsets of variables. In our hypothetical data set with 200 variables, all subjects may be partitioned into 6 clusters in two different ways: the first partition is driven by the first 50 variables, the second is driven by the last 50 variables, and the remaining 100 variables are once again irrelevant. To account for this scenario, users of CKM can follow an iterative procedure: after each step of identifying clusters and selecting signaling variables, the algorithm proceeds to apply CKM to the designated irrelevant variables. To prevent overfitting (i.e., finding clusters and associated signaling variables that are caused by noise only), after each step, theoretical knowledge can be used to confirm the clusters while resampling methods – e.g., bootstrapping and permutation test - can be applied to examine the stability of these clusters. We encourage future research to systematically examine the performance of these strategies in various applications. Second, future studies could investigate how different types of initialization affect the results of CKM. A notable limitation of the current simulation study is that, when initializing the alternating procedure for estimating CKM solutions (i.e., Step 2), we utilized only one rational start, estimated from a procedure inspired by USLPCA, yet we did not consider a multi-start procedure that employs multiple random starts. However, we would also like to point out that, according to Xu et al. (2015), a PCA-guided rational start likely yields comparable performance as a multi-start procedure when estimating KM results. Third, currently, CKM is only able to deal with continuous data with no missing responses. In future research, different imputation methods could be evaluated and compared, resulting in a preferred pre-processing scheme for a CKM analysis. Moreover, an extension of CKM can be developed to tackle mixed types of data (i.e., a combination of nominal, ordinal, and continuous variables).

APPENDICES

2.A. AN ALTERNATIVE FORMULATION OF KM

In this section, our goal is to illustrate that the objective function of a KM analysis could be re-formulated as $\mathbf{argmax_H}(Tr\mathbf{H}'\mathbf{X}\mathbf{X}'\mathbf{H})$, subject to $\mathbf{H}'\mathbf{H} = \mathbf{I}_k$ and an orthogonality constraint imposed on $\mathbf{H}$.

First, we acknowledge that Equation (2.1) could be re-written in

$$\begin{aligned}\mathbf{argmin_C}\sum_{k=1}^{K}\sum_{i\in C^{-1}(k)}||\mathbf{x}_i-\mathbf{m}_k||_2^2 &= \mathbf{argmin_C}(\sum_{i=1}^{N}||\mathbf{x_i}||_2^2-\sum_{k=1}^{K}\mathbf{m}_k\sum_{i\in C^{-1}(k)}(2\mathbf{x}_i-\mathbf{m}_k)')\\ &= \mathbf{argmin_C}(\sum_{i=1}^{N}||\mathbf{x}_i||_2^2-\sum_{k=1}^{K}n_k\mathbf{m}_k\mathbf{m}_{k'}).\end{aligned} \tag{2.14}$$

Per Equation (2.3), $\mathbf{m}_k$ in Equation (2.14) could be further replaced by $\frac{1}{\sqrt{n_k}}\mathbf{h}_k'\mathbf{X}$, resulting in

$$\begin{aligned}\mathbf{argmin_C}(\sum_{i=1}^{N}||\mathbf{x}_i||_2^2-\sum_{k=1}^{K}n_k\mathbf{m}_k\mathbf{m}_{k'}) &= \mathbf{argmin_H}(\sum_{i=1}^{N}\sum_{j=1}^{J}x_{ij}^2-\sum_{k=1}^{K}\mathbf{h}_k'\mathbf{X}\mathbf{X}'\mathbf{h}_k)\\ &= \mathbf{argmin_H}(||\mathbf{X}||_2^2-Tr\mathbf{H}'\mathbf{X}\mathbf{X}'\mathbf{H}).\end{aligned} \tag{2.15}$$

The last part of the equation holds because of the orthogonality of $\mathbf{H}$.

2.B. THE EQUIVALENCE OF THE TWO OPTIMIZATION FORMULATIONS THAT CONCERN KM WITH IRRELEVANT VARIABLES

In the current section, we discuss the equivalence between Equation (2.7) and Equation (2.10).

We apply the equivalence of Equation (2.6) and Equation (2.5) in Equation (2.9) and obtain

$$\mathbf{argmax}_{\hat{\mathbf{H}},\mathbf{g}} Tr\hat{\mathbf{H}}'\mathbf{X}_{-\mathbf{g}}\mathbf{X}_{-\mathbf{g}}'\hat{\mathbf{H}} \Leftrightarrow \mathbf{argmin}_{\hat{\mathbf{H}},\mathbf{P},\mathbf{g}}||\mathbf{X}_{-\mathbf{g}} - \hat{\mathbf{H}}\mathbf{P}'_{-\mathbf{g}}||_2^2.$$

Therefore, Equation (2.7) could be reformulated in

$$\mathbf{argmin}_{\mathbf{g}}||\mathbf{X}_{\mathbf{g}}||_2^2 + \mathbf{argmin}_{\hat{\mathbf{H}},\mathbf{P},\mathbf{g}}||\mathbf{X}_{-\mathbf{g}} - \hat{\mathbf{H}}\mathbf{P}'_{-\mathbf{g}}||_2^2 \Leftrightarrow \mathbf{argmin}_{\hat{\mathbf{H}},\mathbf{P}}||\mathbf{X} - \hat{\mathbf{H}}\mathbf{P}'||_2^2$$

where $\mathbf{P}$ contains V rows of zero entries.

2.C. PROOF FOR PROCEDURES TO UPDATE $\hat{\mathbf{H}}$ AND $\mathbf{P}$

In the current section, we provide detail derivations to support Algorithm 1. We first show the optimization problem $\mathbf{argmin}_{\hat{\mathbf{H}}}||\mathbf{X} - \hat{\mathbf{H}}\mathbf{P}'||_2^2$ subject to $\hat{\mathbf{H}}'\hat{\mathbf{H}} = \mathbf{I}$ has the solution

$$\hat{\mathbf{H}} = \mathbf{U}\mathbf{V}'$$

where $\mathbf{U}$ and $\mathbf{V}$ are obtained from the SVD of $\mathbf{XP}$.

We rewrite the optimization in

$$\begin{aligned} h(\hat{\mathbf{H}}) &= ||\mathbf{X} - \hat{\mathbf{H}}\mathbf{P}'||_2^2 \\ &= tr\mathbf{P}\hat{\mathbf{H}}'\hat{\mathbf{H}}\mathbf{P}' + tr\mathbf{X}'\mathbf{X} - 2tr\mathbf{X}\mathbf{P}\hat{\mathbf{H}}' \\ &= tr\mathbf{P}\mathbf{P}' + \mathbf{X}'\mathbf{X} - 2tr\hat{\mathbf{H}}'\mathbf{X}\mathbf{P}. \end{aligned}$$

Therefore, the minimization problem is equivalent to the maximization problem of $tr\hat{\mathbf{H}}'\mathbf{X}\mathbf{P}$ subject to $\hat{\mathbf{H}}'\hat{\mathbf{H}} = \mathbf{I}$. Such a maximization problem can be addressed with the Kristof Theorem (for a detailed description and proof of the Kristof Theorem, please refer to ten Berge, 1993). More specifically, we realize

$tr\hat{\mathbf{H}}'\mathbf{XP}$ could be rephrased in

$$\begin{aligned} tr\hat{\mathbf{H}}'\mathbf{XP} &= tr\hat{\mathbf{H}}'\mathbf{UDV}' \\ &= tr\mathbf{V}'\hat{\mathbf{H}}'\mathbf{UD} \\ &= tr\mathbf{GD}, \end{aligned}$$

where $\mathbf{XP} = \mathbf{UDV}'$ represents the SVD of $\mathbf{XP}$.

Since $\mathbf{G} = \mathbf{V}'\hat{\mathbf{H}}'\mathbf{U}$ and all of $\mathbf{V}$, $\hat{\mathbf{H}}$, $\mathbf{U}$ are sub-orthonormal matrices (i.e. they can be completed to orthonormal matrices), $\mathbf{G}$ is also a sub-orthonormal matrix. Therefore, according to the Kristof Theorem, $tr\mathbf{GD} \leq tr\mathbf{D}$, and the maxima is reached when $\mathbf{V}'\hat{\mathbf{H}}'\mathbf{U} = \mathbf{I}$. Given the orthonormality of both $\mathbf{U}$ and $\mathbf{V}$, $\hat{\mathbf{H}} = \mathbf{UV}'$.

Now consider the optimization problem $\mathbf{argmin_P}||\mathbf{X} - \hat{\mathbf{H}}\mathbf{P}'||_2^2$ subject to the constraint that V rows in loading matrix $\mathbf{P}$ are exact zeros. The solution of $\mathbf{P}$ is obtained in two steps: (1) calculate $\mathbf{P_0} = \mathbf{X}'\hat{\mathbf{H}}$ and (2) impose zeros on the V rows of $\mathbf{P_0}$ whose sum-of-squares are smallest.

We re-write the optimization problem in

$$\begin{aligned} h(\mathbf{P}) &= ||\mathbf{X} - \hat{\mathbf{H}}\mathbf{P}'||_2^2 \\ &= tr\mathbf{X}'\mathbf{X} + tr\mathbf{P}\hat{\mathbf{H}}'\hat{\mathbf{H}}\mathbf{P}' - 2tr\mathbf{P}\hat{\mathbf{H}}'\mathbf{X} \\ &= Const. + tr\mathbf{P}'\mathbf{P} - 2tr\mathbf{PW} \\ &= Const. + \sum_{j=1}^{J}\sum_{k=1}^{K} p_{jk} - 2\sum_{j=1}^{J}\sum_{k=1}^{K} p_{jk}w_{jk} \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K}(p_{jk} - w_{jk})^2 - \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk}^2 \\ &= \sum_{j=1}^{J}(\sum_{k=1}^{K}(p_{jk} - w_{jk})^2) + Const. - \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk}^2, \end{aligned}$$

where $Const. = ||\mathbf{X}||_2^2$ is a constant and $\mathbf{W} = \mathbf{X}'\hat{\mathbf{H}}$. Note that $\sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk}^2$ is also a constant. Hereby, we derive the solution to $\mathbf{P}$.

3

A Tutorial on Simultaneous Clustering and Variable Selection

To accurately capture the heterogeneity of human behavior, psychologists frequently apply cluster analysis to identify clusters with distinctive behavioral profiles. Cluster analysis is especially useful when dealing with data-intensive studies involving a large number of variables, as the wealth of information covered by these data sets can potentially lead to important discoveries about hitherto unknown subtypes. These applications, however, face two major challenges. To begin, these large-scale data sets are likely to contain irrelevant variables that do not contribute to cluster separation and, in the worst case, may even prevent the accurate recovery of clusters. Second, to avoid false detection of clusters, the findings should be validated with both theory-driven and data-driven methods, but guidance for this validation process is scarce. In response to these two challenges, this tutorial describes a recently proposed method (Cardinality K-means) that allows simultaneous variable selection and clustering, and discusses a framework for cluster validation.

This chapter is submitted for publication as **Yuan, S.**, De Roover, K., Jaime Hermsdorf, & Van Deun, K. (Revise and resubmit). A Tutorial on Simultaneous Clustering and Variable Selection *Advances in Methods and Practices in Psychological Science*

Moreover, the tutorial provides a step-by-step guide to conduct CKM analyses and cluster validation using the R package `CKM` *and ShinyApp* `ClusterViz`*. An illustrative example of clustering citizens based on their political opinions is presented in detail, where annotated R code is also available.*

3.1. INTRODUCTION

Cluster analysis – partitioning subjects into (unobserved) clusters where everyone in the same group shares similar profiles – is arguably one of the most important statistical techniques in many psychology disciplines, such as organizational psychology (e.g., M. Wang & Hanges, 2011), educational psychology (e.g., Hayenga & Corpus, 2010), personality and social psychology (e.g., Neumann et al., 2020), and developmental psychology (e.g., Lonigan et al., 2018). However, the most popular clustering techniques in psychological research (e.g., K-means clustering, latent profile analysis, etc.,) are not fit for large data sets (e.g., > 100 variables; Groeneveld and Rumsfeld, 2016), despite the exponential growth in the use of this type of data sets in many sub-domains of psychology (Harlow and Oswald, 2016; Adjerid and Kelley, 2018; Mõttus et al., 2020; Putka et al., 2018). Examples of such data sets include psychological data sets that contain novel measures (e.g., genes, GPS trackers, digital footprints) or incorporate many potentially important variables to capture the complex nature of the research question under investigation (see Putka et al., 2018, p692). The reason why classical clustering techniques are not fit for large data sets is that they rely on the strict assumption that *all* variables partition the observations well (i.e., all variables are relevant to cluster separation). Consequently, these techniques are unable to filter out irrelevant variables that contribute little or nothing at all to the partition of clusters. While screening out irrelevant variables is an important - yet always neglected - aspect of cluster analysis in general (Steinley & Brusco, 2008a), this task is particularly pertinent for large data sets and exploratory analyses, where many irrelevant variables are expected (Waldherr et al., 2017). For example, when one tries to identify clinical subtypes based on genetic phenotypes or brain activation, it is hard to believe that all genes or brain signals play an active role in separating these subtypes. Here, the data-driven elimination of irrelevant variables not only simplifies the interpretation of cluster results but also benefits the task of cluster recovery, as these irrelevant variables likely obscure the true cluster structure (Brudvig et al., 2019).

Several simultaneous clustering and variable selection techniques (referred to as SCVS techniques hereafter) have been proposed for filtering out irrelevant variables in a data-driven way. The usefulness of this type of techniques has

been well documented in a recent psychiatric study (Y. Zhang et al., 2021). In this study, two subtypes of post-traumatic stress disorder (PTSD) and major depressive disorder (MDD) have been identified based on their distinct patterns of functional connectivity (quantified by power envelope connectivity or PEC): among the total of 3720 PEC features, the authors identified a large subset of irrelevant features that were of similar values in both subtypes (see P6 of the supplementary material in Y. Zhang et al., 2021). Despite the proven usefulness of SCVS techniques, they have been neglected by most psychological studies (c.f., Y. Zhang et al., 2021; Gharani et al., 2021,Postareff et al., 2017). One probable reason is that many early SCVS techniques were computationally prohibitive for large data sets, as they built on comparing cluster solutions for an exceeding number of variable subsets. Fortunately, employing different forms of recently-proposed regularization methods (e.g., Kang et al., 2021) [1], some recently proposed SCVS techniques (for reviews, see Bouveyron et al. (2019) and Raymaekers and Zamar (2020)) are capable of achieving the dual goals of clustering and variable selection without excessive computational demands. [2].

This Tutorial offers a step-by-step guide for researchers who want to employ the SCVS techniques. The complete analytical process consists of five steps roughly divided into three parts: data preprocessing, cluster analysis, and cluster validation (see Figure 3.1 for a summary of the five steps). The cluster analysis part is based on Cardinality K-means or CKM (Yuan et al., 2022), a recently proposed technique that proved to outperform a number of related methods (e.g., Sparse K-means; Witten and Tibshirani, 2010; Sparse Alternate Sums; Arias-Castro and Pu, 2017), yet we note that the other parts of the analytical process are model-agnostic and can be implemented with other SCVS techniques. The cluster validation part discusses and illustrates a number of approaches to validate

[1] In recent years, the idea of regularization has been adopted by many genres of psychological methods, including psychological networks (Epskamp & Fried, 2018), structural equation models (Jacobucci et al., 2016), exploratory factor analysis (Chen, 2021).

[2] We note that clustering techniques dealing with continuous variables can generally be divided into two categories: model-based clustering (e.g., latent profile analysis; Gibson, 1959) and non-model-based clustering (e.g., K-means clustering; Steinley, 2006). A comparison of these two techniques in psychological studies is provided in Steinley and Brusco, 2011b. The methods detailed in our Tutorial use the K-means clustering framework, and we recommend the excellent review provided in Bouveyron et al., 2019 to those readers interested in model-based SCVS techniques

cluster results. Drawn from a recently proposed cluster validation framework (Ullmann et al., 2021), these approaches include (1) visual validation, (2) cluster stability validation with bootstrapping (Hennig, 2007), and (3) cluster replication with a hold-out sample (i.e., splitting the original data set in two, using one half for cluster analysis and the other half for cluster validation). Visual validation is prevalent in psychological studies (for reviews, see Henry et al. (2005) and Clatworthy et al. (2005)), but the other two approaches are rarely applied. Our Tutorial aims to exemplify and promote the use of these other two approaches. To facilitate visual validation on cluster results, we have developed a ShinyApp `ClusterViz`, with which many types of clustering plots can be generated with limited input from the user. Importantly, the concepts and processes of cluster validation as well as the associated software are generally applicable to many clustering methods. To offer hands-on guidance on the five steps of analysis outlined in Figure 3.1, this Tutorial makes use of data sets on political attitudes collected by CentERdata (Tilburg University, the Netherlands) with the R package `CKM` and the ShinyApp `ClusterViz`.

3.2. THE EMPIRICAL EXAMPLE

3.2.1. DISCLOSURES

This tutorial provides two tools to (1) estimate and validate clusters in the presence of a large number of variables with CKM and (2) visualize clusters with a few mouse clicks. The first one is implemented in the R package `CKM` (Yuan et al., 2022) and is available at https://github.com/syuanuvt/CKM. The second one is implemented as a Shiny app and is available at https://syuan.shinyapps.io/ClusterViz/.

3.2.2. BACKGROUND AND DATA DESCRIPTION

In this case study, we aimed to identify clusters of Dutch citizens with different profiles of political attitudes. In the last two decades, with the rise of the radical right (Silva, 2018), Dutch society has become more polarized than ever before

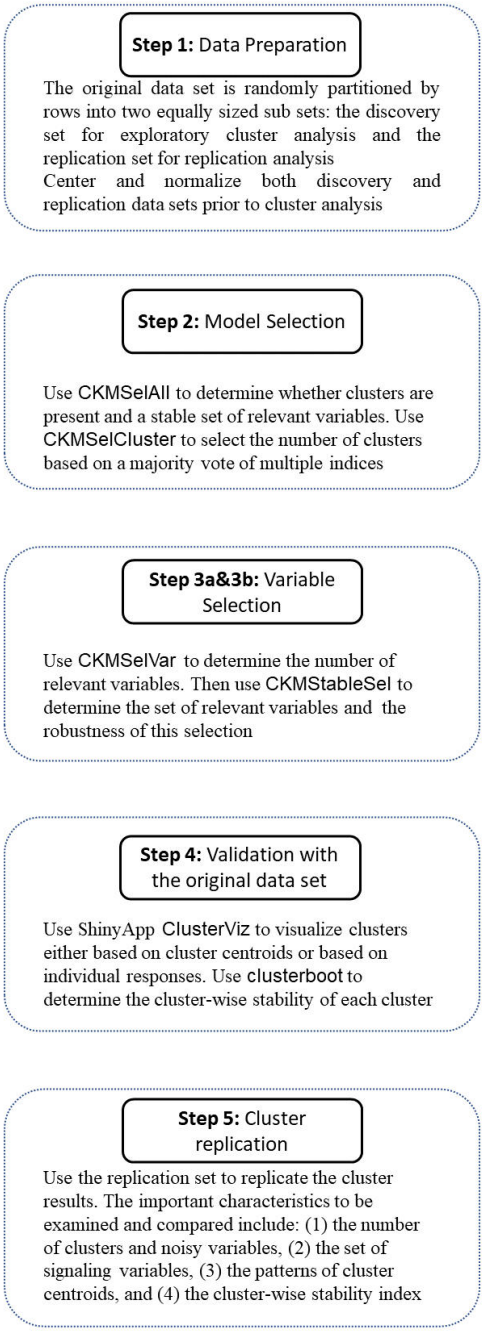


Figure 3.1: The workflow of the cluster analysis presented in the tutorial

(Berning & Schlueter, 2016; Silva, 2018) [3]. Despite the obvious significance of understanding how and to what extent clusters of citizens hold different political attitudes, empirical studies are unfortunately scarce and cover only a limited range of political attitudes (e.g., attitudes towards refugees and migrants; Albada et al., 2021). Here, using a nationally representative sample of Dutch citizens, we extended previous research on opinion profiling by examining (1) the most relevant political attitudes that partition citizens into different clusters and (2) the distinctive response patterns of these clusters. We used a data set from the Longitudinal Internet Studies for the Social Sciences (LISS) panel, collected by CentERdata (Tilburg University, the Netherlands) between December 2020 and March 2021. Only the 5542 observations with non-missing responses on the total of 112 items were retained in the current analysis [4].

3.2.3. STEP 1: PACKAGE SETUP AND DATA PRE-PROCESSING

The package `CKM` can be downloaded and installed from the aforementioned website with the following code.

```
#download and install the dependencies of the CKM package
install.packages("RSpectra")
install.packages("cluster")
install.packages("fpc")
install.packages("NbClust")
install.packages("ggplot2")
#download and install the CKM package and load the package
    onto the current workpath
remotes::install_github("syuanuvt/CKM")
library(CKM)
```

[3] In the Netherlands, the far-right parties such as Partij Voor de Vrijheid (PVV; Party for Freedom) and Forum voor Democratie (FVD; Forum for Democracy) attracted a substantial amount of votes (about 25%) in the recent general elections

[4] We note that many SCVS techniques, for example, CKM used in the current analysis, cannot handle missing responses. Except resorting to a subset consisting of all complete responses, researchers can deal with missing values by ascribing imputation methods such as multiple imputation (Rubin, 2004) before applying SCVS techniques.

Prior to data analysis, a relevant - yet independent - data set should be prepared for replication analysis (hereafter, we term the data set used in the original analysis as the *discovery set* and the one used in the replication analysis as the *replication set*). The replication analysis is essential for highly data-driven methods, such as cluster analysis, since it effectively prevents overly optimistic results (Ullmann et al., 2021). Preferably, the replication analysis would be based on separate data sets; yet, in many research practices, it is very difficult, if not completely impossible, to secure such a separate data set that is independent of but closely related to the original data set. The alternative is to randomly partition the original data set into two equally sized data sets by rows, with one serving as the discovery set while the other as the replication set [5]. Upon the creation of the discovery and replication data sets, the two data sets are further preprocessed by mean-centering and standardization. Overall, the re-sampling and preprocessing can be implemented with the following code (note that the discovery set `train.st` will only be used in Steps 2-4 and the replication set `test.st` will only be used in Step 5).

```
train.indices <- sample(1:nrow(liss),nrow(liss)/2)
liss.train <- liss[train.indices,]
liss.test <- liss[-train.indices,]
train.st <- scale(liss.train)
test.st <- scale(liss.test)
```

3.2.4. STEP 2: SELECTION OF THE NUMBER OF CLUSTERS

In the second step, the number of clusters is to be determined. Given the complexity of selecting the optimal number of clusters (Jain, 2010), we advocate a strategy that considers multiple well-established methods and makes a final decision supported by the majority of these indices (see also Akhanli and Hennig,

[5] This way of dividing the original data set by rows is most appropriate when the aim of cluster analysis is to identify clusters of the general population with distinct patterns of responses. This is also the typical objective of most scientific studies. However, in applied settings, sometimes the purpose of the analysis is to make statements about specific subjects in the original data set (e.g., to classify all subjects into two categories of good performers and poor performers). In this type of analysis, the discovery and replication sets can be created by dividing the original data sets by columns.

2020) [6]. Currently, the package `CKM` employs a total of five methods, yet we acknowledge that other methods can be considered as well. The five methods considered in the package are the Gap statistics (Tibshirani et al., 2001), the elbow point method (Thorndike, 1953), the silhouette index (Rousseeuw, 1987), the prediction strength (Tibshirani & Walther, 2005), and the bootstrapped cluster instability index (Fang & Wang, 2012). These five methods were chosen because they are among the best-performing ones and employ three different strategies for determining the number of clusters. We now outline the basic and intuitive principles of these five methods, and refer to the original work for a more detailed description.

Being the computationally simplest, both the elbow point method and the silhouette index determine the optimal number of clusters based exclusively on the observed data set and are thus referred to as internal validation indices. The elbow point method plots the between-cluster sum-of-squares as a function of the number of clusters and identifies the optimal number of clusters as the turning point (also called the elbow point) where the curve turns from a steep increase to a flat trend. The silhouette index essentially quantifies the difference between the similarity of observations from the same cluster and that of observations from different clusters. A higher value of the silhouette index indicates a stronger cluster pattern with cohesive and separated clusters; thus, the optimal number of clusters is determined so that the silhouette index is maximized. Notably, the silhouette index has been empirically proven to be the most accurate method for selecting the number of clusters among more than 30 internal validation indices (Arbelaitz et al., 2013). The next two methods, the Gap statistic, and the prediction strength, bear the core premise that the optimal number of clusters is the one in which the cluster pattern is the strongest. The strength of the cluster pattern is defined quite differently according to the two methods. The Gap statistic quantifies this strength as the extent to which cluster separation (as quantified by the within-cluster sum of squares) in the observed data set differs from that in the *reference* data sets, which are derived from random permutations of the observed data set per column. The prediction strength method

[6] This strategy, also referred to as the majority voting scheme in the machine learning literature, has also been proposed in Charrad et al. (2014)

partitions the observed data set into two halves and quantifies the similarity of the clusters identified in both halves. In line with their definitions, higher values of the Gap statistic and the prediction strength indicate stronger cluster patterns; therefore, the optimal number of clusters is chosen corresponding to the largest Gap statistic and prediction strength. Last, the bootstrapped cluster instability index was developed based on the notion of cluster stability - the extent to which the identified cluster patterns can be replicated on a bootstrapped replication of the observed data set. Accordingly, the optimal number of clusters should maximize the stability of the cluster solution, quantified by the minimized value of the bootstrapped cluster instability index.

Two major challenges have to be overcome before applying the above five methods to determine the number of clusters. First, the set of irrelevant variables have to be filtered out as much as possible (Brudvig et al., 2019; Yuan et al., 2022). Since the selection of signaling variables, as detailed in Step 3-4, in turn, depends on the number of clusters, Yuan et al. (2022) proposed an approximate strategy that chooses a small set of variables - coined stable variables - that are designated as signaling variables regardless of the number of clusters. Here, `CKM` implements this strategy that consists of the following steps: (1) determine the range of possible cluster numbers, (2) identify the set of signaling variables for each possible cluster number (see Step 3 for details), and (3) use the intersection of the multiple sets derived from (2) as the set of stable variables.

The second challenge is to decide whether the single-cluster solution fits the data well enough (Steinley and Brusco, 2011a). Of the five methods currently implemented in CKM, only the Gap statistic is able to inform the choice between 1 and 2 clusters. Concretely, when the Gap statistic computed from a 2-cluster solution is much larger than the one obtained from a single-cluster solution, we can be convinced that the observed data set contains at least 2 clusters.

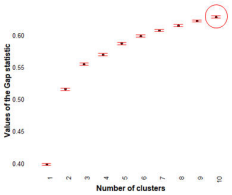
Taken all together, Step 2 can be further divided into two substeps. The first substep involves identifying the set of stable variables and calculating the Gap statistics to determine whether the observed data set contains at least two clusters, whereas the second substep deploys the aforementioned five methods to select the optimal number of clusters from a customized range. We now demonstrate, with our case study, how the package `CKM` can be used to perform these

two substeps. We start by defining the range of the potential number of clusters to be between 1 and 10. The function `CKMSelAll` can then be used to determine the stable set of variables and to compute the Gap statistic associated with each number of clusters.

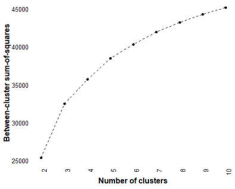
```
sel.gap <- CKMSelAll(train.st, minclust = 1, maxclust = 10)
## select only the stable variables
train.stable <- train.st[,sel.gap$stable.set]
```


Figure 3.2A plots the Gap statistics as a function of the number of clusters as well as the corresponding standard errors (indicated by red bars). Clearly, the Gap statistic resulting from the 2-cluster solution was much larger than the one estimated from the single-cluster solution. This meant that the observed data set likely contained more than 1 cluster. Therefore, the range of the potential number of clusters was reduced to between 2 and 10. In fact, according to Figure 3.2A, the optimal number of clusters that maximized the gap statistic was 10 (as also reported in `sel.gap$opt.cluster`). Another important piece of information extracted from the above analysis was the set of stable variables, which were obtained with the code `sel.gap$stable.set`. The subset of the observed data set (i.e., `train.stable`, which only consists of the stable variables, is used in the second substep to determine the optimal number of clusters for the other four methods, using the function `CKMSelCluster` (note that the minimal number of clusters should be set to 2).

```
sel.major <- CKMSelCluster(train.stable, minclust = 2,
    maxclust = 10, method = "all")
```



A: The gap statistics as a function of the number of clusters



B: The between-cluster sum-of-squares as a function of the number of clusters

Figure 3.2: The gap statistic and the between-cluster sum-of-squares as a function of the number of clusters

Figure 3.2B shows a smooth curve of between-cluster sum-of-squares where the elbow point is not easily distinguishable. Following the principle that, when no clear decision can be made from the graphs, simpler models should be favored, we concluded that the elbow point method suggests a 2-cluster solution. The other three methods selected the optimal number of clusters automatically, which are listed in Table 3.1. Combining these results, we can determine that the optimal number of clusters was equal to 2 (i.e., 4 out of 5 methods selected the 2-cluster solution). Therefore, two clusters were identified that had distinct response profiles in responding to items about political attitudes. In the next step, we determined how many items effectively separated the two clusters.

Table 3.1: The optimal number of clusters determined by the different selection strategies

Selection strategy	Optimal Clusters
`The Gap statistic`	10
`The elbow method`	2
`The prediction strength`	2
`The bootstrapped cluster instability`	2
`The Silhoutte index`	2

3.2.5. STEP 3: SELECTION OF SIGNALING VARIABLES

The main objective of Step 3 is to decide the number of signaling variables conditional upon the number of clusters. This step comprises two substeps: (1) the selection of the *number* of signaling variables with a modified Gap statistic (substep 3A) and (2) the final selection of the set of signaling variables with stability selection (substep 3B) [7]. In substep 3A, a method derived from the original version of the Gap statistic can be applied (Arias-Castro & Pu, 2017; Yuan et al., 2022): the modified Gap statistic is computed for each possible number of signaling variables, and just like the original version of the Gap statistic, this modified Gap statistic indicates the extent to which the cluster pattern in the observed data set differs from those calculated from the reference data sets. Consequently, a larger value of the modified Gap statistic reflects a stronger cluster pattern and the number of signaling variables can be selected as corresponding to the largest value of the Gap statistic.

