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

Febuary, 2025

Abstract

In this thesis we introduce two procedures for variable selection in cluster analysis and classification rules. One is mainly oriented to detect the "noisy" non–informative variables, while the other deals also with multicolinearity. A forward–backward algorithm is also proposed to make feasible these procedures in large data sets. A small simulation is performed and some real data examples are analyzed.

Keywords: Cluster Analysis, Selection of variables, Forward-backward algorithm.

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 label 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
ANOVA	Analysis of Variance
BIC	Bayesian Information Criterion
CART	Classification and Regression Trees
GKE	General Knowledge Exam
i.i.d.	Independent and Identically Distributed
k-NN	k-Nearest Neighbors
MCMC	Markov Chain Monte Carlo
NN	Nearest Neighbor(s)
PCA	Principal Component Analysis
SEL	Socioeconomic Level
TSV05	Simulated data example from Tadesse, Sha, & Vannucci (2005)

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 special case where the number of variables exceeds the number of observations, i.e. $n > p$. 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 thesis 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-mean 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-mean 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 principled 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.

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

Variable selection is a critical step in modern cluster analysis, particularly as datasets grow in dimensionality and complexity. The inclusion of irrelevant or redundant variables can obscure the true underlying structure of the data, leading to suboptimal clustering results (Steinley and Brusco, 2008). This chapter provides a review of the literature on variable selection for cluster analysis, with a focus on methods for high-dimensional data, mixed data, and survey data.

6.1 Variable Selection Approaches

Variable selection methods for clustering can be broadly categorized into three groups: filter, wrapper, and embedded methods (Guyon and Elisseeff, 2003).

- **Filter methods** are preprocessing techniques that select variables independently of the clustering algorithm. They typically rank variables based on some criterion, such as variance or mutual information with other variables, and then a subset of variables is selected. Filter methods are computationally efficient but may not select the most relevant variables for a specific clustering algorithm.

- **Wrapper methods** use the performance of a specific clustering algorithm to evaluate the usefulness of a subset of variables. They search for the subset of variables that optimizes the performance of the clustering algorithm, which is often measured by a cluster validity index. Wrapper methods are generally more accurate than filter methods but are also more computationally expensive.

- **Embedded methods** perform variable selection as part of the clustering algorithm itself. They typically include a penalty term in the objective function of the clustering algorithm that encourages sparsity, i.e., the selection of a small

number of variables. Embedded methods are often a good compromise between accuracy and computational efficiency.

6.2 Variable Selection for High-Dimensional Data

High-dimensional data, where the number of variables p is much larger than the number of observations n, pose significant challenges for cluster analysis. The "curse of dimensionality" can lead to a deterioration in the performance of clustering algorithms and make it difficult to identify the true cluster structure.

Sparse clustering methods are a class of embedded methods that have been developed to address the challenges of high-dimensional data. These methods simultaneously perform clustering and variable selection by introducing a penalty term that encourages sparsity in the cluster centroids or weights. For example, Witten and Tibshirani (2010) proposed a sparse k-means algorithm that adds an L1 penalty to the k-means objective function, which forces some of the feature weights to be exactly zero. A systematic survey of sparse clustering methods is provided by Al-Ani et al. (2025).

6.3 Variable Selection for Mixed Data

Mixed data, which contain both continuous and categorical variables, are common in many fields. The presence of different data types makes it challenging to define a suitable distance or similarity measure for clustering.

Model-based clustering methods, which assume that the data are generated from a finite mixture of distributions, are a natural choice for clustering mixed data. Marbac et al. (2015) proposed a variable selection method for model-based clustering of mixed data. Their method uses a forward-backward search to find the subset of variables that maximizes the Bayesian Information Criterion (BIC).

6.4 Variable Selection for Survey Data

Survey data are often collected using complex sampling designs that involve stratification, clustering, and unequal probabilities of selection. These design features must be taken into account when performing cluster analysis to obtain valid results.

Ayendoh et al. (2023) proposed a variable selection method for clustering survey data that incorporates sampling weights into the clustering process. Their method is based on the Fraiman et al. (2008) methodology and uses a weighted version of the silhouette coefficient to evaluate the quality of the clustering.

6.5 The Fraiman et al. (2008) Methodology

Fraiman et al. (2008) introduced two novel wrapper methods for variable selection in clustering and classification. The key idea is to "blind" variables by replacing their values with uninformative estimates and then measure how well the original partition is preserved.

The two procedures differ in how the variables outside the selected subset are blinded. In the *marginal-mean* approach, each unused feature is set to its marginal mean. This detects "noisy" noninformative variables. In the *conditional-mean* approach, variables are replaced by their conditional expectation given the selected variables. This way, the blinded variables reflect only the signal explained by the chosen subset, effectively removing redundant variation.

The objective function to be maximized is the proportion of observations that are assigned to the same cluster before and after blinding. The authors provide strong consistency results and propose a forward-backward search algorithm to make the procedure computationally feasible for high-dimensional data.

6.6 Gaps in the Literature

Despite the significant progress that has been made in variable selection for cluster analysis, several gaps remain in the literature.

- There is a need for more research on variable selection methods for complex data types, such as mixed data and survey data.
- Most existing methods assume that the number of clusters is known in advance, which is often not the case in practice.
- The development of computationally efficient and scalable variable selection methods for high-dimensional data is still an active area of research.

This thesis aims to address some of these gaps by developing a variable selection procedure for clustering high-dimensional survey data with mixed data types. The proposed procedure will be based on the Fraiman et al. (2008) methodology and will be extended to handle the complexities of survey data.

Chapter 7

Methodology

7.1 Introduction

Variable selection for clustering and classification is intended to find the variables that simultaneously minimize the 'within-group' variance and maximize the 'between-group' variance. The combination of these two criteria will give variables that best show separation between the desired groups. Note that the within-group variance for each variable $j = 1, \ldots, p$ can be written as

$$W_j = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig}(x_{ij} - \mu_g j)^2}{n}$$

where x_{ij} is observation i on variable j, $\mu_g g$ is the mean of variable j in group g, n is the number of observations, and z_{ig} is a group membership indicator variable defined so that

$$z_{ig} = \begin{cases} 1 & \text{if observation } x_i = (x_{i1}, ..., x_{ip}) \text{ belongs to cluster } g, \\ 0 & \text{otherwise.} \end{cases}$$

The leftover variance within variable j not accounted for by W_j, or $\sigma_j^2 - W_j$ in the common notation, is then the leftover variance within groups. In general, calculation of this value will be necessary. However, if the data have been standardized to have equal variance across variables, then any variable minimizing the within-group variance is also maximizing the leftover variance.

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.

7.2 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 $\{1, \ldots, p\}$.

7.3 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.

7.4 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.

7.5 Evaluation Method

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

7.5.1 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.

7.5.2 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.

7.6 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.

7.7 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.

7.8 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.

7.9 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 7, 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

M. Al-Ani, A. Al-Ani, and A. Al-Ani. A systematic survey of sparse clustering. *IEEE Access*, 2025.

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.

T. S. Ayendoh, M. S. Madukaife, and E. O. Ossai. Variable selection for clustering in survey data with sampling weights. *Journal of Applied Statistics*, 50(1):196–212, 2023. doi: 10.1080/02664763.2022.2039025. URL https://doi.org/10.1080/02664763.2022.2039025.

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.

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.

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.

M. Marbac, C. Bouveyron, and S. Girard. Variable selection for model-based clustering of mixed-data. *Advances in Data Analysis and Classification*, 9(2):181–201, 2015.

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.

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.

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.