Iterative Exploration-Driven Sparse SDP Clustering via Thompson Sampling

Jongmin Mun*
Department of Data Sciences and Operations, University of Southern California
Paromita Dubey
Department of Data Sciences and Operations, University of Southern California
and
Yingying Fan
Department of Data Sciences and Operations, University of Southern California

February 12, 2026

Abstract

This paper studies high-dimensional sparse clustering, a combinatorial NP-hard problem arising from the bilinear coupling between cluster assignment and feature selection. We analyze semidefinite programming (SDP) relaxations of K-means and establish minimax separation bounds, demonstrating that these relaxations are theoretically robust to feature over-selection: exact recovery is preserved even in the presence of non-informative features. Leveraging this robustness, we propose a block-coordinate ascent framework that alternates between SDP-based clustering and non-conservative feature selection. To address the tendency of deterministic greedy methods to become trapped in local optima, we formulate the feature selection step as a Thompson sampling bandit problem. This approach introduces adaptive memory by aggregating historical variable-selection outcomes into posterior distributions, and selects features via posterior sampling, enabling stochastic exploration that promotes the inclusion of underexplored features and facilitates escape from local maxima. We establish conditions for consistent variable selection and exact clustering recovery, and extend the method to settings with unknown covariance through a scalable, inverse-free estimation procedure. Numerical experiments demonstrate that the proposed memory-driven approach consistently outperforms state-of-the-art sparse clustering methods.

Keywords: exact recovery; Gaussian mixture models; high-dimensional data; K-means; precision matrix; semidefinite relaxation; sparsity; unsupervised learning; variable selection

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the version that gives all author information*

1 INTRODUCTION

a key message we need to emphasize: sparse support set and population parameters are nuisances for clustering, and they do not need to be perfectly recovered for exact clustering

High-dimensional clustering methods group unlabeled observations when the number of features exceeds or matches the sample size. They are widely applied to high-cost recordings with limited sample sizes such as gene expression (Curtis et al. 2012, He et al. 2021) and neurophysiological signals (Cunningham & Yu 2014, Kadir et al. 2014, Park et al. 2024). They are also used in machine learning tasks such as reinforcement learning (Liu & Frank 2022), privacy protection (Javanmard & Mirrokni 2023), and federated learning (Kim et al. 2024). In many such problems, only a small subset of features is relevant to the clustering. For example, in document clustering, each document contains only a few informative words out of a large corpus (Bing et al. 2020, Wu et al. 2023, Yi et al. 2014). In disease subtype discovery from gene expression data, only a handful of genes out of thousands differ across subtypes (Golub et al. 1999).

If sparsity is present and effectively leveraged, clustering performance improves, as demonstrated in broad applications such as genomics (Lu et al. 2017), neuroimaging data analysis (Mishra et al. 2017, Zhang et al. 2021, Namgung et al. 2024), and financial modeling (Nystrup et al. 2021). However, sparsity in clustering introduces a key challenge of interdependence between the two unsupervised tasks of feature selection and clustering. On one hand, with unknown sparsity, including too many noise features in the clustering algorithm hinders the results. On the other hand, it is challenging to pinpoint the signal features when the true cluster structure is unknown. This interplay motivates two-step or iterative algorithms that alternates between feature selection and clustering, leading to the identification of the true clusters.

1.0.0.1 Sparse Clustering Model

We use the following Gaussian mixture model to study the role of sparsity in clustering. Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be the data matrix with columns $\mathbf{X}_1, \ldots, \mathbf{X}_n$ independently generated as

$$\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}^*) \text{ if } i \in G_k^*, \tag{1}$$

where the non-random sets $G_1^*, \ldots, G_K^*$ partition $\{1, \ldots, n\}$. The number of clusters K is assumed known. The unknown cluster centers are denoted by $\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_K^* \in \mathbb{R}^p$, and $\boldsymbol{\Sigma}^* \in \mathbb{R}^{p\times p}$ is the common covariance matrix. We consider both cases where $\boldsymbol{\Sigma}^*$ is known and where it is unknown but $\boldsymbol{\Omega}^* := (\boldsymbol{\Sigma}^*)^{-1}$ has a sparse structure. The goal is to recover

$$G^* := (G_1^*, \ldots, G_K^*),$$

treating $\boldsymbol{\mu}_k^*$'s and $\boldsymbol{\Sigma}^*$ as nuisance parameters. We assume the existence of an unknown signal feature set

$$S_0 := \bigcup_{1\le k\ne l\le K} \operatorname{supp}\Big(\boldsymbol{\Omega}^*(\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_l^*)\Big) \subset \{1, \ldots, p\}, \tag{2}$$

which encodes the complete clustering signal. This assumption aligns with results from classification of two-class Gaussian mixtures, where the optimal decision boundary maximizing the signal-to-noise ratio is proportional to $\boldsymbol{\Omega}^*(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)$(Fan et al. 2013).

1.1 Previous Methods

1.1.0.1 Previous Two-Step Methods

Conventional dimension reduction, such as PCA (Ghosh & Chinnaiyan 2002, Liu et al. 2003, Tamayo et al. 2007), can fail to capture a sparse clustering signal (Chang 1983). Consequently, many methods adopt models such as (1) and (2) to perform feature selection guided by the clustering signal. They select features using multiple testing based on goodness-of-fit (Jin & Wang 2016) or moments (Azizyan et al. 2013, Verzelen & Arias-Castro 2017), model selection (Maugis et al. 2009, Raftery &

Dean 2006), sparse PCA (Löffler et al. 2022), merging algorithm (Banerjee et al. 2017), or an estimated sparse discriminating direction (defined in Section 3.1; Azizyan et al. 2015). Other relevant works include Bouveyron & Brunet-Saumard (2014), Brodinová et al. (2019), Kadir et al. (2014), Liu et al. (2023). Jin et al. (2017) show that, in certain regimes of signal strength and sparsity, accurate feature selection is not required for accurate clustering. Even et al. (2025) use data thinning (Leiner et al. 2025, Dharamshi et al. 2025) to decouple feature selection from clustering and achieves optimality among polynomial-time clustering algorithms. They further show that the separation required for the success of their method differs from the minimax separation needed for optimal statistical performance without computational constraints.

1.1.0.2 Previous Iterative Methods Iterative methods repeatedly refine cluster assignments and model parameters estimates. Most methods alternate between clustering and regularized estimation of model parameters, such as feature weights (Chakraborty et al. 2023, Huo & Tseng 2017, Li et al. 2018, Tsai & Chiu 2008, Witten & Tibshirani 2010, Yuan et al. 2024), cluster centers (Sun et al. 2012, Yi et al. 2014), and feature rankings (Arias-Castro & Pu 2017, Zhang et al. 2020). Regularized EM algorithms follow the same principle: the E-step updates cluster assignments, while the M-step performs regularized estimation of sparse model parameters (Bouveyron et al. 2007, Dong et al. 2024, Fraley & Raftery 2007, Fu et al. 2021, Guo et al. 2010, Pan & Shen 2007, Wang & Zhu 2008, Xie et al. 2008, Zhou et al. 2009). Notably, Cai et al. (2019) establish a minimax rate for the excess clustering risk based on estimating $\boldsymbol{\beta}^*$ from the training data using ℓ_1-regularized EM. Sparse Bayesian hierarchical models (Tadesse et al. 2005, Yao et al. 2025) and sparse convex clustering (Chakraborty & and Xu 2023, Wang et al. 2018) also follow the iterative paradigm, as their fitting procedures, such as MCMC or ADMM, progressively refine parameter estimates and feature selection.

1.2 Limitations of Existing Approaches and Our Approach

Standard iterative clustering methods, such as Expectation-Maximization (EM), are highly sensitive to local optima due to the non-convex nature of joint clustering and feature selection. Theoretical guarantees for these methods often rely on strong assumptions, such as a "warm start"—an initialization sufficiently close to the true solution (Cai et al. 2019).

To address these limitations, we propose an iterative framework that combines stochastic exploration with sequential learning to escape local optima. Our approach is grounded in the theoretical properties of the semidefinite programming (SDP) relaxation of K-means, which provides a convex formulation that bypasses explicit cluster center estimation and achieves minimax optimality for fixed dimension p (Chen & Yang 2021). We extend this analysis to variable selection by quantifying the sensitivity of SDP K-means to the chosen feature subset. Specifically, we establish a uniform minimax separation bound guaranteeing exact recovery whenever the intersection of the selected set S and the true signal set S_0 satisfies

$$\Delta^2 \gtrsim \log n + \frac{|S| \log p}{n} + \sqrt{\frac{|S| \log p}{n}}. \tag{3}$$

This result shows that the method is robust to feature over-selection: a slight overestimation of S_0 does not compromise clustering performance, provided the true signal features are retained.

Motivated by this insight, we propose an iterative algorithm that alternates between SDP K-means clustering and feature selection, where the feature selection step is formulated as a **Thompson Sampling** bandit problem, introducing two key innovations:

1. **Historical Integration (Memory):** Each feature's utility is modeled as a latent probability θ_j, with a Beta posterior $\text{Beta}(a_j^t, b_j^t)$ updated over iterations. Aggregating past successes and failures stabilizes the optimization and mitigates the influence of

noisy or transient signals.

2. **Active Exploration:** Feature weights are sampled from their posteriors rather than chosen deterministically. This stochasticity allows the algorithm to explore underrepresented or uncertain features, escape poor local optima, and robustly recover the global signal set S_0.

When the covariance structure is unknown, instead of estimating the full covariance, we perform clustering and feature selection using the "innovated transformation" $\mathbf{\Omega}^*\mathbf{X}$ (Fan & Lv 2016). This approach leverages high-dimensional precision matrix estimation in a supervised manner to compute only the minimally required quantities for the transformation, making the procedure computationally efficient.

1.2.0.1 Notations We denote vectors and matrices by boldface letters, and their individual elements by the corresponding non-boldface letters. For a matrix $\mathbf{A}$, single and double subscripts denote columns and (row, column) elements, respectively. A subscript may be a single index or an index set. For vectors not derived from a matrix, a subscript denotes an element or a sub-vector. If the vector or matrix already contains a subscript, we use parentheses to avoid ambiguity, for example, $(\tilde{\boldsymbol{\mu}}_1^*)_S$. We write $\mathbf{A} \geq 0$, $\mathbf{A} > 0$, and $\mathbf{A} \succeq 0$ to denote elementwise nonnegativity, elementwise positivity, and positive semidefiniteness of a matrix, respectively. For a set S, let $|S|$ denotes its cardinality. For an integer $a > 0$, let $\mathbf{1}_a \in \mathbb{R}^a$ denote the *all-one vector*, and $[a]$ denote $\{1, \cdots, a\}$. For a set $G_k \subset [n]$, let $\mathbf{1}_{G_k} \in \{0, 1\}^n$ denote the indicator vector that takes the value 1 at indices in G_k and 0 elsewhere.

2 PROBLEM SETUP AND MOTIVATION

We first introduce SDP K-Means and highlight its limitations in high dimensions through the lens of minimax analysis.