While this strategy of selecting the optimal number of signaling variables works well with small numbers of clusters (< 10) in simulation studies (Arias-Castro & Pu, 2017; Yuan et al., 2022), its accuracy decreases as the number of clusters increases. Yuan et al. (2022) proposed to address this shortcoming by an additional procedure after maximizing the Gap statistic if the optimal number of clusters exceeds 10. Specifically, the additional procedure uses a small grid

[7] Note that a simplified version of Step 3 has already been followed in Step 2 in the detection of the set of stable variables. The simplified version does not include substep 3B that employs an extensive resampling scheme.

of candidate values, constructed from several (e.g., ± 10) neighbors of the optimal value located in the initial step. For each possible value on this grid, the corresponding between-cluster sum-of-squares is estimated. Finally, the optimal number of signaling variables is identified in which the scree ratio, defined for p signaling variables as $\frac{BSS_p - BSS_{p-1}}{BSS_{p+1} - BSS_p}$ is maximized (here BSS_p refers to the between-cluster sum-of-squares in the presence of p signaling variables; see for example De Roover, Ceulemans, Timmerman, Nezlek, et al. (2013) for details). [8]

Substep 3A is implemented in the package `CKM` with the function `CKMSelVar`, which allows users to either include or leave out the additional procedure as described above. In the case study reported here, since the selected number of clusters equals 2 (i.e., < 10), the additional procedure is not necessary. Accordingly, the option "sr" is set to "FALSE" to avoid this extra step.

```
sel.var <- CKMSelVar(train.st, n.cluster = 2, sr = FALSE)
```

The above function picked 53 as the number of signaling variables. In other words, out of the total of 86 variables, 53 were deemed to be signaling variables that effectively separated the two clusters whereas 33 were irrelevant variables upon which clusters were not clearly separated.

Although substep 3A selected a set of signaling variables, this selection might be heavily influenced by sampling variation. Therefore, in substep 3B, a resampling technique called stability selection (Meinshausen and Bühlmann, 2010; Li and Jacobucci, 2021) is employed to account for sampling variation and generate the final set of signaling variables that are consistently picked up across replications. This substep first generates a number of subsamples from the original sample (which, according to the original proposal in Meinshausen and Bühlmann (2010), could be 200 subsamples, each consisting of one-half of the randomly chosen observations). Then, for each subsample, the model selection routine detailed in substep 3A is applied, where the number of irrelevant variables is as-

[8] An alternative strategy is to determine the optimal number of signaling variables through the application of a threshold: given a threshold θ specified by the user (e.g., $\theta = \frac{1}{3}$), the largest p that satisfies $BSS_p < \theta \times (BSS_{p+1} + BSS_{p+2})$ is considered to be the optimal number of signaling variables.

sumed to be half the number of irrelevant variables singled out in substep 3A [9]. Therefore, in the case study reported here, for each of the subsamples, the analysis assumed a total of 16 irrelevant variables and 70 signaling variables. Last, a final set of signaling variables was identified that appeared in more than 95% of the sub-samples. Here, the percentage used to threshold signaling variables is again of subjective nature, determined by the users with their own consideration for the stability of results. In the case study, substep 3B was called by the function `CKMStableSelect` of the package `CKM` where the threshold was set to the default value of 95%:

```
sel.final <- CKMStableSelect(train.st, 2, 33) #here the input
    is the number of noisy variables
signaling.set <- train.st[,sel.final$sig.set]
```

Table 3.2 presents a frequency table of the top 60 most frequently selected items. Two important observations should be noted. First, the top 52 variables were selected very frequently (i.e., at least 95% of the times), indicating that they were to a large degree the crucial variables separating the two clusters. Second, a clear difference was found between the 52^{nd} (item 56, selected 193 times) and 53^{rd} (item 61, selected 188 times) most frequently selected items. This clear discrepancy consolidated the conclusion that a total of 52 items were deemed to be the most significant and robust indicators that separated the two clusters.

Last, conventional K-means analysis can be applied to the subset with only signaling variables (i.e., `signaling.set` in our case study) to obtain the final cluster partition, as follows.

```
cluster.partition <- kmeans(signaling.set, 2, nstart = 100)$
    cluster
```

3.2.6. STEP 4: CLUSTER VALIDATION WITH THE ORIGINAL DATA SET

In our case study, through Steps 2 - 3, we successfully identified two clusters with distinct profiles of political values and the set of 52 signaling items that best sep-

[9] We note that users are free to decide how many irrelevant (or equivalently signaling) variables are to be identified in the subsamples. This ratio, coupled with the threshold of frequency, tends to produce a consistent pattern of the relative importance of each variable in separating the clusters. Here we opted for a convenient value of 1/2.

Table 3.2: The top 60 items that were most frequently selected as relevant variables in sub-samples

Item	Times	Items	Times	Items	Times
1	200	21	200	49	200
2	200	22	200	51	200
3	200	23	200	54	200
4	200	24	200	57	200
5	200	25	200	58	200
6	200	26	200	60	200
7	200	27	200	48	199
8	200	28	200	45	198
9	200	29	200	78	198
10	200	30	200	77	196
11	200	31	200	86	196
12	200	32	200	56	193
13	200	33	200	61	188
14	200	34	200	79	182
15	200	35	200	65	179
16	200	36	200	63	173
17	200	37	200	83	155
18	200	38	200	80	141
19	200	46	200	84	132
20	200	47	200	42	126

arated the two clusters. A key question that remained unanswered was whether the findings were theoretically relevant and methodologically robust. We addressed this question in Steps 4-5, through cluster validation of the discovery set (Step 4) and the replication set (Step 5). We first illustrate how to validate the cluster results of the discovery set.

STEP 4A: VISUAL VALIDATION

The first way to validate cluster results is through visual validation, which examines whether the partition can be translated into easily visible and (theoretically) meaningful patterns in the data set. Data visualization is considered essential for many types of analysis (see Hehman and Xie (2021) for guidelines of data visualization) and it is particularly informative for exploratory analyses, like clustering, since these exploratory analyses often lack clear and well-defined expectations.

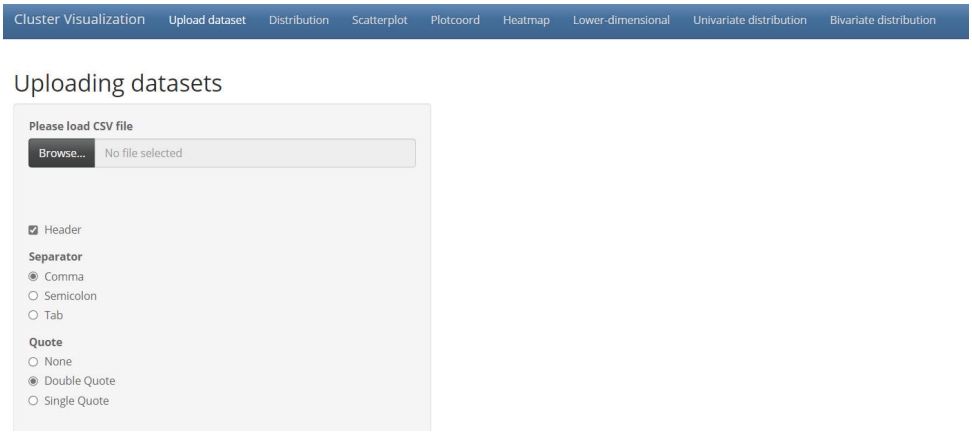


Figure 3.3: A screenshot of the Shinyapp designed for cluster visualization

To remove the technical barriers that prevent such visualizations from being used, we created a ShinyApp called `ClusterViz` that allows users to generate various types of plots with just a few mouse clicks. Figure 3.3 shows a screenshot of `ClusterViz`, which is freely accessible via https://syuan.shinyapps.io/ClusterViz/. Note that ClusterViz is designed in such a way that it is completely separate from the package `CKM` so that users, even without coding experience, can take advantage of the software and apply it to clustering algorithms other than CKM. `ClusterViz` requires a special arrangement of data sets (in csv format) in which a column representing the cluster partition, named "Cluster", is adjacent to the original data set to be visualized. In our case study, this data reformatting can be easily achieved by the following code

```
viz.dataset <- cbind(signaling.set, cluster.partition)
#rename the last column
names(viz.dataset)[ncol(viz.dataset)] <- "Cluster"
```

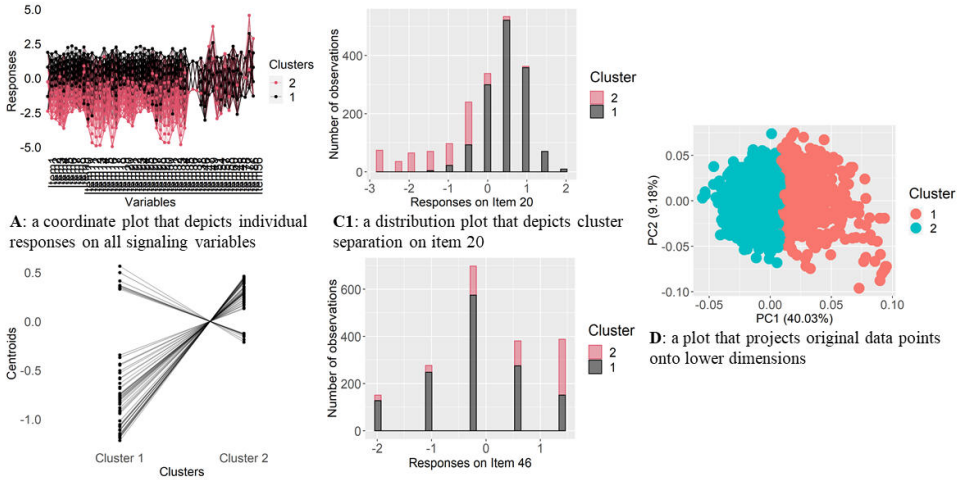
In general, two types of plots can be generated depending on their purpose: plots that directly visualize the original data sets, and plots that visualize cluster centroids. Below, with the example of the case study, we elaborate how to use these plots (shown in Figure 3.4) to inspect cluster partitions.

(i) Plots of individual responses on the signaling items (Figure 3.4A): this plot offers an overview of the differences between the individual response pro-

files. In our case study, to interpret the results, we inspected the content of the signaling items. These items measure three themes: (1) one's confidence and satisfaction with important political and societal institutions (e.g., Dutch Parliament, United Nations) and (2) one's attitudes toward the social integration of immigrants, and (3) one's attitudes on gender equality. The general response patterns of the two identified clusters were markedly different: compared to the subjects classified in Cluster 1, those in Cluster 2 consistently scored lower in terms of confidence and satisfaction with institutions, and they were more opposed to the integration of immigrants, European unification, as well as gender equality. The results, therefore, coincided with political psychology findings that people with a far-right ideology (hereafter referred to as populists), classified to Cluster 2 in the current analysis, are characterized by an anti-establishment stance and extremely conservative views and hold fundamentally different political attitudes compared to mainstreams (e.g., Wood & Gray, 2019). More interestingly, the findings also clearly highlighted that, across the entire spectrum of political themes, the core themes highlighting populists are immigration and European unification as well as gender equality. This novel finding can be further scrutinized in confirmatory studies.

(ii) Cluster centroids plotted on the coordinates (Figure 3.4B): the plot can be considered as a cluster-wise summary of the individual responses. In the case study, once again, the separation of two clusters was clearly visible. In general, this summary plot is particularly informative because It illustrates the variables on which clusters are most clearly discernable.

(iii) Plots of cluster separation on a single variable (Figures 3.4C1 and 3.4C2): both plots offer a closer look at how one or a few variables separate the clusters. In our case study, we focused on two items that separated the two clusters to the greatest degree: (1) item 20: how confident are you for the Dutch parliament? and (2) item 46: Where would you place yourself on a scale from 1 to 5, where 1 means that European unification should go further and 5 means that it has already gone too far? The two items separated the two clusters in distinctive ways: as shown in Figure 3.4C1, compared



A: a coordinate plot that depicts individual responses on all signaling variables

C1: a distribution plot that depicts cluster separation on item 20

B: a coordinate plot that depicts cluster centroids of all relevant variables

C2: a distribution plot that depicts cluster separation on item 46

D: a plot that projects original data points onto lower dimensions

Figure 3.4: The four types of visualizations from a cluster analysis

to non-populists (i.e., Cluster 1), populists (i.e., Cluster 2) were on average much less confident in the Dutch parliament, while, cf. Figure 3.4C2, they were to a much greater degree opposed to European unification.

(iv) Projection of clusters in lower dimensions by principal component analysis (Figure 3.4D): this plot indicates the dimensions that partition the clusters. In our case study, the first principal component explained the majority of the differences between the two clusters. Hence, we can conclude that only one underlying dimension (i.e., populist ideology) effectively distinguished the responses of the two clusters

STEP 4B: STABILITY VALIDATION

Another method for validating the cluster solution is a purely data-driven approach: to examine the stability of each cluster using a clusterwise stability index, as proposed in Hennig (2007). Hennig (2007) argued that, for any cluster result, it is important to examine whether the identified clusters disappear when minor and non-essential changes to the data set (e.g., resampling from the same underlying distribution or adding a small amount of noise) are made. Based on this reasoning, Hennig (2007) developed an index to evaluate the cluster-wise

stability of each cluster estimated from the analysis. To calculate this index in our case study, we used the following code:

```
library(fpc)
validation.cluster.train <- clusterboot(train.st, B=100,
    bootmethod="boot", clustermethod=kmeansCBI, krange=2)
validation.cluster.test <- clusterboot(test.st, B=100,
    bootmethod="boot", clustermethod=kmeansCBI, krange=2)
```


To inspect the indices, we used the command `validation.cluster$bootmean`. The stability indices corresponding to the two clusters were .99 and .98, respectively. These values, according to the guidelines provided in Hennig (2008) (i.e., values below .5 indicate "dissolved clusters", values between .6 and .75 indicate adequate stability, and values larger than .9 represent very stable clusters), indicated that two clusters obtained were very stable clusters.

3.2.7. STEP 5: CLUSTER REPLICATION

Finally, a replication analysis of the replication set is recommended to evaluate the robustness of the findings. In this replication analysis, several aspects of the replicability of the cluster findings are investigated. These include (1) the number of clusters (Step 2), (2) the set of signaling variables (Step 3), (3) the pattern of clusters as depicted in various visualizations (Step 4a), and (4) the internal stability index computed for each cluster (Step 4b). The replication analysis procedure is identical to the discovery analysis, as described in Steps 2 - 4. In our case study, the replication set `test.st` was analyzed in the same way as the discovery set `train.st`. To avoid repetitions, we only report the most important results here.

Once again, the 2-cluster solution emerged as the optimal solution in the replication analysis. However, this analysis identified a total of 59 signaling variables, including 52 variables that were also selected as signaling variables in the discovery set (i.e., `train.st`) plus 7 additional variables. In other words, the 52 signaling variables identified from the analysis of `train.st` can be considered the most robust set of items to differentiate between populists and non-populists in the Netherlands. In terms of interpretation, the same pattern of responses emerged, namely that the populists have less confidence in institutions, and, to

a greater extent, oppose progressive themes such as immigration, European unification, and gender equality. Last, in terms of cluster-wise stability, the replication analysis was also able to obtain highly stable clusters (the stability indices equaled .99 and .96 for the two clusters, respectively).

Overall, the results found in the replication analysis were very consistent with those reported in the discovery analysis in all aspects examined.

3.2.8. SUMMARY

To summarize, our Tutorial walks readers through five major steps for CKM analysis and cluster validation. Some of these steps can be adapted in research practices to address the specific needs of the investigation (e.g., incorporating prior knowledge of the number of clusters).

3.3. DISCUSSION

In this Tutorial, we provided a detailed guide for researchers on (1) how to apply simultaneous clustering and variable selection (SCVS) techniques and, especially, CKM and (2) how to use various methods to systematically validate cluster results. We demonstrated the use of CKM and its validation procedure through a case study of clustering people's political values whereby two clusters were identified and separated by 52 signaling variables. Filtering out these irrelevant variables, which do not contribute to cluster separation, improved the interpretation of cluster results and the accuracy of cluster recovery. To visualize cluster results, we present a novel, user-friendly ShinyApp with a variety of plotting options. The procedure described here can be applied to a variety of SCVS techniques developed within the K-means framework (e.g., Sparse K-means, Witten and Tibshirani, 2010; Sparse Alternate Sums, Arias-Castro and Pu, 2017; Robust and Sparse K-means, Kondo et al., 2016; Hard-thresholding K-means, Raymaekers and Zamar, 2020).

This tutorial is especially useful for cluster analyses with a large number of variables (e.g., ≥ 100), because some of the variables have little theoretical relevance to the clustering task at hand. However, we should point out that the analysis presented here is also very useful in research practice when the data set in question contains a limited set of theoretically relevant variables. In such cases,

this tutorial can assist researchers in determining which variables best separate the clusters. Consider a hypothetical study in which an organizational psychologist aims to identify clusters of employees suffering from various health conditions in order to provide personalized intervention. To this end, the psychologist can collect responses from these employees on 60 items. However, due to the limited scope of the intervention, the psychologist must select the five most important items for determining these employees' assigned interventions. The analysis procedure outlined in this tutorial can be used to choose the five most relevant items (by fixing the number of signaling variables to 5), estimate the cluster partition, and examine whether the results are methodologically stable and theoretically relevant.

The tutorial emphasized the importance of validating clustering results and illustrated three approaches for doing so. We hope that these discussions and illustrations promote and facilitate the use of cluster validation, particularly when the purpose of the study is to directly use the cluster assignment for high-stake decisions (e.g., personnel selection, personalized treatment). For data analysts, it is crucial to be mindful that cluster analysis is always exploratory; therefore, great care should be taken in interpreting and applying the results. A validation analysis is essential to assess the robustness of cluster solutions, and ideally, a confirmatory analysis should be followed to empirically test the conclusions from the previous cluster analysis.

APPENDICES

3.A. APPENDIX

The list of items used in the case study item1 'How satisfied or dissatisfied are you, generally speaking, about what the government has done lately?'
item2 'Confidence: Dutch government'
item3 'Confidence: Dutch parliament'
item4 'Confidence: the legal system'
item5 'Confidence: the police'
item6 'Confidence: politicians'
item7 'Confidence: political parties'
item8 'Confidence: European Parliament'
item9 'Confidence: United Nations'
item10 'Confidence: the media'
item11 'Confidence: the military'
item12 'Confidence: the education system'
item13 'Confidence: healthcare'
item14 'Confidence: science'
item15 'Confidence: the economy'
item16 'Confidence: democracy'
item17 'Confidence: shops/firms that you deal with personally (that you visit in person)'
item18 'Confidence: shops/firms on the Internet'
item19 'Satisfaction: Dutch government'
item20 'Satisfaction: Dutch parliament'
item21 'Satisfaction: the legal system'
item22 'Satisfaction: the police'
item23 'Satisfaction: politicians'
item24 'Satisfaction: political parties'

item25 'Satisfaction: European Parliament'
item26 'Satisfaction: United Nations'
item27 'Satisfaction: the media'
item28 'Satisfaction: the military'
item29 'Satisfaction: the education system'
item30 'Satisfaction: healthcare'
item31 'Satisfaction: science'
item32 'Satisfaction: the economy'
item33 'Satisfaction: democracy'
item34 'Satisfaction: shops/firms that you deal with personally (that you visit in person)'
item35 'Satisfaction: shops/firms on the Internet'
item36 'Parliamentarians do not care about the opinions of people like me'
item37 'Political parties are only interested in my vote and not in my opinion'
item38 'People like me have no influence at all on government policy'
item39 'I am well capable of playing an active role in politics'
item40 'I have a clear picture of the most important political issues in our country'
item41 'Politics sometimes seems so complicated that people like me can hardly understand what is going on'
item42 'Where would you place yourself on the scale below, where 0 means left and 10 means right?'
item43 'Where would you place yourself on a scale from 1 to 5, where 1 means that euthanasia should be forbidden and 5 means that euthanasia should be permitted?'
item44 'Where would you place yourself on a scale from 1 to 5, where 1 means that differences in income should increase and 5 means that these should decrease?'
item45 'Where would you place yourself on a scale of 1 to 5, where 1 means that immigrants can retain their own culture and 5 means that they should adapt entirely?'
item46 'Where would you place yourself on a scale from 1 to 5, where 1 means that European unification should go further and 0 means that it has already gone

too far'
item47 'A working mother's relationship with her children can be just as close and warm as that of a non-working mother.'
item48 'A child that is not yet attending school is likely to suffer the consequences if his or her mother has a job.'
item49 'Overall, family life suffers the consequences if the mother has a full-time job.'
item50 'Both father and mother should contribute to the family income.'
item51 'The father should earn money, while the mother takes care of the household and the family.'
item52 'Fathers ought to do more in terms of household work than they do at present.'
item53 'Fathers ought to do more in terms of childcare than they do at present.'
item54 'It is good if society consists of people from different cultures.'
item55 'It is difficult for a foreigner to be accepted in the Netherlands while retaining his/her own culture.'
item56 'It should be made easier to obtain asylum in the Netherlands.'
item57 'Legally residing foreigners should be entitled to the same social security as Dutch citizens.'
item58 'There are too many people of foreign origin or descent in the Netherlands.'
item59 'People of foreign origin or descent are not accepted in the Netherlands.'
item60 'Some sectors of the economy can only continue to function because people of foreign origin or descent work there.'
item61 'It does not help a neighborhood if many people of foreign origin or descent move in.'
item62 'Married people are generally happier than unmarried people.'
item63 'People that want to have children should get married.'
item64 'A single parent can raise a child just as well as two parents together.'
item65 'It is perfectly fine for a couple to live together without marriage intentions.'
item66 'For a couple that wants to get married, it is good to first start living together.'

item67 'A divorce is generally the best solution if a married couple cannot solve their marital problems.'
item68 'It is all right for a married couple with children to get divorced.'
item69 'Children ought to care for their sick parents.'
item70 'When parents reach old age, they should be able to live with their children.'
item71 'Children that live close by ought to visit their parents at least once a week.'
item72 'Children ought to take unpaid leave in order to care for their sick parents.'
item73 'You can only do what you feel like doing after you have done your duty.'
item74 'If someone wants to enjoy life, he/she must be prepared to work hard for it.'
item75 'I feel happiest after working hard.'
item76 'Work should always come first, even if it means having less leisure time.'
item77 'If she has a baby (a child younger than 1 year).'
item78 'If she has a child that does not yet attend school.'
item79 'After the youngest child starts primary school.'
item80 'After the youngest child starts secondary school.'
item81 'Trade unions should take a much tougher political stance, if they wish to promote the workers' interests.'
item82 'Trade unions should advise their members to vote for those parties that best promote the workers' interests.'
item83 'A woman is more suited to rearing young children than a man.'
item84 'It is actually less important for a girl than for a boy to get a good education.'
item85 'Generally speaking, boys can be reared more liberally than girls.'
item86 'It is unnatural for women in firms to have control over men.'

4

REVEALING SUBGROUPS THAT DIFFER IN COMMON AND DISTINCTIVE VARIATION IN MULTI-BLOCK DATA: CLUSTERWISE SPARSE SIMULTANEOUS COMPONENT ANALYSIS

Social and behavioral studies more and more often yield multi-block data, which consists of novel blocks of data (e.g. data from wearable devices) and traditional blocks of data (e.g. survey data) collected from the same sample. Multi-block data offer researchers valuable insights into complex social mechanisms where several influences act together. Yet, such mechanisms are likely to differ among

This chapter is published as **Yuan, S.**, De Roover, K., Dufner, M., Denissen, J. J., & Van Deun, K. (2021). Revealing subgroups that differ in common and distinctive variation in multi-block data: Clusterwise sparse simultaneous component analysis *Social Science Computer Review*, 39(5), 802-820

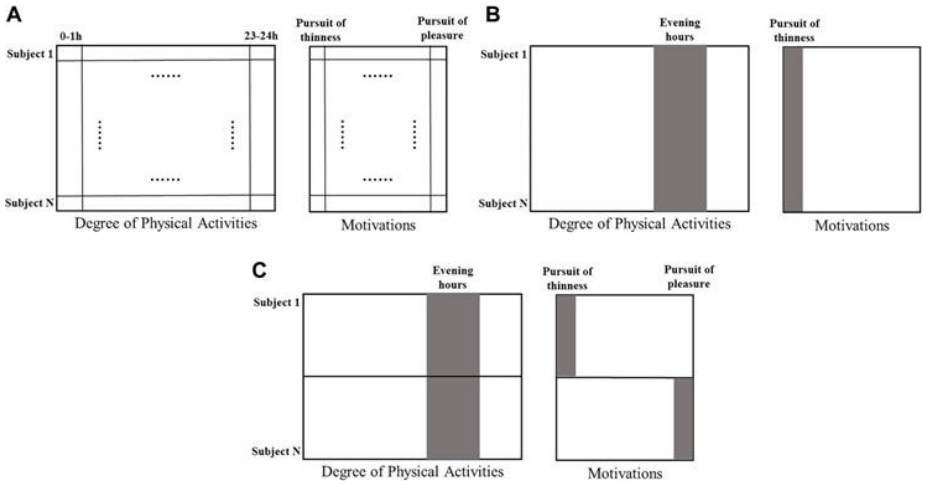
subgroups. Hence, fully revealing the composite mechanisms underlying multiblock data is challenging, since proper clustering analysis of such data requires a method that simultaneously detects the covariation of variables underlying all data blocks and the group differences therein. Additionally, such a method should be able to handle high-dimensional data sets that might include many irrelevant variables. Here we present Clusterwise Sparse Simultaneous Component Analysis (CSSCA), a method that groups the subjects that are driven by the same mechanisms and, at the same time, extracts cluster-specific components that model these mechanisms. By imposing structure constraints, CSSCA further recognizes common mechanisms that underlie all data blocks and distinctive mechanisms that only underlie one or a few data blocks. In extensive simulations, CSSCA delivered convincing results in recovering the clusters and their associated component structures across various conditions. More importantly, CSSCA showed a clear advantage over existing methods when substantial cluster differences in the component structures were present. We demonstrated the usefulness of CSSCA in an application to data stemming from a study on personality.

4.1. INTRODUCTION

Thanks to recent technological developments and the increasing adaption of data-rich research in social and behavioral sciences (Gil de Zuniga & Diehl, 2017), novel types of data such as genetic data, global positioning system coordinates, and social media data are collected more and more often, along with traditional sociodemographic and questionnaire data (Hofferth et al., 2017). Such linked data that contain different types of measurements, collected from the same sample, are labeled multi-block data. In the domain of communication science, for example, Wells and Thorson (2017) proposed and demonstrated a novel method to examine political content flows by linking social media data with survey data. Another example comes from Vargo and Hopp (2017), who identified the connections between individuals' political polarization and the extent of civil conversation with a multi-block of tweets and census data.

Studies based on multi-block data have the potential to advance social and behavioral sciences: it offers opportunities to obtain novel insights into complex social mechanisms where several influencing factors – each of them reflected by a particular data block - act jointly. Let us consider an illustrative example of multi-block data, as depicted in Figure 4.1A, with rows referring to subjects and columns representing variables. The multi-block data consists of two blocks: one block covering self-reported motivations (each column represents one type of motivation) and one block covering participants' degrees of physical activities, measured by wearable devices and aggregated across several time intervals (each column represents the average degree of physical activity per hour). With such multi-block data, health psychologists would be able to investigate how different types of motivations are related to different patterns of physical activities.

Because of a lack of theoretical knowledge about the novel types of data and (or) their linkage with traditional data, exploratory analyses could offer important insights into the structure of the data (Fan et al., 2014). In our illustrative example, appropriate exploratory analyses should detect crucial yet subtle links between motivations, as reflected by some variables in the self-report data block, and patterns of physical activity, as reflected by several variables in the physical activity data block. A potential outcome is illustrated in Figure 4.1B, with the columns marked by grey implying associations between variables from different



Note. The multi-block data includes two data blocks, with column entries referring to variables and row entries to subjects. Panel A illustrates the data structure of multi-block data. Panel B suggests a potential outcome of a data analysis that reveals variables that are associated with the different data blocks (marked with gray color). Panel C visualizes heterogeneity between subjects in the variables that are associated with the data blocks.

Figure 4.1: Graphic visualization of the illustrative example.

data blocks: Figure 4.1B demonstrates that the pursuit of thinness is linked with intensive physical activities during evening hours. In essence, the aim of multi-block data analysis is to identify the common variation – typically implying synergies between variables – underlying all data blocks (De Roover, Timmerman, et al., 2013; Van Deun et al., 2011).

A unique feature of the multi-block data is that, in addition to the common variation, they also contain distinctive variation, which refers to the covariation of variables from one or a few - but not all - data blocks (Lock et al., 2013; Van Deun et al., 2011). Concerning our illustrative example, while the researchers mainly want to detect the common variation (i.e., the covariation of motivations and patterns of physical activities), the multi-block data may also include distinctive variation, such as response styles underlying self-reports of motivations (e.g., Harzing, 2006) and individual baseline levels of physical activity underlying data derived from wearable devices. Hence, to extract the common variation that is typical of interest, it is necessary to partial out the distinctive variation, which, in some cases, might also contain substantive information that is of

interest to the researchers. Because of the presence of both common and distinctive variations in the multi-block data, conventional ways of analyzing such multi-block data might be less desirable. On the one hand, principal component analysis (PCA; Jolliffe, 2002; Meredith & Millsap, 1985) could be applied to the concatenated data blocks to summarize the associations between variables by a few components. However, this analysis would only detect the components explaining the largest (co)variation across the blocks, which are likely to describe a mixture of common and distinctive variations. Therefore, this approach is not able to uncover both common and distinctive variations. On the other hand, one could also first perform a separate analysis (e.g., PCA) on each data block and then integrate the results over all data blocks. However, as pointed out by M. Wang and Hanges (2011), this approach to data analysis has two noteworthy shortcomings: (1) it is likely to omit the important common variation, and (2) its performance deteriorates with increasing disparities between the results of separate analyses.

Recently, some component-based integrative analysis methods, most noticeably JIVE (Lock et al., 2013) and DISCO-SCA (Schouteden et al., 2013), have been introduced and gained substantial popularity in multi-block data analysis. A particularly useful feature of these methods is their capability to effectively discern the common and distinctive variations. These methods have been successfully applied in psychology (e.g., Chawarska et al., 2016; Gu & Van Deun, 2019), neuroscience (e.g., Yu et al., 2017), biology (e.g., Wehrens & Salek, 2019), and medicine (e.g., Sandri et al., 2018), among other research fields. Nevertheless, these methods fail to overcome two additional challenges of multi-block data analysis.

First, multi-block data frequently include data sets of a high-dimensional nature (i.e., an equal or greater number of variables than subjects). With very little theoretical guidance, researchers often, by default, include all information they have gathered in the analysis, leading to a substantial amount of redundant information (Waldherr et al., 2017). This severely hampers the interpretation of the components and makes it intricate to reveal the variables that are most interesting for further investigation, since the components may correlate with a large number of irrelevant variables (Zou et al., 2006). Therefore, methods are needed that can automatically and effectively filter out irrelevant variables. Note that

the common practice of dropping out variables with small loadings (i.e. treating them as zero loadings) yields a sub-optimal solution, as first discussed in Cadima and Jolliffe (1995).

Another challenge of analyzing multi-block data is the heterogeneity among subjects: subgroups may be present in the data that differ in the patterns of covariation (Jung & Wickrama, 2008). In our illustrative example, the association between motivations and degree of physical activity may differ among subjects. For instance, as demonstrated in Figure 4.1C, the degree of physical activity in the evening hours may be associated with the pursuit of thinness among some subjects (the first half) and with the pursuit of pleasure among others (the second half). The presence of subgroups is often not known to the researchers beforehand. Hence, a clustering method is needed that can reveal the subgroups of subjects, with the desired result that (only) participants who belong to the same subgroup have a similar pattern of covariation.

