THE UNIVERSITY OF NIGERIA

POSTGRADUATE PROGRAMME

DEPARTMENT OF STATISTICS

SELECTION OF VARIABLES FOR CLUSTER ANALYSIS WITH APPLICATION TO DHS DATA IN CAMEROON

A Proposal submitted to the Department of statistics in partial fulfillment of the requirements for the award of a PhD in Applied Statistics.

By

AYENDOH TERRENCE SAMA

Reg. No. PG/PHD/23/97892

(M.Sc. in Probability and Statistics)

SUPERVISOR
Dr. M S Madukaife

CO-SUPERVISOR
Dr. E O Ossai

November, 2025

Contents

List of Figures

List of Tables

List of Symbols

$X = (X_1, \ldots, X_p)$	Random vector of p variables
p	Total number of variables
n	Sample size (number of observations)
K	Number of clusters
$f : \mathbb{R}^p \to \{1, \ldots, K\}$	Population partition function (cluster allocation rule)
G_k	k-th cluster region: $G_k = f^{-1}(k)$
$I \subset \{1, \ldots, p\}$	Index set of selected variables
d	Cardinality of the selected subset I ($d < p$)
Y^I	"Blinded" vector: $Y_i^I = X_i$ if $i \in I$, else $E(X_i)$
Z^I	Conditional "blinded" vector: $Z_i^I = X_i$ if $i \in I$, else $E(X_i \mid X_I)$
$h(I)$	Population objective function: fraction of points with unchanged cluster labels when using I
$h_n(I)$	Empirical version of $h(I)$ computed from data
X_j^*	Observation j with blinded variables replaced
$\bar{X}[i]$	Sample mean of variable i
m_n	Size of smallest cluster in a sample
r	Number of nearest neighbors in conditional mean estimation
∂G_k	Boundary of cluster region G_k
$d(x, A)$	Distance from point x to set A
$\mathbb{I}\{\cdot\}$	Indicator function
$\mathbb{P}(\cdot)$	Probability measure
$\mathbb{E}(\cdot)$	Expectation operator

List of Abbreviations

A.M.S.	American Mathematical Society
BIC	Bayesian Information Criterion
k-NN	k-Nearest Neighbors
MCMC	Markov Chain Monte Carlo
NN	Nearest Neighbor(s)
PCA	Principal Component Analysis
TSV05	Simulated data example from Tadesse, Sha, & Vannucci (2005)

Abstract

abstract High-dimensional data presents a significant challenge for cluster analysis, as the presence of noisy and redundant variables can obscure the underlying group structure and degrade the quality of results. This proposal introduces a variable selection framework designed to identify a parsimonious and informative subset of features for clustering. The methodology is based on a "blinding" principle, where the importance of a variable is assessed by measuring the distortion to the cluster partition when its influence is neutralized.

We develop two complementary strategies: one using marginal means to identify and remove non-informative "noisy" variables, and another using conditional means to address multicollinearity by removing redundant variables. To ensure computational feasibility in large datasets, these procedures are embedded within a forward-backward search algorithm. A key contribution of this work is the adaptation of this framework for complex, mixed-type, and weighted survey data, with a specific application to the Demographic and Health Surveys (DHS) for Cameroon.

The proposed methods will be validated through simulation studies and compared against standard techniques like sparse k-means and hierarchical clustering. The expected outcome is a statistically robust, scalable, and interpretable variable selection procedure that improves clustering accuracy and provides clearer insights. For reproducibility and broader use, the final methodology will be implemented as an R package.

Keywords: Cluster Analysis, Variable Selection, High-Dimensional Data, Forward-Backward Algorithm, DHS Data. abstract

Chapter 1

Introduction

1.1 Background

Recent technological developments in big data have made it fairly easy to collect data with high volume, velocity, variety, veracity, value, and variability. Thus these datasets contain a large number of variables within a single study, "large" used here subjectively. Examples of studies on these datasets include:- examinations of genetic influences in organizational psychology (e.g., Chi et al., 2016; Arvey 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., 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 and Meyer-Lindenberg, 2018). A noteworthy advantage of high-dimensional datasets 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 application. In behavioral sciences, one can think of data sets with more than 100 variables (Groeneveld and 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. Thus "high-dimensional" datasets are a case where number of variables (p) exceeds the number of observations (n), i.e. $p > n$. When datasets are high-dimensional, they often contain variables that are either irrelevant, redundant, or contaminated with noise. These non-informative features can mask the true cluster structure, degrade the quality of partitions, and make interpretation difficult.

In the context of cluster analysis – where the intent is to group observations in such a way that those in the same subgroup are similar to each other, using high-dimensional data will likely result in a more accurate estimation of subgroups and (or) a discovery of novel subgroups. In one of the very few reported attempts to cluster high-dimensional datasets, Mothi et al. (2019) combined clinical measures, Behavior Research Methods 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 subtypes of psychoses. Evidently, clustering high-dimensional datasets grants an opportunity to clarify and deepen our understanding of the heterogeneity and true underlying structure of the phenomena in question.