2.1 SDP K-means

The partition $G_1, \ldots, G_K$ maximizing the K-means objective is the solution of the following NP-hard mixed integer program (Peng & Wei 2007, Zhuang et al. 2023):

$$\max_{\mathbf{H} \in \{0,1\}^{n \times K}} \langle \mathbf{X}^\top \mathbf{\Omega}^* \mathbf{X}, \mathbf{H}\mathbf{B}\mathbf{H}^\top \rangle, \text{ s.t. } \mathbf{H}\mathbf{1}_K = \mathbf{1}_n, \tag{4}$$

where the inner product denotes the Frobenius inner product, $\mathbf{B} := \mathrm{diag}(|G_1|, \ldots, |G_K|)$, and each row of $\mathbf{H}$ is a one-hot cluster indicator, formally $H_{i,k} := \mathbb{1}(\mathbf{X}_i \in G_k)$ for $i \in [n]$ and $k \in [K]$. Appendix **??** connects the objective (4) to the log likelihood function in the Gaussian mixture model. Since the change of variable $\mathbf{Z} = \mathbf{H}\mathbf{B}\mathbf{H}^\top$ satisfies $\mathbf{Z}^\top = \mathbf{Z}, \mathbf{Z} \succeq 0, \mathrm{tr}(\mathbf{Z}) = K, \mathbf{Z}\mathbf{1}_n = \mathbf{1}_n$, and $\mathbf{Z} \geq 0$, we can lift $\mathbf{Z}$ into the convex space and relax (4) into a semidefinite program:

$$\max_{\mathbf{Z} \in \mathbb{R}^{n \times n}} \langle \mathbf{X}^\top \mathbf{\Omega}^* \mathbf{X}, \mathbf{Z} \rangle \quad \text{s.t.} \quad \mathbf{Z}^\top = \mathbf{Z},\ \mathbf{Z} \succeq 0,\ \mathrm{tr}(\mathbf{Z}) = K,\ \mathbf{Z}\mathbf{1}_n = \mathbf{1}_n,\ \mathbf{Z} \geq 0. \tag{5}$$

The SDP K-means algorithm recovers cluster labels by applying spectral clustering to the solution of (5). When $\mathbf{\Sigma}^*$ is known to be $\mathbf{I}_p$ and the dimensionality p is fixed, SDP K-means achieves the minimax separation under the exact cluster recovery loss (Chen & Yang 2021), where the separation is defined as in (10) with $S = [p]$. The minimax separation $\log n + \sqrt{(\log n)^2 + (Kp \log n)/n}$ increases with p at a square-root rate. This shows that any clustering method that does not adapt to sparsity, including SDP K-means, may fail in high-dimensional settings where p is comparable to or larger than n, unless the separation also scales quickly at the square-root rate with p — an assumption that can be

very stringent for large p. This motivates us to modify the SDP K-means to incorporate sparsity in high dimensions and explore the corresponding separation rate.

2.2 Incorporating Sparsity via Best Subset SDP K-Means

To handle high-dimensional settings where the ambient dimension p grows with the sample size n, we extend the SDP relaxation in (5) to jointly estimate the cluster structure and the informative feature set S_0. Our approach incorporates feature selection by exploiting the identity

$$\mathbf{X}^\top \mathbf{\Omega}^* \mathbf{X} = \tilde{\mathbf{X}}^\top \mathbf{\Sigma}^* \tilde{\mathbf{X}}, \qquad \tilde{\mathbf{X}} := \mathbf{\Omega}^* \mathbf{X},$$

which allows the clustering objective to be expressed in terms of the covariance matrix $\mathbf{\Sigma}^*$ rather than the precision matrix.

For exposition, assume $\mathbf{\Omega}^*$ is known. Writing the data matrix as

$$\mathbf{X} = \mathbf{M}^* + \mathcal{E}, \qquad \tilde{\mathbf{X}} = \tilde{\mathbf{M}}^* + \tilde{\mathcal{E}},$$

with $\tilde{\mathbf{M}}^* := \mathbf{\Omega}^* \mathbf{M}^*$ and $\tilde{\mathcal{E}} := \mathbf{\Omega}^* \mathcal{E}$, the similarity matrix becomes

$$\mathbf{A} = \tilde{\mathbf{X}}^\top \mathbf{\Sigma}^* \tilde{\mathbf{X}} = (\tilde{\mathbf{M}}^*)^\top \mathbf{\Sigma}^* \tilde{\mathbf{M}}^* + \text{noise terms}.$$

The noise terms include mean-zero cross terms and a term independent of $\tilde{\mathbf{M}}^*$, so the clustering signal is contained entirely in $(\tilde{\mathbf{M}}^*)^\top \mathbf{\Sigma}^* \tilde{\mathbf{M}}^*$.

Let $S_0 \subset [p]$ denote the signal coordinates where $\boldsymbol{\mu}_1^*$ and $\boldsymbol{\mu}_2^*$ differ, and let $N = [p] \setminus S_0$. Since $\boldsymbol{\beta}^*$ is sparse, $\tilde{\mathbf{M}}^*$ admits the same decomposition, and we can write

$$(\tilde{\mathbf{M}}^*)^\top \mathbf{\Sigma}^* \tilde{\mathbf{M}}^* = (\tilde{\mathbf{M}}^*_{S_0})^\top \mathbf{\Sigma}^*_{S_0, S_0} \tilde{\mathbf{M}}^*_{S_0} + \text{terms involving } N,$$

where the first term captures all clustering-relevant information.

Motivated by this decomposition, we propose running SDP-relaxed K-means using the empirical, noisy analogue $\tilde{\mathbf{X}}_S^\top \mathbf{\Sigma}^*_{SS} \tilde{\mathbf{X}}_S$ as the objective, for a candidate feature set S. For

fixed $S \subset \{1, \ldots, p\}$ with $|S| \leq |S_0|$, the SDP clustering problem is

$$\max_{\mathbf{Z} \in \mathbb{R}^{n \times n}} \quad \left\langle \tilde{\mathbf{X}}_S^\top \mathbf{\Sigma}^*_{S,S} \tilde{\mathbf{X}}_S, \mathbf{Z} \right\rangle \text{ s.t. } \mathbf{Z}^\top = \mathbf{Z},\ \mathbf{Z} \succeq 0,\ \mathbf{Z} \geq 0,\ \operatorname{tr}(\mathbf{Z}) = K,\ \mathbf{Z}\mathbf{1}_n = \mathbf{1}_n.$$

The effectiveness of this objective in capturing the clustering signal is evaluated numerically in Appendix **??**. Since this optimization problem forms the backbone of our algorithms, we present it formally as Algorithm 1, which maps two matrices to cluster labels.

To jointly optimize over the partition $\mathbf{Z}$ and the feature set S, we introduce binary indicators $\mathbf{y} \in \{0,1\}^p$, where $y_j = 1$ if feature j is selected. This yields the *Best Subset SDP*:

$$\begin{aligned} \max_{\mathbf{Z}, \mathbf{w}} \quad & \sum_{j=1}^{p} \sum_{l=1}^{p} y_j y_l \Sigma^*_{jl} \langle \tilde{\mathbf{x}}_j \tilde{\mathbf{x}}_l^\top, \mathbf{Z} \rangle \\ \text{s.t.} \quad & \mathbf{Z}^\top = \mathbf{Z}, \quad \mathbf{Z} \succeq 0, \quad \mathbf{Z} \geq 0, \\ & \operatorname{tr}(\mathbf{Z}) = K, \quad \mathbf{Z}\mathbf{1}_n = \mathbf{1}_n, \\ & y_j \in \{0,1\}, \quad \sum_{j=1}^{p} y_j \leq |S_0|. \end{aligned} \tag{6}$$

This formulation is a Mixed-Integer Semidefinite Program (MISDP) with a bilinear coupling between $\mathbf{Z}$ and $\mathbf{y}$. Global optimization is intractable because the optimal feature weights depend on the partition $\mathbf{Z}$, and vice versa. This interdependence persists even under structural simplifications such as a diagonal covariance, where the challenge of simultaneously identifying the active feature set and the cluster structure remains.

To address this, we propose a block coordinate ascent framework that alternates between

Algorithm 1 Subset SDP K-means for $K = 2$: `SDPclust`

Require: Transformed and truncated data matrix $\tilde{\mathbf{X}}_{S,\cdot}$, covariance sub-matrix $\mathbf{\Sigma}_{S,S}$

1: $\hat{\mathbf{Z}} \leftarrow \arg\max_{\mathbf{Z} \in \mathbb{R}^{n \times n}} \langle \tilde{\mathbf{X}}_{S,\cdot}^\top \mathbf{\Sigma}_{S,S} \tilde{\mathbf{X}}_{S,\cdot}, \mathbf{Z} \rangle$

$$\text{s.t. } \mathbf{Z}^\top = \mathbf{Z},\ \mathbf{Z} \succeq 0,\ \operatorname{tr}(\mathbf{Z}) = 2, \mathbf{Z}\mathbf{1}_n = \mathbf{1}_n,\ \mathbf{Z} \geq 0$$

2: $\hat{G}_1, \hat{G}_2 \leftarrow$ spectral clustering result on $\hat{\mathbf{Z}}$

3: Output $\hat{G} = (\hat{G}_1, \hat{G}_2)$

SDP-based clustering and feature selection. This iterative strategy is theoretically justified: a sequence of SDP solutions can achieve exact recovery as long as the feature selection step maintains a sufficient signal-to-noise ratio, even if it is not perfect.

2.3 Sparse SDP on a Fixed Feature Set

In this section, we provide the theoretical foundation for our iterative clustering approach by establishing the minimax properties of SDP K-means under variable selection. Specifically, we analyze the separation conditions required for exact cluster recovery when the algorithm is restricted to a fixed arbitrary feature subset S. Our main result, Theorem 1, proves that exact recovery is achievable with high probability uniformly over all subsets S that retain sufficient signal strength relative to their cardinality. This uniform guarantee is critical: it implies that our iterative algorithm does not need to identify the exact signal set S_0 immediately. Instead, so long as the algorithm traverses a sequence of "good" subsets—those that capture enough signal to outweigh their noise dimensions—it will succeed in recovering the true clusters. We further demonstrate in Theorem 2 that this sufficient condition matches the information-theoretic lower bound in key regimes, confirming that our theoretical framework captures the fundamental limits of sparse clustering.

To simplify technical analysis, we assume that $\mathbf{\Sigma}^*$ is known to be $\sigma^2\mathbf{I}_p$. This is the setting where SDP's exact recovery property has been investigated for fixed p (Chen & Yang 2021). Given any variable subset S, consider SDP K-means restricted to S, which solves the SDP problem (5) over the subset S:

$$\max_{\mathbf{Z}\in\mathbb{R}^{n\times n}} \langle \mathbf{X}_{S,\cdot}^\top \mathbf{X}_{S,\cdot}, \mathbf{Z}\rangle \quad \text{s.t.} \quad \mathbf{Z}^\top = \mathbf{Z},\ \mathbf{Z}\succeq 0,\ \mathrm{tr}(\mathbf{Z}) = K,\ \mathbf{Z}\mathbf{1}_n = \mathbf{1}_n,\ \mathbf{Z}\geq 0, \tag{7}$$

and then obtain cluster labels by applying spectral clustering to the solution. The minimax analysis proceeds in two steps. First, we show that if the clustering signal (or separation) in $S\cap S_0$, defined in (10), exceeds a certain threshold explicitly depending on the sample size

n, ambient dimension p, and subset size $|S|$, then exact recovery of cluster labels is possible with high probability. This high-probability event is established uniformly over all possible subsets S. Second, we identify the minimax optimal regime for our result. Together, these results reveal the role of sparsity in clustering difficulty. They also motivate our iterative algorithm, which incorporates feature selection to address sparsity and solves a sequence of SDPs corresponding to the sequence of selected feature subsets.