To respond to these challenges, we present Clusterwise Sparse Simultaneous Component Analysis (CSSCA), a novel method designed for multi-block data analysis. CSSCA assigns all subjects to mutually exclusive clusters, such that the subjects that belong to the same cluster have the same common and distinct components, while the subjects that belong to different clusters are assumed to vary on different common and distinctive components.

The remainder of the paper is organized into five sections: in Section 4.2, we formally introduce CSSCA and contrast it with several existing methods. The performance of CSSCA and its model selection procedure are evaluated in Section 4.3. The usefulness of CSSCA for applied psychological research is demonstrated in Section 4.4. Finally, the implications, limitations, and a blueprint for future research are elaborated in section 4.5. To increase the accessibility of the method, we have made CSSCA, its model selection procedure, as well as other auxiliary functions, available in the R package `ClusterSSCA`. The package can be downloaded freely from https://github.com/syuanuvt/CSSCA. On the same webpage, we have also provided a step-by-step user guide to facilitate the usage of CSSCA in applied research.

4.2. METHOD

In this section, we present CSSCA by specifying the assumed data generation model and objective function. First, however, we introduce multi-block data from a formal point of view and discuss existing methods that serve as the building blocks of CSSCA.

4.2.1. MULTI-BLOCK DATA

Multi-block data consist of multiple blocks of data containing information about the same group of respondents (Tenenhaus & Tenenhaus, 2014). More formally, each of the L data blocks $X_l(N \times J_l)(l = 1, 2 \ldots, L)$ contains values of N subjects on J_l variables. A popular framework for analyzing multi-block data is simultaneous component analysis (SCA; Kiers & ten Berge, 1989; Van Deun et al., 2009), from which CSSCA originates.

4.2.2. SCA

Similar to PCA, SCA reduces the dimensions of all data blocks simultaneously and results in a few components that maximally account for the total variation across the data blocks. Formally, the SCA model, as proposed in Timmerman and Kiers (2003), is represented in

$$\mathbf{X}_l = \mathbf{T}\mathbf{P}_l^T + \mathbf{E}_l, \tag{4.1}$$

where $\mathbf{T}$ with size $N \times R$ denotes the simultaneous component scores on R components (i.e., $\mathbf{T}$ is assumed to be the same for each of the data blocks), $\mathbf{P}_l$ with size $J_l \times R$ denotes the component loadings of the variables in the l^{th} data block and $\mathbf{E}_l$ with size $N \times J_l$ denotes the error matrix associated with the l^{th} data matrix $\mathbf{X}_l$. For SCA-based methods, usually, all variables are mean-centered and standardized (see Van Deun et al., 2009). To identify the solution, Equation 4.1 is made subject to suitable constraints, for example, a principal axis orientation in combination with the orthogonality of the component scores: $\mathbf{T}^T\mathbf{T} = \mathbf{I}$. The objective of SCA is to minimize the sum-of-squares of residuals, given by Equation 4.2 as follows,

$$\mathbf{argmin}_{\mathbf{T},\mathbf{P}^{con}} ||\mathbf{X}^{con} - \mathbf{T}\mathbf{P}^{con^T}||_2^2, \tag{4.2}$$

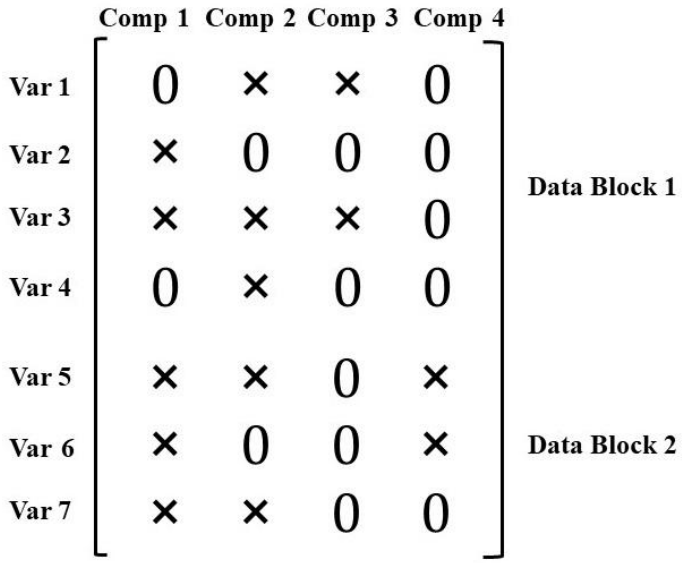
subject to $\mathbf{T}^T\mathbf{T} = \mathbf{I}$, where $\mathbf{X}^{con}$ of size $N \times J$ denotes the concatenated data matrix (J equals the total number of variables across all data blocks) and $||\mathbf{X}^{con} - \mathbf{T}\mathbf{P}^{con^T}||_2^2$ denotes the square of the Frobenius norm of ($\mathbf{X}^{con} - \mathbf{T}\mathbf{P}^{con^T}$). $\mathbf{P}^{con}$ with size $J \times R$ is the concatenated component loading matrix.

As pointed out in Van Deun et al. (2011), SCA fails to appropriately address two of the most important challenges of multi-block data analysis. First, the interpretation of the resulting components is daunting as it is based on the contributions of all variables. Second, the components obtained by SCA do not account for the block structure; in particular, they do not separate the common and distinctive sources of variation. Solutions have been proposed to address the two drawbacks of SCA, resulting in SSCA (i.e. sparse SCA) with common and distinctive components.

SSCA WITH COMMON AND DISTINCTIVE COMPONENTS

To tackle the first challenge of automatic variable selection and thus to ease the interpretation of components, especially in dealing with high-dimensional data sets, regularization has been used to shrink some component loadings to (exact) zero (hereafter these entries will be called sparseness-induced zero loadings), leading to SSCA (Van Deun et al., 2011). Different forms of regularization can be used in SSCA; here in developing CSSCA, we adopt the l_0 norm regularization (also known as a cardinality constraint). This constraint fixes the number of zero elements in the loading matrices to a pre-defined number with a range between 0 and $J \times R$ and thereby allows fixing the proportion of zero loadings (called the level of sparsity hereafter and indicated by $spar()$).

To approach the challenge of discerning common and distinctive variations, Schouteden et al. (2013) has proposed DISCO-SCA that determines the status of components (i.e. common or distinctive components) through rotations. To further avoid post-hoc rotations, Gu et al. (2019) directly imposed zero loadings in a structured way, which results in an unambiguous status of each component. Specifically, to define a distinctive component, except for the variables of the block(s) that the component is supposed to underlie, all other loadings on this distinctive component are fixed to zero (this type of zero loadings are hereafter called the distinctiveness-induced zero loadings). We illustrate this structure



Note. The components are represented in columns, while the variables are indicated in rows. The first two components are defined as common components while the third and fourth components are distinctive components pertaining to block 1 and block 2, respectively. Var $j = j^{th}$ variable, Comp $r = r^{th}$ component.

Figure 4.2: An example of common and distinctive components in a concatenated loading matrix.

for a loading matrix that includes both sparseness-induced zero loadings and distinctiveness-induced zero loadings in Figure 4.2. The depicted loading matrix includes 7 variables (rows) from 2 data blocks (the first data block has 4 variables while the second has 3) and 4 components (columns), and zero loadings are denoted by "0" while non-zero loadings are denoted by "×". In the figure, the first two components are sparse common components, since they are associated with variables from both data blocks. The third component, with all non-zero loadings associated with variables in data block 1, is a sparse distinctive component that pertains to block 1. In the same vein, the fourth component can be regarded as a sparse distinctive component that pertains to block 2.

Formally, for the analysis of SSCA with common and distinctive components, the objective function is described in

$$\mathbf{argmin}_{\mathbf{T},\mathbf{P}^{con}} ||\mathbf{X}^{con} - \mathbf{T}\mathbf{P}^{con^T}||_2^2, \tag{4.3}$$

subject to (i) $\mathbf{T}^T\mathbf{T} = \mathbf{I}$, (ii) $spar(\mathbf{P}^{con}) = S$, where S is a pre-defined number between 0 and 1 that indicates a pre-defined level of sparsity of the loading matrices, and (iii) distinctiveness-induced zero loadings are pre-specified in $\mathbf{P}^{con}$ to impose common and distinctive components.

4.2.3. CLUSTERWISE SPARSE SIMULTANEOUS COMPONENT ANALYSIS

CSSCA extends SSCA to account for heterogeneity in the mean structure and the component structure. Specifically, instead of assuming that the same component loading matrix pertains to all subjects, a few loading matrices are assumed to underlie the multi-block data, where each applies to a particular subgroup of subjects. CSSCA aims to detect these subgroups (also called clusters) and their associated mean structures and component structures.

MODEL AND OBJECTIVE FUNCTION

Formally, the cluster-specific model of CSSCA on the level of the concatenated data is given by

$$\mathbf{X}_k^{con} = \boldsymbol{\mu}_k^{con} + \mathbf{T}_k\mathbf{P}_k^{con^T} + \mathbf{E}_k (k = 1, \ldots, K), \tag{4.4}$$

subject to (i) $\mathbf{T}_k^T\mathbf{T}_k = \mathbf{I}$ and $\mathbf{T}_k'\mathbf{1} = \mathbf{0}$, (ii) $Spar(\mathbf{P}_k^{con}) = S$, and (iii) distinctiveness-induced zero loadings are pre-specified in $\mathbf{P}_k^{con}$ to impose common and distinctive components. In Equation 4.4, $\mathbf{X}_k^{con}(N_k \times J)$, $\mathbf{T}_k(N_k \times R)$ and $\mathbf{P}_k^{con}(J \times R)$ denote the concatenated data, the component score matrix, and the component loading matrix of Cluster k, respectively, while $\boldsymbol{\mu}_k^{con}(N_k \times J)$ with all identical rows representing the mean structure of Cluster k. Note that, in addition to these constraints, CSSCA assumes the same number of common components and also the same structure of distinctive components for each of the clusters. In other words, the method assumes that the dimensions of the loading matrices as well as the positions of distinctiveness-induced zero loadings are identical across clusters. This is because we would like to keep CSSCA a simple method in terms of model

selection and interpretation. The objective function of CSSCA is presented in

$$\textbf{argmin}_{\boldsymbol{\mu}_k, \mathbf{T}_k, \mathbf{P}_k^{con}} \mathbf{P}^{con} \sum_{k=1}^{K} ||\mathbf{X}_k^{con} - \boldsymbol{\mu}_k^{con} - \mathbf{T}_k \mathbf{P}_k^{con^T}||_2^2, \tag{4.5}$$

subject to (i) $\mathbf{T}_k^T \mathbf{T}_k = \mathbf{I}$ and and $\mathbf{T}_k' \mathbf{1} = \mathbf{0}$, (ii) $spar(\mathbf{P}_k^{con}) = S$, and (iii) pre-specified distinctiveness-induced zero loadings in $\mathbf{P}_k^{con}$.

4.2.4. RELATED METHODS

A number of related dimension-reduction-based clustering methods have been developed for the analysis of single-block data: e.g., Reduced K-means (RKM; Stute & Zhu, 1995), Factorial K-means (RKM; Vichi & Kiers, 2001), and Subspace K-means (Timmerman et al., 2013). As argued in the introduction, the clustering analyses carried out on the concatenated data set fail to distinguish the common and distinctive components. Thus, they are less desirable in the analysis of multi-block data.

Recently, some clustering methods for multi-block data have been proposed in the field of bioinformatics. In their systematic review, D. Wang and Gu (2016) classified all these methods into two categories, and they demonstrated the advantages of the methods with a direct integrative clustering strategy, which, instead of first performing a separate clustering analysis on each data block and then integrating all partitions, accounts for all data blocks simultaneously. CSSCA, with its simultaneous dimension reduction of all data blocks, clearly falls into this category.

Among the clustering methods that also employ the direct integrative approach, iCluster (R. Shen et al., 2009) is a popular choice and it also lays the basis for several succeeding methods, including the low-rank approximation clustering method (LRAcluster; Wu et al., 2015) and Joint and Individual Clustering (JIC; Hellton and Thoresen, 2016). In essence, iCluster projects the high-dimensional data onto a lower-dimensional subspace by summarizing the common variation over multiple data blocks into several latent variables. Subsequently, iCluster utilizes K-means clustering to obtain cluster assignments from the resulting latent variables. iCluster (and other mean-level-based methods) differs from CSSCA in that iCluster does not actively model the underlying covariance structures per-

taining to each cluster, while CSSCA allows the covariance structures to differ across clusters. In this respect, CSSCA offers an important extension to the existing methods.

We can briefly conclude that CSSCA is the only clustering method available so far, that, in the context of multi-block data analysis, partitions subjects based on both mean structures and covariance structures.

4.2.5. ALGORITHM AND MODEL SELECTION

ALGORITHM

Starting from a random partition of the subjects, the CSSCA algorithm obtains an SSCA solution for each of the initial clusters. Subsequently, the procedure iterates over a loop in which the subjects are re-assigned one by one: for each subject, the SSCA solution is obtained for each of the $K-1$ potential re-assignments and the subject is assigned to the cluster with which the total loss is minimized (implying that the total loss is guaranteed to be non-increasing for each update of the cluster membership). After a complete iteration of re-assigning all subjects, the algorithm starts the next iteration if and only if (1) the total decrease in loss value of the current iteration is larger than a pre-defined value, and (2) the number of iterations is smaller than a pre-defined maximum. Since the algorithm may result in local optima, a multi-start procedure is used (e.g., De Roover, Ceulemans, Timmerman, and Onghena, 2013; Timmerman et al., 2013). Using pseudocode, we present in Algorithm 3 the algorithm of CSSCA (see Section 4.A of the appendix). Embedded in Algorithm 3 is the iterative procedure to estimate the cluster-specific SSCA solution, which applies an alternating strategy first proposed in Gu and Van Deun (2019). In essence, Algorithm 4 iteratively optimizes $\mathbf{P}_k^{con}$ conditional on $\mathbf{T}_k$, and optimizes $\mathbf{T}_k$ conditional on $\mathbf{P}_k^{con}$, using well-established optimization routines. The procedure is detailed in Algorithm 4 in the appendix (Section 4.B), also in the form of pseudocode. Similar to Algorithm 3, Algorithm 4 yields a non-increasing sequence of loss values throughout the iterations, and thus guarantees to converge to a fixed point. In Section 4.C of the appendix we report some technical details on the generation of the starting partitions and on some additional requirements of the model parameters. More

information about the implementation of the CSSCA algorithm can be found in the R package `ClusterSSCA`.

MODEL SELECTION

To run the CSSCA algorithm, the actual model parameters (e.g. the number of clusters and the level of sparsity) need to be specified. In practice, however, researchers only have limited or no knowledge of the true values of these parameters. To facilitate the application of CSSCA, we propose a model selection procedure to determine the number of clusters and the level of sparsity with the best balance between model fit (i.e. the total loss) and model complexity.

Wilderjans et al. (2013) showed that a sequential model selection strategy may have several advantages. Adapted to solving the model selection problem of CSSCA, the sequential strategy includes two steps: (1) picking the optimal number of clusters K and (2) determining the optimal level of sparsity S, given the selected K. To illustrate our model selection procedure, assume that K and S are selected from ascending candidate sets (K_1, K_2, …, K_U) and (S_1, S_2, …, S_V), respectively. In the first step, as illustrated in Equation 4.6, the conditional scree ratio $sr(K_u|S_v)$ is computed for each possible pair of K_u (K_1, K_2, …, K_U), and S_v (S_1, S_2, …, S_V)

$$sr(K_u|S_v) = \frac{\frac{Loss(K_{u-1},S_v)-Loss(K_u,S_v)}{K_u-K_{u-1}}}{\frac{Loss(K_u,S_v)-Loss(K_{u+1},S_v)}{K_{u+1}-K_u}}, \tag{4.6}$$

with $Loss(K_u, S_v)$ referring to the total loss resulting from the CSSCA analysis with the number of clusters set to K_u and the level of sparsity to S_v. Afterwards, for each possible value of K_u ($K_u = K_2, \ldots, K_U - 1$), the average conditional scree ratio $sr(K_u)$ is computed by averaging $sr(K_u|S_v)$ over all possible values of S_v ($S_v = S_1, S_2, \ldots, S_V$), as

$$sr(K_u) = \frac{\sum_v sr(K_u|S_v)}{V}. \tag{4.7}$$

The optimal number of clusters K_{opt} is then determined by maximizing $sr(K_u)$. In the second step, conditional on the optimal number of clusters K_{opt}, the conditional scree ratio $sr(Sv|K_{opt})$ can be calculated for each $S_v(S_v = S_2, \ldots, S_V - 1)$, as shown in

$$sr(S_v|K_{opt}) = \frac{\frac{Loss(K_{opt},S_{v+1})-Loss(K_{opt},S_v)}{S_{v+1}-S_v}}{\frac{Loss(K_{opt},S_v)-Loss(K_{opt},S_{v-1})}{S_v-S_{v-1}}}. \tag{4.8}$$

Again, we select the optimal level of sparsity by maximizing the conditional scree ratio.

It is important to note that, according to Equations 4.6 and 4.8, $sr(K_u|S_v)$ is not defined if $K_u = K_1$ (minimum) or $K_u = K_U$ (maximum) and $sr(S_v|K_{opt})$ is not defined when $S_v = S_1$ (minimum) or $S_v = S_V$ (maximum). Therefore, the sequential approach does not allow for the selection of the minimal and maximal values of K and S.

4.3. SIMULATION STUDIES

To investigate the performance of CSSCA and its model selection procedure, we conducted two simulation studies. In simulation study 1, the performance of CSSCA given the correct number of clusters and level of sparsity was evaluated and compared with the performance of iCluster in various conditions. The proposed model selection procedure for CSSCA was examined in simulation study 2.

4.3.1. SIMULATION STUDY 1

DESIGN

Three model characteristics that were expected to have relatively small impacts on the clustering performance were kept constant: the number of data blocks $L = 2$, the number of common components $R_c = 2$, and the number of distinctive components in each data block $R_l = (1, 1)$.

The following eight factors were used to create the various conditions:

1 The number of variables J_l: low-dimensional condition ($J_l = (15, 15)$) and high-dimensional condition ($J_l = (15, 50)$). Hence, the total number of variables J was 30 in low-dimensional conditions and 65 in high-dimensional conditions.

2 The number of clusters K: small ($K = 2$) and large ($K = 4$).

3 The cluster size N_k: small ($N_k = 50$ or 30, dependent on Factor 4) and large ($N_k = 100$ or 60, dependent on Factor 4).

4 The equality of cluster size. In the equality conditions, each cluster contained 50 or 100 subjects while in the inequality conditions, one cluster contained 30 or 60 subjects, and the rest contained 50 or 100 subjects (see Factor 3).

5 The level of sparsity S (of loading matrices): low (.3), medium (.5), and high (.7).

6 The proportion of the structural variance accounted for by the mean structure b: small (.1), medium (.5), and large (.9). Since the mean structure is assumed to be equal for all subjects of the same cluster, b also represents cluster differences in the mean structures. Excluding the variance accounted for by the noise structures, the remaining $1-e$ of the total variance (which is also called structural variance hereafter) can be decomposed into variance caused by cluster differences in the mean structures and in the component structures. As such, $b(1-e)$ of the total variance can be attributed to cluster differences in the mean structures. Mathematically, $b = \frac{\sum_{k=1}^{K} tr(\boldsymbol{\mu}_k' \boldsymbol{\mu}_k)}{\sum_{k=1}^{K} tr(\boldsymbol{\mu}_k' \boldsymbol{\mu}_k) + \sum_{k=1}^{K} tr(\mathbf{P}_k \mathbf{T}_k' \mathbf{T}_k \mathbf{P}_k')}$

7 The proportion of the total variance accounted for by the noise structure, or the noise level of the data, e: low (.1), medium (.2), and high (.3). Mathematically, $e = \frac{tr(\mathbf{E}_k' \mathbf{E}_k)}{\sum_{k=1}^{K} tr(\boldsymbol{\mu}_k' \boldsymbol{\mu}_k) + \sum_{k=1}^{K} tr(\mathbf{P}_k \mathbf{T}_k' \mathbf{T}_k \mathbf{P}_k') + tr(\mathbf{E}_k' \mathbf{E}_k)}$

8 The average congruence level ϕ of cluster-specific loadings: low (approximately .2) and high (approximately .53). Here, congruence is measured by the average Tucker congruence (Haven & ten Berge, 1977; Tucker, 1951) between the cluster-specific loadings across all pairs of clusters.

In total, the full factorial design of the eight factors resulted in 2 x 2 x 2 x 2 x 3 x 3 x 3 x 2 = 864 conditions. In each condition, we generated 40 replications. Hence, a total of 34,560 data sets were created and analyzed. The data generation procedure is detailed in the appendix (Section 4.D).

RESULTS AND DISCUSSION

Over all 34,560 data sets, the average execution time of CSSCA was 774 seconds, or around 13 minutes. In the simulation, the maximal size of the data sets was

400 rows by 65 columns when $K = 4$, $N_k = 100$, and $J = 65$. For each of these data sets, CSSCA spent an average of 1900 seconds, or around 31 minutes. Overall, taking into consideration that the computation speed can be greatly improved by the parallel computation function available in the R package `ClusterSSCA`, the execution time of CSSCA should be considered acceptable for applied research.

The main indicator of the clustering performance is the accuracy of cluster recovery, i.e., how well the partition produced by CSSCA recovers the true partition (i.e., the partition used to generate the data sets). A widely-used measure of cluster recovery is the Adjusted Rand Index (ARI; Hubert and Arabie, 1985). ARI takes values between 0 and 1, with 0 indicating that the overlap between the two cluster partitions is at the chance level and 1 suggesting a complete overlap between the two cluster partitions. In the current study, the estimated ARI of the recovered cluster partition and true cluster partition was used as the indicator of cluster recovery.

We expected that two of the eight factors – the mean-level cluster differences b and the noise level of the predictor blocks e – would have the strongest impact on the cluster recovery of CSSCA. First, a larger b means that the component structure accounts for a smaller proportion of the structural variance, and, in extreme cases, can become very small compared to the error variance (e.g., the component structure accounts for 7% of the total variance while the noise accounts for 30%). In such cases, it can be expected that the component structure is masked by the noise. Second, a larger e results in the fact that the true data structure is masked by a larger amount of noise, and the true cluster partition is therefore more difficult to be recovered.

The results of the simulation studies fit the expectations well. We found that both b and e were indeed among the most influential factors, and that a better recovery of the clusters, averaged across replications, was obtained when (1) b was smaller (ARI = .997 when $b = .1$, ARI = .999 when $b = .5$, and ARI = .986 when $b = .9$), and (2) e was smaller (ARI = 1 when $e = .1$, ARI = .999 when $e = .2$, and ARI = .984 when $e = .3$). The average cluster recovery of CSSCA as a function of the other six factors is reported in the appendix (Section 4.E).

To further investigate the effects of the interactions between b and e on the cluster recovery of CSSCA, we examine the average ARI in cross-tabulation of

the two factors, as illustrated in Table 4.1. In all conditions, the average ARI between the resulting partitions and true partitions was above .95. Thus, in general, CSSCA yielded an adequate cluster recovery, according to the widely-adopted criterion proposed in Steinley (2004). When the proportion of mean-level differences was low or medium, CSSCA recovered the true clusters exceptionally well (i.e. ARI > .99), even with relatively noisy data. Table 4.1 also reveals that the worst ARI (ARI = .96) is obtained for the combination of large b and large e, as expected.

Table 4.1: The means (and the standard deviations between brackets) of ARI between the true cluster partitions and the CSSCA-recovered cluster partitions in various conditions

	Proportion of mean-level differences: low (b = .1)	Proportion of mean-level differences: medium (b = .5)	Proportion of mean-level differences: high (b = .9)
Noise level: low (e = .1)	1(0)	1(0)	1(0)
Noise level: medium (e = .2)	1(0)	1(0)	.99(.04)
Noise level: high (e = .3)	.99(.02)	.99(.02)	.96(.18)

Figure 4.3 illustrates the results of the comparison between the average cluster recovery of CSSCA and that of iCluster for various levels of b (i.e. the proportion of structural variance accounted for by the mean structure). Clearly, when the structural variance mainly pertained to the component structure (i.e. b = .1), CSSCA drastically outperformed iCluster with an ARI of .997 compared to only .105 for iCluster. In line with our expectations, CSSCA demonstrated an overwhelming advantage in terms of cluster recovery when the component structure is the predominant source of variation. The superior performance of CSSCA persisted when the cluster differences in the mean structures and component structures contributed equally to the structural variance (i.e. b = .5), where CSSCA achieved an average ARI of almost 1 (.999 to be exact) while iCluster obtained an average ARI of .93. When b equaled .9, CSSCA could still recover the clusters

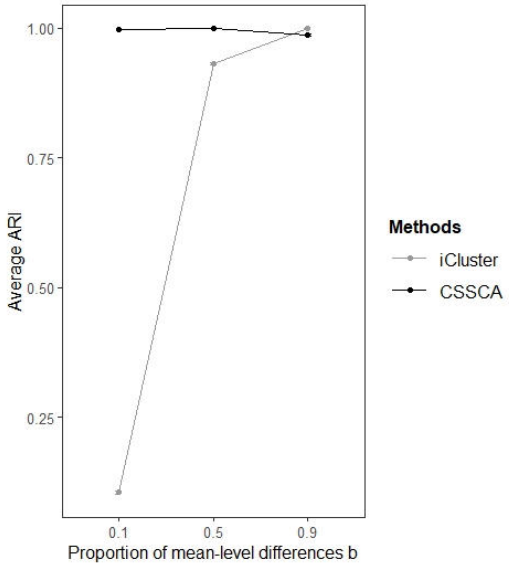


Figure 4.3: The means and the 95% confidence intervals of the Adjusted Rand Index between the true cluster partitions and the recovered cluster partitions of CSSCA and iCluster with different values of b

very well (average ARI = .986), but the clusters obtained with iCluster were more accurate (average ARI = 1). In general, CSSCA has demonstrated consistent and convincing performance in terms of cluster recovery across all conditions. While CSSCA achieved a good cluster recovery even in unfavorable conditions, iCluster did not perform better than the chance level in the most difficult condition.

We also measured the correspondence between the estimated cluster-specific loading matrices and the true loading matrices that were used to generate the data, which we quantified by the goodness-of-cluster-loading-recovery statistic (GOCL, see De Roover, Ceulemans, and Timmerman, 2012). GOCL was calculated by first obtaining Tucker's congruence coefficients between the corresponding components of the true and estimated loading matrices and then averaging across all components and clusters. Since iCluster only detects mean-level cluster differences, GOCL is not available for the iCluster results. With an average GOCL equaling .95 over all data sets, CSSCA appeared to perform very well in recovering the cluster-specific loading matrices. Since the recovery of the loading matrices largely depends on the recovery of the cluster partitions, we

would expect the factors b and e to be also important in predicting CSSCA's performance in recovering the cluster-specific loading matrices. A cross-tabulation of the average GOCL in the function of these two factors is presented in Table 4.2. Similarly, the average GOCL reached its lowest value (GOCL = .745) for the combination of a large b and large e, i.e., the condition where the true component structure was severely masked by the noise.

Table 4.2: The means and the standard deviations (in brackets) of GOCL between the true loading matrices and the CSSCA-recovered loading matrices in various conditions

	Proportion of mean-level differences: low(b = .1)	Proportion of mean-level differences: medium (b = .5)	Proportion of mean-level differences: high (b = .9)
Noise level: low (e = .1)	.99(.02)	.99(.01)	.96(.03)
Noise level: medium (e = .2)	.99(.02)	.98(.02)	.89(.06)
Noise level: high (e = .3)	.98(.02)	.97(.03)	.77(.11)

4.3.2. SIMULATION STUDY 2

We evaluated the accuracy of the model selection procedure in simulation study 2. From simulation study 1, it was clear that the two most influential factors determining CSSCA's performances were (1) the proportion of mean-level cluster differences b, and (2) the error level e. Both factors were retained in simulation study 2 (note that in this study, e had two levels: e = .15 or .3). Since the level of sparsity S and the number of clusters K are to be selected, they have also been added as varying factors in simulation study 2. The true level of sparsity S_{true} is either .3 or .7, and the true number of cluster K_{true} is either 2 or 4. In total, 576 data sets were created. We executed the model selection procedure for all data sets with K being selected from [1, 2, 3, 4, 5, 6, 7], and S from [.2, .3, .4, .5, .6, .7, .8]. Over all 576 data sets, both K and S were correctly selected in 194 data sets (33.68%). In 298 data sets (51.73%), only K_{true} (but not S_{true}) was cor-

rectly selected, while in 14 data sets (2.43%), only S_{true} (but not K_{true}) was successfully selected. Overall, the proposed model selection procedure performed reasonably well in recovering the number of clusters (K_{true} was successfully recovered in 492, or 85.42%, of the data sets). This procedure, however, was less successful in determining the level of sparsity, where it only succeeded for a total of 208 data sets (36.11%). It is important to note, however, that in most cases the selected level of sparsity differed from the actual level only by a small margin of 0.1. Furthermore, we found that the model selection process of CSSCA was more successful (i.e., both K_{true} and R_{true} were selected correctly) when (1) b was of small-to-medium size (46.35% when $b = .1$ versus 47.40% when $b = .5$ versus 7.29% when $b = .9$) and (2) e was small (36.10% when $e = .15$ versus 31.25% $e = .3$). The condition with a large proportion of mean-level cluster differences (i.e. $b = .9$) and a high level of noise (i.e. $e = .3$), again, was the most challenging condition with the least successful rate

4.4. APPLICATION

To demonstrate the usefulness of CSSCA, we present an analysis of personality data from Dufner et al. (2015). As part of a large-scale investigation on motive dispositions, the multi-block data – consisting of a total of 171 subjects - contained one block of self-reported scores on motive dispositions, and one block of observers' ratings on participants' nonverbal behavior in dyadic interviews. The first data block contained a total of six sum scores of the self-reported scales with three of them indicating the power motive while the other three indicating the affiliation motive. The second data block included observers' ratings on participants' eighteen types of videotaped non-verbal behaviors (see Table 4.3 for a full list of coded behaviors). A detailed description of the procedures and measurements is available in Hagemeyer et al. (2016).

Our analysis attempted to explore the associations between motive dispositions and nonverbal behaviors and to detect subgroup differences therein. [1]

[1] note that in the current analysis, we were only interested in how motive dispositions and nonverbal behaviors related differently in the two clusters (i.e., the respective subspace structures - but not the mean structures - of the two clusters). For other research applications where the mean structures are also of interest, the R Package `ClusterSSCA` also provides detailed results of the cluster centroids.

Previous studies that tried to reveal the connections between the nonverbal behaviors and the two types of motives drew inconclusive and even contradictory conclusions (see Hall et al., 2005), probably because of the oftentimes ambiguous meanings of nonverbal behaviors (e.g. Vrij et al., 2010). We postulate that such contradictory findings might hint at the existence of subgroups, since people belonging to different subgroups may exhibit different nonverbal behaviors that express their motives.