Although research that exploits high-dimensional datasets to identify subgroups is promising, it also comes with challenges. One of the most compelling challenges, as stressed by a number of scholars (e.g., Yarkoni and Westfall, 2017; Waldherr et al., 2017; Bzdok and Meyer-Lindenberg, 2018), is that these data sets may comprise a large amount of "irrelevant variables" (Fowlkes and Mallows, 1983). They are variables that do not separate clusters well and therefore do not define cluster structure. These irrelevant variables may hinder subgroup discovery by masking the cluster structure under investigation (Steinley and Brusco, 2008). Therefore, a cluster analysis should effectively recover the cluster structure while simultaneously filtering out irrelevant variables.

Traditional clustering methods such as hierarchical clustering, k-means, and k-medoids operate under the implicit assumption that all variables contribute equally to the definition of clusters. In reality, certain variables may carry no meaningful discriminatory power, while others may be strongly correlated with more informative ones, introducing multicollinearity. This redundancy can distort the clustering space by overemphasizing specific dimensions and inflating the apparent significance of certain patterns. The challenge, therefore, lies in identifying a subset of variables that adequately explains the cluster structure, while discarding those that add little or no value to the partition.

Dimension reduction techniques like Principal Component Analysis (PCA) have been employed to address high dimensionality, but these approaches create linear combinations of variables that can be difficult to interpret in substantive terms. Variable selection methods, by contrast, aim to retain a subset of the original variables, preserving interpretability while improving clustering performance.

In this study, we propose a pair of statistical procedures specifically suited for variable selection in clustering and classification. The first method targets the identification of "noisy" non-informative variables by substituting their values with a global measure such as the marginal mean, effectively "blinding" their influence. The second method extends this concept to address multicollinearity by replacing a variable's values with conditional expectations based on the selected subset, thereby preserving local dependency structures while removing redundant information. Both procedures are grounded in the principle that a good subset of variables is one that preserves, as closely as possible, the cluster allocations obtained using the full set of variables.

This work is organized as follows:- We present a synthesis of the body of work done on the problem of variable selection in clustering, where we point out the most notable studies along with gaps that we intend to fill.

We also discuss the proposed procedure for solving the problem of variable selection in clustering in the methodology section. We also present the dataset which we will use for simulation in this section. Lastly, we outline the results we expect to obtain by the end of this study, in the last section.

Chapter 2

Statement of the Problem

Cluster analysis has become a cornerstone of data analysis, providing a means to uncover latent structures and natural groupings in complex datasets. Its utility spans multiple disciplines, including biology, social sciences, marketing, image processing, and bioinformatics. Yet, despite its broad applicability, the reliability and interpretability of clustering results are often compromised when the datasets under investigation contain a large number of variables, many of which may be irrelevant, redundant, or noisy. In such high-dimensional data, the inclusion of non-informative variables can obscure the underlying structure of the data, distort similarity measures, and lead to unstable or misleading clustering solutions. This problem is compounded by the well-known "curse of dimensionality", whereby distances between observations become less discriminative as dimensionality increases, causing clusters to appear less distinct and making the task of identifying meaningful partitions substantially more challenging.

The challenge is not merely one of computational efficiency, although high dimensionality does indeed increase the computational burden of most clustering algorithms. Rather, the core of the problem lies in the fact that many variables contribute little or no meaningful information to the definition of clusters, while others may actively mislead the clustering process. In practice, different subsets of variables may hold relevance for different cluster structures within the same dataset, and methods that treat all variables equally risk diluting or entirely masking those structures. Furthermore, redundancy among variables can create implicit weighting effects, whereby certain dimensions exert disproportionate influence on the resulting clusters, not because they contain more information, but because their repeated patterns amplify their perceived importance.

The difficulty is exacerbated by the absence of external supervision in cluster analysis. Unlike in supervised learning, where model performance can be measured against known labels, clustering typically operates without a predefined ground truth, making it far less straightforward to evaluate the role and significance of each variable. Traditional model selection criteria such as the Bayesian Information Criterion (BIC) or the Gap Statistic, while widely used for determining the optimal number of clusters, are themselves sensitive to the presence of irrelevant or redundant variables. As a result, even when advanced model-based or non-parametric clustering algorithms are applied, the inclusion of extraneous features can lead to both overestimation and underestimation of the number of clusters, as well as a degradation in the clarity of the discovered structures.

Over the years, numerous methods have been proposed to address the problem of variable selection for clustering, ranging from forward and backward selection strategies to penalised likelihood approaches, sparse clustering frameworks, and Bayesian variable selection models. While these methods represent important progress, they remain constrained by limitations related to scalability, assumptions about data distribution, or their ability to detect locally relevant variables. There is still no universally accepted method that balances accuracy, interpretability, and computational feasibility, particularly for large and complex datasets where variable relevance may vary across subspaces of the data.