We begin by specifying the problem setting. For each parameter tuple $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{H})$, with $\boldsymbol{\mu}_k \in \mathbb{R}^p$ cluster centers and $\mathbf{H} \in \{0,1\}^{n\times K}$ cluster assignment matrix, let $\mathbb{P}_{\boldsymbol{\theta}}$ be the corresponding joint distribution of n independent Gaussian random vectors. Each random vector has mean given by the assigned cluster center and covariance $\sigma^2\mathbf{I}_p$. For all $1 \le k \ne l \le K$, we impose two restrictions on $\boldsymbol{\mu}_k - \boldsymbol{\mu}_l$. First, it has at most s nonzero entries. Second, its squared ℓ_2 norm exceeds a specified threshold Δ^2. These conditions define the following Gaussian mixture distribution class:

$$\Theta(s, \Delta, K) := \mathcal{M}(s, \Delta, K) \times \mathcal{H}(K), \tag{8}$$

$$\text{where } \mathcal{M}(s, \Delta, K) := \Big\{(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) : \boldsymbol{\mu}_k \in \mathbb{R}^p, \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|_2^2 \ge \Delta^2, \text{ for all } 1 \le k \ne l \le K, \cup_{k,l} \operatorname{supp}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l) = S_0 \subset [p], |S_0| \le s\Big\},$$

$$\text{and } \mathcal{H}(K) := \Big\{\mathbf{H} = (H_{i,k}) \in \{0,1\}^{n\times K} : \sum_{k=1}^{n} H_{i,k} = 1 \text{ for all } i = 1, \ldots, n\Big\}.$$

For a given clustering algorithm $\hat{\mathbf{H}} : \mathbb{R}^{p\times n} \to \mathcal{H}(K)$, we assess its performance using the worst-case probability of exact cluster recovery over the distribution class $\Theta(s, \Delta, K)$:

$$\sup_{\boldsymbol{\theta}\in\Theta(s,\Delta,K)} \min_{\pi\in\mathcal{S}_K} \mathbb{P}_{\boldsymbol{\theta}}\Big(\hat{\mathbf{H}}(\mathbf{X}) \ne \mathbf{H}\Big), \tag{9}$$

where $\mathcal{S}_K$ is the set of label permutations of $[K]$. For brevity, we omit $\min_{\pi\in\mathcal{S}_K}$ henceforth.

Given a problem instance $\boldsymbol{\theta}^* = (\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_k^*, \mathbf{H}^*) \in \Theta(s, \Delta, K)$, the true cluster assignment $\mathbf{H}^*$ has one-to-one correspondence to the block-diagonal matrix $\mathbf{Z}^* := \mathbf{H}^*\mathbf{B}^*(\mathbf{H}^*)^\top =$

$\sum_{k=1}^{K} |G_k^*|^{-1} \mathbf{1}_{G_k^*} \mathbf{1}_{G_k^*}^\top$, up to label permutations. Let the clustering signal (or separation) of a feature subset S be defined as

$$\Delta_{S\cap S_0}^2 := \min_{1\le k\neq l\le K} \|(\boldsymbol{\mu}_l^* - \boldsymbol{\mu}_k^*)_{S\cap S_0}\|_2^2. \tag{10}$$

We next define the collection of feature subsets with sufficiently strong signal:

$$\begin{aligned} &\mathcal{S} := \Big\{S : S \subset [p], |S| \le \sqrt{p}, \Delta_{S\cap S_0}^2 \gtrsim \bar{\Delta}_S^2\Big\}, \\ &\text{where } \bar{\Delta}_S^2 := \sigma^2\left(\log n + \frac{|S|\log p}{m} + \sqrt{\frac{|S|\log p}{m}}\right), \\ &\text{and } m := 2 \min_{1\le k\neq l\le K} \left\{\left(\frac{1}{|G_k^*|} + \frac{1}{|G_l^*|}\right)^{-1}\right\}. \end{aligned} \tag{11}$$

Theorem 1 gives the separation required for SDP K-means restricted to S to achieve exact recovery:

Theorem 1. *Assume that there exists a universal constant C_1 such that $m \ge C_1 n/\log n$. Also assume $|G_k^*| \ge 2$ for all $k \in [K]$. Let $\hat{\mathbf{Z}}(S)$ be the solution of (7) corresponding to the subset $S \in \mathcal{S}$. Then*

$$\mathbb{P}\Big(\hat{\mathbf{Z}}(S) = \mathbf{Z}^*, \forall S \in \mathcal{S}\Big) \ge 1 - C_3 K/n,$$

where $C_3 > 0$ is some constant.

The proof of Theorem 1 is presented in Appendix **??**. The stated probability in Theorem 1 holds uniformly over all feature subsets in $\mathcal{S}$. This ensures the validity of iterative algorithms that solve a sequence of SDPs corresponding to the sequence of selected feature subsets: once the algorithm lands on a 'good' S, that step ensures accurate clustering. The assumption $m \ge C_1 n/\log n$ prevents overly imbalanced clusters. Such conditions are essential for theoretical guarantees in clustering and are similarly imposed in previous works (Azizyan et al. 2013, Cai et al. 2019, Chen & Yang 2021, Chen & Zhang 2024, Jin & Wang 2016, Löffler et al. 2021, Verzelen & Arias-Castro 2017, Giraud & Verzelen 2019, Royer 2017, Even et al. 2025). By definition of m as a harmonic average, the restriction on m implies a

restriction on the cluster size $K \lesssim \log n$, which allows a known K to diverge slowly with n. This restriction is milder than most of the previous literature, which often requires K bounded above by a constant (Verzelen & Arias-Castro 2017, Jin & Wang 2016).

Next, we show the optimality of result in Theorem 1.

Theorem 2. *Consider the regime* $(s \log p)/n = o(1)$. *For* $n \geq 2$, *no clustering rule can guarantee exact recovery in the worst case over the class*

$$\overline{\Theta} := \Theta\Big(s, C_4\sigma^2 \log n, K\Big),$$

with C_4 *some small positive constant. In other words,*

$$\inf_{\hat{\mathbf{H}}:\mathbb{R}^{p\times n}\to\mathcal{H}(K)} \sup_{\boldsymbol{\theta}\in\overline{\Theta}} \mathbb{P}_{\boldsymbol{\theta}}\Big(\hat{\mathbf{H}}(\mathbf{X}) \neq \mathbf{H}\Big) \geq C_5,$$

with C_5 *a positive constant.*

The proof of Theorem 2 is presented in Appendix **??**. Theorem 2 implies that no clustering rule can provide a worst-case guarantee of exact recovery across all subsets S whose signal strength is of order $\sigma^2 \log n$. Under the regime of $(|S| \log p)/m \lesssim \log n$, the separation condition in Theorem 1 (i.e., lower bound for $\Delta^2_{S\cap S_0}$) is of order $\log n$ and thus matches the bound in Theorem 2 up to a constant, demonstrating the minimax optimality of Theorem 1 result.

Theorems 1 and 2 also indicate that, in the sparse regime where $|S_0| \ll p$, feature selection can be critical for clustering accuracy. To see this, consider the simple scenario where $K = 2$, $|G_1^*| = |G_2^*|$ and $(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)_{S_0} = \mu_0 \mathbf{1}_{|S_0|}$. Then Theorems 1 and 2 prove that for any set S satisfying $|S| \lesssim \sqrt{p}$ and $|S| \lesssim (n \log n)/\log p$, exact recovery is possible if and only if

$$|S \cap S_0|\mu_0^2 \gtrsim \log n + \frac{|S| \log p}{n} + \sqrt{\frac{|S| \log p}{n}} \asymp \log n.$$

This condition becomes very stringent if S is a very noisy estimate of S_0 (i.e., missing too many signal variables). It also suggests that a slight overestimate of S_0 may not be very

damaging. These motivate us to design an algorithm that incorporates some conservative feature selection method into clustering.

Despite the importance of feature selection, an important message that we want to convey through Theorems 1 and 2 is that, recovering the exact sparse set is not essential for accurate clustering. Instead, identifying some sparse feature set that has sufficiently large signal strength is key for good clustering performance. Such a "good" feature set can be any S in $\mathcal{S}$, which naturally motivate us to consider the iterative algorithm to be presented in the next subsection, where we do not aim at recovering any fixed single S in $\mathcal{S}$ but focusing on getting some set in it.

3 Iterative Block Coordinate Optimization Framework

To address the computational intractability of the best-subset SDP in (6), and motivated by our theoretical results showing that iterative SDP solutions are statistically robust to imperfect feature selection, we propose a block coordinate optimization framework which alternates between feature selection and clustering. For simplicity of exposition, we assume that $\mathbf{\Omega}^*$ is known. The framework initializes by applying spectral clustering to the full feature set to obtain an initial cluster assignment $\hat{G}^0$. Subsequently, the framework alternates between the following blocks:

- **Selection block:** Given the current cluster assignment $\hat{G}^{t-1}$, the integer-constrained selection problem reduces to evaluating feature importance. Each feature is i) scored based on its observed mean difference between the clusters and ii) selected according to the score. We denote the resulting feature set by $\hat{S}^t$.
- **Clustering block:** With the feature set $\hat{S}^t$ fixed, we (i) apply the transformation $\tilde{\mathbf{X}} = \mathbf{\Omega}^*\mathbf{X}$ and (ii) solve the subset SDP clustering problem (Algorithm 1). This yields

the updated estimate:

$$\hat{G}^t = \texttt{SDPclust}(\tilde{\mathbf{X}}_{\hat{S}^t,\cdot}, \mathbf{\Sigma}^*_{\hat{S}^t,\hat{S}^t}).$$

The updated cluster estimates are then fed into the subsequent feature selection step.

These two blocks are illustrated as large dotted boxes in Figure 1. The specific form of the proposed algorithm depends on the particular procedures implemented within these two blocks.

Within the selection block, one must specify both a feature importance scoring rule and a feature selection rule. We consider two approaches: (i) greedy screening and (ii) Thompson sampling. Figure 1 provides a schematic overview of the two algorithms, shown on the left and right sides of the selection block, respectively. Section 3.1 first introduces a greedy screening algorithm as a simple and effective baseline. Section 3.2 identifies its main limitation: the premature exclusion of weak but genuine signals due to multiple testing, a phenomenon that becomes pronounced in ultra-high-dimensional regimes. To address this issue, Section 3.3 propose a Thompson sampling–based feature update that incorporates controlled exploration, improving robustness when early screening decisions are unstable.

We first present these algorithms under the known covariance setting, which covers the scenarios considered in Theorems 1 and 2. Extensions to the unknown covariance case are provided in Section 3.5. Figure 1 offers a schematic overview of both settings, illustrated on the left and right sides of each algorithm within the clustering block, respectively. For clarity, we focus on the case $K = 2$, though the methods naturally generalize to larger K.

3.1 Greedy Screening via Permutation Testing

This section introduces the greedy screening rule for evaluating feature importance within the selection block in the framework of Section 3. Given the cluster estimates from the previous iteration, the selection block consists of two substeps:

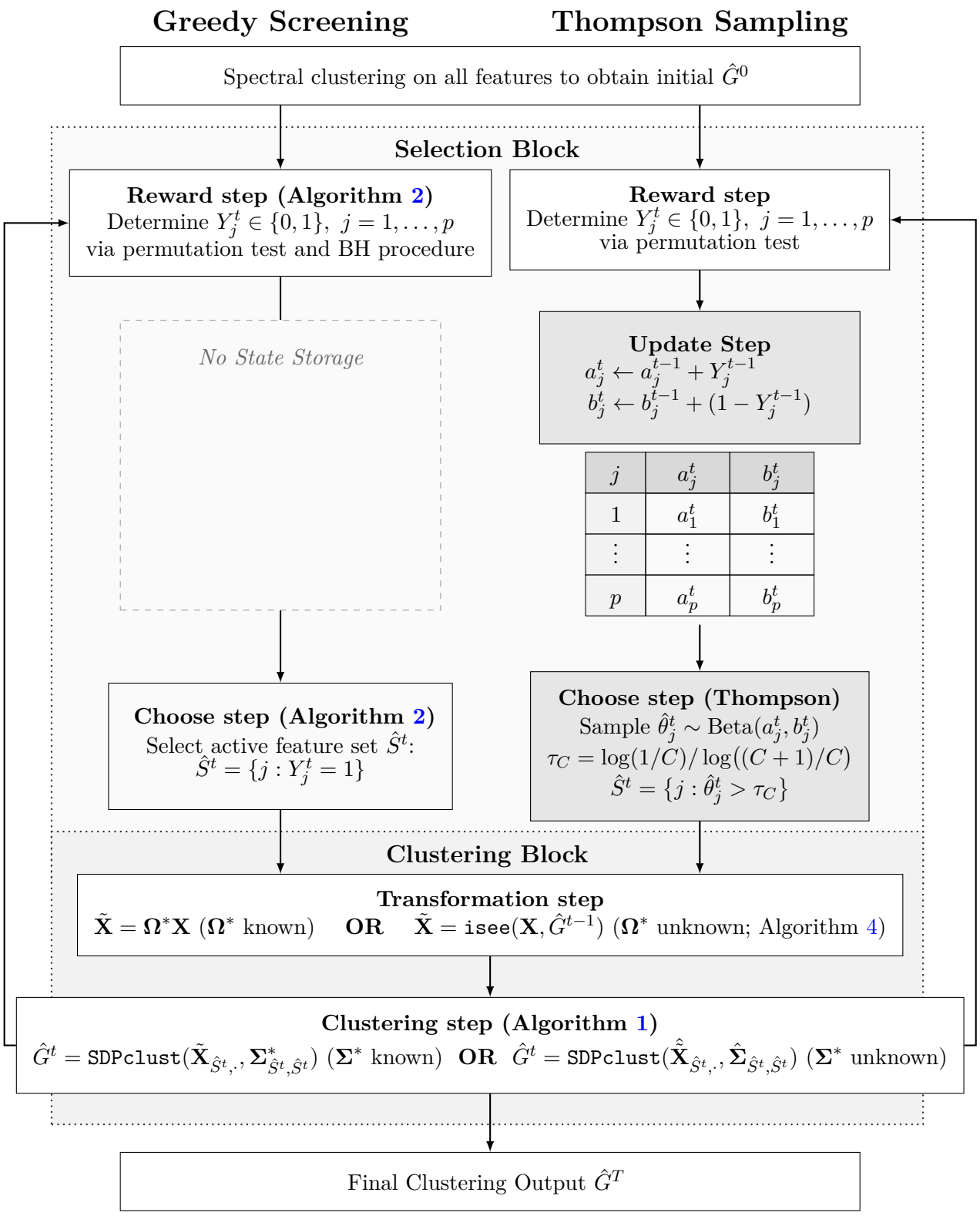


Figure 1: Schematic comparison of the block coordinate ascent algorithms.

- **Reward step.** Each feature is rewarded (or scored) with Y_j^t based on its cluster separation, evaluated via a permutation test with false discovery rate control.

- **Choose step.** All features with $Y_j^t = 1$ are deterministically selected.

Given the straightforward nature of the choose step, our discussion focuses primarily on the probabilistic reward step. The complete selection block is presented in Algorithm 2, and the full iterative procedure is illustrated schematically in the left panel of Figure 1.

The success of feature selection hinges on signal strength. In the related but different problem of classification, it has been recognized (Fan et al. 2013, 2015) that the innovated transformation $\tilde{\mathbf{X}} = \boldsymbol{\Omega}^*\mathbf{X}$ maximally boosts the signal strength of important features for a general precision matrix that may not be identity matrix. See Fan et al. (2013) for detailed arguments and optimality study for the innovated transformation in Gaussian classification. Our algorithms work with the innovated transformation $\tilde{\mathbf{X}}$ (or its estimate when $\boldsymbol{\Omega}^*$ unknown, in Section 3.5).

As argued in Cai et al. (2019), Gaussian clustering is closely related to classification, where the optimal decision rule is the Bayes rule defined by the linear boundary $\mathbf{x}^T\boldsymbol{\beta}^*$, with $\boldsymbol{\beta}^* := \boldsymbol{\Omega}^*(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)$. The magnitudes of the entries in $\boldsymbol{\beta}^*$ directly reflect feature importance. If the true cluster assignment G^* were known, $\boldsymbol{\beta}^*$ could be accurately estimated by the sample mean difference of the innovated data vectors. However, since the clusters are unknown, we iteratively estimate $\boldsymbol{\beta}^*$ based on the current cluster assignment $\hat{G}^{t-1}$.

Rather than relying on analytical concentration inequalities to determine a cutoff threshold—which often require strict distributional assumptions and precise knowledge of nuisance parameters—we adopt a robust, non-parametric permutation test. At iteration t, we define the observed test statistic for feature j as the magnitude of the mean difference between

the current estimated clusters:

$$T_j^{obs} = \left| (\bar{\tilde{\mathbf{X}}}_{\hat{G}_1^{t-1}} - \bar{\tilde{\mathbf{X}}}_{\hat{G}_2^{t-1}})_j \right|. \tag{12}$$

To assess the significance of this statistic, we approximate the null distribution under the hypothesis of no association between the feature and the cluster labels. We generate B random permutations of the cluster labels, denoted as

$$\left\{ \left(\pi_b(\hat{G}_1^{t-1}), \pi_b(\hat{G}_2^{t-1}) \right) \right\}_{b=1}^{B},$$

preserving the cluster sizes. For each permutation b, we compute the null statistic $T_j^{(b)}$ and derive the empirical p-value for feature j:

$$p_j^t = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1} \left(T_j^{(b)} \geq T_j^{obs} \right).$$

Since we test p features simultaneously, a fixed significance level α is insufficient due to the multiplicity problem. Instead, we employ the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR). Given that the cluster assignments in early iterations are merely rough estimates, stringent error control (e.g., FDR ≤ 0.05) is practically overly conservative and may discard weak but informative signals necessary for refinement. Consequently, we adopt a relaxed FDR target of $q = 0.4$, which our simulations suggest effectively balances signal recovery with noise suppression.

Let $j_{(1)}, \ldots, j_{(p)}$ denote the indices of the features sorted by their p-values in ascending order, such that $p_{j_{(1)}}^t \leq \cdots \leq p_{j_{(p)}}^t$. We determine the rejection threshold index k^* as:

$$k^* = \max \left\{ k : p_{j_{(k)}}^t \leq \frac{k}{p} q \right\}.$$

Formally, the active set for the next iteration is updated as:

$$\hat{S}^t := \left\{ j_{(1)}, \ldots, j_{(k^*)} \right\}. \tag{13}$$

This data-driven approach allows the algorithm to adaptively identify features that exhibit strong separation relative to the current noise level, bypassing the need for manual tuning of theoretical constants.

Algorithm 2 Greedy Feature Screening (Selection Block)

Require: Transformed data $\tilde{\mathbf{X}}$, Clusters $\hat{G}^{t-1}$, Permutations B, FDR target q

1: $T_j^{obs} \leftarrow |(\bar{\tilde{\mathbf{X}}}_{\hat{G}_1} - \bar{\tilde{\mathbf{X}}}_{\hat{G}_2})_j|$ for $j = 1, \ldots, p$ ▷ Compute observed statistics

2: **for** $b = 1, \ldots, B$ **do** ▷ Generate null distribution

3: Generate permuted cluster labels to obtain $\pi_b(\hat{G}_1^{t-1}), \pi_b(\hat{G}_2^{t-1})$

4: $T_j^{(b)} \leftarrow |(\bar{\tilde{\mathbf{X}}}_{\pi_b(\hat{G}_1^{t-1})} - \bar{\tilde{\mathbf{X}}}_{\pi_b(\hat{G}_2^{t-1})})_j|$ for $j = 1, \ldots, p$

5: $p_j \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(T_j^{(b)} \geq T_j^{obs})$ for all j ▷ Compute empirical p-values

6: Sort p-values such that $p_{(1)} \leq \cdots \leq p_{(p)}$

7: $k^* \leftarrow \max\{k : p_{(k)} \leq \frac{k}{p} q\}$ ▷ Benjamini-Hochberg threshold

8: $Y_j^t \leftarrow 1$ for $j = j_{(1)}, \ldots, j_{(k^*)}$; $Y_j^t \leftarrow 0$ for $j \neq j_{(1)}, \ldots, j_{(k^*)}$ ▷ Reward step

9: **return** $\hat{S}^t \leftarrow \{j \in [p] : Y_j^t = 1\}$ ▷ Selection step

3.2 Limitations of Greedy Screening and the Need for Exploration

Block coordinate optimization by greedy screening (Algorithm 2) provides a simple and computationally attractive baseline which alternates between feature selection and updating the clustering by solving an SDP K-means instance on the selected features. Empirically, this greedy update can be effective when the initial clustering is already moderately accurate and the marginal feature signals are strong. However, in the high-dimensional regimes of interest, the greedy mechanism can stall, and motivates a feature-update rule with *memory* and *active exploration*.

The feature screening step relies on the observed mean differences T_j^{obs} in (12) computed under the current estimated labels. When the estimated cluster label $\hat{G}^{t-1}$ is inaccurate,

the effective signal in T_j^{obs} can be substantially attenuated, making the permutation test underpowered even for truly informative coordinates. This is problematic because our fixed-subset SDP theory (Section 2.3) shows that exact recovery requires that feature subset S contain enough total signal, that is, $|S \cap S_0|$ must be large enough, while also indicating that moderate over-selection is often far less damaging than missing signal variables (see Theorems 1–2 and the discussion thereafter). Greedy screening can therefore fail precisely in the regime where a conservative bias toward retaining/revisiting borderline coordinates would be beneficial.

The greedy update $\hat{S}^t$ in (13) depends only on the current iteration's p-values. In particular, a coordinate that repeatedly appears *nearly significant* under noisy cluster assignments does not accumulate evidence across iterations; it is treated from scratch each time. This lack of memory can yield a self-reinforcing loop of poor labels $\Rightarrow$ low test power $\Rightarrow$ under selection of S_0 $\Rightarrow$ and poor SDP estimated labels. While one can partially mitigate this by using a permissive FDR target (we use $q = 0.4$), increasing q also increases false positives and can inflate $|\hat{S}^t|$, potentially violating the cardinality regime (e.g. $|\hat{S}^t| \leq \sqrt{p}$) under which uniform SDP recovery on the good-set family $\mathcal{S}$ is guaranteed.

The above considerations suggest that the selection block should: (i) *remember* historical evidence about each feature's usefulness, and (ii) *occasionally re-select* uncertain/borderline features even if they are not currently significant, so that improved labels can later validate them. This motivates replacing the memoryless greedy update (13) by a Thompson-sampled feature update that maintains feature-wise posteriors and performs randomized selection based on posterior draws. Crucially, the clustering subroutine `SDPclust` (Algorithm 1) and the permutation-test feedback remain unchanged; only the feature scoring and selection rules are upgraded. Section 3.3 develops this Thompson-sampling refinement and Section 3.4 presents its theoretical guarantees.

3.3 Thompson-Sampled Feature Updates

Without applying FDR control via the BH procedure, the outcome of the permutation test can be interpreted, conditional on a data instance sampled from model (1), as a Bernoulli random variable. Its success probability depends both on the feature index j and on the currently estimated cluster id partition $G = (G_1, G_2)$. Formally, we have the following definition of Bernoulli reward.

we need to unify notations in this and the next few subsections. I made an atmmept on it

Definition 3 (Bernoulli reward)**.** *Fix any randomized testing rule $\varphi = \{\varphi_j\}_{j=1}^p$ such that for any partition $G = (G_1, G_2)$ and any feature index j, $\varphi_j(G; U) \in \{0, 1\}$ is measurable in the data and an auxiliary random seed U (e.g., the random permutations used to select features). Given a partition G, define the corresponding reward as $Y_j(G) := \varphi_j(G; U)$, $j \in [p]$. We write the arm mean conditional on $\mathbf{X}$ and G as $\theta_j(G) := \mathbb{P}\Big(Y_j(G) = 1 \,\Big|\, \mathbf{X}, G\Big)$, where the probability is taken over the test randomization U.*

If at each iteration t, the algorithm uses i.i.d. test randomization U^t, then conditional on $(\mathbf{X}, G)$ the rewards $\{Y_j^t(G)\}_{t \geq 1}$ are i.i.d. Bernoulli$(\theta_j(G))$ for each fixed j.

Let $G^* = (G_1^*, G_2^*)$ denote the oracle partition. Recall Definition 3. The oracle arm mean is $\theta_j^{\text{orc}} := \theta_j(G^*)$. In this subsection we use *individual* permutation tests (without FDR control), so θ_j^{orc} is feature-wise.