We performed CSSCA on the multi-block data that consisted of the self-reported scores on motive dispositions and the expert ratings on non-verbal behaviors. The multi-block data was column-wise centered and re-scaled such that the sum-of-squares of each variable equaled 1. To choose the appropriate model, the proposed model selection procedure was used with the number of clusters selected from 1, 2,..., 8, and the level of sparsity from .1, .2,..., .9. Furthermore, we fixed the number of common components to two (i.e., the two types of motives), and the number of distinctive components to one per block (i.e., response styles in the first block while specific coding patterns in the second block) According to the results of model selection, the average scree ratio achieved its highest value when the number of clusters equaled 3, and, conditional on three clusters, the scree ratio was maximized with the level of sparsity equaling .4. We, therefore, inspected the CSSCA solution with 3 clusters and 40% zero loadings. The estimated common component loading matrices of the three clusters, which are of particular interest as they imply associations between motive dispositions and nonverbal behaviors, are presented in Table 4.3.

Table 4.3: The loading matrices of the common component for the three detected clusters in the personality data set.

	Cluster 1 ($N_1 = 68$)		Cluster 2 ($N_2 = 58$)		Cluster 3 ($N_3 = 45$)	
	Component 1	Component 2	Component 1	Component 2	Component 1	Component 2
Explicit power motive XMS	0	0	-.19	-.10	.21	-.10
Explicit power motive PRF	-.31	0	-.17	0	.21	0
Explicit power motive UMS	-.20	.12	-.19	0	.21	0
Explicit affiliation motive XMS	.27	0	-.15	0	0	.18
Explicit affiliation motive PRF	0	-.28	-.21	.18	-.11	.47
Explicit affiliation motive UMS	0	-.27	-.13	.20	-.18	.48
Wiggle	0	-.30	0	0	0	0
Overall gesture	0	-.44	-.37	0	0	0
Brash gesture	0	-.42	-.40	0	0	0
Gaze oriented to an experimenter	.30	-.21	-.18	.17	.35	0
Nod	.41	-.24	0	0	.13	.11
Shakes head	0	-.28	0	.13	0	0
Smile	0	0	0	.31	0	0
Friendly laugh	-.22	0	0	.42	0	.18
Relaxed voice	-.24	0	.40	.18	.16	0

Loud voice	-.39	0	0	.09	.11	-.10
Interrupts experimenter	-.35	-.60	0	0	.15	0
Tries to create a pleasant atmosphere	0	0	0	.56	0	.28
Mentions other persons	0	0	.18	0	.29	.54
Behaves Friendly	0	0	0	.50	0	.27
Tries to dominate the conversation	-.40	-.29	-.18	0	.49	0
Boasts	-.14	-.17	0	0	.62	-.17
Appears self-secure	-.35	-.21	0	.18	.40	0
Touches self	0	-.17	-.26	0	0	0

To illustrate the interpretation of the table, we consider the two common components of Cluster 3. While Component 1 correlates with both power and affiliation motives, Component 2 is primarily related to the affiliation motive, as evidenced by zero or small loadings on the self-reported measurements of the power motive. Component 1 indicates that the following non-verbal behaviors are positively related to the self-reported power motives and negatively related to the self-reported affiliation motives: gazing towards the experimenter, mentioning other persons, trying to dominate the conversation, boasting, appearing to be self-secure, nodding, expressing relaxed and loud voice, and interrupting the experimenter. Among these non-verbal behaviors, the first five behaviors appear to be more closely related to self-reported motives because of their relatively high loadings. In the same vein, the loadings of Component 2 indicate that, for subjects in Cluster 3, the affiliation motive is most strongly related to trying to create a pleasant atmosphere, mentioning other persons, and behaving friendly. Moreover, from Table 4.3, we can also infer that the correlations between the two motives and the non-verbal behaviors are indeed different for different clusters. For example, although for all three clusters Component 2 primarily relates to the affiliation motive, it is also clear from the component loadings that the affiliation motive is linked to different sets of nonverbal behaviors for the three clusters, although with a large overlap between the sets of Cluster 2 and Cluster 3. Overall, the application shows that the CSSCA approach to data analysis can reveal interesting insights into inter-individual differences in the concerted action of attitudinal, emotional, and behavioral indicators.

4.5. GENERAL DISCUSSION

Applied researchers more and more often make use of multiple blocks of data to obtain insight into complex relations between those factors that influence behavior, often involving novel types of data that consist of a large number of variables. As discussed here, understanding the subtle relations that exist between these influencing factors and their concerted action effectively means revealing those variables that co-vary across data blocks. We also discussed that heterogeneity in such joint influences can be expected which necessitates the detection of unknown subgroups for which these underlying common sources of variation

show up in different sets of linked variables. As argued, to identify unknown clusters and extract cluster-specific components, two challenges of multi-block data analysis should be addressed: (1) the high-dimensionality of the data sets makes the interpretation of the common components infeasible, and (2) multi-block data sets might include distinctive variation underlying one or a few data blocks, which should be set apart from the common variation.

In the current paper, we introduced CSSCA as a novel clustering method for multi-block data analysis. This method not only accounts for cluster differences in the mean structures but also for differences in the covariance structures. Furthermore, the two challenges are tackled by automatic variable selection and simultaneously estimating both distinctive and common variations. CSSCA partitions the subjects in such a way that subjects belonging to the same cluster possess the same set of components and cluster centroids. Through two simulation studies, CSSCA successfully recovered clusters and component structures across various conditions. More importantly, CSSCA clearly outperformed iCluster, a popular clustering method that solely detects cluster differences in the mean structures, especially in the presence of substantial cluster differences attributed to the component structures. We further proposed and verified a model selection procedure to select the number of clusters and level of sparsity. Last, we demonstrated in our illustrative analysis how CSSCA could be applied to exploratory research and how this new analysis could bring about novel insights. Concerning the application of CSSCA, we would like to stress that we expect CSSCA to also perform well in analyzing data sets with a large number of subjects (e.g. social network data), despite not being formally tested in the current paper. This is because, with a larger cluster size, the cluster-specific components could probably be estimated more accurately; as a result, in each update, the subject has a better chance to be assigned to the best cluster.

We propose several future directions for CSSCA. First, we believe that the optimization procedure and the implementation of CSSCA could still be improved to speed up the CSSCA analysis. This is especially important when dealing with data sets of large size. Second, we found that CSSCA was slightly less accurate and stable in comparison to iCluster when cluster differences mainly pertained to mean-level differences. Future research could, therefore, seek to discover a

model selection procedure to determine whether cluster differences are mainly differences in the component structures or in the mean structures. If indeed the latter fits the observed data better, one could instead apply iCluster to obtain more accurate partitions.

Although the current model selection procedure allows a data-driven selection of the number of clusters and level of sparsity, to successfully implement CSSCA, researchers are still required to specify the number of common and distinctive components a priori. Nevertheless, further incorporation of component selection tools could surely offer more freedom in the analysis. We refer to two approaches that have been proposed and validated in the existing literature to select the pattern of the components in the SCA-based methods (see Gu et al., 2019, for technical details). The first approach detects the number of total components with the VAF method and determines the status (i.e. common or distinctive) of each component with the DISCO-SCA method. The second approach, called the PCA-GCA method (Smilde et al., 2017), first applies PCA to determine the number of components in each data block and then applies GCA to determine the number of common components.

The current version of CSSCA can only deal with continuous data without missing values. Future research could extend the CSSCA framework to analyze categorical and mixed data types and to handle missing values (Stacklies et al., 2007). For example, a useful strategy in dealing with categorical variables is to treat each of these variables as a host of dummy variables, much like the strategy implemented in the categorical PCA algorithm (CATPCA; Linting et al., 2007).

Last, the currently proposed model selection procedure, by design, prohibits the selection of the smallest possible value. As a result, the solution of one cluster (i.e. no subgroups exist) and (or) that of non-sparse loading matrices can never be selected. Not being able to select a one-cluster solution is actually a well-known problem of many deterministic clustering methods (Milligan, 1996). Some remedies have been provided to solve this issue, for instance, the Lower Bound Technique (LBT) in the context of K-means clustering (Steinley & Brusco, 2008b). We encourage future research to address this issue in the context of CSSCA.

APPENDICES

4.A. DETAILS ABOUT THE CSSCA ALGORITHM (ALGORITHM 3)

4.B. DETAILS ABOUT THE SPARSE DISCO-SCA ALGORITHM (ALGORITHM 4)

4.C. TECHNICAL MINUTIAE OF ALGORITHM 3

4.C.1. THE STARTING PARTITIONS OF THE ALGORITHM

As explained in the main text, to reduce the probability of ending in a local minimum (instead of a global one), we utilize a multi-start procedure in the algorithm. The multiple starting partitions are created on the basis of the partitioning results of the two other clustering methods: Clusterwise PCA and iCluster.

Clusterwise PCA is in principle a simplified version of CSSCA that is applied to the concatenated data set. The major difference between Clusterwise PCA and CSSCA is that the former estimates non-sparse loading matrices and does not distinguish between common and distinctive components. We have noticed that similar approaches have been proposed in statistics (for example McWilliams and Montana, 2014), despite notable differences in estimation procedures. Because of its model configuration, we expect the resulting cluster recoveries of Clusterwise PCA to be fairly similar to the true clusters when the component structure accounts for a large proportion of the total variance (e.g., $b = 10\%$). The algorithm of Clusterwise PCA is also implemented in the package `ClusterSSCA`.

On the other hand, as argued in the main text, iCluster partitions the observations mainly based on the mean structure. Therefore, we expect the resulting cluster recovery of iCluster to be similar to the true clusters when the proportion of mean-level differences is large (e.g.,$b = 90\%$). The algorithm of iCluster is provided in the R package `iCluster` (H. Shen & Huang, 2008; R. Shen et al., 2012).

To ensure that CSSCA performs well at different levels of b, we generate the starting partitions of CSSCA based on the results of both Clusterwise PCA and

Algorithm 3: CSSCA

Input: the concatenated data matrix $\mathbf{X}(N \times J)$, the number of clusters K, and the number of starts Q, the number of sparseness-induced zeros Z (in the loading matrix), and the position vector $\mathbf{w}_{dis}$ that indicates the positions of the distinctiveness-induced zeros.

Initialize the converge rate ϵ, the maximal number of iterations allowed $Iter_{max}$ and the minimal total loss $Loss_{global}$ (initially $Loss_{global}$ is set to an arbitrary large number.)

Column-wise mean-center and standardize X

for $q \in (1,2,...,Q)$ **do**

- **Initialize** the $\mathbf{h}_q$, with its i^{th} element $\mathbf{h}_q[i]$ representing the cluster assignment of observation i (see 4.C for details)
- **Initialize** the total loss of the current start $Loss_q$ to an arbitrarily large number, the total loss resulted from an intermediate iteration $Loss_{current}$ as well as the (initial) count of iterations $Iter_q$ to 0
- **while** $(Loss_q - Loss_{current} > \epsilon)$ & $(Iter_q < Iter_{max}$ **do**
  - **Update** $Iter_q = Iter_q + 1$
  - **if** $Iter_q == 1$ **then**
    - **for** $k \in (1,2,...,K)$ **do**
      - **Estimate** the Sparse DISCO-SCA solution with multiple random starts and estimate $Loss_{q_k}$, $\mathbf{T}_{q_k}$ and $\mathbf{P}_{q_k}$
    - $Loss_{current} = \sum_k Loss_{q_k}$
  - $Loss_q = Loss_{current}$
  - **for** $i \in (1,2,...,N)$ **do**
    - $c = \mathbf{h}_q[i]$
    - **for** $k \in (1,2,...,K)$ **do**
      - **if** $k == c$ **then**
        - **Estimate** the Sparse DISCO-SCA solution with a single start $\mathbf{P}_{q_k}$ and estimate $Loss_{q_s}$, $\mathbf{T}_{q_s}$ and $\mathbf{P}_{q_s}$
        - $Loss_{current_k} = Loss_{current}$
      - **if** $k \neq c$ **then**
        - **Estimate** the Sparse DISCO-SCA solution with a single start $\mathbf{P}_{q_k}$ and estimate $Loss_{q_b}$, $\mathbf{T}_{q_b}$ and $\mathbf{P}_{q_b}$
        - $Loss_{current_k} = Loss_{current} - Loss_{q_k} - Loss_{q_c} + Loss_{q_s} + Loss_{q_b}$
    - **Assign** $\mathbf{h}_q[i] = b$ where $Loss_{current_b} = \min_k Loss_{current_k}$
    - **Update** $Loss_{current} = Loss_{current_b}$ as well as the score and loading matrices of the newly formulated clusters (i.e. Cluster s and b)
- **if** $Loss_{current} < Loss_{global}$ **then**
  - **Update** $Loss_{global} = Loss_{current}$ and the corresponding cluster partitions, and score and loading matrices

Algorithm 4: Sparse DISCO-SCA

Input: the concatenated data matrix $\mathbf{X}(N \times J)$, the number of clusters K, and the number of starts Q, the number of sparseness-induced zeros Z (in the loading matrix), and the position vector $\mathbf{w}_{dis}$ that indicates the positions of the distinctiveness-induced zeros.

Initialize the converge rate ϵ, the maximal number of iterations allowed $Iter_{max}$ and the minimal total loss $Loss_{global}$(initially $Loss_{global}$ is set to an arbitrary large number.)

Column-wise mean-center X

for $q \in (1, 2, ..., Q)$ **do**

  Initialize the loss of the current start $Loss_q$ to an arbitrarily large number, the loss calculated at a specific iteration $Loss_{current}$ to 0, and the number of iterations that have been executed $Iter_q = 0$

  if *loading matrix is expected to be randomly generated* **then**

    Initialize loading matrix $\mathbf{P}_q$: each entry p_{kr} is generated from a uniform distribution $\mathbf{U}[-1, 1]$

  while $(Loss_q – Loss_{current}) > \epsilon$ *and* $(Iter_q < Iter_{max})$ **do**

    Update $Iter_q = Iter_q + 1$

    if $Iter_q > 1$ **then**

      Update $Loss_q = Loss_{current}$

    Perform singular value decomposition $\mathbf{P}_q^T\mathbf{X}^T = \mathbf{U}\boldsymbol{\sigma}\mathbf{V}^T$

    Update $\mathbf{T}_q$: $\mathbf{T}_q = \mathbf{V}\mathbf{U}^T$

    Update $\mathbf{P}_q$: (1) $\mathbf{P}_q = \mathbf{X}^T\mathbf{T}_q$, (2) impose the distinctiveness-induced zeros on $\mathbf{P}_q$ based on $\mathbf{w}_{dis}$, (3) Impose the sparseness-induced zeros on $\mathbf{P}_q$: the smallest Z loadings of $\mathbf{P}_q$ (except for the distinctive-induced zero loadings) are set to zero

  If$Loss_{current} = Loss_{global}$ **Update** $Loss_{global} = Loss_{current}$ and **update** the corresponding score and loading matrices

iCluster. More specifically, two of the starts (which are called "user-specified starts") are cluster partitions produced by Clusterwise PCA (which is labeled $\mathbf{h}_c$; we further name its resulting partition $\mathbf{h}_{cr}$) and iCluster (which is labeled $\mathbf{h}_i$; we further name its resulting partition $\mathbf{h}_{ir}$). The other starts (which are called "semi-random starts") are generated by randomly changing the cluster memberships of a certain amount of observations in $\mathbf{h}_c$ or $\mathbf{h}_i$. When the similarity between $\mathbf{h}_c$ and $\mathbf{h}_{cr}$ is larger than the similarity between $\mathbf{h}_i$ and $\mathbf{h}_{ir}$, the component structure is probably more important, therefore, $\mathbf{h}_c$ is used to generate the semi-random starts; otherwise, $\mathbf{h}_c$ is used to generate the semi-random starts.

4.C.2. THE STARTING PARTITIONS OF THE ALGORITHM

Some restrictions concerning model parameters apply to CSSCA. First, during the iterations, the number of observations of every cluster should always be larger than the number of components. If this condition is not met, the CSSCA analysis with the package `ClusterSSCA` will automatically cease estimation following the current starting partition, and the total loss associated with the current starting partition will be set at an invalid value. We should also note that the failure to meet the restriction indicates that the number of clusters is potentially overestimated.

4.D. DATA GENERATION PROCEDURE

A partition vector $\mathbf{h}$ with size N was first generated to represent the true cluster partitions. The i^{th} element of $\mathbf{h}$, i.e., $\mathbf{h}_i$ indicates the cluster membership of observation i.

Equation 4.4 in the main text expresses the observed data matrix $\mathbf{X}^{con}$ as the addition of three parts: the component structure part $\mathbf{X}_{comp}$, the mean structure part $\mathbf{X}_{mean}$, and the noise part $\mathbf{E}$. In simulations, the average variance of the variables was 1, among which e was attributable to $\mathbf{E}$ and a total of $1-e$ to $\mathbf{X}_{comp}$ and $\mathbf{X}_{mean}$. Subsequently, a fraction b of the remaining variance was further attributed to the mean structure and $1-b$ to the component structure (note that as the average variance of all variables was equal to 1, the average variance attributable to the component structure was $(1-e)(1-b)$). In what follows, the data generation procedures for the three parts are detailed.

To construct the component structure part, for each cluster k, the *component score matrix* (with dimension $N_k \times R$) was generated as follows: (1) each entry was initially sampled from the univariate standard normal distribution, (2) the resulting matrix was column-wise mean-centered and (3) to ensure that the component scores were orthonormal, the Gram-Schmidt orthonormalization was applied to each score matrix. To set the variance – rather than the sum-of-squares – of each component equal to 1, we then multiplied each entry of the score matrices by the square root of the corresponding cluster size.

We then constructed the *component loading matrices* (with dimension $J \times R$), where we first generated a component loading matrix for each cluster, and then imposed distinctiveness-induced zeros and sparseness-induced zeros, as follows.

First, a different procedure was used to create the non-sparse version of the loading matrices in the high-congruence and low-congruence conditions.

- For the low-congruence condition, each element in the cluster-specific loading matrices was obtained initially by uniformly sampling from the range of -1 to 1. Subsequently, the resulting matrix was rescaled such that the sum of squares of each row equaled 1.

- For the high-congruence condition, in addition to the cluster-specific matrices generated as described above, a common base matrix was also generated. The entries of the common base matrices were also uniformly sampled from the range of -1 to 1. Afterward, these matrices were re-scaled such that the sum-of-squares of each row equaled .7 for the common base matrix and .3 for the cluster-specific matrices. The final cluster-specific loading matrices were then obtained by adding the common base matrix to the cluster-specific matrices.

Second, the distinctiveness-induced zero loadings were introduced to the cluster-specific loading matrices, in order to structure the distinctive components, as shown in Figure 4.2. More specifically, for the $l^{th}(l = 1,2)$ distinctive component, the loadings of the variables that did not belong to the l^{th} data block were set to zero.

Third, the sparseness-induced zero loadings were also imposed on the cluster-specific loading matrices. The number of the sparseness-induced zero loadings Z was jointly determined by the level of sparsity S, the block-specific number of variables and components. In the current simulation, Z equaled $3 \times J \times S$. We selected Z of the remaining non-zero entries in each cluster-specific loading matrices – after imposing the distinctive-induced zero loadings in the previous step – and imposed these entries to zero. For the low-congruence condition, the Z positions in each cluster-specific loading matrix were selected randomly. For the high congruence condition, about $70\% \times Z$ (or more concretely, the largest positive integer that is smaller than $70\% \times Z$) zero positions were randomly selected and were identical across all clusters while the remaining zero positions were selected randomly for each cluster.

Finally, we re-scaled the cluster-specific loading matrices such that the average sum of squares of each row equaled $(1–b)(1–e)$.

For each cluster, the component structure part of the data was constructed by multiplying its score matrix and the transpose of its loading matrix. $\mathbf{X}_{comp}$ was then created by stacking together vertically the cluster-specific component-structure part according to the cluster assignment of each observation.

To quantify the degrees of similarities between the resulting cluster-specific loading matrices, we computed Tucker's congruence coefficient φ for each pair of the corresponding components and averaged them across all components and clusters. Formally, φ between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as their normalized inner product: $\frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x}}\sqrt{\mathbf{y}'\mathbf{y}}}$, according to Tucker (1951). In the simulated data sets, the average congruence coefficients equaled .18 (SD = .07) in the low congruence conditions and .53 (SD = .01) in the high congruence conditions. As the first step in the creation of $\mathbf{X}_{mean}$, the $K \times J$ cluster centroids matrix $\mathbf{M}$ was created where each row $k(k \in 1,2,\ldots,K)$ represented the centroids of cluster k. Each entry in $\mathbf{M}$ was randomly sampled from the univariate uniform distribution $\mathbf{U}(-1,1)$. We then created a preliminary version of the mean structure $\mathbf{X}_{mean_{pre}}$ by multiplying $\mathbf{h}$ and $\mathbf{M}$. Subsequently, we re-scaled each column of $\mathbf{X}_{mean_{pre}}$ such that the variance equaled $b \times (1-e)$ in the resulting mean-structure part $\mathbf{X}_{mean}$.

Last, the each entry of the error matrix $\mathbf{E}$ was randomly sampled from a univariate normal distribution $\mathbf{N}(0,\sqrt{e})$

The final concatenated data was constructed by adding together $\mathbf{X}_{comp}$, $\mathbf{X}_{mean}$ and $\mathbf{E}$.

4.E. SUPPLEMENTARY REPORT ON THE CLUSTER RECOVERY OF CSSCA

In addition to the results reported in the section Simulation Studies of the original article, we report hereafter the average clustering accuracy of CSSCA as a function of the other six factors. We found that, on average, CSSCA resulted in better cluster recovery when (1) the total number of variables J was larger (ARI = 1 when J = 65 and ARI = .99 when J = 30), (2) the number of clusters K was larger (ARI = .997 when K = 4 compared to ARI = .991 when K = 2), (3) the cluster size N_k was larger (ARI = .996 when the largest cluster includes 100 observations and ARI = .992 when the largest cluster includes 50 observations), (4) the cluster sizes were identical across all clusters (ARI = .995 when all clusters have the equal number of observations compared to ARI = .993 when all clusters have an unequal number of observations), (5) the congruence between the cluster-specific loading matrices φ was lower (ARI = .997 when φ = about .2 , and ARI = .992 when φ = about .55), and (6) the level of sparsity S was larger (ARI = .995 when S = .5 or .7, and ARI = .985 when S = .3)

5

CLUSTERWISE SIMULTANEOUS COVARIATES REGRESSION: A NOVEL METHOD THAT BALANCES PREDICTION AND INTERPRETATION WITH HIDDEN SUBGROUPS

Many behavioral scientists base their predictive analyses on multiple data sets that are gathered from different sources (e.g., self-reports, media fingerprints) and (or) that measure different psychological constructs (e.g., attitudes, traits, values). The objective of this type of analysis is usually not limited to developing an accurate prediction model, but also involves providing clear insight in terms of how predictors individually and jointly relate to outcomes. Since there is often heterogeneity regarding these relationships, predictive analyses should account for this heterogeneity by identifying clusters with different predictor-outcome relationships. Un-

This chapter is based on **Yuan, S.**, De Roover, K., & Van Deun, K. (Resubmitted). Clusterwise Simultaneous Covariates Regression: A Novel Method that Balances Prediction and Interpretation with Hidden Subgroups. *Behavior Research Methods*

fortunately, existing methods allowing for this heterogeneity suffer from important drawbacks (e.g., failure to deal with a high degree of multicollinearity). To address these drawbacks, this study presents a novel method, called Clusterwise Simultaneous Component Regression or CSCR. Inspired by Principal Covariates Regression and incorporating the clusterwise approach, CSCR is a generic method that encompasses several other methods as special cases. In two simulation studies, we found that CSCR outperformed four existing methods in terms of predictive accuracy and cluster recovery and that the proposed model selection procedure recovered the parameters of the CSCR model well. In our illustrative analysis, we demonstrate that CSCR can yield additional insights into how one's attitude towards immigrants is affected by personality traits and personal values for different clusters of observations. Last, we provide some practical guidance on conducting predictive analyses.

5.1. INTRODUCTION

Behavioral scientists working on predictive analyses are often confronted with heterogeneous samples: the subjects on which prediction models are built come from unknown clusters, and to fully capture the unique characteristics of these clusters, each cluster may require a different prediction model. For example, an HR researcher who wants to predict leadership effectiveness has to take into account the different types of leadership profiles where effective outcomes (e.g., improved team performance; Doucet et al., 2015) are achieved through different sets of behaviors (Gandolfi & Stone, 2018). In other words, for each (unknown) cluster of leaders, a unique prediction model is required to infer leadership effectiveness from leadership behaviors. To accommodate such modeling needs, prediction methods should automatically detect hidden clusters of subjects and simultaneously build cluster-specific prediction models. One of the most popular methods belonging to this category is Clusterwise Regression (CR; Späth, 1979), which simultaneously assigns subjects into clusters and estimates cluster-specific regression models. However, since CR is based on Ordinary Least Squares (OLS) regression, it suffers from the two well-known drawbacks of OLS when dealing with a large number of predictors. First, the predictive accuracy of CR is significantly compromised when the notorious problems of multicollinearity and (or) overfitting occur (Yuan, Kroon, et al., 2021). Second, the unique and incremental contribution of each predictor estimated in CR becomes very difficult to interpret when the number of predictors is large.

A potential remedy to address the two drawbacks is Principal Covariates Regression (PCovR; De Jong and Kiers, 1992), which simultaneously summarizes the predictors into a limited number of components (called predictive components hereafter) and regresses the outcome on these components (Vervloet et al., 2016). Predicting from components not only reduces model instability caused by sampling variation (Kiers & Smilde, 2007), but also describes the joint contributions of several related predictors, which are arguably more interpretable than the incremental contributions of individual predictors (e.g., leadership effectiveness can be better inferred from leadership styles, composed from several leadership behaviors; Nanjundeswaraswamy and Swamy, 2014). In this paper, we incorporate PCovR and the clusterwise approach (e.g., DeSarbo and Cron, 1988; De

Roover, Ceulemans, Timmerman, Vansteelandt, et al., 2012; De Roover, Ceulemans, Timmerman, and Onghena, 2013; Durieux and Wilderjans, 2019) into a novel method, called Clusterwise Simultaneous Covariates Regression (CSCR), which identifies hidden clusters as well as cluster-specific predictive components. As the name implies, CSCR originates from Simultaneous Component Analysis (SCA; Kiers and ten Berge, 1994), and, just like SCA, reduces the dimensions of multi-block data simultaneously and summarizes the observed variables in a few components. Here, the multi-block data refers to a few homogeneous data blocks, where each data block covers conceptually different constructs (e.g., a block of self-reported leader behaviors and a block of self-reported leader personalities) and (or) consists of conceptually similar constructs with diverse instruments (e.g., a block of self-reported leader behaviors and a block of leader behaviors measured through textual data). Last, as has become clearer in Sections 5.2 and 5.3, CSCR does not rely on the outcome to determine the cluster assignment of new observations, thus avoiding contaminating cluster assignment with the information contained in the outcome - another critical drawback of CR, first discussed in Brusco et al. (2008). Below, we briefly introduce how CSCR extends PCovR to estimate regression models per cluster and how it deals with multi-block data sets.

PCovR, the predecessor of CSCR, was proposed in response to the aforementioned shortcomings of OLS (i.e., inability to handle multicollinear predictors; strong tendency to overfitting; poor interpretability in the presence of many predictors) and established itself as one of the most popular models to accurately produce predictions and clearly describe joint contributions of predictors (Vervloet et al., 2015). CSCR extends PCovR first by capturing underlying heterogeneity in terms of predictive components and (or) regression coefficients with a simultaneous cluster analysis. For this purpose, a clusterwise approach with an iterative algorithm is employed to simultaneously detect clusters and the associated prediction models.

In addition to modeling unobserved clusters, CSCR also extends PCovR to the setting of multi-block data and allows to identify the predictive components indicative of the joint contributions of variables from different data blocks. The structure of multi-block data and the investigation of underlying joint contribu-

tions are prevalent in behavioral research. In the preceding example of predicting leader effectiveness, HR researchers may want to determine the joint influences of leaders' personalities (as indicated by a block of personality scores) and behaviors (as indicated by a block of behavior scores). Unfortunately, PCovR in its original form does not properly handle multi-block predictors: due to the fact that the components identified by PCovR are subject to rotational freedom, these (rotated) components can become predictive components associated with all data blocks (called *common predictive components*) or components associated with one or a few data blocks (called *distinctive predictive components*) or a combination of the two (Schouteden et al., 2014). In our preceding example, the estimates of PCovR do not necessarily translate into the joint influences of leaders' personalities and behaviors in determining their effectiveness (this type of joint contribution is in many cases the focus of the research), because the identified components can also indicate the distinctive contribution pertaining only to personalities or behaviors. To address this challenge of analyzing multi-block data, S. Park et al. (2021) extended PCovR to incorporate both common and distinctive predictive components, resulting in a method known as common and distinctive covariates regression (or CD-CovR). Essentially, CD-CovR imposes blocks of zero loadings onto a loading matrix that represents the relationships between predictors and predictive components: the components involving such zero blocks represent distinctive predictive components because they are not related to the variables from those data blocks with which the associated loadings are zero. In our previous example, the distinctive predictive component pertaining to leaders' personality traits contains zero loadings on all leadership behaviors.

To benefit from the interpretative and predictive advantages of PCovR and to further account for hidden clusters as well as common and distinctive components, our proposed CSCR method effectively incorporates the clusterwise approach into the CD-CovR model. As such, CSCR is unique in its ability to simultaneously detect common and distinctive predictive components as well as cluster differences therein. Furthermore, CSCR is a generic method that comprises PCovR, CD-CovR, and the clusterwise version of PCovR as its special cases. We note that the idea of integrating the clusterwise approach into component-based methods has also been adapted in two related methods, namely Cluster-

wise Simultaneous Component Analysis (CSCA; Yuan, De Roover, Dufner, et al., 2021) and Principal Covariates Clusterwise Regression (PCCR; Wilderjans et al., 2017). However, CSCR differs from CSCA because the latter identifies clusters and components without predicting outcomes, whereas CSCR differs from PCCR in terms of their different objectives: while PCCR deals with multi-level data sets and finds clusters at the level of existing groups, CSCR deals with single-level data sets and finds clusters at the level of individuals. As far as we know, the only methods that serve a similar purpose as CSCR are Clusterwise Multi-block Partial Least Squares (CW-MBPLS) and Clusterwise Multiblock Redundancy Analysis (CW-MBRA), both have been proposed by Bougeard et al. (2018) and both adopt a PLS-based framework. These authors concluded from a set of simulation studies that CW-MBPLS was preferable to CW-MBRA since it yielded better results in the presence of high multicollinearity between predictors. Therefore, only CW-MBPLS is considered further as a competing method in the current study. Theoretical and empirical comparisons between CSCR and all these related methods are described further in Sections 5.2 and 5.3, respectively.