The problem is thus twofold: first, to define in precise terms what constitutes an "informative" variable for clustering in both global and local contexts; and second, to design a robust and computationally efficient method for identifying such variables without prior knowledge of the true cluster structure. Existing methods, while valuable, are constrained by assumptions about data distribution, sensitivity to noise, and limited scalability. Moreover, most approaches fail to adequately capture locally relevant variables—those that are essential for identifying clusters in subspaces of the data but not across the dataset as a whole.

Addressing this problem is of both theoretical and practical importance. A solution would not only improve the accuracy and stability of clustering results, but also enhance their interpretability for domain experts, reduce computational costs, and increase the reliability of decisions derived from clustered data. This study seeks to address this gap by developing a robust procedure for variable selection in clustering that balances accuracy, interpretability, and scalability, while also accounting for local variations in variable relevance within high-dimensional datasets.

Chapter 3

Aim and Objectives of the Study

The problem of variable selection in cluster analysis remains both practically important and methodologically challenging. This research aims to develop a procedure for variable selection for clustering of high-dimensional data like the DHS data, that achieves a balance between computational efficiency and selection accuracy. Then we empirically validate the proposed approach on real-world datasets, like the DHS dataset for Cameroon, demonstrating improvements in interpretability, robustness, and clustering performance when compared with other standard methods like k-means and hierarchical clustering techniques. This will contribute to more accurate, interpretable, and computationally feasible clustering solutions in high-dimensional data.

Main Objective

The main objective of this study is *to develop a procedure which effectively and efficiently identifies and selects variables that truly contribute to the underlying cluster structure—while filtering out irrelevant or misleading features—to improve clustering accuracy, interpretability, and computational feasibility*

Specific Objectives

1: To review existing variable selection methods for clustering and identify limitations when applied to high dimensional data.

2: To develop variable selection strategies suitable for mixed-type, survey-weighted DHS variables.

3: To evaluate proposed methods through simulation and application to one or more Cameroon DHS datasets, as well as compare it with standard clustering procedures such as k-means and hierarchical clustering algorithms.

4: To develop an implementation of the proposed method as an R package for reproducibility and ease of use.

Chapter 4

Significance of the Study

The problem of selecting relevant variables in cluster analysis is of both theoretical and practical importance in the era of high-dimensional data. As data collection technologies become more advanced and accessible, researchers in fields as diverse as genomics, finance, education, and energy analytics are confronted with datasets containing hundreds or thousands of variables. In such contexts, the presence of irrelevant, noisy, or redundant variables not only undermines the accuracy of clustering procedures but also obscures the interpretability of the resulting group structures. The ability to identify a parsimonious subset of variables that retains the essential information for clustering is therefore a crucial step toward producing reliable, interpretable, and computationally efficient results.

The methodology to be employed in this study offers an elegant and statistically grounded approach to this challenge. By "blinding" non-informative variables using marginal or conditional means, it becomes possible to directly assess their contribution to the clustering process, without relying on arbitrary heuristics or black-box transformations. The conditional mean extension further addresses the pervasive issue of multicollinearity, allowing the procedure to disentangle redundancy from genuine variable importance. The integration of a forward–backward search algorithm ensures that these methods remain feasible and effective even in large-scale applications where exhaustive search would be computationally prohibitive.

The significance of this research extends beyond methodological innovation. From a practical standpoint, effective variable selection enhances the interpretability of cluster solutions, enabling domain experts to link clusters to meaningful real-world constructs. In applied settings such as healthcare diagnostics, market segmentation, and educational assessment, this interpretability can inform targeted interventions, policy decisions, and strategic planning. Furthermore, by reducing dimensionality without sacrificing essential information, the proposed approach lowers computational costs, making sophisticated clustering analyses more accessible to practitioners with limited computational resources.

From a scientific perspective, the study contributes to bridging the gap between theoretical advances in statistical methodology and their application to real-world problems. It demonstrates that rigorous, statistically consistent procedures can be adapted into tools that are both usable and insightful for practitioners across disciplines. Ultimately, the outcomes of this work have the potential to influence best practices in unsupervised learning, encouraging the adoption of variable selection as a standard step in high-dimensional clustering workflows.

Chapter 5

Scope of the Study

This study focuses on the development, adaptation, and evaluation of statistical procedures for variable selection in cluster analysis. Specifically, the research will examine two complementary approaches: the marginal mean "blinding" method, designed to detect and remove non-informative noisy variables, and the conditional mean method, aimed at addressing multicollinearity and redundancy. These methods will be integrated with a forward–backward search algorithm to ensure computational feasibility in high-dimensional contexts. The primary application domain will be on high-dimensional datasets like DHS data obtained from the National Institute of Statistics in Cameroon. This will enable the demonstration of the methods' generality and robustness.

Simulated data may be used in this study. Performance will be evaluated in terms of classification agreement with the full-variable clustering, reduction in dimensionality, computational cost, and interpretability of the resulting variable subsets.