To achieve feature selection, we define the population-level utility as the sum of the these arm means, while penalizing weak features. Specifically, for a given set $S \subset [p]$, motivated by Liu & Ročková (2023), we define the *oracle* utility function as follows with penalty formalized through a regularization parameter $C > 0$:

$$r_C^{\text{orc}}(S) := \sum_{j \in S} \left\{ \theta_j^{\text{orc}} \log\left(\frac{C+1}{C}\right) + \log C \right\}. \tag{14}$$

Including a feature j in S increases the utility $r_C^{\mathrm{orc}}(S)$ iff $\theta_j^{\mathrm{orc}} > \tau_C$, where

$$\tau_C := \frac{\log(1/C)}{\log\left(\frac{C+1}{C}\right)}. \tag{15}$$

We define the oracle target subset

$$S_C^\star := \arg\max_{S\subseteq[p]} r_C^{\mathrm{orc}}(S) = \{j \in [p] : \theta_j^{\mathrm{orc}} > \tau_C\}, \tag{16}$$

where the last equality follows from additivity under individual tests.

Here, C governs the trade-off between sparsity and signal capture: larger values of C lower the threshold τ_C, promoting exploration and the inclusion of larger feature subsets, whereas smaller values of C enforce sparsity. Liu & Ročková (2023) fix C such that $\tau_C = 0.5$; we generalize this approach by considering multiple values of C to explore different sparsity–signal trade-offs.

Now we formally introduce the Thompson sampling screening rule for evaluating feature importance within the selection block in the framework of Section 3. In implementation, G^* and θ_j^{orc}'s are unknown. In our iterative algorithm they can be replaced with the estimates from the previous iteration. At iteration t, we compute an estimated partition $\hat{G}^t$ from the SDP step and evaluate rewards using $\hat{G}^t$:

$$Y_j^t = Y_j(\hat{G}^t) = \varphi_j(\hat{G}^t; U^t), \qquad \mathbb{E}\left[Y_j^t \mid \mathbf{X}, \hat{G}^t\right] = \theta_j(\hat{G}^t).$$

When $\hat{G}^t = G^*$ (up to label permutation), the algorithmic means coincide with the oracle means: $\theta_j(\hat{G}^t) = \theta_j(G^*) = \theta_j^{\mathrm{orc}}$. This "oracle coupling" is formalized in a later section and is the key link between the non-oracle algorithm and the oracle analysis.

The algorithm maintains Beta posteriors $\mathrm{Beta}(a_j^t, b_j^t)$ for each feature, where a_j^t and b_j^t track the accumulated counts of significant (reward $Y_j^t = 1$) and non-significant (reward $Y_j^t = 0$) tests, respectively. At iteration t:

- **Reward& Update step.** Compute the reward Y_j^t using permutation test. Then update

$$a_j^t = a_{j,t-1} + Y_{j,t-1}, \qquad b_j^t = b_{j,t-1} + (1 - Y_{j,t-1}), \qquad j \in \hat{S}^t,$$

and keep $(a_j^t, b_j^t) = (a_{j,t-1}, b_{j,t-1})$ for $j \notin \hat{S}^t$.

- **Choose step.** Sample $\hat{\theta}_j^t \sim \text{Beta}(a_j^t, b_j^t)$ and set $\hat{S}^t = \{j : \hat{\theta}_j^t > \tau_C\}$.

- **Cluster step.** Run the SDP clustering subroutine on $\hat{S}^t$ to obtain $\hat{G}^t$.

double check the index of t The full iterative clustering procedure is illustrated schematically in the right panel of Figure 1.

Since Assumption 2 posits the existence of a specific constant C that ensures identifiability, we adopt a grid search strategy: we execute the algorithm across a range of C values and select the solution that maximizes the clustering objective. This approach is empirically justified by Figure 6, which demonstrates that for a fixed signal set S_0, there exists a specific choice of C that induce threshold τ_C capable of effectively distinguishing the posterior θ estimates of signal features from noise.

3.4 Theoretical Guarantee

Now we present theoretical guarantee of the proposed bandit-based algorithm. We start with introducing an important assumption.

3.4.1 Oracle coupling between SDP exact recovery and Thompson sampling

We introduce Assumption 2 below which provides an *oracle-coupling* argument that explicitly leverages the uniform SDP exact-recovery event of Theorem 1. The key observation is that, on the SDP-good event, whenever the selected feature set belongs to the *good-set family* $\mathcal{S}$, the SDP step returns the true partition; therefore the subsequent testing step produces

oracle rewards whose success probabilities are *context-free*.

Throughout this subsection we focus on $K = 2$ for clarity, as in Algorithm 3. Let G_1^*, G_2^* denote the true partition, and let $\mathbf{Z}^*$ denote the corresponding cluster matrix. Recall the good-set family $\mathcal{S}$ in (11), and define the SDP-good event

$$\mathcal{E}_{\text{SDP}} \;:=\; \Big\{ \widehat{\mathbf{Z}}(S) = \mathbf{Z}^*,\ \forall S \in \mathcal{S} \Big\}. \tag{17}$$

By Theorem 1, $\mathbb{P}(\mathcal{E}_{\text{SDP}}) \geq 1 - C_3 K/n$.

Lemma 4. *Assume $\mathcal{E}_{\text{SDP}}$ holds. Consider any iteration t of Algorithm 3 in which the selected feature set $\widehat{S}^t$ belongs to $\mathcal{S}$. Then the SDP step returns $\widehat{\mathbf{Z}}(\widehat{S}^t) = \mathbf{Z}^*$ and, consequently, the partition produced by Algorithm 1 satisfies $(\widehat{G}_1^t, \widehat{G}_2^t) = (G_1^*, G_2^*)$ up to label permutation. In particular, for every $j \in \widehat{S}^t$, $Y_j^t = Y_j(\widehat{G}^t) = Y_j(G^*) = Y_j^{\text{orc}}$, and hence $\theta_j(\widehat{G}^t) = \theta_j^{\text{orc}}$.*

Proof. On $\mathcal{E}_{\text{SDP}}$, we have $\widehat{\mathbf{Z}}(S) = \mathbf{Z}^*$ for every $S \in \mathcal{S}$. Therefore if $\widehat{S}^t \in \mathcal{S}$ then $\widehat{\mathbf{Z}}(\widehat{S}^t) = \mathbf{Z}^*$. The clustering step in Algorithm 1 yields $(\widehat{G}_1^t, \widehat{G}_2^t)$ equal to (G_1^*, G_2^*) up to label permutation. Since $Y_j(\cdot)$ in Definition 3 depends on data only on the induced grouping of indices, the resulting rewards agree with oracle rewards. □

The lemma above shows that *whenever* the selected feature set lies in $\mathcal{S}$ (and the SDP-good event holds), the bandit feedback is generated by *context-free* oracle means $\{\theta_j^{\text{orc}}\}_{j=1}^p$. We therefore impose signal strength condition directly at the oracle level with $G = G^*$.

Assumption 1 (Oracle margin)**.** *Fix $C > 0$ and define τ_C as in (15). Assume there exists $\alpha \in (0, 1 - \tau_C)$ such that*

$$\min_{j \in S_0} \theta_j^{\text{orc}} \;\geq\; \tau_C + \alpha, \qquad \max_{j \notin S_0} \theta_j^{\text{orc}} \;\leq\; \tau_C - \alpha. \tag{18}$$

Assumption 1 relaxes the corresponding condition in Liu & Ročková (2023) in the sense that it is not imposed for arbitrary misclustered partitions. Rather, it requires a separation

of testing success probabilities only under the true labels. Lemma 4 then transfers this separation to any iteration whose selected set lies in $\mathcal{S}$ on the SDP-good event.

3.4.2 Thompson Sampling Guarantees for Feature Selection

need to remove Assumption 2 2 and adapt the proof using Assumption 1

Assumption 2 (Strong Identifiability)**.** *Assume there exists a constant $C > 0$ and a separation margin $\alpha \in (0, 1-\tau_C)$, and feature subset $S^\star$ which satisfy the following conditions:*

- *For all signal features $i \in S_C^\star$, the inequality $\theta_i(S) > \tau_C + \alpha$ holds for any subset S containing i.*
- *For all noise features $i \notin S_C^\star$, the inequality $\theta_i(S) < \tau_C - \alpha$ holds for any subset S containing i.*

Under this assumption, the set $S^\star$ is the unique maximizer of the regularized objective function:

$$S_C^\star = \arg\max_S r_C(S) = \arg\max_S \sum_{j \in S} \left\{ \theta_j^*(S) \log\left(\frac{C+1}{C}\right) + \log C \right\}.$$

The quality of the sequence of feature subsets selected by our algorithm, denoted by $\{\hat{S}_t\}_{t=1}^T$, is measured by total expected regret over a time horizon T. This is defined as:

$$\mathrm{Reg}^t = \mathbb{E}\left[\sum_{t=1}^T \left\{ r_C(S_C^\star) - r_C(\hat{S}_t) \right\}\right] = \sum_{t=1}^T \underbrace{\mathbb{E}\left[r_C(S_C^\star) - r_C(\hat{S}_t) \right]}_{:=\Delta_{\hat{S}_t}} = \sum_{t=1}^T \Delta_{\hat{S}_t}.$$

The expectation $\mathbb{E}[\cdot]$ accounts for the inherent stochasticity of the bandit framework, specifically arising from two sources:

- **Posterior Sampling:** The subset selection $\hat{S}_t$ depends on $\hat{\theta}_{i,t}$, which is a random draw from the posterior $\mathrm{Beta}(a_{i,t-1}, b_{i,t-1})$ distribution.
- **Reward Mechanism:** The feedback $Y_{i,t}$ is a Bernoulli random variable generated by the randomized permutation tests during the Reward Step.

Theorem 5 (Regret Bound for General TVS)**.** *Under Assumption 2, the cumulative regret of the TVS policy with regularization constant $C \in (0,1)$ scales logarithmically with the time horizon T:*

$$Reg^t \leq \Delta_{\max}\left[\frac{8p\log T}{\alpha^2} + |S^\star| J_{crit}\left(\frac{1}{(1-\tau_C)\alpha} + \frac{1+\tau_C}{1-\tau_C}\right) + p\left(2+\frac{4}{\alpha^2}\right) + |S^\star|\mathcal{R}(\alpha,\tau_C)\right],$$

where p is the total number of features, $J_{crit} = \max\left\{\frac{8}{\alpha}, \frac{(1+\tau_C)^2}{\tau_C(1-\tau_C)} - 1\right\}$, and $\mathcal{R}(\alpha,\tau_C)$ is a constant remainder independent of T given by:

$$\mathcal{R}(\alpha,\tau_C) = \frac{e^{-\alpha^2 J_{crit}/4}}{(1-e^{-\alpha^2/4})^2} + \frac{2\tau_C^2}{\alpha^2(1-e^{-\alpha^2})} + \frac{10-8\tau_C}{(1-\tau_C)(1-e^{-\alpha^2/2})}.$$

The proof of Theorem 5 is provided in Appendix **??**. This regret bound theorem in turn implies the variable selection consistency:

Theorem 6 (Variable Selection Consistency)**.** *Assume there exist $S^\star$, C, α that satisfy Assumption 2. Then the Bandit-based Thompson Sampling policy identifies the true feature set S^* almost surely:*

$$\mathbb{P}\left(\liminf_{t\to\infty} \hat{S}_t = S^*\right) = 1. \tag{19}$$

The proof of Theorem 6 is provided in Appendix **??**. The following corollary synthesizes our minimax separation analysis with the consistency properties of the Thompson Sampling mechanism to provide a clustering consistency guarantee for the proposed algorithm.