The remainder of this article is organized as follows. In Section 5.2, we provide a detailed description of CSCR and explain its relations to similar methods. Section 5.3 presents two simulation studies in which the performance of CSCR is evaluated and compared to the related methods. We then illustrate the application of CSCR in an empirical analysis, presented in Section 5.4, in which self-reported personality types and values are used to predict one's attitude towards immigrants. Finally, in Section 5.5, we discuss the implications and limitations of CSCR, and a blueprint for future research. For users who want to use CSCR or the closely related, clusterwise version of PCovR, a package `CSCR` can be obtained from https://github.com/syuanuvt/CSCR.

5.2. METHOD

Tables 5.1 and 5.2 respectively list the most important acronyms and mathematical notations that are used throughout the article. In general, matrices are denoted by bold upper-case letters (e.g., $\mathbf{X}$), vectors by bold lower-case letters (e.g., $\mathbf{y}$), and scalars by italic lowercase letters (e.g., k). The superscript ^{T} indicates matrix transposition, and the subscript $_k$ refers to cluster k. Furthermore, for

Table 5.1: Cross-reference table of the full name and abbreviation for each method

Acronyms	Full names	Reference
CSCR	Clusterwise Simultaneous Covariates Regression	The current study
CR	Clusterwise Regression	Späth, 1979
PCovR	Principal Covariates Regression	De Jong and Kiers, 1992
SCA	Simultaneous Component Analysis	Kiers and ten Berge, 1994
CD-CovR	Common and Distinctive Covariates Regression	S. Park et al., 2021
CSCA	Clusterwise Simultaneous Component Analysis	Yuan, De Roover, Dufner, et al., 2021
CW-MBPLS	Clusterwise Multiblock Partial Least Squares	Bougeard et al., 2018
iCluster	Integrative Clustering Algorithm	R. Shen et al., 2009

simplicity, we assume a univariate outcome (indicated by $\mathbf{y}$) throughout the article, although CSCR is also applicable to multivariate outcomes. Throughout the article, we assume that both $\mathbf{X}$ and $\mathbf{y}$ are column-wise mean centered. Note that when the variances of the predictors differ significantly, it is recommended to rescale the variables to unit variance.

5.2.1. MODEL

PCOVR MODEL

Because CSCR extends PCovR, we now first formally introduce the statistical model of PCovR. According to PCovR, $\mathbf{X}$ can be decomposed by

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}^{(\mathbf{X})}, \tag{5.1}$$

with $\mathbf{E}^{(\mathbf{X})}$ indicating the matrix of residuals. Simultaneously, PCovR models $\mathbf{y}$ as a linear function of the component scores $\mathbf{T}$,

$$\mathbf{y} = \mathbf{T}\boldsymbol{\beta} + \mathbf{e}^{(\mathbf{y})}, \tag{5.2}$$

with $\boldsymbol{\beta}$ containing R regression weights contained in the vector and $\mathbf{e}^{(\mathbf{y})}$ denoting the vector of residuals.

Table 5.2: Glossary of mathematical notations

Notation	Type and size	Description
N	a scalar	the total number of observations
J	a scalar	the total number of variables
$\mathbf{X}$	a matrix of size $N \times J$	the concatenated predictor block
$\mathbf{y}$	a vector of size N	the (univariate) outcome block
K	a scalar	the total number of clusters
l	a scalar	the total number of data blocks
N_k	a scalar	the number of observations in cluster k
R	a scalar	the total number of components
R_{com}	a scalar	the number of common components
R_{disl}	a scalar	the number of distinctive components pertaining to data block l
$\mathbf{T}_k$	a matrix of size $N_k \times R$	the component score matrix of cluster k
$\mathbf{P}_k$	a matrix of size $J \times R$	the component loading matrix of cluster k
$\boldsymbol{beta}_k$	a vector of size R	the regression weights of cluster k
$\mathbf{m}_k$	a vector of size J	the cluster centroids across the predictor blocks for cluster k
u_k	a scalar	the intercept of cluster k
α	a scalar	the coefficient that balances the importance of reconstructing $\mathbf{X}$ versus predicting $\mathbf{y}$
$\mathbf{g}$	a vector of size N	the partitioning vector that assigns each observation to the corresponding cluster (e.g., $\mathbf{g}(i) = k$)

To find components that simultaneously minimize the reconstruction error in $\mathbf{X}$ and the prediction error in $\mathbf{y}$, PCovR seeks to minimize

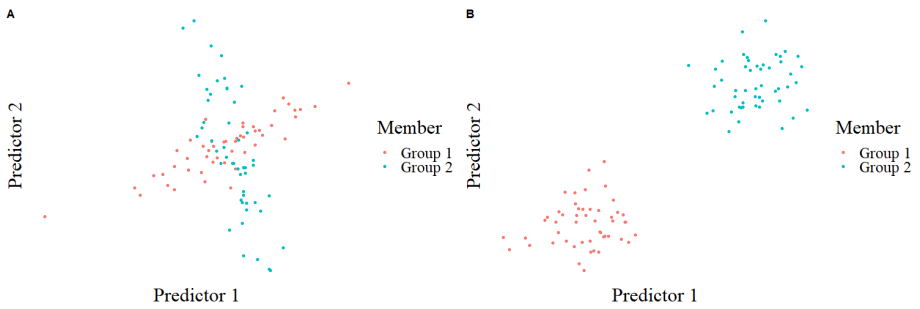
$$L = \alpha \frac{||\mathbf{X} - \mathbf{T}\mathbf{P}^T||_2^2}{||\mathbf{X}||_2^2} + (1-\alpha)\frac{||\mathbf{y} - \mathbf{T}\boldsymbol{\beta}||_2^2}{||\mathbf{y}||_2^2}, \tag{5.3}$$

with $||.||_2^2$ indicating the sum of squared elements of a vector or a matrix (i.e., the squared Euclidean norm of a vector or the squared Frobenius norm of a matrix) and α ranging from 0 to 1. α determines the relative weights of the two parts in the loss function Equation (5.3); in other words, α balances the importance of reconstructing $\mathbf{X}$ versus predicting $\mathbf{y}$. When $\alpha = 1$, the sole purpose of PCovR is to reconstruct $\mathbf{X}$ and it coincides with Principal Component Regression or PCR. Alternatively, when $\alpha = 0$, the sole purpose of PCovR is to predict $\mathbf{y}$ and a special case of PCovR with one component is equivalent to multiple linear regression (S. Park et al., 2021). We refer interested readers to Vervloet et al. (2015) for a detailed discussion of how the choice of α affects the attainment of the two objectives (i.e., prediction versus interpretation) of PCovR. To identify the optimal solution of Equation (5.3), De Jong and Kiers (1992) suggested to impose an orthogonal constraint on $\mathbf{T}$, such that $\mathbf{T}^T\mathbf{T} = \mathbf{I}$. Note that this constraint does not completely resolve the indeterminacy of the solution, as $\mathbf{T}$ is still subject to rotational freedom.

THE CSCR MODEL

So far, PCovR has been presented in its original form where it is applied to the concatenated predictor block. However, the original PCovR model suffers from two limitations: (1) it ignores the multi-block structure of the predictors and gives no insight into the joint and individual sources of variation (i.e., the common and distinctive predictive components); (2) it does not account for heterogeneous subgroups with different components and (or) regression models.

The first limitation is addressed in CD-PCovR by imposing a block structure on the components. Specifically, distinctive components are modeled by imposing zero loadings on all variables but those from the block(s) the component is supposed to underlie. Common components are not subject to this constraint so they remain associated with all variables.



Note. the red and blue dots are data points from the two clusters. Subplot A depicts the case where the two clusters differ primarily in terms of the component structure, while Subplot B depicts the case where the clusters differ primarily in terms of the mean structure.

Figure 5.1: The visualization of the two types of clustering differences concerning multi-block predictors

The second limitation, namely the strict assumption that the same model applies to all observations, can be effectively addressed with the clusterwise approach. Here, the relevant methods to build on are CSCA and CR that partition observations into clusters based on components or regression models.

To illustrate the different types of cluster differences that should be accounted for, we consider a hypothetical sample consisting of 2 clusters - each with 50 data points - and 3 variables (2 predictors from the same data block and 1 outcome). Figures 5.1 and 5.2 visualize the two types of cluster differences with respect to the predictor block and the outcome, respectively. We first discuss Figure 5.1. While the two clusters (shown in red and in blue, respectively) in Figure 5.1A have the same centroids and different within-cluster components (the axes of these components are depicted in lines), the clusters in Figure 5.1B only differ in terms of their centroids.

While the original PCovR model for reconstructing $\mathbf{X}$, as illustrated in Equation (5.1), fails to account for these cluster differences, the addition of a clusterwise approach, first adapted in CSCA (Yuan, De Roover, Dufner, et al., 2021), allows for cluster-specific component scores and loadings. In addition, as discussed above, following CD-PCovR, blocks of zero loadings can be further imposed structurally on cluster-specific loading matrices to indicate common and distinctive predictive components. Mathematically, considering the simplest case with two data blocks, to reflect common and distinctive components, the com-

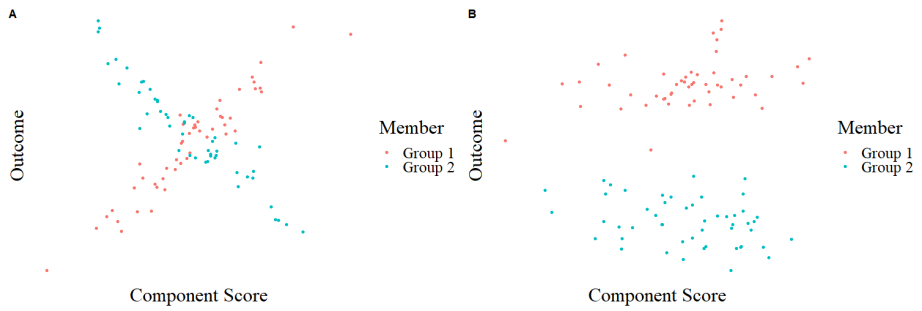
ponent score matrix $\mathbf{T}$ can be defined as $\begin{bmatrix}\mathbf{T}_{com} & \mathbf{T}_1 & \mathbf{T}_2\end{bmatrix}$, where $\mathbf{T}_{com}$ are scores of the common components and $\mathbf{T}_1$ and $\mathbf{T}_2$ are component scores pertaining to data block 1 and 2, respectively. The component loading matrix $\mathbf{P}$ can be defined as $\begin{bmatrix}\mathbf{P}_{com} & \mathbf{P}_{dis1} & 0 \\ & 0 & \mathbf{P}_{dis2}\end{bmatrix}$, where $\mathbf{P}_{com}$ are scores of the common components while $\mathbf{P}_{dis1}$ and $\mathbf{P}_{dis2}$ are specific to data blocks 1 and 2, respectively. This way of defining the structure of common and distinctive components can be easily extended to conditions with more than two data blocks. Taken together, the clusterwise model for reconstructing $\mathbf{X}$ can be written as

$$\mathbf{X}_k = \mathbf{1}_{N_k}\mathbf{m}_k^T + \mathbf{T}_k\mathbf{P}_k^T + \mathbf{E}_k^{(\mathbf{X})}, \tag{5.4}$$

and is subject to the condition that, for each k, $\mathbf{T}_k^T\mathbf{T}_k = \mathbf{I}$. Furthermore, predefined blocks of zeros are imposed on $\mathbf{P}_k$ to model common and distinctive components (see our previous example of structuring $\mathbf{P}$). The fact that all score and loading matrices are of equal size corresponds to the assumption of CSCR that the same number of components pertain to all clusters. This assumption, which has also been imposed in similar methods (e.g., Timmerman et al., 2013; Wilderjans et al., 2017; Yuan, De Roover, Dufner, et al., 2021) is in place because of our goal to keep the model selection process as efficient as possible. Without this assumption, the number of components has to be selected for each cluster; this selection procedure drastically increases computational demand, especially with a large number of clusters.

We now turn to Figure 5.2, which illustrates the two types of cluster differences with respect to the relationship between component scores and the outcome. In Figure 5.2A, observations from the two clusters lie around the two regression lines with the same centroids. In contrast, the two clusters in Figure 5.2B have such a large mean difference in the outcome that this difference alone permits accurate separation of the two clusters.

Again, these important cluster differences cannot be accounted for in the original PCovR model for predicting $\mathbf{y}$, as illustrated in Equation (5.2). However, the clusterwise approach provides a useful solution, as first adapted in CR (Späth, 1979). More specifically, with u_k denoting the cluster-specific intercept and $\boldsymbol{\beta}_k$ the cluster-specific regression weights, both of cluster k, the clusterwise model



Note. the red and blue dots are data points from the two clusters. Subplot A depicts the case where the two clusters differ primarily in terms of the component structure, while Subplot B depicts the case where the clusters differ primarily in terms of the mean structure.

Figure 5.2: The visualization of the two types of clustering differences concerning outcomes

for predicting $\mathbf{y}$ can be written as

$$\mathbf{y}_k = \mathbf{1}_{N_k} u_k + \mathbf{T}_k \boldsymbol{\beta}_k + \mathbf{e}_k^{(y)}. \tag{5.5}$$

Equations (5.4) and (5.5) are the essential models of the novel CSCR method. We have two comments on the models. First, the cluster-specific mean structure $\mathbf{m}_k$ and u_k are indispensable ingredients in Equations (5.4) and (5.5), as the users typically do not know in advance which observation belongs to which cluster. Second, for applications with single-block predictors, one can use a special form of the CSCR model where all components are common components.

THE OBJECTIVE FUNCTION

Estimation of the CSCR model is based on minimizing the following least squares criterion, given that the total number of clusters equals K:

$$L = \frac{\alpha}{||\mathbf{X}||_2^2} \sum_{k=1}^{K} ||\mathbf{X}_k - \mathbf{1}_{N_k} \mathbf{m}_k^T - \mathbf{T}_k \mathbf{P}_k^T||_2^2 + \frac{(1-\alpha)}{||\mathbf{y}||_2^2} || \sum_{k=1}^{K} ||\mathbf{y}_k - \mathbf{1}_{N_k} u_k - \mathbf{T}_k \boldsymbol{\beta}_k||_2^2, \tag{5.6}$$

and is subject to: (1) for each k, $\mathbf{T}_k^T \mathbf{T}_k = \mathbf{I}$, and (2) $\mathbf{P}_k$ in the form of $\begin{bmatrix} \mathbf{P}_{com} & \mathbf{P}_{dis1} & 0 \\ & 0 & \mathbf{P}_{dis2} \end{bmatrix}$ with pre-specified blocks of zeros loadings. Here, the value of α is fixed to .95, because, according to previous research (e.g., Heij et al., 2007; Wilderjans et al.,

2017; Vervloet et al., 2016), the predictive accuracy of PCovR and related methods is optimal when α is between .95 and .99 and, within this range, a smaller α leads to a better balance of prediction and interpretation. Note that the obtained solutions (i.e., the cluster-specific score and loading matrices as well as the sets of regression weights) are not necessarily unique because each set of components of the same status (i.e., common components, distinctive components for the first data block, etc,) are subject to rotational freedom (also see Schouteden et al., 2013 and Schouteden et al., 2014). The rotational freedom does not affect predictions made by CSCR but does affect how within-cluster component structures are to be interpreted. Following the suggestions in Schouteden et al. (2013) and Schouteden et al. (2014), we recommend applying a VARIMAX rotation (Kaiser, 1959) within each set of components to identify the solution.

5.2.2. ALGORITHM

Here, we present an alternating algorithm to minimize (5.6) with a prespecified number of clusters K, common components R_{com} and distinctive components $R_{dis1}...R_{disI}$. We defer the discussion on how to determine these parameters to section 5.2.3. In essence, the algorithm - outlined in the following two paragraphs and detailed in Algorithm 5 - closely follows the procedure used in CR and CSCA where, given initial values, the cluster memberships and the cluster-specific components and regression weights are updated iteratively.

To initialize Algorithm 5, a starting partition to determine the initial clusters should be provided. A multi-start procedure (De Roover, Timmerman, et al., 2013; Timmerman et al., 2013) with semi-rational starts is strongly recommended, in order to avoid local optima. These semi-rational initial partitions are created by randomly swapping the cluster assignments of roughly 20% of the observations from one of the two initial partitions - resulting from a Gaussian Mixture Model (GMM) analysis and a CSCA analysis on multiple blocks of predictors. As illustrated in Yuan, De Roover, Dufner, et al. (2021), CSCA recovers clusters well when differences in cluster-specific components are the main source of cluster differences (i.e., Figure 5.1A), whereas GMM effectively recovers clusters when cluster differences are primarily attributed to differences in cluster centroids (i.e., Figure 5.1B). Therefore, to account for both types of cluster

differences, the partitions resulting from both methods are used to create semi-rational starting partitions.

After having obtained the initial partition of the observations, the cluster-specific predictors are column-wise mean-centered. Subsequently, the cluster-specific score matrices, loading matrices, and regression weights are updated independently for each cluster through an iterative procedure, which is adapted from a PCA-like decomposition of the augmented matrix $\begin{bmatrix}\mathbf{y} & \mathbf{X}\end{bmatrix}$, first proposed in Heij et al. (2007). More specifically, for each cluster k, the score matrix $\mathbf{T}_k$ is updated conditional on the loading matrix $\mathbf{P}_k$, the regression weights $\boldsymbol{\beta}_k$ and the means $\mathbf{u}_k$ while $\mathbf{P}_k$, $\boldsymbol{\beta}_k$, and $\mathbf{u}_k$ are updated conditional on $\mathbf{T}_k$. The iterative procedure for updating the scores, loadings, and regression parameters stops when the change in the loss value calculated from Equation (5.6) is smaller than a pre-defined threshold. Then, the cluster memberships for each observation are updated in turn. For this purpose, the observation under consideration is provisionally combined with each cluster and, for each combination, the optimal model is estimated following the iterative procedure described above. The observation is subsequently assigned to the cluster that minimizes the corresponding loss value. The updating of the cluster memberships only stops when the total reduction of the loss value for one full iteration of updating all observations is less than a pre-specified value (or when the algorithm reaches the maximal number of iterations). The iterative procedure proposed here has the property that - with each step - the loss, as calculated from Equation (5.6), does not increase and eventually converges to a stationary point. The proof that the loss is non-increasing when updating $\mathbf{P}_k$, $\boldsymbol{\beta}_k$, and $\mathbf{T}_k$ is provided in Gu and Van Deun (2019).

Algorithm 5: The CSCR Algorithm

Define: the convergence criterion $conv$, and the maximal number of iterations b_{max}.
Perform a CSCA analysis and a GMM analysis on $\mathbf{X}$, resulting in $\mathbf{g_0^1}$ and $\mathbf{g_0^2}$
for $v \leftarrow 1$ **to** 3 **do**
  for $s \leftarrow 1$ **to** 2 **do**
    Generate the partition $\mathbf{g}$ by randomly swapping the cluster memberships of 20% of the observations in $\mathbf{g_0^s}$
    for $k \leftarrow 1$ **to** K **do**
      Per $\mathbf{g}$, determine $\mathbf{X}_k$, and then calculate $\mathbf{m}_k = \frac{1}{N_k}\mathbf{H}\mathbf{X}_k$ and $\mathbf{X}_k^* = (\mathbf{I} - \frac{1}{N_k}\mathbf{H})\mathbf{X}_k$, where $\mathbf{H}$ is an $N_k \times N_k$ matrix of ones
      Initialize $\mathbf{T}_k$: $\mathbf{T}_k \leftarrow \mathbf{R}_k$ where $\mathbf{R}_k$ consists of the first k left singular vectors of $\mathbf{X}_k^*$
      Initialize the number of iterations: $b \leftarrow 0$, the difference in loss: $\Delta L \leftarrow Inf$
      while $b < b_{max}$ *and* $\Delta L < conv$ **do**
        Update $\mathbf{P}_k$: (1) $\mathbf{P}_k \leftarrow {\mathbf{X}_k^*}^T \mathbf{T}_k$ and (2) impose distinctiveness-induced zeros on $\mathbf{P}_k$ to match the pattern of common and distinctive components
        Update u_k and $\boldsymbol{\beta}_k$: regress $\mathbf{y_k}$ on each column of $\mathbf{T}_k$; u_k is updated by the resulting intercepts and $\boldsymbol{\beta}_k$ is updated by the coefficients of scores
        Update $\mathbf{T}_k$: $\mathbf{T}_k \leftarrow \mathbf{W}\mathbf{V}^T$ where $\mathbf{W}$ and $\mathbf{V}$ are respectively the matrices that contain left and right singular vectors of $\mathbf{Z} = \frac{\alpha||\mathbf{X}||_2^2}{\alpha||\mathbf{X}||_2^2+(1-\alpha)||\mathbf{y}||_2^2}\mathbf{X}_k^*\mathbf{P}_k + (1 - \frac{\alpha||\mathbf{X}||_2^2}{\alpha||\mathbf{X}||_2^2+(1-\alpha)||\mathbf{y}||_2^2})(\mathbf{y_k} - u_k)\boldsymbol{\beta}_k^T$
        $b = b + 1$; Calculate L_b per (5.6) and $\Delta L \leftarrow L_b - L_{b-1}$
      end
      Record the loss value for cluster k: $L_k = L_b$
    end
    Calculate the sum of loss values for the current partition: $L^v = \sum_{k=1}^{K} L_k$
    Initialize $b \leftarrow 0$ and $\Delta L \leftarrow L^v$
    while $b < b_{max}$ *and* $\Delta L < conv$ **do**
      $b = b + 1$
      for $i \leftarrow 1$ **to** N **do**
        Obtain the current assignment of i: $k_0 \leftarrow \mathbf{g}(i)$
        for $k'(k' \in 1, 2, ..., K)$ *where* $k' \neq k_0$ **do**
          Generate a trial partition $\mathbf{g}'$: $\mathbf{g}' \leftarrow \mathbf{g}$, $\mathbf{g}'(i) \leftarrow k'$
          Conduct a full Simultaneous Covariates Regression analysis (see Part 1) on each cluster gathered from $\mathbf{g}'$, resulting in $L^{v\prime}$ and the set of estimates (i.e., $\mathbf{P}_k$, $\mathbf{T}_k$, u_k and $\boldsymbol{\beta}_k$)
          If $L^{v\prime} < L^v$, $L^v \leftarrow L^{v\prime}$, $\mathbf{g} \leftarrow \mathbf{g}'(i)$ and update the set of estimates
        end
      end
    end
    $L^v \leftarrow L_b$; $\Delta L \leftarrow L_b - L_{b-1}$
  end
  If $L^v < L$, $L \leftarrow L^v$ and update the whole set of estimations accordingly
end

5.2.3. MODEL SELECTION

In CSCR, the values of a set of hyper-parameters have to be determined: the number of clusters K, common components R_{com} and distinctive components $R_{dis1}...R_{disI}$ for each data block. Typically, there is limited information about suitable values for these parameters. Here, the proposed model selection procedure consists of five steps and uses three widely applied criteria: the Bayesian Information Criterion (step 1), the predictive accuracy in a holdout sample (steps 3 and 5), and the scree ratio test (steps 2 and 4). These five model selection steps are described and motivated below. Note that, following previous studies (e.g., Gu and Van Deun, 2019; Schouteden et al., 2014), we also proposed a two-part model selection procedure to recover common and distinctive components - the first part determines the total number of components and the second decides the status (i.e., common or distinctive component pertaining to one of the data blocks) of these components. The proposed model selection strategy is evaluated in simulation study 2 (see Section 5.3.2)

- In step 1, a GMM is used to analyze the concatenated block of predictors $\mathbf{X}$; a preliminary value of K (denoted as K_0) is obtained by optimizing the BIC value of the model. In this way, step 1 offers a good initial guess of K that accounts for differences in the cluster centroids pertaining to the predictor blocks.

- In step 2, conditional on K_0, the scree ratio test, one of the most widely used criteria for selecting the number of components (e.g., De Roover, Ceulemans, and Timmerman, 2012, Wilderjans and Ceulemans, 2013), is applied to provisionally determine the total number of components R. Specifically, the scree ratio is calculated for $r \in [2,3,...,R_{max}-1]$ (where R_{max} denotes the maximum possible value of R), as follows:

$$sr_{r|K_0} = \frac{L_{r-1,K_0} - L_{r,K_0}}{L_{r,K_0} - L_{r+1,K_0}}, \tag{5.7}$$

  where L_{r,K_0} denotes the loss obtained from estimating a CSCR model with K_0 and r. An initial value for R, denoted by R_0, is obtained by selecting the value that maximizes the scree ratio defined in Equation (5.7).

- In step 3, conditional on R_0, the optimal value of K, denoted by K_{opt}, is determined using a prediction based approach. Specifically, a randomly selected training sample, consisting of 80% of the observations, is used to train the CSCR model for each candidate value of K. Subsequently, the trained model is applied to the holdout sample, consisting of the remaining 20% of the observations, to determine the predictive accuracy. After calculating the predictive accuracy for all possible values of K, K_{opt} is set to the value that maximizes the accuracy. According to Putka et al. (2018) and Yuan, Kroon, et al. (2021), this prediction-based approach avoids overfitting and promotes generalization. In addition, this prediction-based approach offers the flexibility to select from the full set of candidate values, so it overcomes an important drawback of the scree ratio test, namely the first and last values are impossible to be selected.

- In step 4, the procedure of step 2 is applied again to determine the optimal value of R, referred to as R_{opt}. Here, K_{opt} is used as the number of clusters, and when K_{opt} equals K_0, step 4 can be skipped.

- Finally, in step 5, the numbers of common and distinctive components are determined using the prediction-based approach with a holdout sample. Given the total number of components R_{opt}, blocks of zero loadings are imposed on $\mathbf{P}_k$ ($k \in (1,2,...,K_{opt})$ to reflect common and distinctive components and the structure of the components that maximizes the predictive accuracy is selected as the optimal solution.

5.2.4. RELATED METHODS

CSCR is a flexible method with a three-fold objective: to predict an outcome from multi-block data, to describe common and distinctive predictive components, and to account for the heterogeneity of the observations. Thanks to its generality, CSCR encompasses PCovR and CR as special cases and is closely related to CD-CovR and CSCA, as discussed in the introduction.

The three-fold aim of CSCR can be achieved in four alternative ways. First, CR can be applied to the concatenated data (hereafter referred to as the CR approach). However, as discussed in the introduction, compared to CSCR, the CR

approach has two important drawbacks: (1) it fails to deal with high-dimensional data sets, and (2) it is more prone to over-fitting because it can be affected by multicollinearity and the information contained in the outcome likely contaminates the prediction models (Brusco et al., 2008).

The second approach (called CSCA+LR) first applies CSCA to multi-block data sets to identify clusters of observations with the same common and distinctive components; in the second step, the resulting cluster-specific components are regressed on the outcome. We expect that CSCR outperforms CSCA+LR in the following scenarios. First, when some of the components explain a significant amount of variance in the predictors but only little variance in the outcome), CSCA+LR is more likely to pick up these components because the dimension reduction step of CSCA+LR does not account for the outcome block at all. This may result in missing the predictive components and reduced predictive accuracy. Second, when cluster-specific loading matrices are highly congruent, CSCR is also expected to outperform CSCA+LR, since CSCR uses both components and regression weights to identify clusters while CSCA+LR only uses the former.

The third approach is also a two-step approach. In the first step, iCluster (R. Shen et al., 2009) is used to partition the observations based on multi-block predictors. In the second step, CD-CovR is applied per cluster, yielding cluster-specific regression models. This approach is therefore termed iCluster+CD-CovR. The first step identifies clusters solely on the basis of mean structure. As a result, this step is able to recover the two clusters in Figure 5.1B well, but may not clearly separate the two clusters in Figure 5.1A. Thanks to its ability to recover cluster-specific components, CSCR is likely to outperform iCluster+CD-CovR in terms of predictive accuracy in cases where the components account for most cluster differences.

The last approach, CW-MBPLS, extends Multiblock Partial Least Squares (MB-PLS; Wold, 1984) to account for clusters of observations that differ in both components and regression weights. When comparing PLS-based methods and PCovR-based methods, previous studies (e.g., Gvaladze et al., 2021, S. Park et al., 2021, Kiers and Smilde, 2007) have consistently shown that, although PLS based methods sometimes yielded a better in-sample fit than PCovR based methods, the former were generally more prone to overfitting and, as a result, yielded lower

out-of-sample predictive accuracy than the latter. We expect the same pattern in the comparison between CSCR and CW-MBPLS: CSCR generally outperforms CW-MBPLS with respect to out-of-sample predictive accuracy.

5.3. SIMULATION STUDIES

Two simulation studies were conducted to examine the performance of CSCR as well as the proposed model selection procedure. In simulation study 1, the set of parameters - namely K, R_{com}, and R_{dis} - were regarded as known, and the performance of CSCR was compared with the aforementioned four competing methods (CR, CSCA+LR, iCluster+CD-CovR, and CW-MBPLS) in terms of predictive accuracy and cluster recovery. In simulation study 2, the values of the set of parameters were assumed to be unknown and we evaluated the extent to which the proposed model selection strategy recovered these parameters.

5.3.1. SIMULATION STUDY 1

DESIGN

In simulation study 1, each simulated data set consisted of a total of 120 observations on two predictor blocks (each containing 10 variables) and one outcome variable. The 120 observations were randomly assigned to 3 clusters of equal size. We then randomly partitioned each simulated data set into a training set (90 observations) and a test set (30 observations). This procedure ensures that both the training and test sets contain clusters of unequal size.

A number of parameters were fixed in data generation. The number of predictive components was fixed to 4. The four components consist of two common components and two distinctive ones. While both of the common components are associated with all 20 variables, the two distinctive ones are associated with 10 variables in one of the two predictor blocks. Besides these fixed parameters, a total of five factors were systematically manipulated in the following way:

- the relative size of the four clusters: equal (each of the three clusters contained 40 observations) or unequal (the three clusters contained 30, 40, and 50 observations, respectively)

Table 5.3: Cluster-specific regression weights used in simulation study 1

Regression Weights	Clusters	Common Component 1	Common Component 2	Distinctive Component 1	Distinctive Component 2
large differences in regression weights	cluster 1	.5	.5	.5	.5
	cluster 2	−.5	−.5	−.5	−.5
	cluster 3	.1	.7	−.1	−.7
small differences in regression weights	cluster 1	.5	.5	.5	.5
	cluster 2	−.5	.5	−.5	.5
	cluster 3	.1	.7	−.1	−.7

- the proportion of cluster differences in predictors that is explained by between-cluster mean structures ($b_x = \frac{\sum_{k=1}^{K} trace(\mathbf{m}_k'\mathbf{m}_k)}{\sum_{k=1}^{K} trace(\mathbf{m}_k'\mathbf{m}_k)+\sum_{k=1}^{K} tr(\mathbf{P}_k\mathbf{T}_k'\mathbf{T}_k\mathbf{P}_k')}$): 10%, 50%, or 90%. Therefore, the cluster-specific components explained $(1 - b_x)$ of the cluster differences in predictors.