However, several limitations should be acknowledged. First, the methods under consideration depend on the initial clustering obtained from the full set of variables. If the initial clustering is poor due to inappropriate choice of clustering algorithm, distance measure, or number of clusters, the variable selection process may propagate these deficiencies. Second, while the conditional mean method addresses multicollinearity, it requires reliable estimation of conditional expectations, which in turn demands sufficiently large sample sizes; performance may deteriorate when the sample size is small relative to the number of variables. Third, the forward–backward search algorithm, though more efficient than exhaustive search, may still be computationally intensive for extremely high-dimensional data, particularly when coupled with nonparametric conditional estimation. Finally, the evaluation of variable importance in clustering inherently lacks a ground truth in unsupervised settings, meaning that conclusions must often be drawn from indirect measures such as agreement indices or stability analyses.

Within these parameters, the study aims to provide a rigorous assessment and practical adaptation of variable selection methods for clustering, contributing both to methodological development and to the growing need for interpretable and efficient unsupervised learning in high-dimensional data analysis.

Chapter 6

Literature Review

6.1 The Challenge of Variable Selection in Clustering

The proliferation of high-dimensional datasets across diverse scientific disciplines presents both unprecedented opportunities and significant analytical challenges. In fields ranging from genomics and bioinformatics to social sciences and marketing, researchers are increasingly confronted with data characterized by a large number of variables (p) relative to the number of observations (n). While such rich datasets offer the potential for deeper insights, they simultaneously introduce complexities that can severely undermine the performance of traditional statistical methods, particularly in unsupervised learning tasks like cluster analysis.

The core of this issue is often referred to as the "curse of dimensionality," a term first coined by Bellman (1961) in the context of dynamic programming. Bellman observed that the volume of a data space expands exponentially as the number of dimensions increases, causing the data to become increasingly sparse. In clustering, this phenomenon has profound and detrimental consequences. As dimensionality grows, the distance between any two points in a high-dimensional space tends to become almost indistinguishable from the distance between any other two points. This concentration of distances undermines the effectiveness of distance-based clustering algorithms (e.g., k-means, hierarchical clustering), which fundamentally rely on proximity measures to form groups (Hartigan, 1975).

Furthermore, high-dimensional datasets rarely consist solely of informative variables. Instead, they are typically a mixture of variables with distinct roles, which must be properly identified. Following the foundational work by John et al. (1994), these roles can be formally categorized as:

- **Relevant Variables:** Variables that are essential for describing the underlying cluster structure. Their removal would degrade the quality of the resulting partition.

- **Redundant Variables:** Variables that, while potentially relevant, provide information that is already captured by one or more other relevant variables. They are conditionally independent of the cluster partition given the set of relevant variables. Including them can over-represent certain features and distort the clustering solution.

- **Irrelevant (Noise) Variables:** Variables that contain no information about the cluster structure. These variables can obscure the true groupings, degrade the performance of clustering algorithms, and complicate the interpretation of results (Steinley and Brusco, 2008).

The primary challenge of variable selection in cluster analysis is, therefore, to develop a systematic and computationally efficient procedure to distinguish between these variable types, retaining the relevant ones while discarding the irrelevant and redundant ones. An effective selection strategy is paramount for improving cluster accuracy, enhancing the interpretability of the results, and ensuring computational feasibility.

6.2 A Taxonomy of Variable Selection Methods

To address the challenge of variable selection, a wide array of methods has been developed. A comprehensive and widely adopted classification system organizes these techniques into three main families based on how they interact with the clustering algorithm: **Filter**, **Wrapper**, and **Embedded** methods (Guyon and Elisseeff, 2003).

- **Filter Methods:** These methods function as a preprocessing step, selecting variables independently of the chosen clustering algorithm. They rely on intrinsic characteristics of the data to rank or score variables, using statistical measures such as variance, entropy, or correlation. For example, a simple filter might remove variables with very low variance, under the assumption that they are unlikely to contribute to cluster separation. While computationally efficient and algorithm-agnostic, filter methods ignore the potential interactions between variables and the specific objective function of the clustering algorithm, which may lead to the selection of a suboptimal subset of variables.

- **Wrapper Methods:** In this approach, the clustering algorithm is treated as a "black box," and its performance is used to evaluate the quality of different candidate subsets of variables. A search strategy (e.g., forward selection, backward elimination, or a genetic algorithm) is used to navigate the space of all possible variable subsets. For each subset, the clustering algorithm is executed, and the quality of the resulting partition is assessed using a performance metric, such as an internal validation index (e.g., Silhouette score) or a criterion like the Bayesian Information Criterion (BIC) (Dean and Raftery, 2006). Although wrapper methods are often more accurate because they directly optimize for the performance of a specific clustering algorithm, they are computationally intensive and carry a risk of overfitting to the data.