Corollary 7 (Exact Cluster Recovery)**.** *Suppose there exists a feature subset $S^\dagger \subseteq [p]$ and constants C and α (not necessarily unique) such that the following conditions are satisfied:*

1. ***Total Signal:*** *The subset $S^\dagger$ satisfies the geometric requirements for exact recovery established in Theorem 1:*

$$|S^\dagger| \leq \sqrt{p} \quad \textit{and} \quad \Delta^2_{S^\dagger \cap S_0} \gtrsim \sigma^2\left(\log n + \frac{|S^\dagger|\log p}{m} + \sqrt{\frac{|S^\dagger|\log p}{m}}\right). \tag{20}$$

2. ***Individual Signal:*** *The problem parameters satisfy Assumption 2, ensuring that the informative feature set is identifiable via the posterior inclusion threshold* τ_C.

Then, in the isotropic Gaussian setting, the ***Block Coordinate Ascent with Thompson Sampling*** *algorithm achieves exact cluster recovery almost surely as* $t \to \infty$.

3.4.3 Exact clustering recovery by the iterative algorithm

Assumption 3. *Let* $\{\hat{S}_t\}_{t\geq 1}$ *be the sequence produced by Algorithm 3. Assume the event*

$$\mathcal{E}_{\text{tail}} := \left\{ \exists\, T_0 < \infty \text{ such that } \hat{S}_t \in \mathcal{S} \ \forall t \geq T_0 \right\} \tag{21}$$

holds with probability at least $1 - \delta_{\text{tail}}$ *for some* $\delta_{\text{tail}} \in [0, 1)$.

Assumption 3 is analogous to standard "basin of attraction" assumptions used in iterative nonconvex procedures.

Theorem 8. *Assume Theorem 1, Assumptions 1 and 3. Then on the event* $\mathcal{E}_{\text{SDP}} \cap \mathcal{E}_{\text{tail}}$, *Algorithm 3 satisfies the following:*

(i) *For* $t \geq T_0$, *the rewards satisfy* $\theta_i^*(\hat{S}_t) = \theta_i^{\text{orc}}(\hat{S}_t)$ *for all* $i \in \hat{S}_t$. *In particular, Assumption 2 holds on the tail (for* $t \geq T_0$*) with the same* $(C, \alpha, S_C^\star)$ *as in Assumption 1.*

(ii) *Applying Theorem 6 to the time-shifted process* $\{\hat{S}_t\}_{t\geq T_0}$ *yields*

$$\mathbb{P}\Big(\liminf_{t\to\infty} \hat{S}_t = S_C^\star \;\Big|\; \mathcal{E}_{\text{SDP}} \cap \mathcal{E}_{\text{tail}} \Big) = 1. \tag{22}$$

(iii) *If, moreover, the limiting set satisfies* $S_C^\star \in \mathcal{S}$, *then on* $\mathcal{E}_{\text{SDP}} \cap \mathcal{E}_{\text{tail}}$, $\hat{\mathbf{Z}}(\hat{S}_t) = \mathbf{Z}^*$ *for all sufficiently large* t, *and hence we obtain the exact cluster recovery up to label permutation.*

Proof. On $\mathcal{E}_{\text{tail}}$, there exists $T_0 < \infty$ such that $\hat{S}_t \in \mathcal{S}$ for all $t \geq T_0$. On $\mathcal{E}_{\text{SDP}}$, Lemma 4 applies at every $t \geq T_0$, yielding $\theta_i^*(\hat{S}_t) = \theta_i^{\text{orc}}(\hat{S}_t)$, proving (i). Combining (i) with

Assumption 1 shows that the strong identifiability inequalities required in Assumption 2 hold for all $t \geq T_0$. Since the Beta parameters (a_{i,T_0}, b_{i,T_0}) are strictly positive, Theorem 6 applies to the time-shifted process, proving (ii). Finally, if $S_C^\star \in \mathcal{S}$, then by (ii) we have that for all sufficiently large t, the selected set $\hat{S}_t$ converges to $S_C^\star$ (in the liminf sense) and lies in $\mathcal{S}$ by $\mathcal{E}_{\text{tail}}$; hence $\hat{\mathbf{Z}}(\hat{S}_t) = \mathbf{Z}^*$ for all large t on $\mathcal{E}_{\text{SDP}}$, proving (iii). □

depending on space, we may move this to appendix We now give a fully explicit sufficient condition for Assumption 1 in a standard Gaussian setting. For simplicity we state it for a parametric z-test.

Proposition 1. *Assume the (known covariance) Gaussian mixture model (1) with $K = 2$ and let $n_k := |G_k^*|$, and m as in (11) so that $(1/n_1 + 1/n_2) = 2/m$. Let $\tilde{\mathbf{X}} = \mathbf{\Omega}^*\mathbf{X}$ and define the oracle mean-difference statistic $D_j^{\text{orc}} := \left(\bar{\tilde{\mathbf{X}}}_{G_1^*} - \bar{\tilde{\mathbf{X}}}_{G_2^*}\right)_j$. Consider the following z-test at nominal level $u \in (0,1)$:*

$$\varphi_j(G^*) \;=\; \mathbb{1}\left(\frac{|D_j^{\text{orc}}|}{\sqrt{\Omega_{jj}^*(1/n_1 + 1/n_2)}} \geq z_{1-u/2}\right), \qquad z_q := \Phi^{-1}(q)$$

with Φ the standard Gaussian CDF. Then:

(a) *If $j \notin S_0$ (equivalently $(\mathbf{\Omega}^*(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*))_j = 0$), then $\theta_j^{\text{orc}} = u$.*

(b) *If $j \in S_0$ and we define the standardized effect size*

$$\gamma_j \;:=\; \frac{|(\mathbf{\Omega}^*(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*))_j|}{\sqrt{\Omega_{jj}^*(1/n_1 + 1/n_2)}} \;=\; \sqrt{\frac{m}{2}} \cdot \frac{|(\mathbf{\Omega}^*(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*))_j|}{\sqrt{\Omega_{jj}^*}},$$

then the oracle success probability satisfies the explicit lower bound

$$\theta_j^{\text{orc}} \;=\; \mathbb{P}\Big(|N(\gamma_j, 1)| \geq z_{1-u/2}\Big) \;\geq\; 1 - \Phi\Big(z_{1-u/2} - \gamma_j\Big).$$

Consequently, if we choose $u \leq \tau_C - \alpha$ and assume

$$\min_{j \in S_0} \gamma_j \;\geq\; z_{1-u/2} + \Phi^{-1}(\tau_C + \alpha), \tag{23}$$

then Assumption 1 holds.

Proof. Under the model, conditional on G^*, D_j^{orc} is Gaussian with mean $(\boldsymbol{\Omega}^*(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*))_j$ and variance $\Omega_{jj}^*(1/n_1 + 1/n_2)$. Standardization yields a $N(\gamma_j, 1)$ distribution. Part (a) follows because under $\gamma_j = 0$ the two-sided z-test has exact level u. Part (b) follows by discarding the lower tail: $\mathbb{P}(|N(\gamma_j, 1)| \geq z) \geq \mathbb{P}(N(\gamma_j, 1) \geq z) = 1 - \Phi(z - \gamma_j)$. If $u \leq \tau_C - \alpha$, then all noise features satisfy $\theta_j^{\mathrm{orc}} = u \leq \tau_C - \alpha$. If (23) holds, then for every signal feature $j \in S_0$,

$$1 - \Phi(z_{1-u/2} - \gamma_j) \;\geq\; 1 - \Phi\Big(-\Phi^{-1}(\tau_C + \alpha)\Big) \;=\; \tau_C + \alpha,$$

establishing the oracle margin (18). □

3.5 Unknown Covariance Case

We now address the unknown covariance scenario within the clustering block by adapting the estimation strategy proposed in Fan & Lv (2016). A primary advantage of this approach is that it circumvents the need to estimate the full precision matrix $\boldsymbol{\Omega}^*$. Focusing specifically on the transformation step illustrated in Figure 1, at the t-th iteration, we construct an estimate of the innovated data matrix $\tilde{\mathbf{X}}$ directly, based on the current cluster assignments $\hat{G}^{t-1}$. This estimated matrix is subsequently utilized as the input for the `SDPcluster` procedure (Algorithm 1).

The direct estimation of $\tilde{\mathbf{X}}$ is done via the Innovated Scalable Efficient Estimation (ISEE) procedure (Fan & Lv 2016), originally developed for high-dimensional sparse precision matrix estimation. We modify it to selectively estimate only the required quantities for our algorithm, avoiding full precision matrix recovery, and use its theory to set the threshold for feature selection. Section 3.5.1 summarizes the ISEE procedure adapted as a subroutine in our method, followed by the full algorithm in Section **??**.

3.5.1 Innovated Scalable Efficient Estimation (ISEE) Subroutine

The ISEE procedure is motivated by two key observations. First, the transformed vector $\tilde{\mathbf{X}}_i = \mathbf{\Omega}^* \mathbf{X}_i \in \mathbb{R}^p$ remains Gaussian, with $\mathrm{Cov}(\tilde{\mathbf{X}}_i) = \mathbf{\Omega}^* \mathbf{\Sigma}^* \mathbf{\Omega}^* = \mathbf{\Omega}^*$. Without loss of generality, assume p is an even integer, consider a subset $A \subset [p]$ of size 2, and assume $i \in G_1^*$. We consider the following decomposition of $\tilde{\mathbf{X}}_{A,i}$ as the sum of deterministic mean part and mean-zero random noise vector:

$$\tilde{\mathbf{X}}_{A,i} = (\tilde{\boldsymbol{\mu}}_1^*)_A + \tilde{\mathcal{E}}_{A,i}, \tag{24}$$

Using the the conditional distribution property of multivariate Gaussian and Shur complement, we derive the following regression relationship between sub-vectors of $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_1^*, \mathbf{\Sigma}^*)$:

$$\underbrace{\mathbf{X}_{A,i}}_{\text{response} \in \mathbb{R}^{|A|}} = \underbrace{(\boldsymbol{\mu}_1^*)_A + (\mathbf{\Omega}_{A,A}^*)^{-1} \mathbf{\Omega}_{A,A^c}^* (\boldsymbol{\mu}_1^*)_{A^c}}_{:= (\boldsymbol{\alpha}_1)_A \in \mathbb{R}^{|A|} \text{ (intercept)}} - \underbrace{(\mathbf{\Omega}_{A,A}^*)^{-1} \mathbf{\Omega}_{A,A^c}^*}_{\text{slope} \in \mathbb{R}^{|A| \times (p-|A|)}} \underbrace{\mathbf{X}_{A^c,i}}_{\text{predictor} \in \mathbb{R}^{p-|A|}} + \underbrace{\mathbf{E}_{A,i}}_{\text{residual} \in \mathbb{R}^{|A|}}, \tag{25}$$

where $\mathbf{E}_{A,i} \sim \mathcal{N}\big(0, (\mathbf{\Omega}_{A,A}^*)^{-1}\big)$. Leveraging this relationship, the quantities in (24) can be further expressed as follows (Fan & Lv 2016):

$$(\tilde{\boldsymbol{\mu}}_1^*)_A = \mathbf{\Omega}_{A,A}^* (\boldsymbol{\alpha}_1)_A \quad \text{and} \quad \tilde{\mathcal{E}}_{A,i} = \mathbf{\Omega}_{A,A}^* \mathbf{E}_{A,i}. \tag{26}$$

Appendix **??** provides detailed derivations for (25) and (26). Since $p - |A|$ can be comparable or larger than n, one may want to leverage the existing regularization method for fitting the regression model (25) to obtain sample estimates of $\mathbf{\Omega}_{A,A}^*$, $\mathbf{E}_{A,i}$ and $(\boldsymbol{\alpha}_1)_A$, which further yields plug-in estimate of $\tilde{\mathbf{X}}_{A,i}$. This process can be repeated for all $A \in \mathcal{A}$ with $\mathcal{A}$ a partition set of $[p]$ consisting of sets of size 2. By combining estimates for $\tilde{\mathbf{X}}_{A,i}$ for all $A \in \mathcal{A}$, we can form an estimate of the full matrix $\tilde{\mathbf{X}}_{G_1^*} \in \mathbb{R}^{|G_1^*| \times p}$.