- the proportion of cluster differences in outcomes explained by cluster-specific intercepts ($b_y = \frac{\sum_{k=1}^{K} trace(u_k'u_k)}{\sum_{k=1}^{K} trace(u_k'u_k)+\sum_{k=1}^{K} tr(\boldsymbol{beta}_k\mathbf{T}_k'\mathbf{T}_k\boldsymbol{beta}_k')}$): 10%, 50%, or 90%. Therefore, the weighted sum of components explained $(1 - b_y)$ of the cluster differences in outcomes.

- the proportion of noise in the predictor blocks (e_x): 20% or 30%

- the proportion of noise in the outcome (e_y): 10% or 30%

- the congruence level ϕ of the cluster-specific loading matrices, quantified by the average Tucker congruence (Haven and ten Berge, 1977; Tucker, 1951) between the cluster-specific loadings across all pairs of clusters: low (approximately .7) and high (approximately .85)

- the similarity between cluster-specific regression weights associated with the four predictive components: low or high; also see Table 5.3 for the values of these regression weights

A full factorial design that crossed the four factors was used, resulting in a total of $2 \times 3 \times 3 \times 2 \times 2 \times 2 \times 2 = 288$ conditions. In each condition, a total of 25 data

sets were generated and analyzed with CSCR and the aforementioned four related methods, namely CR, CSCA+LR, iCluster+CD-CovR, and CW-MBPLS. Therefore, in total 7200 data sets were generated and analyzed.

For CSCR, a prerequisite for making predictions is to identify the cluster assignment of each new observation in the test set and compute their respective component scores. Here, to determine the cluster assignment of each observation, we calculate its distance from the within-cluster component subspace of each cluster and assign it to the closest cluster. When calculating the distance between observations and clusters, we adopt an important assumption (also see Yuan, De Roover, Dufner, et al. (2021) for a similar assumption) that the addition of a new observation does not alter $\mathbf{P}_k$; therefore, when a new observation $\mathbf{x}$ is paired with cluster k, its component score equals $(\mathbf{x}-\mathbf{m_k^T})\mathbf{P}_k(\mathbf{P}_k^T\mathbf{P}_k)^{-1}$. The final component score of this observation is determined by which cluster it is assigned to.

All analyses were carried out in the software R (R core team, 2013), and we used the following packages for our estimation: for CSCR, we used the package `CSCR` that we have developed specifically for the current research; for CR, the package `flexmix` (Grün & Leisch, 2007); for CSCA+LR, the package `ClusterSSCA` (Yuan, De Roover, Dufner, et al., 2021) and some basic functions in base R; for iCluster+SCD-CovR, the package `iCluster` (R. Shen et al., 2012) and a self-developed function inspired by Gu and Van Deun (2019) and S. Park et al. (2021); and, last, for CW-MBPLS, the package `mbclusterwise` (Bougeard et al., 2018). In addition to CSCR and the four competing methods introduced above, we have also conducted a baseline test in our simulation study 1. The baseline test utilized an intuitive and simple approach: it used the average score of the outcome calculated from the training data as the predicted value for all new observations. Despite its simplicity, this prediction approach at times outperforms more complex and sophisticated prediction methods (Campbell & Thompson, 2008).

The results of these methods were compared on two main performance metrics: cluster recovery and predictive accuracy. Here, the degree of cluster recovery is quantified by the Adjusted Rand Index (ARI; Hubert and Arabie, 1985): When ARI reaches 1, the obtained partition is considered perfectly consistent with the true partition, while an ARI of 0 indicates that the cluster recovery is

only at the chance level. Meanwhile, the predictive accuracy is quantified by the mean squared error (MSE): a lower MSE indicates a better predictive accuracy.

DATA GENERATION PROCEDURE

We constructed each data set following the procedures outlined in Yuan, De Roover, Dufner, et al. (2021) and Wilderjans et al. (2017), as detailed below. First, a cluster indicator vector $\mathbf{g}$ was generated to represent the true cluster partition and $\mathbf{g}_i$ denotes the cluster assignment of observation i.

Second, we generated the concatenated block of predictors $\mathbf{X}$. According to Equations (5.4) and (5.5), it is the sum of three parts: $(\mathbf{T}_k\mathbf{P}_k^T)$, $(\mathbf{1}_{N_k}\mathbf{m}_k^T)$, and $(\mathbf{E}_k^{(\mathbf{X})})$ for each cluster $k \in (1,2,...,K)$. Without loss of generality, we set the average variance to 1 across all predictors in $\mathbf{X}$. Consequently, the average variance of predictors in the above three parts is, respectively, $(1-e_x)(1-b_x)$, $(1-e_x)(b_x)$, and e_x. We now elaborate on how to create each of these three parts. To create the first part, we constructed $\mathbf{T}_k$ in four steps: (1) each element was initially sampled from the standard normal distribution, (2) the resulting matrix was column-wise mean-centered and standardized, (3) the matrix was further orthogonalized such that $\mathbf{T}_k^T\mathbf{T}_k = \mathbf{I}$, and (4) each entry of $\mathbf{T}_k$ was multiplied by the square root of N_k. Next, $\mathbf{P}_k$ was generated in the following three steps. First, a common matrix and K cluster-specific matrices were generated whose elements were uniformly sampled from $\mathbf{U}(-1,1)$. We then re-scaled these matrices in accordance with the designated congruence level: in the low-congruence condition, the sum-of-squares of each row of the common matrix was re-scaled to .75 and that of the cluster-specific matrices was re-scaled to .25; in the high-congruence condition, the two values were set to .9 and .1, respectively. Adding the common matrix to the cluster-specific matrices resulted in the non-sparse version of the loading matrix for each cluster. Second, blocks of zero loadings were imposed on these non-sparse loading matrices to structure distinctive components, following the definition and operationalization detailed in section 5.2.1. Last, these loading matrices were re-scaled such that the average sum-of-squares of each row equaled $(1-e_x)(1-b_x)$. $\mathbf{P}_k^T$ was then multiplied by $\mathbf{T}_k$ to create the first part of $\mathbf{X}$. The generation of $\mathbf{1}_{N_k}\mathbf{m}_k^T$ was as follows: for each cluster k, $\mathbf{m}_k$ was generated with each entry sampled randomly from a uniform distribution $\mathbf{U}(-1,1)$; in accordance with the cluster partition as indicated in $\mathbf{g}$, $\mathbf{m}_k$ was assigned to

each row associated with cluster k, and all rows were then aggregated to form the complete mean structure $\mathbf{M}$; $\mathbf{M}$ was then re-scaled to ensure the average variance across its predictors was $(1-e_x)b_x$. Last, each entry in $\mathbf{E}^{(\mathbf{X})}$ was generated from a standard normal distribution, and the resulting matrix was then column-wise centered and re-scaled so that its average variance across all columns equaled e_x.

Last, we generated the outcome vector, which was again the sum of three parts: $\mathbf{T}_k\boldsymbol{\beta}_k$, u_k, and $\mathbf{e}^{(\mathbf{y})}$. The first part was generated per cluster by multiplying $\mathbf{T}_k$ and $\boldsymbol{\beta}_k$ and then mean-centering the multiplication; without loss of generalizability, we used σ to denote the variance of this part. u_k was created in the following three steps: (1) the K initial values of cluster-specific intercepts were randomly sampled from $\mathbf{U}(-1,1)$; (2) these K intercepts were distributed to each observation according to their cluster assignment; (3) the intercepts were centered and re-scaled such that their variance was $\sigma\frac{b_y}{1-b_y}$. Last, $\mathbf{e}^{(\mathbf{y})}$ was simulated in the following two steps: (1) each of the N entries of $\mathbf{e}^{(\mathbf{y})}$ was sampled from a standard normal distribution, and (2) the resulting matrix was centered and re-scaled such that its variance equaled $\sigma\frac{e_y}{(1-e_y)(1-b_y)}$.

RESULTS AND DISCUSSION

Since a satisfactory cluster recovery underpins the proper interpretation of cluster structures and cluster-specific components and is a prerequisite for accurate predictions, we first examine which of the seven factors used in simulation study 1 had a significant impact on cluster recovery for CSCR and the four competing methods[1]. Three of the seven factors manipulated in simulation study 1 had a significant impact on the extent to which CSCR was able to recover the clusters accurately. More specifically, CSCR recovered the clusters significantly better in the presence of (1) a medium-to-large proportion of cluster differences in the predictor blocks pertaining to differences in the cluster-specific components, i.e., when b_x = 50%, ARI = .97 and when b_x = 10%, ARI = .86 (as a reference: when b_x = 90%, ARI = .26), (2) a small portion of random error was added to the predictor blocks, i.e., when e_x = 20%, ARI = .76 (as a reference: when e_x = 30%, ARI = .63), and (3) the average congruence of cluster-specific loadings was low, i.e,

[1] Appendix 5.A provides a full descriptive table summarizing the cluster recovery and predictive accuracy for all methods at different levels of the seven factors.

when ϕ = .7, ARI = .74 (as a reference: when ϕ = .85, ARI = .66). The other four factors (i.e., the size of the clusters, the similarity of cluster-specific regression weights, the portion of random error added to the outcome block (e_y), and the portion of cluster differences in the outcome block pertaining to differences in the weighted sum of components b_y) had minimal effect on the extent of cluster recovery achieved by CSCR - indeed, for each of these four factors, the differences in the average ARI resulting from the different conditions were less than .05. Interestingly, the same results pattern can be observed for the other four methods as well: three of the seven factors (i.e., ϕ, b_x, and e_x as described above for CSCR) significantly affect the accuracy of each method in recovering clusters. Therefore, we further inspect the individual and joint effects of these three factors on the relative performances of the six methods [2] in terms of cluster recovery and predictive accuracy.

Figure 5.3 shows how the average ARI of each of the five methods varies with b_x and e_x (subplot A), and with b_x and ϕ (subplot B). Clearly, CSCR emerged as the winner in all conditions when b_x = 10% (CSCR: ARI = **.86**; iCluster+CD-CovR: ARI = .03; CSCA +LR: ARI = .77; CR: ARI = .01; CW-MBPLS: ARI = .04) or b_x = 50% (CSCR: ARI = **.97**; iCluster+CD-CovR: ARI = .72; CSCA +LR: ARI = .59; CR: ARI = .01; CW-MBPLS: ARI = .22). When b_x = 90%, however, the two-step approach iCluster+CD-CovR recovered the clusters best (CSCR: ARI = .26; iCluster+CD-CovR: ARI = **.99**; CSCA +LR: ARI = .27; CR: ARI = .00; CW-MBPLS: ARI = .10). The better performance of iCluster+CD-CovR over CSCR when b_x = 90% was expected and in line with the findings in Yuan, De Roover, Dufner, et al. (2021) where, in the absence of an outcome block, iCluster outperformed CSCA when b_x = 90%. This is because, in these scenarios, the total variance of the predictors explained by the underlying components is equal to or even smaller than the total variance explained by the random error. As a result, CSCR tends to incorrectly treat the added random error as meaningful co-variation attributed to the cluster-specific components. We now inspect the joint effects of the aforementioned three factors (i.e., ϕ, b_x, and e_x) on the relative performance of the competing methods. According to Figure 5.3, compared to the four competing methods, CSCR is most advantageous when b_x = 50% with e_x = 20% (CSCR: ARI

[2] For the baseline test, we only evaluate its predictive accuracy.

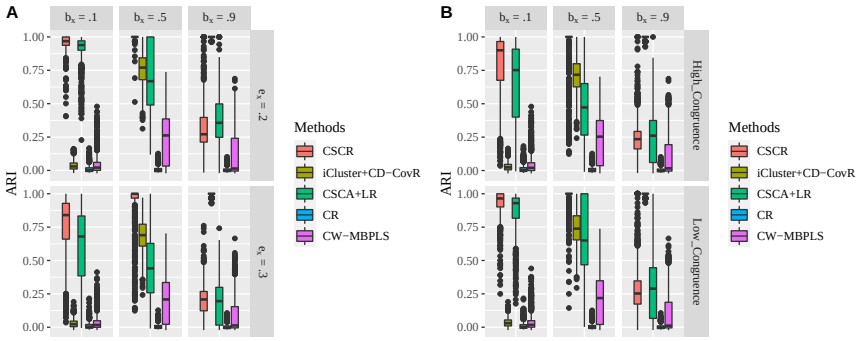


Figure 5.3: The extent of cluster recovery (indicated by ARI) by CSCR, iCluster+CD-CovR, CSCA+LR, CR and CW-MBPLS

= **.99**; iCluster+CD-CovR: ARI = .76; CSCA +LR: ARI = .70; CR: ARI = .01; CW-MBPLS: ARI = .24) and when $b_x = 50\%$ with $\phi = .7$ (CSCR: ARI = **.99**; iCluster+CD-CovR: ARI = .73; CSCA +LR: ARI = .67; CR: ARI = .01; CW-MBPLS: ARI = .20). More importantly, the results also suggest that as long as $b_x \neq 90\%$, even under the most difficult conditions, CSCR led to an average ARI of at least .75 (i.e., ARI = .79 when $b_x = 10\%$ with $\phi = .85$; ARI = .76 when $b_x = 10\%$ with $e_x = 30\%$). This level of cluster recovery is considered excellent according to the widely-adopted criterion proposed by Steinley (2004).

The above patterns about the relative performances of the five methods in terms of cluster recovery are generally consistent with the patterns regarding predictive accuracy, as shown in Figure 5.4. Here too, on average, CSCR yielded the best predictive accuracy when $b_x = 10\%$ (CSCR: MSE = **2.37**; iCluster+CD-CovR: MSE = 5.69; CSCA+LR: MSE = 2.47; CR: MSE = 5.01; CW-MBPLS: MSE = 5.75; Baseline: MSE = 5.61) or $b_x = 50\%$ (CSCR: MSE = **1.68**; iCluste+CD-CovRr: MSE = 2.60; CSCA+LR: MSE = 2.08; CR: MSE = 3.64; CW-MBPLS: MSE = 5.75; Baseline: MSE = 5.62), while iCluster+CD-CovR predicted the new observations most accurately when $b_x = 90\%$ (CSCR: MSE = 3.08; iCluster+CD-CovR: MSE = **2.16**; CSCA+LR: MSE = 2.38; CR: MSE = 3.62; CW-MBPLS: MSE = 3.98; Baseline: MSE = 5.63). These findings once again confirm that CSCR excels when $b_x = 50\%$ or $b_x = 10\%$ while the two-step approach combining iCluster and CD-CovR was superior when $b_x = 90\%$. Similarly and as expected, the predictive accuracy of CSCR was better with a lower ϕ and e_x. As shown in Figure 5.4, we also compared

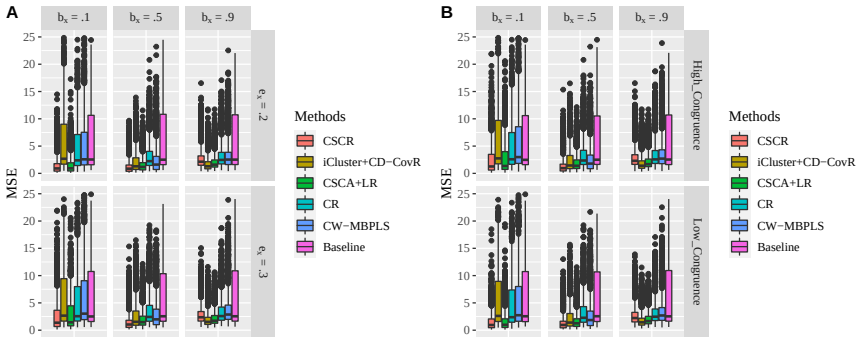


Figure 5.4: The extent of predictive accuracy (indicated by MSE) by CSCR, iCluster+CD-CovR, CSCA+LR, CR, CW-MBPLS, and the baseline test

the predictive accuracy of CSCR and competing methods with the baseline test under different conditions. Figure 5.4 illustrates that CSCR - along with CSCA+LR and, to a lesser extent, iCluster+CD-CovR - consistently outperformed the baseline test in terms of predictive accuracy across all conditions, while CR and CW-MBPLS exhibited larger prediction error compared to the baseline test in some of the conditions. Furthermore, CSCR convincingly outperformed the baseline test even in the most challenging situations (when $b_x = 90\%$ with $\phi = .85$: MSE = 3.17 for CSCR and MSE = 5.62 for the baseline test; when $b_x = 90\%$ and $e_x = 30\%$: MSE = 3.11 for CSCR and MSE = 5.66 for the baseline test). These results proved the stability of CSCR in terms of cluster recovery and predictive accuracy across conditions.

5.3.2. SIMULATION STUDY 2

DESIGN AND PROCEDURE

The purpose of simulation study 2 was to examine the performance of the proposed model selection method for CSCR. All parameters were set to the same values as in simulation study 1, with the following exceptions. First, to increase the scope of this study, the number of common and distinctive components (namely the number of common components R_{com}, the number of distinctive components for the first block R_{dis1} and the second block R_{dis2}) and the number of clusters K were varied in simulation study 2. More specifically, K was set to either 3 or 4, and the set (R_{com}, R_{dis1}, and R_{dis2}) could take one of the following

four sets of values: (2,0,0), (1,1,0), (3,0,0), (1,1,1). Second, ϕ was fixed to a high congruence level of .85. This decision was made because (1) the patterns of relative performances of the different methods were very similar under conditions with low versus high congruence, and (2) the use of a relatively high level of congruence allowed us to test the performance of CSCR in challenging conditions. Third, as reported above, some of the factors manipulated in simulation study 1 (e.g., cluster sizes, e_y, etc.) barely affect how well CSCR recovered the clusters. Therefore, we have fixed some of these factors in order to make simulation study 2 scalable: specifically, all clusters were assumed to be of equal size, and e_y was fixed to .1.[3] Forth, as e_x was one of the most important factors in simulation study 1, we increased the potential range of e_x to (20%, 30%, and 40%) in order to examine the performance of the model selection procedure in a more systematic way. Last, in simulation study 2, we set both $b_{\mathbf{X}}$ and $b_{\mathbf{y}}$ to either .1 or .5 (but not .9). This was because the selection of the number of components was less relevant when the mean structure explained most of the cluster differences in the predictor and outcome blocks - here, successful predictions would be possible even without an accurate estimate of the number of components. To summarize, six factors were systematically manipulated in simulation study 2:

- the proportion of noise in the blocks of predictors (e_x): 20%, 30% or 40%
- the number of clusters (K): 3 or 4
- the number of common components, and distinctive components for each predictor block (R_{com}, R_{dis1}, R_{dis2}): (2,0,0), (1,1,0), (3,0,0) or (1,1,1)
- the proportion of cluster differences in the scores on the *predictors* explained by the between-cluster mean structure (b_x): 0.1 or 0.5
- the proportion of cluster differences in the scores on the *outcome* explained by the between-cluster mean structure (b_y): 0.1 or 0.5
- the similarity between cluster-specific regression weights that were associated with the four predictive components: low or high; see Table 5.3 for the values of these regression weights

[3] Note that the similarity of cluster-specific regression weights was not fixed, because we expected that this factor might affect the process of selecting the correct number of clusters.

Table 5.4: The Percentage of replications where the parameters (i.e., K, R, and R_{com}) were recovered successfully

e_x	b_x	b_y	K	R	R_{com}
0.2	0.1	0.1	96%	97%	55%
		0.5	93%	95%	53%
	0.5	0.1	79%	77%	28%
		0.5	79%	79%	29%
0.3	0.1	0.1	89%	96%	50%
		0.5	87%	94%	47%
	0.5	0.1	61%	56%	31%
		0.5	60%	56%	26%
0.4	0.1	0.1	40%	86%	38%
		0.5	45%	85%	38%
	0.5	0.1	46%	48%	27%
		0.5	44%	54%	25%

A factorial design in which these factors were fully crossed resulted in a total of $3 \times 2 \times 4 \times 2 \times 2 \times 2 = 192$ conditions. Each condition included 30 replications. Therefore, a total of 5760 data sets were generated and analyzed with CSCR and the proposed model selection procedure. Both K and R were chosen from positive integers between 2 and 6. Furthermore, simulation study 2 followed the same procedures as simulation study 1 to generate the simulated data sets and to partition the training and test data sets.

Since the aim of simulation study 2 was to examine the performance of the proposed model selection procedure, we recorded the proportion of replications that successfully recovered K, R, and R_{com}. It is worth noting that, in a small-scale simulation, we found that successful retrieval of both K and R had a strong impact on the cluster recovery and predictive accuracy of CSCR. However, this did not apply to R_{com}: as long as R was successfully recovered, how well the number of common components was identified only had a limited impact on cluster recovery and predictive accuracy.

RESULTS AND DISCUSSION

Table 5.4 summarizes the percentage of replications where the true values of K, R, and R_{com} were successfully recovered. In general, the application of the proposed model selection procedure yielded high success rates in retrieving the true values of K and R (the average recovery rate across all conditions was 68% for K and 77% for R). The average success rate for retrieving the less important R_{com}, however, was 37%.

Despite the overall satisfactory performance of the proposed model selection procedure, its success became limited with the addition of a relatively large amount of random noise to the predictor blocks (i.e., when e_x = 40%). Indeed, a higher value of e_x significantly reduced the success rate of retrieving K (when e_x = .4: 44%; when e_x = .3: 74%; when e_x = .2: 87%), R (when e_x = .4: 68%; when e_x = .3: 76%; when e_x = .2: 87%), and R_{com}(when e_x = .4: 32%; when e_x = .3: 38%; when e_x = .2: 41%). The significant impact of e_x on the performance of the model selection procedure was in line with our finding that e_x was one of the most important parameters for CSCR analysis. Furthermore, corresponding to the findings of simulation study 1, b_x had a strong impact on the performance of the model selection procedure: a smaller b_x increased the success rate of retrieving K (when b_x = .1: 75%; when b_x = .5: 61%), R (when b_x = .1: 92%; when b_x = .5: 62%), and R_{com} (when b_x = .1: 47%; when b_x = .5: 28%).

5.3.3. SUMMARY

The two simulation studies reported above provided important insights into the relative performance of CSCR compared to the competing methods, and the extent to which the proposed model selection procedure successfully recovered the parameters of the CSCR model.

Simulation study 1 reported the superior performance of CSCR in both cluster recovery and predictive accuracy, as long as a substantial part of cluster differences in the predictor blocks were accounted for by cluster-specific components (i.e., b_x = 10% or b_x = 50%). When differences in these components explained only a small fraction of cluster differences (b_x = 90%), however, the two-step approach combining iCluster with CD-CovR proved to be the best method. Overall,

if the primary goal is to achieve high predictive accuracy while gaining some insights into individual and joint effects of predictors, CSCR is a promising method.

The results of simulation study 2 confirmed that the proposed model selection procedure can effectively retrieve the values of the two most important parameters: the number of clusters K and the total number of components R. The success rate of retrieving the true number of common components R_{com} was also acceptable, although lower than that of retrieving K or R.

5.4. ILLUSTRATIVE APPLICATION

We illustrate the usefulness of CSCR in an empirical analysis where survey respondents' attitudes towards immigrants (referred to as ATI hereafter) were predicted based on their self-reported personality traits and personal values. Previous research in personality and social psychology has shown that both personality traits and personal values are important determinants of ATI (Gallego and Pardos-Prado, 2014; Ackermann and Ackermann, 2015; Dinesen et al., 2016; (Davidov & Meuleman, 2012)). Among the five personality traits known as the Big Five (i.e., Extraversion, Openness, Conscientiousness, Agreeableness, and Neuroticism; McCrae and Costa Jr, 1989), only openness was found to be consistently predictive of ATI (Gallego and Pardos-Prado, 2014; Ackermann and Ackermann, 2015; Dinesen et al., 2016). Meanwhile, based on Schwartz's framework of personal values (Schwartz, 1992), Davidov and Meuleman (2012) found that valuing security was one of the most important determinants of ATI. Therefore, in the current illustration, we predicted ATI using one block of items measuring trait openness and one block of items measuring the personal value of security.

Although existing literature revealed the effects of openness and security on ATI in separate studies, an important question remains open as to whether openness and security jointly affect ATI, and, if so, how. Furthermore, given that both openness (Christensen et al., 2019) and security (Cieciuch & Schwartz, 2012) are fairly broad concepts that encompass various facets, it is reasonable to argue that different predictive mechanisms, separately or jointly, may apply to different subgroups of respondents. Therefore, the proposed CSCR method allows analysts to achieve the dual goal of predicting ATI while describing the predictive mechanisms and group differences therein; this may potentially give insights

that have not been obtained so far as suitable methods lacked. In addition, unlike most previous research that analyzed the personality and personal values at the level of constructs (e.g., openness and security as two dimensions) relying on factor scores (extracted from a factor analysis) or composite scores, CSCR allows for an analysis at the item level using item scores as predictors. As such, CSCR responds to the recent call in personality psychology for item-level analysis to examine the unique predictive contribution of items, beyond that of dimensions (e.g., Mõttus et al., 2017; Mõttus et al., 2019).

The data set used in the current study comes from the Longitudinal Internet Studies for the Social Sciences (LISS) panel, administrated by CenERdata (Tilburg University, the Netherlands). The panel collects responses from a representative sample of Dutch citizens and covers a wide range of measures, including but not limited to personality, political values, economic conditions, household composition, etc. A detailed discussion of the panel is available in Scherpenzeel (2018). Here we use the fourth wave of data, collected in 2018. More specifically, we picked 10 items measuring openness (adapted from the International Personality Item Pool; Goldberg et al., 2006) and 4 items measuring security value (adapted from the Rokeach Value Survey; Beatty et al., 1985). These 10 and 4 items, respectively, form the two blocks of predictor variables. For openness, respondents were instructed to indicate how accurately the 10 statements describe themselves on a 5-point Likert scale; for security, respondents were asked about the importance of each of the 4 security values in their lives, using a 7-point Likert scale. More specifically, the 10 statements pertaining to openness include (1) I have a rich vocabulary, (2) I have difficulty understanding abstract ideas (reverse coding), (3) I have a vivid imagination, (4) I am not interested in abstract ideas (reverse coding), (5) I have excellent ideas, (6) I do not have a good imagination (reverse coding), (7) I am quick to understand things, (8) I use difficult words, (9) I spend time reflecting on things, and (10) I am full of ideas. The 4 items about security are (1) responsible, (2) family security, (3) national security, and (4) inner harmony. Last, the dependent variable was the mean of 5 items measuring ATI (developed by the researchers involved in the panel; $\alpha = .76$; see Appendix 5.B for details of the items).

For illustrative purposes, the current analysis used a subset of 300 observations in which there were no missing values in the responses to the 19 selected items. We applied CSCR to the pre-processed data set in which each variable was mean-centered and standardized, as suggested in section 5.2. More specifically, the five-step model selection procedure described above was used to select the parameters (i.e., the number of clusters K, the total number of components R, the number of common components R_{com}, and the numbers of distinctive components pertaining to each block R_{dis1} and R_{dis2}). Both K and R were selected from all integers between 1 and 8, while the values of R_{com}, R_{dis1} and R_{dis2} were determined with the condition $R_{com} + R_{dis1} + R_{dis2} = R$.

The model selection procedure for CSCR used the following values as the optimal values for the parameters: 2 for both K and R, 1 for both R_{com} and R_{dis1}, and 0 for R_{dis2}. We report in Table 5.5 the cluster centroids of the two clusters across all 14 variables. Table 5.6 describes the component structures of the two clusters with each value indicating the correlation between the corresponding component and variable (note that because the components in the CSCR model are orthogonal, these correlations can be interpreted as ordinary correlations), and Table 5.7 contains the cluster-specific regression weights, which indicated the strengths of the connections between the components and outcome.

Table 5.5 reveals how the two clusters separate from each other in terms of cluster centroids. participants in Cluster 1 (versus Cluster 2) scored lower on average on Items 2 (understand abstract ideas), 4 (be interested in abstract ideas), and 6 (have a good imagination) from data block 1 and Item 4 (inner harmony is important) from data block 2, while scored higher on Item 2 (family security is important) from data block 2. Combining these results, we can conclude that participants in Cluster 2 (versus Cluster 1) were more attentive to inner feelings and imaginations, while participants in Cluster 1 (versus Cluster 2) focused more on family security.

The regression weights in Table 5.7 reveal another striking difference between the two clusters: while both components are predictive of ATI for respondents from Cluster 1, only Component 2 - but not Component 1 - is predictive for respondents from Cluster 2. We further interpret the cluster-specific components by inspecting the correlation between the components and variables in Table 5.6.

Table 5.5: Cluster centroids of the two clusters in the illustrative sample

Items	Cluster 1 ($N = 95$)	Cluster 2 ($N = 205$)
o1	0.21	-0.10
o2	**-0.28**	**0.13**
o3	0.08	-0.04
o4	**-0.27**	**0.12**
o5	0.13	-0.06
o6	**-0.44**	**0.21**
o7	-0.06	0.03
o8	0.12	-0.05
o9	0.01	-0.01
o10	0.06	0.21
sec1	0.12	-0.06
sec2	**0.35**	**-0.16**
sec3	-0.02	0.01
sec4	**-0.51**	**0.24**

Note. o1-o10 represent the 10 items from the openness scale and sec1-sec4 represent the 4 items from the security value scale. The reported values are the centroids of the two clusters, with values in bold indicating the items separating the two clusters to the greatest extent.

Table 5.6: Correlations between the components and the variables in the two-cluster solution

	Cluster 1		Cluster 2	
	Comp1	Comp2	Comp1	Comp2
o1	0.02	0.55	0.01	0.68
o2	0.16	0.8	-0.1	0.69
o3	-0.5	0.02	0	0.78
o4	0.29	0.62	0.08	0.65
o5	-0.01	0.29	-0.1	0.59
o6	0.35	-0.61	-0.11	0.73
o7	0.36	0.58	-0.22	0.56
o8	0	0.42	0.37	0.46
o9	-0.02	0.6	-0.19	0.2
o10	-0.4	0.17	-0.13	0.73
sec1	-0.24	0	-0.65	0
sec2	-0.46	0	-0.85	0
sec3	-0.44	0	-0.82	0
sec4	-0.96	0	-0.57	0

Note. The terms *Comp1* and *Comp2* refer to the first and second components for each cluster, respectively. The reported values indicate the correlation between the corresponding components and variables.

Table 5.7: Cluster-specific regression weights for the illustrative sample

Clusters	Intercepts	Comp1	Comp2
Cluster 1	−.08	.19	3.97
Cluster 2	.04	−.02	3.09

Note. The terms *Comp1* and *Comp2* were used to refer to the first and second components for each cluster, respectively.