- **Embedded Methods:** These methods integrate the variable selection process directly into the training of the clustering model itself. This is typically achieved by adding a penalty term to the clustering objective function that penalizes model complexity. This penalty encourages sparsity by shrinking the contributions of less important variables toward zero, effectively performing variable selection and clustering simultaneously. Prominent examples include sparse k-means (Witten and Tibshirani, 2010) and sparse mixture models. Embedded methods offer a compromise between the efficiency of filters and the performance of wrappers, as they account for variable interactions and the clustering objective in a computationally feasible manner.

This taxonomy provides a crucial framework for navigating the vast literature on variable selection and for situating the contribution of any new proposed methodology. The following sections will delve deeper into specific examples from these families, critically evaluating their strengths and weaknesses.

6.3 Wrapper Methods: Evaluating Subsets

Wrapper methods epitomize the principle of optimizing variable selection for a specific learning task. In the context of clustering, this means the clustering algorithm itself is used as the core component of the evaluation function to score candidate subsets of variables. The process is iterative: a search algorithm proposes a subset of variables, the clustering algorithm is run on this subset, and the quality of the resulting partition is evaluated using a predefined criterion. This score guides the search toward more promising subsets.

The primary advantage of the wrapper approach is its directness: it assesses variable subsets based on how well they perform for the ultimate clustering objective, thereby accounting for the interaction between the feature subset and the learning algorithm. The search strategy is a critical component, as an exhaustive search of all $2^p - 1$

possible subsets is computationally intractable for even a moderate number of variables p. Common search heuristics include:

- **Forward Selection:** Starting with an empty set, variables are iteratively added one at a time, selecting the variable that produces the greatest improvement in the clustering criterion at each step.
- **Backward Elimination:** Starting with the full set of variables, the algorithm iteratively removes the variable whose removal results in the smallest degradation (or largest improvement) of the clustering criterion.
- **Stepwise Selection:** A combination of forward and backward steps, which allows for both adding and removing variables at different stages to escape local optima.

A prominent and influential example of the wrapper method in a model-based clustering context is the work of Dean and Raftery (2006). They proposed a method for selecting variables in Gaussian mixture models where the goal is to find the subset of variables that leads to the best clustering, as measured by the Bayesian Information Criterion (BIC). Their algorithm performs a forward-backward (stepwise) search. At each step, it considers either adding a variable to the current subset or removing one. The change that results in the largest increase in the BIC (which balances model fit with complexity) is accepted, and the process continues until no change can improve the BIC. This approach directly links variable selection to the objective of finding a parsimonious, well-fitting statistical model of the data's cluster structure.

While originally formulated for model-based clustering, the wrapper principle is broadly applicable. Similar strategies have been adapted for partitional algorithms like k-means, where the evaluation criterion is often an internal validation index (e.g., average silhouette width) or a measure of cluster stability (Steinley and Brusco, 2008).

Critical Analysis of Wrapper Methods

The main strength of wrapper methods is their potential for high performance. By customizing the variable subset to the specific biases and assumptions of the chosen clustering algorithm, they are more likely to find a subset that yields a better-quality partition compared to filter methods.

However, this approach has significant drawbacks. The most notable is its **high computational cost**. Because the clustering algorithm must be run for every single subset evaluated, the process can be extremely slow, especially with a large number of variables or a computationally expensive clustering algorithm. Furthermore, there is a considerable **risk of overfitting**. The method may select a variable subset that performs exceptionally well on the given data and the chosen evaluation criterion but does not generalize well to new data. The search process is also susceptible to getting trapped in **local optima**, meaning the final selected subset may be highly dependent on the starting point and the search heuristic, with no guarantee of finding the globally optimal subset. Despite these limitations, wrapper methods remain a powerful and conceptually straightforward approach to variable selection in clustering.

6.4 Embedded Methods: Integrating Selection and Clustering

In contrast to the computationally intensive nature of wrapper methods, embedded methods offer a more integrated and efficient solution by incorporating the variable selection process directly into the model-fitting procedure. This approach is particularly powerful in high-dimensional settings, as it simultaneously performs clustering and identifies the most informative variables in a single, unified framework. Embedded techniques typically achieve this by introducing a penalty term into the clustering objective function, which penalizes model complexity and encourages sparsity. This forces the contributions of irrelevant variables toward zero, effectively selecting a parsimonious set of features that best explains the cluster structure.

6.4.1 Penalized and Sparse Methods

A major class of embedded methods is based on penalization, drawing inspiration from successful techniques in supervised learning like the Lasso (Least Absolute Shrinkage and Selection Operator). These methods modify the objective function of a clustering algorithm (e.g., the within-cluster sum of squares in k-means or the log-likelihood in model-based clustering) by adding a penalty proportional to the size of the variable weights or contributions.

A seminal work in this area is the sparse k-means algorithm proposed by Witten and Tibshirani (2010). Their method adds an L_1 penalty to the standard k-means objective function, which has the effect of making the feature weights sparse. By tuning a penalty parameter, the algorithm can select a subset of variables that contribute to the clustering, while giving a weight of exactly zero to all other variables. This results in a more interpretable clustering where groups are defined only by the selected features.