A similar expression holds for $i \in G_2^*$, with $\boldsymbol{\mu}_1^*$ and $\boldsymbol{\alpha}_1$ replaced by $\boldsymbol{\mu}_2^*$ and $\boldsymbol{\alpha}_2$, respectively. Following the same steps as described above, we can obtain the estimate of the submatrix $\tilde{\mathbf{X}}_{G_2^*} \in \mathbb{R}^{|G_2^*| \times p}$.

Two issues remain to be addressed. First, the true cluster labels are unknown. Second, we need an implementable ISEE subroutine. Since our algorithm is iterative by nature, we use the cluster assignments in the current iteration as the proxy for true cluster labels to implement the ISEE subroutine, and then update the estimates for $\tilde{\mathbf{X}}$ and cluster assignments in the next iteration. We next explain the ISEE subroutine for a given pair of cluster assignment (G_1, G_2). jongmin:Revising here

The ISEE subroutine is formally defined as follows. The feature indices $[p]$ is partitioned into non-overlapping subsets $A_1, \ldots, A_L$, each of size 2 if p is even. For odd p, we set the size of A_L to 3. The estimation procedure is then applied for each feature subset, and the results are aggregated. Without loss of generality, we describe the steps for the subset A_1. Given the previous cluster estimates $\hat{G}_1^{t-1}$ and $\hat{G}_2^{t-1}$, we treat $\{(\mathbf{X}_{A_1^C,i}, \mathbf{X}_{A_1,i}) : i \in \hat{G}_1^{t-1}\}$ as (predictor, response) pair and apply off-the-shelf high-dimensional regression method (for example, lasso; Tibshirani 1996) to estimate the intercept $(\boldsymbol{\alpha}_1)_{A_1}$ and residuals $\{\mathbf{E}_{A_1,i}\}_{i \in \hat{G}_1^{t-1}}$. Similarly, using $\{(\mathbf{X}_{A_1^C,i}, \mathbf{X}_{A_1,i} : i \in \hat{G}_2^t\}$, we estimate the corresponding $(\boldsymbol{\alpha}_2)_{A_1}$ and residuals $\{\mathbf{E}_{A_1,i}\}_{i \in \hat{G}_2^{t-1}}$ for cluster 2. To estimate $\boldsymbol{\Omega}^*_{A_1,A_1}$, we exploit the fact that the residuals $\mathbf{E}_{A_1,i} \sim \mathcal{N}\Big(0, (\boldsymbol{\Omega}^*_{A_1,A_1})^{-1}\Big)$, and use the inverse of the estimated sample covariance matrix of $\mathbf{E}_{A_1,i}$ as an estimator. Then applying (26), we obtain the estimates for $(\tilde{\boldsymbol{\mu}}_1^*)_{A_1}$, $(\tilde{\boldsymbol{\mu}}_2^*)_{A_1}$, and $\{\tilde{\mathcal{E}}_{A_1,i}\}_{i=1}^n$. Accordingly, using (24), we can estimate $\{\tilde{\mathbf{X}}_{A,i}\}_{i=1}^n$. Repeating this procedure for all subsets $A_1, \ldots, A_L$ and stacking the results yields estimates of $\tilde{\boldsymbol{\mu}}_1^*$, $\tilde{\boldsymbol{\mu}}_2^*$ and $\tilde{\mathbf{X}}$. In summary, we have the output $\hat{\tilde{\mathbf{X}}}^t = \texttt{isee}(\mathbf{X}, \hat{G}^{t-1})$ in the t-th iteration. The full procedure is detailed in Algorithm 4.

Algorithm 4 ISEE Transformation Subroutine (`isee`)

Require: Data $\mathbf{X} \in \mathbb{R}^{p \times n}$, Cluster assignments $\hat{G} = (\hat{G}_1, \hat{G}_2)$

Partition $[p]$ into disjoint subsets $A_1, \ldots, A_L$ of size 2 or 3

for $\ell = 1, \ldots, L$ **do**

  for $k \in \{1, 2\}$ **do** ▷ Estimate parameters for each cluster

    Regress $\mathbf{X}_{A_\ell}$ on $\mathbf{X}_{A_\ell^c}$ using samples $i \in \hat{G}_k$ (e.g., via Lasso)

    Obtain: Intercept $\hat{\boldsymbol{\alpha}}_{(k)}$ and residuals $\{\hat{\mathbf{E}}_{A_\ell, i} : i \in \hat{G}_k\}$

  $\hat{\mathbf{E}}_{A_\ell} \leftarrow [\hat{\mathbf{E}}_{A_\ell, i}]_{i=1}^{n}$ ▷ Pool residuals across clusters

  $\hat{\boldsymbol{\Omega}}_{A_\ell A_\ell} \leftarrow \left(\frac{1}{n}\hat{\mathbf{E}}_{A_\ell}\hat{\mathbf{E}}_{A_\ell}^\top\right)^{-1}$ ▷ Estimate block precision

  for $i = 1, \ldots, n$ **do** ▷ Reconstruct innovated data

    Let k be the cluster index of sample i

    $\hat{\tilde{\mathbf{X}}}_{A_\ell, i} \leftarrow \hat{\boldsymbol{\Omega}}_{A_\ell A_\ell}\left(\hat{\boldsymbol{\alpha}}_{(k)} + \hat{\mathbf{E}}_{A_\ell, i}\right)$

return $\hat{\tilde{\mathbf{X}}} \leftarrow [\hat{\tilde{\mathbf{X}}}_{A_1}^\top, \ldots, \hat{\tilde{\mathbf{X}}}_{A_L}^\top]^\top$

4 NUMERICAL STUDIES

For simulations, the data are generated from Gaussian distributions with two symmetric clusters:

$$\mathbf{X}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*),\ i = 1, \ldots, n/2, \quad \text{and} \quad \mathbf{X}_i \overset{i.i.d.}{\sim} \mathcal{N}(-\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), i = n/2 + 1, \ldots, n. \quad (27)$$

We fix $S_0 = [10]$ and increase the ambient dimension p. We assign equal magnitude to all entries of $\boldsymbol{\mu}^*_{S_0}$, with the magnitude chosen so that the signal strength, measured Mahalanobis distance $\Delta_{\boldsymbol{\Sigma}^*} = 2\sqrt{(\boldsymbol{\mu}^*)^\top \boldsymbol{\Omega}^* \boldsymbol{\mu}^*} = \sqrt{(\boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}^* \boldsymbol{\beta}^*}$, remains constant as p increases. We evaluate performance using the mis-clustering rate, averaged over 200 independent replications. We set the number of iteration as $T = 100$, but used early stopping criterion. The details on this criterion, code implementation, and further information required for

Table 1: Summary of Parameter Settings for (27). $i, j \in [p]$.

Covariance	p	n	Separation
known ($\mathbf{\Sigma}^* = \mathbf{I}_p$)	50-5000	200	$\Delta^2 = 4^2, 5^2$
unknown ($\Omega^*_{i,i} = 1$, $\Omega^*_{i,i+1} = \Omega^*_{i+1,i} = \rho$, $\Omega^*_{i,j} = 0$ o.w., $\rho \in \{0.2, 0.45\}$)	50-400	500	$\Delta^2_{\mathbf{\Sigma}^*} = 3^2, 4^2$

replicating our results are provided in Appendix **??**.

4.1 Simulation under Known Covariance

We evaluate Algorithm 2 under the setting of line 1 of Table 1. We first compare non-sparse clustering methods, including spectral clustering (Han et al. 2023), hierarchical clustering, and SDP-relaxed K-means with all features (Chen & Yang 2021), with Algorithm 2 initialized by each of them. Here, initialization refers to a rough cluster assignment $\hat{G}_1(0), \hat{G}_2(0)$ used to guide the first feature selection step. Results in Figures 2-(a) and 2-(b) show that while the non-sparse methods fail to adapt to sparsity, Algorithm 2, when properly initialized, is robust to high-dimensional noise and effectively adapts to sparsity. Even moderately good initializations are sufficient, as evidenced by the improved performance of Algorithm 2 when initialized with spectral or SDP K-means clustering. In contrast, poor initializations can hinder the clustering performance, as reflected in the weaker performance of Algorithm 2 initialized by hierarchical clustering. Next, Figures 2-(c) and 2-(d) highlight the competitiveness of our approach against existing two-step and iterative sparse clustering approaches: influential feature PCA (IFPCA; Jin & Wang 2016), sparse alternate similarity clustering (SAS; Arias-Castro & Pu 2017), sparse K-means (SKM; Witten & Tibshirani 2010) and CHIME (Cai et al. 2019). Details of these baseline methods are provided in Appendix **??**. Appendix **??** presents results under non-isotropic covariances.

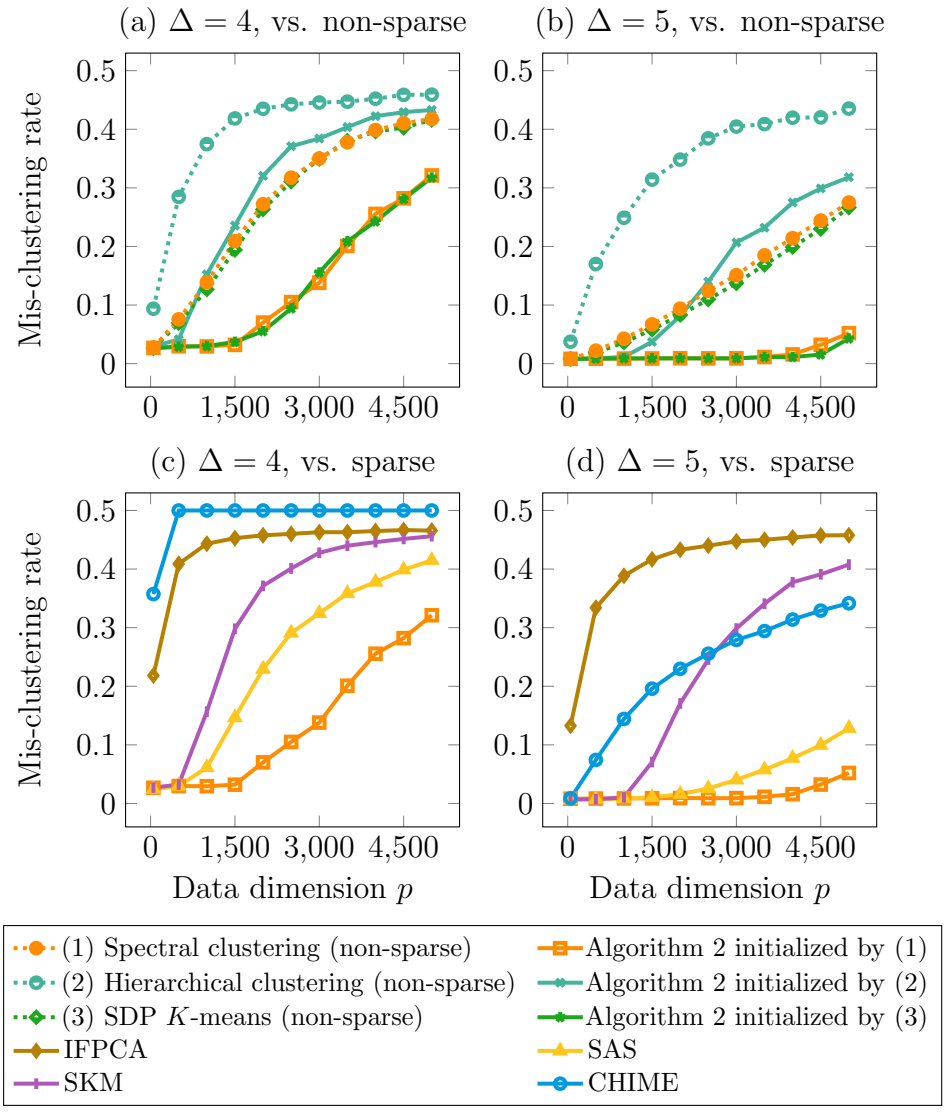


Figure 2: Mis-clustering rates of Algorithm 2 compared to non-sparse and sparse baselines. The dotted lines (non-sparse methods) can also be interpreted as the mis-clustering rates of Algorithm 2 immediately after initialization, before any iteration. The settings are provided in line 1 of Table 1.