We first discuss the interpretation of Component 1 for Cluster 1. This component is a common component that correlates with items from both predictor blocks. It correlates most strongly with the fourth item of security ($r = -.96$; inner harmony); it also correlates negatively with the third item of openness ($r = -.5$; have a vivid imagination), as well as the second (family security) and the third item of security (national security). Therefore, this component can be considered representative of respondents' imagination about how immigrants affect the security of their own, that of their families, and that of society. For respondents classified in Cluster 1, this is one of the (relatively minor) causes of their ATI. Interestingly, with the corresponding regression weight of -.02 (see Table 5.7), Component 1 is hardly predictive of ATI for respondents from Cluster 2. For both clusters, Component 2, the distinctive component that is only related to items measuring openness, is positively related to ATI (see Table 5.7; for Cluster 1, B = 4.0, for Cluster 2, B = 3.1). However, a closer inspection of the correlations between the openness items and the component scores of Component 2 in Table 5.6 shows that Component 2 should be interpreted differently for the two clusters. For Cluster 1, the component scores correlate positively with all items except item 3 ($r = .02$; have a vivid imagination) and item 6 ($r = -.61$; do not have a good imagination (reverse coded)). Therefore, Component 2 for Cluster 1 includes all facets of openness except the facet "active imagination". This fits very well with the above finding that Component 1 for Cluster 1 can be interpreted as a joint, negative effect of active imagination and the degree to which they value security. However, for Cluster 2, the component score of Component 2 correlates positively with all items, including items 3 and item 6. As such, the factors predicting ATI for respondents classified in Cluster 2 include all facets of openness (also note that Component 2 is the only predictive factor of ATI for Cluster 2).

Taking the above two pieces of findings together, our CSCR analysis yielded novel insights into the different joint contributions of openness and security on ATI for the two clusters: for one cluster that placed relatively high importance on inner feelings, abstract thoughts, and imaginations (i.e., Cluster 2), all facets of openness were positively related to ATI. For another cluster with relatively low levels of abstractions and imagination but a greater emphasis on family security (i.e., Cluster 1), "active imagination", as one of the facets pertaining to trait open-

ness, did not directly relate to ATI. However, it is important to note that CSCR is of an exploratory nature and that the above findings in no way reflect causal relationships between predictors and outcomes. We encourage researchers to conduct confirmatory analyses to validate these results.

5.5. DISCUSSION

In many behavioral studies, in order to incorporate different predictors to make accurate predictions, researchers more and more frequently employ multi-block data sets, which contain data collected from different sources and (or) measure distinct constructs. In these applications, it is desirable to pursue two goals simultaneously: (1) predict the outcome as accurately as possible, and (2) describe the individual and joint contributions from the diverse set of predictors. Furthermore, given the heterogeneous nature of human behavior, it is likely that these contributions differ across the various clusters. Successful detection of the clusters and their corresponding (individual and joint) contributions is a prerequisite for accurate prediction and interpretation. To achieve this goal, the current paper introduces a novel technique, called Clusterwise Simultaneous Component Regression, or CSCR, that simultaneously identifies clusters and recovers individual and joint contributions for each subgroup. Simulation study 1 confirmed that CSCR outperformed all competing methods when cluster differences in predictors were largely due to differences in within-cluster components. Furthermore, our proposed model selection procedure for CSCR was positively evaluated in simulation study 2. In the empirical illustration, we showed how CSCR provided additional insights into behavioral research by predicting respondents' attitudes towards immigrants (ATI) from a block of items measuring openness and a second block measuring security. The CSCR analysis identified two clusters of respondents where people in the first cluster were relatively more attentive to family security and those in the second cluster focused more on inner feelings and imaginations. Moreover, for those in the first cluster, their attitudes toward immigrants were negatively related to the joint effects of active imagination and insecurity and positively related to other facets of openness. For those in the second cluster, however, all facets of openness were positively related to attitudes toward immigrants. These results extend current theories by propos-

ing the potential heterogeneity of the relationships between traits/values and personal attitudes. Although this finding is exploratory in nature, it highlights how CSCR could contribute to both predictive analysis and theory development. From a broader perspective, CSCR is one of the very few methods that naturally blends the "person-centered approach" (i.e., describing simultaneously the underlying clusters) and the "variable-centered approach" (i.e., identifying the most important predicting forces summarized from all predictors). The integration of the person-centered and variable-centered approaches is particularly useful because (1) it responds to the recent call for such methods (Morin et al., 2018) and provides applied researchers with novel ways to detect and understand heterogeneity in their behavioral studies, and (2) it partly addresses the concerns raised in Brusco et al. (2019) about how to detect heterogeneous subgroups (i.e., clusters) in which variables interact in different ways.

An important question in the applied setting is how to choose the most appropriate (prediction) method from CSCR and the series of competing methods. Following recommendations in Grimmer et al. (2021) and Yuan, Kroon, et al. (2021), we recommend a data-driven, model-agnostic approach to model selection: rather than settling for one specific prediction model (prior to data analysis), researchers are encouraged to apply a full set of candidate models to empirical data sets and pick up the model with a relatively high predictive accuracy while enjoying reasonable interpretability. This strategy aligns very well with the "no free lunch theorem", which states that there is no model that fits every data set. In determining the set of candidate models, researchers can consult previous methodological papers that introduce and compare different prediction methods, as well as empirical papers that apply these prediction methods in their applications. They can then decide whether or not to include a specific method based on factors such as: (1) whether the assumptions of the method hold in the analysis (e.g., regression models assume normally distributed residuals), (2) whether the method yields relatively good predictive accuracy in previous studies, and (3) whether the results of the method are interpretable. As discussed in the introduction, point (3) is especially prominent when predictive mechanisms are of interest; the proposed CSCR is one of the few prediction methods that pro-

duce highly interpretable results (see Molnar, 2020 for an overview of other interpretable prediction methods).

When interpreting the predictive models based upon CSCR, users should keep in mind that any interpretation should be of an explanatory nature. This means that these results should not be used directly as conclusive evidence for or against theories or used directly to advise high-stake decisions. Instead, confirmatory analyses should be conducted to validate the results derived from CSCR.

The current study is not without limitations. First, in the current CSCR model, the number of components is restricted to be equal across all clusters. However, as we noted in the Methods section, this assumption may be too stringent to reflect the reality of some empirical studies and the violation of this assumption can largely complicate the interpretation of the results. Therefore, we will consider the possibility of relaxing this assumption and allowing different clusters to have a different number of components, while finding a way to retain computational feasibility. Second, as discussed in the Method section, the current method fixes the value of α, while, in a fully flexible model, α can be treated as a tuning parameter whose value is chosen in a data-driven manner. Third, the current version of CSCR is arguably computationally demanding. To develop a more efficient algorithm, future research can potentially integrate newly proposed optimization routines (e.g., Erichson et al., 2020) that solve the Sparse Principal Component Analysis problem in a more efficient manner.

We propose two future directions for the further development of CSCR. First, sparsity can be introduced into the CSCR model to account for high-dimensional data sets. For example, a cardinality constraint (Yuan, De Roover, Dufner, et al., 2021) or a lasso penalty (Gu & Van Deun, 2019) can be introduced to impose some loadings to be exactly zero. Second, the current study only compared the performances of different prediction methods with simulated data sets. We encourage future studies to compare these prediction methods in various empirical data sets to gain a solid understanding of their relative predictive accuracy.

APPENDICES

5.A. SUMMARY OF SIMULATION STUDY 1

See Table 5.8

5.B. DETAILS ABOUT ITEMS

Attitudes to Immigrants: What is your opinion on the following statements?

- It is good if society consists of people from different cultures
- Legally residing foreigners should be entitled to the same social security as Dutch citizens
- There are too many people of foreign origin or descent in the Netherlands (reverse coding)
- Some sectors of the economy can only continue to function because people of foreign origin or descent work there
- It does not help a neighborhood if many people of foreign origin or descent move in (reverse coding)

Table 5.8: Percentage of data sets resulting in a successful recovery of K, R, and R_{com}

Parameters	Values	CSCR		iCluster+CD-CovR		CSCA+LR		CR		CW-MBPLS		Baseline	
		MSE	ARI	MSE	ARI	MSE	ARI	MSE	ARI	MSE	ARI	NSE	ARI
Cluster sizes	30,40,50	2.40	.69	3.51	.58	2.30	.54	4.06	.01	4.29	.12	5.60	/
	40,40,40	2.35	.71	3.46	.58	2.32	.54	4.12	.01	4.32	.12	5.65	/
b_x	.1	2.37	.86	5.69	.03	2.47	.77	5.01	.01	5.75	.04	5.61	/
	.5	1.68	.97	2.60	.72	2.08	.59	3.64	.01	3.18	.22	5.62	/
	.9	3.08	.26	2.16	.99	2.38	.27	3.62	.00	3.98	.09	5.63	/
b_y	.1	1.03	.71	1.17	.58	.94	.55	1.7	.01	1.6	.13	1.43	/
	.5	1.38	.70	1.79	.59	1.32	.53	2.36	.01	2.34	.12	2.60	/
	.9	4.72	.68	7.51	.58	4.67	.54	8.22	.00	8.99	.10	12.80	/
e_x	.2	2.13	.76	3.37	.60	1.99	.66	3.85	.01	3.94	.13	5.62	/
	.3	2.62	.63	3.61	.57	2.63	.42	4.33	.01	4.67	.11	5.63	/
e_y	.1	1.40	.71	2.57	.58	1.51	.56	2.94	.01	2.83	.16	4.96	/
	.3	3.45	.68	4.50	.58	3.19	.52	5.37	.00	5.94	.07	6.28	/
ϕ	.75	2.19	.74	3.37	.59	2.09	.61	4.02	.01	4.24	.11	5.63	/
	.9	2.56	.66	3.60	.58	2.52	.47	4.16	.01	4.37	.13	5.61	/
similarity of $\boldsymbol{\beta}_K$	low	2.35	.59	3.42	.57	2.30	.43	4.01	.01	4.24	.11	5.63	/
	high	2.40	.60	3.56	.57	2.32	.44	4.18	.01	4.36	.11	5.61	/

Note. MSE quantifies the prediction error of the test set, while ARI quantifies the accuracy of cluster recovery of the training set. The baseline test was used for prediction only; therefore, no ARI was reported for the baseline test.

6

DISCUSSION

In the previous four chapters of this dissertation, we presented three novel statistical methods (i.e., CKM, CSSCA, and CSCR; see Table 6.1 for a brief summary of their functionalities) and computational tools (i.e., three R packages that implemented the three methods and a ShinyApp for cluster visualization) for detecting underlying clusters in behavioral data sets with complex cluster structures (e.g., data sets that contain a substantial proportion of irrelevant variables and (or) that consist of variables from different sources). In this section, we discuss how researchers can choose from different methods (the three methods developed here but also other related methods), as well as the opportunities these methods and computational tools offer for behavioral research in the data-rich era.

6.1. MODEL SELECTION AND CLUSTER VALIDATION

In practice, researchers who want to use the three proposed methods (i.e., CKM, CSSCA, and CSCR) in their research must make many decisions during the analysis process. Arguably one of the most important decisions is to determine which clustering methods are best suited to the analysis at hand. One way to make this decision is to consider the characteristics of these methods and the criteria they use to define clusters (see Table 6.1) and select the one that fits best the purpose of the analysis. For example, when the data set under consideration contains an

Table 6.1: A summary of the three methods described in the dissertation

Methods	Criteria to define clusters	Multi-block analysis	Variable selection	Prediction
CKM	Between-cluster mean structures only	No	Yes	No
CSSCA	Between-cluster mean structures and within-cluster component structures	Yes	Yes	No
CSCR	Between-cluster mean structures, within-cluster component structures and regression weights	Yes	No	Yes

exceeding amount of variables, CKM may be useful in generating a first impression of the cluster profiles, because (1) it is less computationally demanding than more complex methods such as CSSCA and CSCR, and (2) it is able to select a subset of variables that, compared to the traditional K-means algorithm, greatly eases interpretation. In contrast, when the aim is to determine which combinations of health behaviors best predict obesity, CSCR should be considered first because it identifies within-cluster components and optimizes prediction models within each cluster.

Although considering the fit between the clustering methods and the goal of the analysis can be very informative, it does not always provide definite answers. For example, researchers may be completely unaware of the underlying cluster structure of the data set and thus unable to determine whether within-cluster component structures should be accounted for in addition to between-cluster mean structures (e.g., whether CKM or CSSCR should be preferred). Cluster validation, as described in Chapter 3, can be a promising, post-hoc approach to inform these decisions. More specifically, researchers are advised to take a model-agnostic approach (also see Grimmer et al., 2021 and Yuan, Kroon, et al., 2021), in which they first employ all clustering methods and then determine the most suitable one(s) - as well as the optimal cluster assignments - using theory-guided validation and stability validation. First, theory-guided validation can be applied to examine whether the patterns of the between-cluster mean structures and within-cluster component structures can be understood from a theoretical

perspective. Failure to understand these results strongly suggests a mismatch between the methods and the data set under consideration. Second, as detailed in Chapter 3, a bootstrap-based cluster stability index can be used to infer the stability of the obtained cluster solutions. An unstable cluster structure indicates that the identified clusters cannot be replicated upon re-sampling and (or) are highly susceptible to random noise. Future research could use a combination of simulation studies and empirical analysis to determine whether the model-agnostic approach is successful in selecting the best clustering methods for the analysis.

6.2. CHALLENGES AND FUTURE DIRECTIONS

Hopefully, it is clear that the novel methods and computational tools presented in this dissertation offer new insights and opportunities for identifying heterogeneity in behavioral data sets with complex cluster structures.

Admittedly, the reported methods can be further developed and improved in many important ways. Below, we detail three aspects in which these methods can be further improved.

6.2.1. A MODEL-BASED CLUSTERING APPROACH TO VARIABLE SELECTION AND COMMON AND DISTINCTIVE COMPONENTS IDENTIFICATION

We believe that the novel methods proposed in this dissertation can be flexibly combined with other methods to inspire the creation of novel techniques for social science research. A particularly interesting prospect would be to incorporate the regularization approach for variable selection (see Chapters 2 and 4) and for common and distinctive component identification (see Chapters 4 and 5) into other clustering methods, for example, soft partitioning methods (i.e., assigning each subject to each cluster with a certain probability; also known as model-based clustering). Although the hard partitioning methods described in the dissertation (i.e., assigning each subject to one of the clusters with 100% certainty) possess many important advantages, such as computational efficiency and unambiguous cluster assignments, they tend to yield unfavorable clustering accuracy when clusters overlap to a large extent (Vermunt, 2011). In addition, for K-means specifically, another unsatisfactory characteristic is that it gives the

same weight to all variables in determining the underlying cluster structure; in other words, K-means adopts a highly unrealistic assumption that all variables contribute equally to cluster separation (Magidson & Vermunt, 2002). Although CKM proposed in this dissertation partly addresses this issue by distinguishing between signaling variables - which do effectively separate clusters - and irrelevant variables - which do not separate clusters at all, CKM still assumes that *all* signaling variables contribute to cluster separation to the same extent. One way to relax this stringent assumption is to incorporate the regularization approach into weighted K-Means (as implemented in Witten and Tibshirani, 2010 and Kondo et al., 2016), where each variable is assigned a specific weight indicating its unique contribution to cluster separation and the weights regularized to zero correspond to irrelevant variables. From a model-based clustering perspective, however, the two shortcomings mentioned above (i.e., failure to handle clusters with large overlaps and the stringent assumption that all variables separate clusters equally well) can be addressed simultaneously and naturally, thus offering a fascinating alternative to the hard partitioning approaches used throughout the dissertation. Below, we discuss two proposals to integrate the regulation approach into the model-based clustering methods, such as Gaussian Mixture Models (GMM; Vermunt and Magidson, 2002, McLachlan et al., 2019, McNicholas, 2016) and Mixtures of Factor Analyzers (MFA; McLachlan et al., 2003; Hinton and Ghahramani, 1997; Andrews and McNicholas, 2012).

First, a hybrid CKM-GMM analysis - applying GMM directly to a subset of signaling variables selected from CKM (see Chapters 2 and 3 for details) - can be considered as a model-based clustering algorithm with the additional support of variable selection. While the application of CKM in the first step offers an efficient way to filter out irrelevant variables, using GMM - instead of KM - in the second step brings greater flexibility and accuracy in recovering clusters. This hybrid approach is related to the methods proposed in Pan and Shen (2007) and S. Wang and Zhu (2008) which both accomplished variable selection within the GMM framework by penalizing the mean structure in the log-likelihood function (in fact, Pan and Shen (2007) proposed to penalize with the l_1 norm and Z. Zhang et al. (2009) with the l_∞ norm). Compared to these two methods that perform simultaneous variable selection and clustering by optimizing one single likeli-

hood criterion, this hybrid method may be more computationally efficient (because the variable selection step implemented in CKM relies on a much simpler criterion), but less accurate (because some of the signaling(irrelevant) variables identified within the KM framework may be irrelevant(signaling) for GMM).

Second, the idea of discerning common and distinctive variations can be potentially combined with the MFA model to create a new method that finds two types of clusters: common clusters pertaining to all data blocks and distinctive clusters pertaining to one or only a few data blocks. First proposed in Hinton and Ghahramani (1997) and later developed in McLachlan et al. (2003), MFA effectively assumes that the observed data set comes from a mixture of K factor analyzers - K being the number of clusters - and simultaneously finds the clusters and locally reduces the dimensions for each cluster with a likelihood function. Since MFA transforms the original variable space into cluster-specific subspaces (indicated by cluster-specific factors), MFA is able to handle a large number of variables with adequate efficiency. However, the clusters estimated by MFA are always separated by subspaces that span all variables. Therefore, when dealing with multi-block data sets, MFA cannot identify distinctive clusters separated by subspaces underlying one or only a few data blocks. To address this shortcoming of MFA, a two-step approach can be potentially useful: in the first step, the regularized Simultaneous Component Analysis (Gu & Van Deun, 2016, 2019) - the method discussed extensively in Chapter 4 - can be deployed to divide the multi-block data sets into two parts: one part with only common variation and the other part with only distinctive variation; in the second step, MFA can be applied to these two parts separately to recover common and distinctive clusters, respectively.

6.2.2. DEVELOPMENTS OF MORE FLEXIBLE VERSIONS OF THE METHODS

The three methods discussed in the dissertation can be improved to provide greater flexibility for analysts. These improvements include (1) increasing the type of data sets they can deal with and (2) lifting some of the model constraints. Currently, the three proposed methods can only deal with continuous variables with no missing values. In order to provide greater flexibility to users, the methods can be extended by building on the discrete PCA model (Kolenikov, Angeles,

et al., 2004) or the mix PCA model (Anderson-Bergman et al., 2018) to accommodate discrete and (or) mixed types of data. Note that the most commonly used approach to deal with discrete data in PCA, namely translating the variables with multiple categories into a set of dummy variables (Filmer & Pritchett, 2001), is found to be unreliable because it introduces spurious correlations (Kolenikov, Angeles, et al., 2004). Other more complicated ways of handling discrete variables include, for example, CATegorical Principal Component Analysis (CATPCA; Linting et al., 2007; Linting and van der Kooij, 2012) that assigns numeric values through a process of optimal scaling. Future studies can potentially evaluate how this interesting approach fares for the newly proposed CSSCA and CSCR methods. Furthermore, various missing data imputation methods proposed for Principal Component Analysis (e.g., Josse et al., 2011; Malan et al., 2020) can be examined, compared, and extended to serve as a data-prepossessing step prior to the application of CKM, CSSCA, or CSCR.

Another important limitation of these methods is that they impose rather stringent model constraints. Specifically, CKM defines signaling variables as those that effectively separate *all* clusters, but a more realistic and flexible definition of these signaling variables should be those that effectively separate *any* pair of clusters. To develop this less constrained model, CKM can build on different versions of SPCA that do not restrict variables to have zero loadings on all components. Moreover, CSSCA and CSCR currently impose the constraint that all clusters have the same number of (common and distinctive) components, which may be too stringent in behavioral studies. To lift this constraint, future research can develop new methods that use an extensive model selection procedure to determine the optimal number of components for each cluster independently. However, it should be noted that these additional model selection steps can drastically increase the computational burden of the original methods.

6.2.3. STRATEGIES TO IMPROVE COMPUTATIONAL EFFICIENCY

The three methods (i.e., CKM, CSSCA and CSCR) and their associated model selection procedures involve repeated model estimation for a large number of re-samples and (or) several multi-start procedures. As a result, these estimations can be time-consuming for applied researchers and eventually hinder the

dissemination of the methods. To improve the computational feasibility of the methods described in the dissertation, we present two potential solutions.

First, from a modeling perspective, the convex clustering approach can be integrated into the three proposed methods to provide more efficient solutions for model estimation. Convex clustering, as proposed in Pelckmans et al. (2005) and Lindsten et al. (2011), aims to address the notorious problem inherited by many hard partitioning methods, namely that cluster initialization has a strong impact on the clustering results and that globally optimized solutions are difficult to achieve. To address this issue, instead of assigning each observation directly to a cluster, convex clustering assigns each observation to a point called "cluster centroid" and later puts observations into the same cluster if their cluster centroids are close enough; by doing so, convex clustering effectively transforms the optimization process into a convex minimization problem (for technical details, please refer to Pelckmans et al., 2005). Consequently, convex clustering enjoys the highly desirable property of always finding the unique solution corresponding to the global minimum of the loss functions. Therefore, the deployment of convex clustering effectively avoids multiple initializations and improves the computational efficiency of algorithms. Recently, convex clustering has been extended to incorporate the sparseness approach (B. Wang et al., 2018). Based on these previous studies on convex clustering, future research can redefine the optimization criteria of the proposed methods and turn the objectives of data analysis into convex optimization problems.

Another strategy for increasing the computational efficiency of the described methods is to program them in a more efficient programming language (e.g., C++) and then embed the code in a parallel computational system. The parallel system synchronizes the computational power of multiple high-performance computing facilities. With this system, the re-sampling procedures can be distributed among these computing facilities and the computational speed can be increased by a factor of 10 or even 100.

ACKNOWLEDGEMENTS

When applying for the NWO research talent grant, which eventually made this thesis possible, I dared to refer to my supervisory and collaborative team “a dream team”. What I did not know at the time was that this would prove to be an understatement. With Katrijn and Kim, I probably have the most supportive and helpful supervisors one could ever imagine. Recognized as respectful figures in the field of clustering and component analysis, they are, of course, able to answer all the research questions that I may have encountered and provide important guidance when results turned out to be not as expected. Importantly, their profound knowledge and research experience have not translated into a strong research agenda in which their students would have limited room to explore (read: make mistakes); quite the opposite, I have been very lucky to be granted almost unlimited freedom to carry out research projects to my liking. Outside of their professional lives as professors, they are simply super fun people with whom one would always like to be around. Indeed, it was not an accident if you heard loud laughter coming from the conference room where we held our regular meetings. Until, of course, the Covid-pandemic. The beginning of the pandemic in the spring 2020 was especially difficult for me, as I lived in the middle of the Chinese narrative, where the pandemic was almost a doomsday, and the (initial) Dutch narrative, where the virus was nothing more than a flu. Again, fortunately, I have two of the most supportive mentors who fully understood the struggles I might have, allowed me to work, for an extended period of time, from home in China, and, above all, offered me a lot of psychological and emotional support. This thesis is certainly not the last journey of the dream team, and I much look forward to working together for many years to come. An anecdote: I believe I might be the mysterious mascot of my supervisors: during the final stage of my thesis, Katrijn became the head of the department and Kim won two highly competitive grants.

Besides my daily supervisors, my promotor Jeroen and indeed the entire MTO department have been an important part of this amazing PhD journey. Jeroen

is such a sharp person that, even though he was not a frequent in our research meetings, in occasions when he did attend my research presentations, he could easily come up with challenging questions and important suggestions. The entire MTO department has really made me feel at home with regular research sandwiches, monthly drinks, countless organizations of study groups and journal clubs, amazing department trips, and above all, an extremely fun and supportive PhD communities. Special thanks to the core members of the Cool People club: although our journal clubs always turned into casual chats and our planned side projects never worked out, the group defined the notion of academic family for me; for that I will always be grateful. One of my academic family members unfortunately misses my defense, because he prefers Caribbean carnivals to a boring, lengthy presentation; arguably, he is the one with whom I have spent more time with than my girlfriend, as we have shared the magnificent sixth floor – a.k.a. away from the crowds – office for over three years. I am also grateful to my collaborators inside and outside Tilburg University, who have made this academic journey all the more memorable.

With these amazing peopl, surely, I stared my Ph.D. journey with great ease. Another factor enabling this smooth start is our "Tilburg Homies"; we have developed strong and lasting bonds ever since our master studies here at Tilburg. Understandably, some homies have left the cosmopolitan city of Tilburg and embarked on their adventures in the UK or the US; our bond, however, has never been weakened. Indeed, our frequent chats over the years have been one of the great witnesses of our development from insecure master students to ones who have made and are still making small, yet unneglectable, contributions to science and the society by and large.

One of the most important supports for my PhD journey, and indeed my life in general, has always been the most amazing Chinese community residing in the Netherlands. The company of these important and happy people ensured that I never faced any cultural adaption issues – I could comfortably stay within this circle of friends without worrying too much about the many challenges in adapting to a new culture. Perhaps more importantly, the friendships, supports, and chats I shared with them constantly reminds me that I am not alone in my

pursuit of understanding the complexity of human nature and that I can safely and wholeheartedly share all my joys and sorrows with these like-minded people.

Together with some of these Chinese friends, we initiated two small social enterprises – a blog and, more recently, a podcast – that advocate rigorous social science methodologies to researchers, and introduce useful psychological findings to the general population, respectively. The creation and operation of these two baby projects fulfills my long-lasting dream of realizing research impact on a larger scale. Gradually, I realized that they have another existential meaning to me: they made the problem of lacking timely rewards – an issue almost inherent to the scientific careers – much less pressing. Even in a low period when most manuscripts were turned down, the outputs of these toy enterprises always cheered me up.

As a traveler, I find myself almost implausible to settle down in one place for a long period of time. This is perhaps the reason that, while studying at Tilburg University, I lived in Tilburg, Amsterdam and Rotterdam for more than one years. In Amsterdam and Rotterdam, too, I have been incredibly lucky to be in the company of amazing people. We started an almost bi-weekly hotpot and Chinese food group in Amsterdam; the group, later termed "Play Play group", has been active until now. Also in Amsterdam, but much more recently, I discovered the fun of playing Padel and appreciating Korean culture with some of my best friends. In Rotterdam (and sometimes Delft), we enjoyed brunch, played tennis and table tennis, celebrated Chinese New Year, drank through weekdays and weekends, and were regulars at hotpot restaurant, bubble tea shops, and board game stations. With this fantastic group of people, not a day goes by without fun activities and deep conversations.

During the final stage of my PhD journey, the best thing I can never dream of became a reality: I was accepted into the Leadership and Management section in Amsterdam Business School, opening a new chapter of my academic career as an independent researcher. The best thing: my new colleagues are just as friendly, supportive and fun as those in the MTO department. Although I have only been working there since February 2022, almost 8 months before this acknowledgement is written, I have already been involved in so many funny and impactful research and educational activities. Together with other colleagues at Amster-

dam People Analytics Centre (APAC), we are doing our best to advance the synergies between the HR industry and academia, and between data sciences and HR theories. Indeed, the most fulfilling life is the one with a grand mission.

Last but certainly not least, I can never have made the journey this far without the support of THE most important people in my life: my girlfriend and my parents. They are the shelter to the traveler, the answer to all questions, and the reason for all beautiful things. With them, I can confidently and proudly claim that even if the world turns upside down, there will always be someone with me. LOVE YOU.

BIBLIOGRAPHY

Ackermann, K., & Ackermann, M. (2015). The big five in context: Personality, diversity and attitudes toward equal opportunities for immigrants in switzerland. *Swiss Political Science Review, 21*(3), 396–418.

Adachi, K., & Trendafilov, N. T. (2016). Sparse principal component analysis subject to prespecified cardinality of loadings. *Computational Statistics, 31*(4), 1403–1427.

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist, 73*(7), 899.

Akhanli, S. E., & Hennig, C. (2020). Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Statistics and Computing, 30*(5), 1523–1544.

Albada, K., Hansen, N., & Otten, S. (2021). Polarization in attitudes towards refugees and migrants in the netherlands. *European Journal of Social Psychology, 51*(3), 627–643.

Anderson-Bergman, C., Kolda, T. G., & Kincher-Winoto, K. (2018). Xpca: Extending pca for a combination of discrete and continuous variables. *arXiv preprint arXiv:1808.07510*.

Andrews, J. L., & McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing, 22*(5), 1021–1029.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243–256.

Arias-Castro, E., & Pu, X. (2017). A simple approach to sparse clustering. *Computational Statistics & Data Analysis, 105*, 217–228.

Arvey, R. D., Li, W.-D., & Wang, N. (2016). Genetics and organizational behavior. *Annual Review of Organizational Psychology and Organizational Behavior, 3*, 167–190.

Beatty, S. E., Kahle, L. R., Homer, P., & Misra, S. (1985). Alternative measurement approaches to consumer values: The list of values and the rokeach value survey. *Psychology & Marketing*, *2*(3), 181–200.

Berning, C. C., & Schlueter, E. (2016). The dynamics of radical right-wing populist party preferences and perceived group threat: A comparative panel analysis of three competing hypotheses in the netherlands and germany. *Social Science Research*, *55*, 83–93.

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The annals of statistics*, *44*(2), 813–852.

Bougeard, S., Abdi, H., Saporta, G., & Niang, N. (2018). Clusterwise analysis for multiblock component methods. *Advances in Data Analysis and Classification*, *12*(2), 285–313.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, *71*, 52–78.

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science: With applications in r* (Vol. 50). Cambridge University Press.

Brudvig, S., Brusco, M. J., & Cradit, J. D. (2019). Joint selection of variables and clusters: Recovering the underlying structure of marketing data. *Journal of Marketing Analytics*, *7*(1), 1–12.

Brusco, M. J., & Cradit, J. D. (2001). A variable-selection heuristic for k-means clustering. *Psychometrika*, *66*(2), 249–270.

Brusco, M. J., Cradit, J. D., Steinley, D., & Fox, G. L. (2008). Cautionary remarks on the use of clusterwise regression. *Multivariate Behavioral Research*, *43*(1), 29–49.

Brusco, M. J., Steinley, D., Hoffman, M., Davis-Stober, C., & Wasserman, S. (2019). On ising models and algorithms for the construction of symptom networks in psychopathological research. *Psychological Methods*, *24*(6), 735.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230.

Cadima, J., & Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle compenents. *Journal of applied Statistics*, *22*(2), 203–214.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of statistical software*, *61*, 1–36.

Chawarska, K., Ye, S., Shic, F., & Chen, L. (2016). Multilevel differences in spontaneous social attention in toddlers with autism spectrum disorder. *Child development*, *87*(2), 543–557.

Chen, J., Patil, K. R., Weis, S., Sim, K., Nickl-Jockschat, T., Zhou, J., Aleman, A., Sommer, I. E., Liemburg, E. J., Hoffstaedter, F., et al. (2020). Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: An international machine learning study. *Biological psychiatry*, *87*(3), 282–293.

Chen, J. (2021). A bayesian regularized approach to exploratory factor analysis in one step. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–11.

Chi, W., Li, W.-D., Wang, N., & Song, Z. (2016). Can genes play a role in explaining frequent job changes? an examination of gene-environment interaction from human capital theory. *Journal of Applied Psychology*, *101*(7), 1030.

Chipman, H., & Tibshirani, R. (2006). Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, *7*(2), 286–301.