In the context of model-based clustering, Pan and Shen (2007) introduced a penalized likelihood approach for Gaussian mixture models. They added a penalty term to the log-likelihood function that penalizes the estimation of the mean parameters. By shrinking some of the mean parameters to be equal across clusters, their method effectively implies that the corresponding variables do not contribute to the separation of those clusters. This approach elegantly combines the statistical rigor of mixture models with the variable selection capabilities of penalization.

6.4.2 Bayesian Methods

An alternative and powerful embedded approach is offered by Bayesian variable selection. In this framework, the relevance of each variable is treated as a random variable to be estimated from the data. This is typically accomplished by introducing a binary latent indicator variable for each feature, where the indicator denotes whether the variable is relevant to the clustering process or not.

The work of Tadesse et al. (2005) provides a prominent example of this approach. They developed a Bayesian variable selection method for model-based clustering of high-dimensional data, such as gene expression profiles. Their model uses Markov Chain Monte Carlo (MCMC) methods to explore the posterior distribution of the latent indicator variables. The posterior probability of a variable's indicator being "on" provides a natural and intuitive measure of its importance for the clustering structure. This allows for a probabilistic assessment of variable relevance, which can be more nuanced than the hard-thresholding of penalized methods.

Critical Analysis of Embedded Methods

Embedded methods represent a significant advance in variable selection for clustering, offering a compelling balance between statistical elegance, computational efficiency, and performance. Their primary advantage is their ability to handle high-dimensional data more effectively than wrapper methods. By integrating selection into the model, they avoid the costly iterative search over all possible subsets. The resulting models are often more parsimonious and interpretable.

However, these methods are not without their challenges. The performance of penalized methods often depends on the careful tuning of one or more penalty parameters, which can be a non-trivial task. Cross-validation is often used, but it adds to the computational burden. Bayesian methods, while providing rich probabilistic output, can be computationally intensive and may require expertise to diagnose convergence and set appropriate priors. Furthermore, both penalized and Bayesian methods are based on specific model assumptions (e.g., Gaussian mixtures), and their performance may be compromised if these assumptions are violated. Despite these considerations, embedded methods are a cornerstone of modern variable selection for clustering, providing a powerful toolkit for researchers dealing with complex, high-dimensional data.

6.5 Gaps, Challenges, and Synthesis

The review of filter, wrapper, and embedded methods reveals a clear trade-off between computational efficiency, statistical rigor, and flexibility. Filter methods are fast and simple but are disconnected from the clustering objective, risking suboptimal variable selection. Wrapper methods directly optimize for clustering performance but at a high computational cost and with a significant risk of overfitting. Embedded methods offer an elegant and efficient compromise by integrating selection into the model-fitting process, but their performance is often tied to specific model assumptions and requires careful parameter tuning.

Despite the considerable progress, several key challenges remain in the field of variable selection for clustering. A primary challenge is the development of methods that are simultaneously **statistically robust**, **computationally scalable**, and **highly interpretable**. Many existing methods excel in one or two of these areas but fall short in the third. For instance, complex Bayesian models may offer statistical robustness but can be difficult for non-experts to implement and interpret.

Furthermore, most existing variable selection methods are designed for simple, continuous data and may not be directly applicable to the complex data structures found in real-world applications. A significant gap exists in methods that can handle:

- **Mixed-type data:** Datasets containing a combination of continuous, categorical, and ordinal variables.
- **Weighted data:** Survey data, like the Demographic and Health Surveys (DHS), often include sampling weights that must be incorporated to produce representative results.
- **High-dimensional data with complex correlation structures:** The presence of intricate dependencies between variables can challenge many selection algorithms.

This synthesis of the literature highlights the need for a variable selection method that is not only grounded in a clear statistical principle but is also flexible enough to be adapted to the messy, complex data often encountered in applied research. The ideal method should be intuitive, provide a clear justification for its selections, and be capable of handling the specific characteristics of datasets like the DHS.

6.6 Justification for the Proposed Methodology

The preceding analysis of the literature reveals a critical need for a variable selection methodology that is not only robust and scalable but also transparent and adaptable to complex data. The proposed research addresses this need by adopting and extending the "blinding" procedure introduced by Fraiman et al. (2008). This approach, which can be viewed as a distinct and highly intuitive type of wrapper method, is uniquely suited to overcome the limitations of existing techniques.

The Fraiman et al. method is grounded in a clear and interpretable statistical principle: a variable's importance is measured by the impact its removal has on the final cluster partition. Instead of physically removing the variable, its influence is neutralized by replacing its values with a non-informative substitute. This makes the method's logic transparent and easy to communicate.

Crucially, the methodology provides a neat conceptual separation for tackling the two primary challenges in variable selection:

1. **Identifying Noise:** By replacing a variable's values with its marginal mean, we can assess whether it contributes anything more than random noise. If the clustering structure remains largely unchanged, the variable is deemed irrelevant.