Christensen, A. P., Cotter, K. N., & Silvia, P. J. (2019). Reopening openness to experience: A network analysis of four openness to experience inventories. *Journal of Personality Assessment*, *101*(6), 574–588.

Cieciuch, J., & Schwartz, S. H. (2012). The number of distinct basic values and their structure assessed by pvq–40. *Journal of personality assessment*, *94*(3), 321–328.

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., & Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British journal of health psychology*, *10*(3), 329–358.

Davidov, E., & Meuleman, B. (2012). Explaining attitudes towards immigration policies in european countries: The role of human values. *Journal of Ethnic and Migration Studies*, *38*(5), 757–775.

Davis, C., Zai, C. C., Adams, N., Bonder, R., & Kennedy, J. L. (2019). Oxytocin and its association with reward-based personality traits: A multilocus genetic profile (mlgp) approach. *Personality and Individual Differences*, *138*, 231–236.

De Jong, S., & Kiers, H. A. (1992). Principal covariates regression: Part i. theory. *Chemometrics and Intelligent Laboratory Systems*, *14*(1-3), 155–164.

De Roover, K., Ceulemans, E., & Timmerman, M. E. (2012). How to perform multiblock component analysis in practice. *Behavior Research Methods*, *44*(1), 41–56.

De Roover, K., Ceulemans, E., Timmerman, M. E., Nezlek, J. B., & Onghena, P. (2013). Modeling differences in the dimensionality of multiblock data by means of clusterwise simultaneous component analysis. *Psychometrika*, *78*(4), 648–668.

De Roover, K., Ceulemans, E., Timmerman, M. E., & Onghena, P. (2013). A clusterwise simultaneous component method for capturing within-cluster differences in component variances and correlations. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 81–102.

De Roover, K., Ceulemans, E., Timmerman, M. E., Vansteelandt, K., Stouten, J., & Onghena, P. (2012). Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychological methods*, *17*(1), 100.

De Roover, K., Timmerman, M. E., Mesquita, B., & Ceulemans, E. (2013). Common and cluster-specific simultaneous component analysis. *PLoS One*, *8*(5), e62280.

DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, *5*(2), 249–282.

Dinesen, P. T., Klemmensen, R., & Nørgaard, A. S. (2016). Attitudes toward immigration: The role of personal predispositions. *Political Psychology*, *37*(1), 55–72.

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on Machine learning*, 29.

Doucet, O., Fredette, M., Simard, G., & Tremblay, M. (2015). Leader profiles and their effectiveness on employees' outcomes. *Human performance*, *28*(3), 244–264.

Dufner, M., Arslan, R. C., Hagemeyer, B., Schönbrodt, F. D., & Denissen, J. J. (2015). Affective contingencies in the affiliative domain: Physiological assessment, associations with the affiliation motive, and prediction of behavior. *Journal of Personality and Social Psychology*, *109*(4), 662.

Durieux, J., & Wilderjans, T. F. (2019). Partitioning subjects based on high-dimensional fmri data: Comparison of several clustering methods and studying the influence of ica data reduction in big data. *Behaviormetrika*, *46*(2), 271–311.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological methods*, *23*(4), 617.

Erichson, N. B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N., & Aravkin, A. Y. (2020). Sparse principal component analysis via variable projection. *SIAM Journal on Applied Mathematics*, *80*(2), 977–1002.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293–314.

Fang, Y., & Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, *56*(3), 468–477.

Feldman, R., Monakhov, M., Pratt, M., & Ebstein, R. P. (2016). Oxytocin pathway genes: Evolutionary ancient system impacting on human affiliation, sociality, and psychopathology. *Biological Psychiatry*, *79*(3), 174–184.

Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data—or tears: An application to educational enrollments in states of india. *Demography*, *38*(1), 115–132.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, *78*(383), 553–569.

Friedman, J. H., & Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(4), 815–849.

Gallego, A., & Pardos-Prado, S. (2014). The big five personality traits and attitudes towards immigrants. *Journal of Ethnic and Migration Studies*, *40*(1), 79–99.

Gandolfi, F., & Stone, S. (2018). Leadership, leadership styles, and servant leadership. *Journal of Management Research*, *18*(4), 261–269.

Gharani, P., Ray, S., Aruru, M., & Pyne, S. (2021). Differential patterns of social media use associated with loneliness and health outcomes in selected socioeconomic groups. *Journal of technology in behavioral science*, 1–10.

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika*, *24*(3), 229–252.

Gil de Zuniga, H., & Diehl, T. (2017). Citizenship, social media, and big data: Current and future research in the social sciences. *Social Science Computer Review*, *35*(1), 3–9.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, *40*(1), 84–96.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, *24*, 395–419.

Groeneveld, P. W., & Rumsfeld, J. S. (2016). Can big data fulfill its promise? *Circulation: Cardiovascular Quality and Outcomes*, *9*(6), 679–682.

Grün, B., & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in r. *Computational Statistics & Data Analysis*, *51*(11), 5247–5252.

Gu, Z., de Schipper, N. C., & Van Deun, K. (2019). Variable selection in the regularized simultaneous component analysis method for multi-source data integration. *Scientific reports*, *9*(1), 1–21.

Gu, Z., & Van Deun, K. (2016). A variable selection method for simultaneous component based data integration. *Chemometrics and Intelligent Laboratory Systems*, *158*, 187–199.

Gu, Z., & Van Deun, K. (2019). Regularizedsca: Regularized simultaneous component analysis of multiblock data in r. *Behavior research methods*, *51*(5), 2268–2289.

Guerra-Urzola, R., Van Deun, K., Vera, J. C., & Sijtsma, K. (2021). A guide for sparse pca: Model comparison and applications. *psychometrika*, *86*(4), 893–919.

Gvaladze, S., Vervloet, M., Van Deun, K., Kiers, H. A., & Ceulemans, E. (2021). Pcovr2: A flexible principal covariates regression approach to parsimoniously handle multiple criterion variables. *Behavior Research Methods*, 1–21.

Hagemeyer, B., Dufner, M., & Denissen, J. J. (2016). Double dissociation between implicit and explicit affiliative motives: A closer look at socializing behavior in dyadic interactions. *Journal of Research in Personality*, *65*, 89–93.

Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological bulletin*, *131*(6), 898.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447.

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International journal of cross cultural management*, *6*(2), 243–266.

Hastie, T., Tibshirani, R., & Wainwright, M. (2019). *Statistical learning with sparsity: The lasso and generalizations*. Chapman; Hall/CRC.

Haven, S., & ten Berge, J. M. (1977). *Tucker's coefficient of congruence as a measure of factorial invariance: An empirical study*. Psychologische Instituten der Rijksuniversiteit Groningen.

Hayenga, A. O., & Corpus, J. H. (2010). Profiles of intrinsic and extrinsic motivations: A person-centered approach to motivation and achievement in middle school. *Motivation and Emotion*, *34*(4), 371–383.

Hehman, E., & Xie, S. Y. (2021). Doing better data visualization. *Advances in Methods and Practices in Psychological Science*, *4*(4), 25152459211045334.

Heij, C., Groenen, P. J., & van Dijk, D. (2007). Forecast comparison of principal component regression and principal covariate regression. *Computational statistics & data analysis*, *51*(7), 3612–3625.

Hellton, K. H., & Thoresen, M. (2016). Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics*, *17*(3), 537–548.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271.

Hennig, C. (2008). Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. *Journal of multivariate analysis*, *99*(6), 1154–1176.

Henry, D. B., Tolan, P. H., & Gorman-Smith, D. (2005). Cluster analysis in family psychology research. *Journal of Family Psychology*, *19*(1), 121.

Hinton, G. E., & Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1358), 1177–1190.

Hofferth, S. L., Moran, E. F., Entwisle, B., Aber, J. L., Brady, H. E., Conley, D., Cutter, S. L., Eckel, C. C., Hamilton, D., & Hubacek, K. (2017). Introduction: History and motivation.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2007). David bioinformatics resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, *35*(suppl_2), W169–W175.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193–218.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural equation modeling: a multidisciplinary journal*, *23*(4), 555–566.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, *31*(8), 651–666.

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of business research*, *70*, 338–345.

Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is romantic desire predictable? machine learning applied to initial romantic attraction. *Psychological science*, *28*(10), 1478–1489.

Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.

Josse, J., Pagès, J., & Husson, F. (2011). Multiple imputation in principal component analysis. *Advances in data analysis and classification*, *5*(3), 231–246.

Jung, T., & Wickrama, K. A. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and personality psychology compass*, *2*(1), 302–317.

Kaiser, H. F. (1959). Computer program for varimax rotation in factor analysis. *Educational and psychological measurement*, *19*(3), 413–420.

Kang, I., Yi, W., & Turner, B. M. (2021). A regularization method for linking brain and behavior. *Psychological Methods*.

Kiers, H. A., & Smilde, A. K. (2007). A comparison of various methods for multivariate regression with highly collinear variables. *Statistical Methods and Applications*, *16*(2), 193–228.

Kiers, H. A., & ten Berge, J. M. (1989). Alternating least squares algorithms for simultaneous components analysis with equal component weight matrices in two or more populations. *Psychometrika*, *54*(3), 467–473.

Kiers, H. A., & ten Berge, J. M. (1994). Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure. *British Journal of mathematical and statistical psychology*, *47*(1), 109–126.

Kolenikov, S., Angeles, G. et al. (2004). The use of discrete data in pca: Theory, simulations, and applications to socioeconomic indices. *Chapel Hill: Carolina Population Center, University of North Carolina*, *20*, 1–59.

Kondo, Y., Salibian-Barrera, M., & Zamar, R. (2016). Rskc: An r package for a robust and sparse k-means clustering algorithm. *Journal of Statistical Software*, *72*(1), 1–26.

Krzanowski, W. J., & Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 23–34.

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, *103*(9), 1449–1477.

Lebart, L., Morineau, A., & Piron, M. (1995). *Statistique exploratoire multidimensionnelle* (Vol. 3). Dunod Paris.

Li, X., & Jacobucci, R. (2021). Regularized structural equation modeling with stability selection. *Psychological Methods*.

Lindsten, F., Ohlsson, H., & Ljung, L. (2011). Clustering using sum-of-norms regularization: With application to particle filter output computation. *2011 IEEE Statistical Signal Processing Workshop (SSP)*, 201–204.

Linting, M., Meulman, J. J., Groenen, P. J., & van der Koojj, A. J. (2007). Nonlinear principal components analysis: Introduction and application. *Psychological methods*, *12*(3), 336.

Linting, M., & van der Kooij, A. (2012). Nonlinear principal components analysis with catpca: A tutorial. *Journal of personality assessment*, *94*(1), 12–25.

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, *7*(1), 523.

Lonigan, C. J., Goodrich, J. M., & Farver, J. M. (2018). Identifying differences in early literacy skills across subgroups of language-minority children: A latent profile analysis. *Developmental psychology*, *54*(4), 631.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., et al. (2012). Cluster: Cluster analysis basics and extensions. *R package version*, *1*(2), 56.

Magidson, J., & Vermunt, J. (2002). Latent class models for clustering: A comparison with k-means. *Canadian journal of marketing research*, *20*(1), 36–43.

Malan, L., Smuts, C. M., Baumgartner, J., & Ricci, C. (2020). Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. *Nutrition Research*, *75*, 67–76.

McCrae, R. R., & Costa Jr, P. T. (1989). Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, *57*(1), 17–40.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, *6*, 355–378.

McLachlan, G. J., Peel, D., & Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, *41*(3-4), 379–388.

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification*, *33*(3), 331–373.

McWilliams, B., & Montana, G. (2014). Subspace clustering of high-dimensional data: A predictive approach. *Data Mining and Knowledge Discovery*, *28*(3), 736–772.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473.

Meredith, W., & Millsap, R. E. (1985). On component analyses. *Psychometrika*, *50*(4), 495–507.

Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. *Clustering and classification* (pp. 341–375). World Scientific.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Morin, A. J., Bujacz, A., & Gagné, M. (2018). Person-centered methodologies in the organizational sciences: Introduction to the feature topic.

Mothi, S. S., Sudarshan, M., Tandon, N., Tamminga, C., Pearlson, G., Sweeney, J., Clementz, B., & Keshavan, M. S. (2019). Machine learning improved classification of psychoses using clinical and biological stratification: Update from the bipolar-schizophrenia network for intermediate phenotypes (b-snip). *Schizophrenia research*, *214*, 60.

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*(3), 474.

Mõttus, R., Sinick, J., Terracciano, A., Hřebičková, M., Kandler, C., Ando, J., Mortensen, E. L., Colodro-Conde, L., & Jang, K. L. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *117*(4), e35.

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., et al. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. *European Journal of Personality*, *34*(6), 1175–1201.

Nanjundeswaraswamy, T. S., & Swamy, D. R. (2014). Leadership styles. *Advances in management*, *7*(2), 57.

Neumann, C. S., Kaufman, S. B., ten Brinke, L., Yaden, D. B., Hyde, E., & Tsykayama, E. (2020). Light and dark trait subtypes of human personality–a multi-study person-centered approach. *Personality and Individual Differences*, *164*, 110121.

Nishimura, Y., Martin, C. L., Vazquez-Lopez, A., Spence, S. J., Alvarez-Retuerto, A. I., Sigman, M., Steindler, C., Pellegrini, S., Schanen, N. C., Warren, S. T., et al. (2007). Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Human molecular genetics*, *16*(14), 1682–1698.

Pan, W., & Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of machine learning research*, *8*(5).

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934.

Park, S., Ceulemans, E., & Van Deun, K. (2021). Sparse common and distinctive covariates regression. *Journal of Chemometrics*, *35*(2), e3270.

Pelckmans, K., De Brabanter, J., Suykens, J. A., & De Moor, B. (2005). Convex clustering shrinkage. *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*.

Postareff, L., Mattsson, M., Lindblom-Ylänne, S., & Hailikari, T. (2017). The complex relationship between emotions, approaches to learning, study success and study progress during the transition to university. *Higher education*, *73*(3), 441–457.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689–732.

Qiu, L., Chan, S. H. M., & Chan, D. (2018). Big data in social and psychological science: Theoretical and methodological issues. *Journal of Computational Social Science*, *1*(1), 59–66.

Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, *101*(473), 168–178.

Raymaekers, J., & Zamar, R. H. (2020). Regularized k-means through hard-thresholding. *arXiv preprint arXiv:2010.00950*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Sandri, B. J., Kaplan, A., Hodgson, S. W., Peterson, M., Avdulov, S., Higgins, L., Markowski, T., Yang, P., Limper, A. H., Griffin, T. J., et al. (2018). Multiomic molecular profiling of lung cancer in copd. *European Respiratory Journal*, *52*(1).

Scherpenzeel, A. C. (2018). ""true" longitudinal and probability-based internet panels: Evidence from the netherlands. *Social and behavioral research and the internet* (pp. 77–104). Routledge.

Schouteden, M., Van Deun, K., Pattyn, S., & Van Mechelen, I. (2013). Sca with rotation to distinguish common and distinctive information in linked data. *Behavior research methods*, *45*(3), 822–833.

Schouteden, M., Van Deun, K., Wilderjans, T. F., & Van Mechelen, I. (2014). Performing disco-sca to search for distinctive and common information in linked data. *Behavior research methods*, *46*(2), 576–587.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology* (pp. 1–65). Elsevier.

Serang, S., Jacobucci, R., Brimhall, K. C., & Grimm, K. J. (2017). Exploratory mediation analysis via regularization. *Structural equation modeling: a multidisciplinary journal*, *24*(5), 733–744.

Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, *99*(6), 1015–1034.

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., Ladanyi, M., & Sander, C. (2012). Integrative subtype discovery in glioblastoma using icluster. *PloS one*, *7*(4), e35236.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, *25*(22), 2906–2912.

Silva, B. C. (2018). Populist radical right parties and mass polarization in the netherlands. *European Political Science Review*, *10*(2), 219–244.

Smilde, A. K., Måge, I., Naes, T., Hankemeier, T., Lips, M. A., Kiers, H. A., Acar, E., & Bro, R. (2017). Common and distinct components in data fusion. *Journal of Chemometrics*, *31*(7), e2900.

Späth, H. (1979). Algorithm 39 clusterwise linear regression. *Computing*, *22*(4), 367–373.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., & Selbig, J. (2007). Pcamethods—a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, *23*(9), 1164–1167.

Steinley, D. (2004). Properties of the hubert-arable adjusted rand index. *Psychological methods*, *9*(3), 386.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1–34.

Steinley, D., & Brusco, M. J. (2008a). A new variable weighting and selection procedure for k-means cluster analysis. *Multivariate Behavioral Research*, *43*(1), 77–108.

Steinley, D., & Brusco, M. J. (2008b). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, *73*(1), 125.

Steinley, D., & Brusco, M. J. (2011a). Choosing the number of clusters in k-means clustering. *Psychological methods*, *16*(3), 285.

Steinley, D., & Brusco, M. J. (2011b). Evaluating mixture modeling for clustering: Recommendations and cautions. *Psychological Methods*, *16*(1), 63.

Stute, W., & Zhu, L. (1995). Asymptotics of k-means clustering based on projection pursuit. *Sankhyā: The Indian Journal of Statistics, Series A*, 462–471.

Sun, D., van Erp, T. G., Thompson, P. M., Bearden, C. E., Daley, M., Kushan, L., Hardt, M. E., Nuechterlein, K. H., Toga, A. W., & Cannon, T. D. (2009). Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: Classification analysis using probabilistic brain atlas and machine learning algorithms. *Biological psychiatry*, *66*(11), 1055–1060.

ten Berge, J. M. (1993). *Least squares optimization in multivariate analysis*. DSWO Press, Leiden University Leiden.

Tenenhaus, A., & Tenenhaus, M. (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of operational research*, *238*(2), 391–403.

Thorndike, R. L. (1953). Who belongs in the family. *Psychometrika*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511–528.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Timmerman, M. E., Ceulemans, E., De Roover, K., & Van Leeuwen, K. (2013). Subspace k-means clustering. *Behavior research methods*, *45*(4), 1011–1023.

Timmerman, M. E., & Kiers, H. A. (2003). Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika*, *68*(1), 105–121.

Tseng, G. C. (2007). Penalized and weighted k-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics*, *23*(17), 2247–2255.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (tech. rep.). Educational Testing Service Princeton Nj.

Ullmann, T., Hennig, C., & Boulesteix, A.-L. (2021). Validation of cluster analysis results on validation data: A systematic framework. *arXiv preprint arXiv:2103.01281*.

Van Deun, K., Smilde, A. K., Van Der Werf, M. J., Kiers, H. A., & Van Mechelen, I. (2009). A structured overview of simultaneous component based data integration. *Bmc Bioinformatics*, *10*(1), 1–15.

Van Deun, K., Wilderjans, T. F., Van Den Berg, R. A., Antoniadis, A., & Van Mechelen, I. (2011). A flexible framework for sparse simultaneous component based data integration. *BMC bioinformatics*, *12*(1), 1–17.

Vargo, C. J., & Hopp, T. (2017). Socioeconomic status, social capital, and partisan polarity as predictors of political incivility on twitter: A congressional district-level analysis. *Social Science Computer Review*, *35*(1), 10–32.

Vermunt, J. K. (2011). K-means may perform as well as mixture model clustering but may also be much worse: Comment on steinley and brusco (2011).

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis*, *11*(89-106), 60.

Vervloet, M., Kiers, H. A., Van den Noortgate, W., Ceulemans, E., et al. (2015). Pcovr: An r package for principal covariates regression. *Journal of Statistical Software*, *65*(8), 1–14.

Vervloet, M., Van Deun, K., Van den Noortgate, W., & Ceulemans, E. (2016). Model selection in principal covariates regression. *Chemometrics and Intelligent Laboratory Systems*, *151*, 26–33.

Vichi, M., & Kiers, H. A. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, *37*(1), 49–64.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological science in the public interest*, *11*(3), 89–121.

Waldherr, A., Maier, D., Miltner, P., & Günther, E. (2017). Big data, big noise: The challenge of finding issue networks on the web. *Social Science Computer Review*, *35*(4), 427–443.

Waldman, D. A., Wang, D., & Fenters, V. (2019). The added value of neuroscience methods in organizational research. *Organizational Research Methods*, *22*(1), 223–249.

Wang, B., Zhang, Y., Sun, W. W., & Fang, Y. (2018). Sparse convex clustering. *Journal of Computational and Graphical Statistics*, *27*(2), 393–403.

Wang, D., Ding, C., & Li, T. (2009). K-subspace clustering. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 506–521.

Wang, D., & Gu, J. (2016). Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quantitative Biology*, *4*(1), 58–67.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, *97*(4), 893–904.

Wang, M., & Hanges, P. J. (2011). Latent class procedures: Applications to organizational research. *Organizational Research Methods*, *14*(1), 24–31.

Wang, S., & Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics*, *64*(2), 440–448.

Wehrens, R., & Salek, R. (2019). *Metabolomics: Practical guide to design and analysis*. CRC Press.

Wells, C., & Thorson, K. (2017). Combining big data and survey techniques to model effects of political content flows in facebook. *Social Science Computer Review*, *35*(1), 33–52.

Wilderjans, T. F., & Ceulemans, E. (2013). Clusterwise parafac to identify heterogeneity in three-way data. *Chemometrics and Intelligent Laboratory Systems*, *129*, 87–97.

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). Chull: A generic convex-hull-based model selection method. *Behavior research methods*, *45*(1), 1–15.

Wilderjans, T. F., Gaer, E. V., Kiers, H. A., Van Mechelen, I., & Ceulemans, E. (2017). Principal covariates clusterwise regression (pccr): Accounting for multicollinearity and population heterogeneity in hierarchically organized data. *psychometrika*, *82*(1), 86–111.

Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, *105*(490), 713–726.

Wold, S. (1984). Three pls algorithms according to sw. *Report from the symposium MULTDAST (multivariate analysis in science and methodology)*, 26–30.

Wood, M. J., & Gray, D. (2019). Right-wing authoritarianism as a predictor of pro-establishment versus anti-establishment conspiracy theories. *Personality and Individual Differences*, *138*, 163–166.

Wu, D., Wang, D., Zhang, M. Q., & Gu, J. (2015). Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: Application to cancer molecular classification. *BMC genomics*, *16*(1), 1–10.

Xu, Q., Ding, C., Liu, J., & Luo, B. (2015). Pca-guided search for k-means. *Pattern Recognition Letters*, *54*, 50–55.

Yamashita, N., & Adachi, K. (2020). A modified k-means clustering procedure for obtaining a cardinality-constrained centroid matrix. *Journal of Classification*, *37*(2), 509–525.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Yu, Q., Risk, B. B., Zhang, K., & Marron, J. (2017). Jive integration of imaging and behavioral data. *NeuroImage*, *152*, 38–49.

Yuan, S., De Roover, K., Dufner, M., Denissen, J. J., & Van Deun, K. (2021). Revealing subgroups that differ in common and distinctive variation in multiblock data: Clusterwise sparse simultaneous component analysis. *Social Science Computer Review*, *39*(5), 802–820.

Yuan, S., De Roover, K., & Van Deun, K. (2022). Simultaneous clustering and variable selection: A novel algorithm and model selection procedure. *Behavior Research Methods*.

Yuan, S., Kroon, B., & Kramer, A. (2021). Building prediction models with grouped data: A case study on the prediction of turnover intention. *Human Resource Management Journal*.

Zhang, Y., Wu, W., Toll, R. T., Naparstek, S., Maron-Katz, A., Watts, M., Gordon, J., Jeong, J., Astolfi, L., Shpigel, E., et al. (2021). Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography. *Nature biomedical engineering*, *5*(4), 309–323.

Zhang, Z., Dai, G., & Jordan, M. I. (2009). A flexible and efficient algorithm for regularized fisher discriminant analysis. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 632–647.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, *15*(2), 265–286.

SUMMARY

Large-scale data sets with a large number of variables become increasingly available in behavioral research. Encompassing a wide range of measurements and indicators, they provide behavioral scientists with unprecedented opportunities to synthesize different pieces of information so that novel - and sometimes subtle - clusters can be identified and recovered. Furthermore, because the wide range of variables may be derived from different data sources (e.g., self-reports, genetic data, brain signals, etc.), novel clusters can also emerge from the joint effects of the extensive and diverse set of variables. Clearly, the ever-increasing size and complexity of behavioral data sets is a treasure trove for behavioral scientists working on revealing crucial heterogeneity in the population.

However, just like finding jewelry in a treasure trove is never an effortless task, to accurately recover the complex cluster structures hidden in these large data sets, two major challenges should be overcome. First, a considerable number of variables may be completely irrelevant to the hidden cluster structure. These irrelevant variables may hinder the successful detection and recovery of hidden clusters and complicate the interpretation of cluster structures. Therefore, they have to be completely filtered out during the data analysis process. Second, when integrating variables from diverse data sources, it is always desirable to discern between variable covariations underlying all data sources (defined as the common variation) and those underlying a single or only a few data sources (defined as the distinctive variations). Only by completely disentangling these two types of covariations can we obtain a precise and direct understanding of which joint and (or) individual forces effectively differentiate the clusters.

This dissertation developed new statistical models and computational tools to address the two challenges mentioned above, and in doing so, offers new opportunities to detect complex cluster structures and better understand heterogeneity in behavioral studies.

Chapter 2 developed a new method called Cardinality K-means or CKM. CKM partitions observations based on the between-cluster mean structures and therefore enjoys excellent computational efficiency, even in the presence of a large number of variables. More importantly, CKM finds hidden clusters while automatically detecting a set of variables best separating these clusters. Therefore, CKM can be particularly useful for exploratory studies aimed at detecting clusters in novel types of data (e.g., GPS data, transaction data), which, most often than not, contain a large number of variables with little theoretical guidance on the relevance of these variables. The solid performance of CKM in terms of cluster recovery and variable selection was confirmed through three simulation studies and CKM has proven to outperform a suite of competing methods. This chapter also proposed a novel model selection procedure that determined the number of clusters based on a subset of variables that are constantly classified as crucial variables for cluster separation. This strategy outperformed the traditional approach of determining the number of clusters with all variables. An R package `CKM` was designed that implemented CKM and the novel model selection strategy.

Chapter 3 extended the work in Chapter 2 by offering a detailed and accessible guide for researchers to use simultaneous clustering and variable selection (SCVS) methods in their research. Specifically, this chapter elaborated on five key steps that have to be taken for the SCVS techniques and presented readers with an empirical example of clustering subjects based on their political attitudes. The five steps can be classified into three parts: data preprocessing, cluster analysis, and cluster validation. A unique contribution of this chapter was a comprehensive discussion and demonstration of cluster validation. Three types of cluster validation approaches were addressed in the chapter, namely visual validation, cluster stability validation, and cluster replication. Last but not least, to facilitate visual validation of cluster results, this chapter offered a Shinyapp that can generate many types of visualizations with minimal user input.

Chapter 4 developed Clusterwise Sparse Simultaneous Component Analysis (CSSCA), a method that performs automatic variable selection and distinguishes between common and distinctive (co)variations when dealing with variables coming from a wide variety of data sources. More specifically, building

on a few previous methods, CSSCA defines clusters in such a way that only the observations from the same cluster have the same between-cluster mean structure and within-cluster component structure. In two simulation studies, CSSCA recovered hidden clusters and identified the associated cluster-specific component structures with high accuracy. Furthermore, CSSCA outperformed a popular competing method (i.e., iCluster), especially when a large proportion of cluster differences were attributed to differences in these within-cluster component structures. Finally, CSSCA proved useful in an illustrative example where different measures of personality were connected. To better disseminate the CSSCA method, an R package `ClusterSCA` was developed.

Chapter 5 extended Chapter 4 by considering predictive analysis and developed a new method called Clusterwise Simultaneous Component Regression or CSCR. CSCR was developed with the dual goal of predicting outcomes as accurately as possible while simultaneously offering a clear interpretation of which within-cluster components are important to prediction. In other words, CSCR extracts components with the guidance of the outcomes, thus ensuring that these components are not merely good summaries of predictors but also useful ingredients for prediction. As far as we know, CSCR is the only method that combines two desirable features of prediction: (1) for each cluster, the dimensions of the predictors are largely reduced to prevent over-fitting, and (2) regression models are estimated per cluster to account for heterogeneity. Like CSSCA, CSCR can handle a diverse set of variables from different sources, thanks to its capability to discern common and distinctive (co)variations. Two simulation studies reported in this chapter demonstrated the excellent performance of CSCR in terms of cluster recovery and predictive accuracy. Notably, the performances of CSCR were considerably better than the performances of other competing methods across a large number of conditions. Finally, in the illustration where participants' attitudes toward immigrants were inferred from their personalities and values, the application of CSCR resulted in additional insights into the different predictive mechanisms for different clusters of observations. In addition, an R package `CSCR` was developed for the estimation with CSCR.

Concluding, in Chapter 6, we presented some thoughts on the practical applications of the methods developed in this dissertation and shed light on future research directions.

List of Publications

1. **Yuan, S.**, De Roover, K., Dufner, M., Denissen, J. J., & Van Deun, K. (2021). Revealing subgroups that differ in common and distinctive variation in multi-block data: Clusterwise sparse simultaneous component analysis *Social Science Computer Review*, 39(5), 802-820. https://doi.org/10.1177/0894439319888449

2. **Yuan, S.**, De Roover, K., & Van Deun, K. (in press). Simultaneous Clustering and Variable Selection: a Novel Algorithm and Model Selection Procedure. *Behavior Research Methods* https://doi.org/10.3758/s13428-022-01795-7

3. **Yuan, S.**, Kroon, B., & Kramer, A. (2021). Building prediction models with grouped data: A case study on the prediction of turnover intention. *Human Resource Management Journal* https://doi.org/10.1111/1748-8583.12396

4. **Yuan, S.**, De Roover, K., & Van Deun, K. (Resubmitted). Clusterwise Simultaneous Covariates Regression: A Novel Method that Balances Prediction and Interpretation with Hidden Subgroups. *Behavior Research Methods*

5. **Yuan, S.**, De Roover, K., Jaime Hermsdorf, & Van Deun, K. (Revise and resubmit). A Tutorial on Simultaneous Clustering and Variable Selection *Advances in Methods and Practices in Psychological Science*

6. Knappert, L., Van Dijk, H., **Yuan, S.**, Engel, Y., van Prooijen, J. W., & Krouwel, A. (2021). Personal contact with refugees is key to welcoming them: An analysis of politicians' and citizens' attitudes towards refugee integration. *Political Psychology*, 42(3), 423-442. https://doi.org/10.1111/pops.12705

7. Dong, M., Spadaro, G., **Yuan, S.**, Song, Y., Ye, Z., & Ren, X. (2021). Self-Interest Bias in the COVID-19 Pandemic: A Cross-Cultural Comparison between the United States and China. *Journal of Cross-Cultural Psychology* https://doi.org/10.1177/00220221211025739

8. Zhang, Z., Yao, X., **Yuan, S.**, Deng, Y., & Guo, C. (2021). Big five personality influences trajectories of information seeking behavior. *Personality and Individual Differences*, 173, 110631. https://doi.org/10.1016/j.paid.2021.110631