2. **Identifying Redundancy:** By using a more sophisticated blinding—replacing a variable's values with its conditional mean given a subset of other variables—we can determine if its information is already captured by that subset. This directly addresses the problem of multicollinearity.

A key advantage of this framework is its **flexibility**. It is not tied to a specific clustering algorithm. The blinding procedure can be paired with any clustering method, from k-means to hierarchical or model-based approaches. This allows the selection process to be tailored to the algorithm best suited for the data at hand.

Therefore, this research frames the Fraiman et al. procedure as an ideal choice for analyzing complex survey data like the DHS. The novel contribution of this work will be to adapt, rigorously evaluate, and extend this powerful but less-commonly used method. We will develop strategies to handle the mixed-type and weighted nature of the DHS data within this framework, addressing a significant gap in the current literature. By doing so, we aim to provide a practical, robust, and interpretable solution for variable selection in a context where it is sorely needed.

Chapter 7

Methodology

This study is based on statistical procedures for variable selection in clustering. The core idea is to identify the smallest subset of variables that preserves, as closely as possible, the clustering structure obtained when using the full set of variables. The methods rely on the concept of "blinding"—replacing the values of certain variables so as to neutralize their influence—and on a quantitative measure of how much the resulting partition differs from the original one. This chapter describes the methods, underlying mathematical formulation, and algorithmic implementation that we will apply.

Two complementary blinding strategies will be implemented:

1. **Marginal Mean Blinding** — aimed at detecting and removing non-informative noisy variables.

2. **Conditional Mean Blinding** — designed to detect and remove both noisy and redundant (multicollinear) variables.

A forward–backward search algorithm will be employed to make the procedures computationally feasible for high-dimensional data.

7.1 Mathematical Notation

Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ be a p-dimensional random vector with joint distribution P. Let n independent realizations $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ constitute the dataset. A clustering procedure applied to the dataset produces a *partition function*:

$$f : \mathbb{R}^p \to \{1, 2, \ldots, K\}$$

assigning each observation to one of K clusters. The corresponding partition of the space is denoted by $\mathcal{G}_k = f^{-1}(k)$, $k = 1, \ldots, K$.

The goal is to find a subset of variables $I \subset \{1, 2, \ldots, p\}$, with $|I| = d < p$, such that the clustering assignments produced using only variables in I closely match those from the full set $\{1, \ldots, p\}$.

7.2 Blinding Procedures

To evaluate the contribution of each variable, we consider two complementary blinding strategies:

Marginal Mean Blinding

Each variable X_j to be blinded is replaced by its marginal mean, effectively removing its variability. This neutralizes the influence of variables that add random noise without contributing to cluster separation.

Conditional Mean Blinding

Each variable X_j to be blinded is replaced by its conditional mean given the remaining variables. This removes not only noise but also redundant information arising from strong correlations or multicollinearity among variables.

Both procedures allow us to measure the degree to which removing a variable alters the clustering structure, thus quantifying its informativeness.

7.3 Variable Selection Algorithm

An exhaustive search over all subsets of size d is computationally infeasible when p is large. To address this, we adopt a *forward–backward search strategy*:

- **Forward Step:** Iteratively add variables that most improve the preservation of the original clustering.
- **Backward Step:** Remove variables that, once included, are found to be redundant or uninformative.
- **Stopping Rule:** The algorithm terminates when no further improvement can be achieved without violating a predefined efficiency threshold (e.g., proportion of preserved cluster labels).

This iterative approach ensures scalability while balancing accuracy and computational cost.

7.4 Evaluation Method

The methodology will be validated through a combination of simulation studies and real-world applications.

7.4.1 Simulation Studies

Simulation experiments will be conducted under varying conditions to test robustness:

- Different levels of noise variables.
- Sample size variation (small, medium, large n).
- Increasing dimensionality (p).

These experiments will allow us to evaluate the ability of the methodology to recover informative variables under controlled scenarios.

7.4.2 Real-world Applications

The framework will be applied to high-dimensional real datasets such as the **Demographic and Health Survey (DHS)** dataset containing numerous socioeconomic and demographic variables.

7.5 Performance Metrics

Performance will be assessed using the following criteria:

- **Clustering preservation:** Proportion of labels preserved between the full-variable and reduced-variable partitions.
- **Subset size:** Number of variables retained relative to p.
- **Computational efficiency:** Runtime and memory usage of the algorithm.
- **Interpretability:** Practical relevance and interpretability of the selected variables for domain experts.

7.6 Comparative Benchmarking

The proposed methodology will be benchmarked against the following existing approaches:

- Sparse k-means clustering.
- Hierarchical clustering

Statistical tests will be applied to evaluate differences in performance across methods.

7.7 Justification of Method Choice

The chosen methodology offers a balance of *theoretical soundness, interpretability, and scalability.*

- **Marginal Mean Blinding** provides a simple and computationally efficient baseline for detecting noisy variables.
- **Conditional Mean Blinding** addresses redundancy, a more challenging but critical issue in high-dimensional data.
- The **forward–backward search algorithm** ensures practical feasibility without sacrificing performance.

Together, these methods form a stable and adaptable approach to variable selection in clustering, capable of handling real-world datasets.

7.8 Summary

This chapter has presented the proposed methodology for variable selection in clustering. By using blinding procedures, forward–backward search, and rigorous evaluation strategies, the research seeks to develop a framework that is accurate, interpretable, and computationally efficient.

Chapter 8

Expected Results

This chapter outlines the anticipated outcomes of the proposed study on variable selection for clustering. Based on the methodology described in Chapter **??**, the research is expected to produce results that advance both the theoretical understanding and practical implementation of variable selection in high-dimensional clustering.

8.1 Theoretical Contributions

- A theoretical framework demonstrating how blinding techniques (marginal mean and conditional mean) capture noise and redundancy in different ways.
- Development of a scalable forward–backward search algorithm for variable selection that avoids exhaustive search while retaining accuracy.
- Formal analysis of the algorithm's computational complexity and efficiency in high-dimensional settings.

From controlled simulation experiments, the proposed methodology is expected to show:

- **High preservation rates:** The reduced-variable clustering partitions will closely match those obtained from the full set of variables, with minimal loss of information.
- **Effective noise elimination:** Marginal mean blinding will successfully identify and remove purely random variables.
- **Redundancy handling:** Conditional mean blinding will outperform marginal blinding in detecting and eliminating highly correlated or multicollinear variables.

- **Robustness across conditions:** Performance will remain stable under varying levels of noise, correlation structures, dimensionality (p), and sample size (n).

When applied to real datasets such as Demographic and Health Survey (DHS) data and other high-dimensional data, the methodology is expected to yield:

- **Smaller, interpretable subsets of variables** that are meaningful to domain experts.
- **Improved cluster stability and clarity**, leading to partitions that are more robust and interpretable.
- **Computational efficiency**, reducing runtime and memory requirements compared to full-variable clustering.

Against existing approaches (e.g., sparse k-means, hierarchical clustering with selection, the proposed framework is expected to demonstrate:

- **Competitive or superior clustering accuracy**, particularly in datasets with strong redundancy or high noise.
- **Better balance between accuracy and interpretability**, due to the clear role of blinding procedures.
- **Improved scalability**, enabling application to datasets with thousands of variables.

The anticipated findings will have broader significance:

- **For theory:** Establishing blinding-based selection as a principled and generalizable approach to variable selection in clustering.
- **For practice:** Providing researchers and practitioners in fields such as public health, social sciences, and bioinformatics with a reliable tool for high-dimensional data exploration.
- **For computation:** Contributing an efficient algorithm that can be adapted into software packages for widespread use.

8.2 Summary

In summary, the proposed research is expected to deliver:

- A principled framework for variable selection in clustering.
- An efficient and scalable algorithm that integrates marginal and conditional blinding.
- Demonstrated effectiveness through both simulation studies and real-world applications.
- Enhanced accuracy, interpretability, and computational feasibility compared to existing methods.

These results will lay the foundation for more reliable and interpretable clustering in high-dimensional data, advancing both the methodology and its applications.

Bibliography

R. D. Arvey, W.-D. Li, and N. Wang. Genetics and organizational behavior. *Annual Review of Organizational Psychology and Organizational Behavior*, 3(1):167–190, 2016.

R. E. Bellman. *Adaptive control processes: a guided tour*. Princeton university press, 1961.

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

W. Chi, W.-D. Li, N. Wang, and Z. Song. 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, 2016.

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

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

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

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

R. Fraiman, A. Justel, and M. Svarc. Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association*, 103(483):1294–1303, 2008.

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

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

J. A. Hartigan. *Clustering Algorithms.* John Wiley & Sons, 1975.

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

G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994*, pages 121–129. Elsevier, 1994.

S. S. Mothi, M. Sudarshan, N. Tandon, C. Tamminga, G. Pearlson, J. Sweeney, B. Clementz, and M. S. Keshavan. Machine learning improved classification of psychoses using clinical and biological stratification: update from the bipolar-schizophrenia network for intermediate phenotypes (b-snip). *Schizophrenia research*, 214:60–69, 2019.

W. Pan and X. Shen. Penalized model-based clustering with application to variable selection. *Journal of machine learning research*, 8:1145–1164, 2007.

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

D. Steinley and M. J. Brusco. Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, 73(1):125–144, 2008.

D. Sun, T. G. van Erp, P. M. Thompson, C. E. Bearden, M. Daley, L. Kushan, M. E. Hardt, K. H. Nuechterlein, A. W. Toga, and T. D. Cannon. 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, 2009.

M. G. Tadesse, N. Sha, and M. Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470): 602–617, 2005.

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

D. A. Waldman, D. Wang, and V. Fenters. The added value of neuroscience methods in organizational research. *Organizational Research Methods*, 22(1):223–249, 2019.

D. M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

T. Yarkoni and J. Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122, 2017.