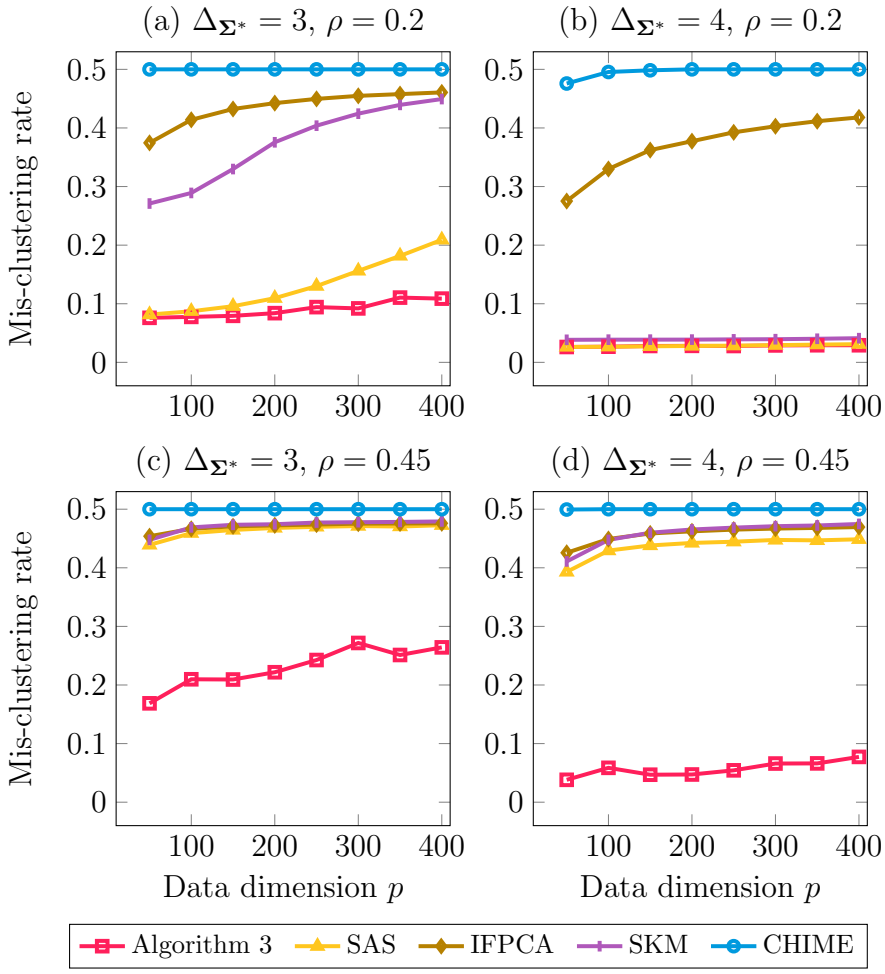


Figure 3: Mis-clustering rates of Algorithm **??**, compared to other sparsity-aware baseline methods under unknown covariance scenario with varying conditional correlations. The settings follow the line 2 of Table 1.

We further include the bandit-based variant of Algorithm 2 in the comparison. In the same simulation setting, the bandit approach outperforms all other methods, highlighting the benefit of stochastic exploration and memory integration in navigating the high-dimensional feature space (Figure 4).

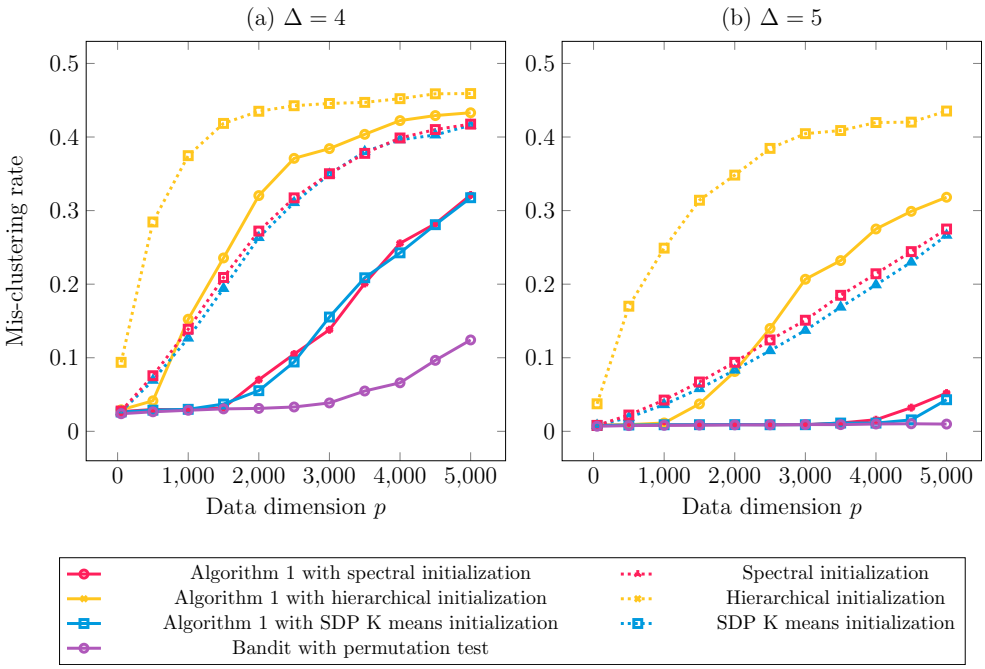


Figure 4: Bandit

4.2 Simulation under Unknown Covariance

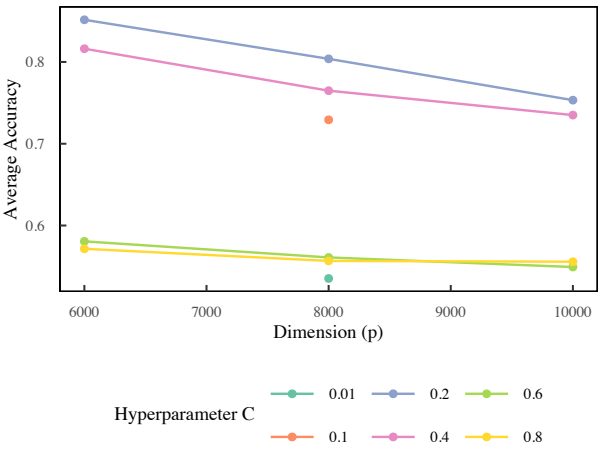
Algorithm **??** is evaluated in the setting with a chain-graph sparse precision matrix with conditional correlation $\rho \in \{0.2, 0.4\}$, following the setup described in line 2 of Table 1. The results in Figure 3 show that our method outperforms other sparsity-aware clustering approaches. Appendix **??** demonstrates the effectiveness of bypassing full precision matrix estimation. It also illustrates Algorithm **??** with an alternative feature selection threshold, demonstrating that an exact threshold is unnecessary and constant factors in thresholds are not a big concern. Appendix **??** demonstrates the convergence of the iterative algorithm and shows that slight overestimation of S_0 does not affect clustering performance.

4.3 Bandit Simulation

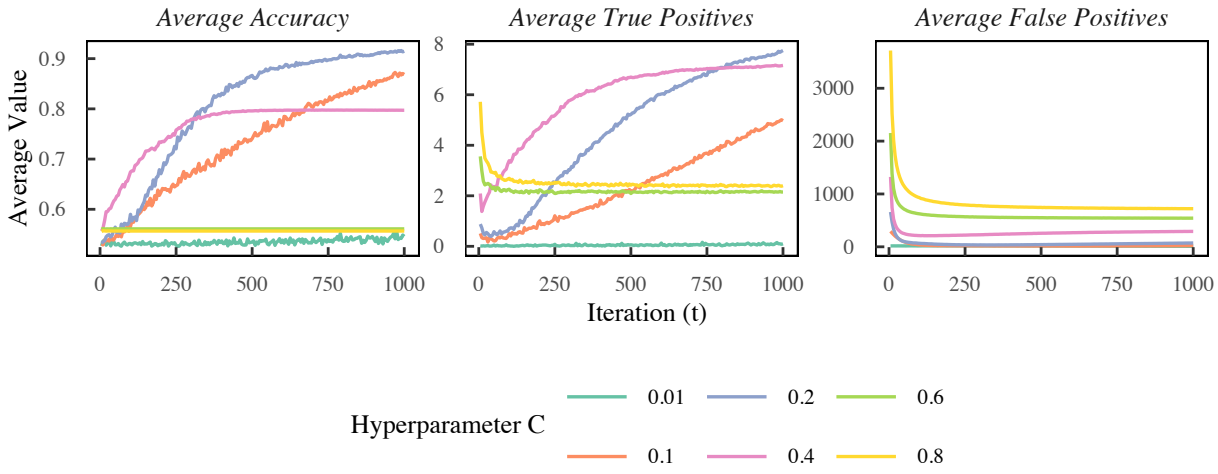
We investigate the sensitivity of the model to C using a symmetric two-cluster Gaussian mixture ($p \in [6000, 10000]$, ℓ_2-separation of 4, $|S_0| = 10$). While Liu & Ročková (2023) suggests $C \approx 0.618$ ($\tau_C = 0.5$), we explore a broader range $C \in \{0.01, \ldots, 0.8\}$ to modulate the exploration-exploitation trade-off. As shown in Figure 5a, extreme values degrade performance, with $C = 0.2$ yielding the best results in this high-dimensional setting. Figure 5b elucidates the convergence dynamics for $p = 8000$: a permissive $C = 0.8$ results in early stagnation with high false positives, while a restrictive $C = 0.1$ suppresses both signal and noise, preventing improvement. An intermediate $C = 0.2$ strikes the necessary balance, starting conservatively but successfully accumulating signal features over iterations to achieve superior overall performance.

4.4 Application to Real Datasets

We apply Algorithm 2 to leukemia subtype clustering using gene expression data ($n = 45$, $p = 3571$; Golub et al. 1999) and Algorithm **??** to clustering pre-processed MNIST images of digits 1 and 7 ($n = 1000$, $p = 784$). The leukemia dataset, originally created by Golub et al. (1999), comprises 72 gene expression profiles of 3871 genes, obtained from bone marrow and peripheral blood samples of patients with two types of leukemia: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). This dataset was previously analyzed by Jin & Wang (2016), and we used the data file available from their repository. To assess statistical performance, we performed stratified subsampling of 45 samples, repeating this procedure 100 times to compute the average clustering accuracy. The MNIST dataset consists of 28×28 grayscale images of digits from 1 to 10. We trained a k-sparse autoencoder (Makhzani & Frey 2013) on 30000 samples spanning all digits, using a sparsity parameter of $k = 20$. The trained model was then applied to 8640 test images of digits '1'and '7',



(a) Enter Caption



(b) Performance trajectories across varying levels of the hyperparameter C...

Figure 5: Main caption describing both plots.

converting them into 784-dimensional sparse representations. We applied stratified subsampling by selecting $n = 1000$ images from the 8640 test samples, applied the clustering algorithms to distinguish digits '1'and '7', repeated the process 100 times, and reported the average clustering accuracy. We compared Algorithm **??** with Sparse K-means (SKM; Witten & Tibshirani 2010), Influential Feature PCA (IFPCA; Jin & Wang 2016), and Sparse Alternate Similarity clustering (SAS; Arias-Castro & Pu 2017). We do not report

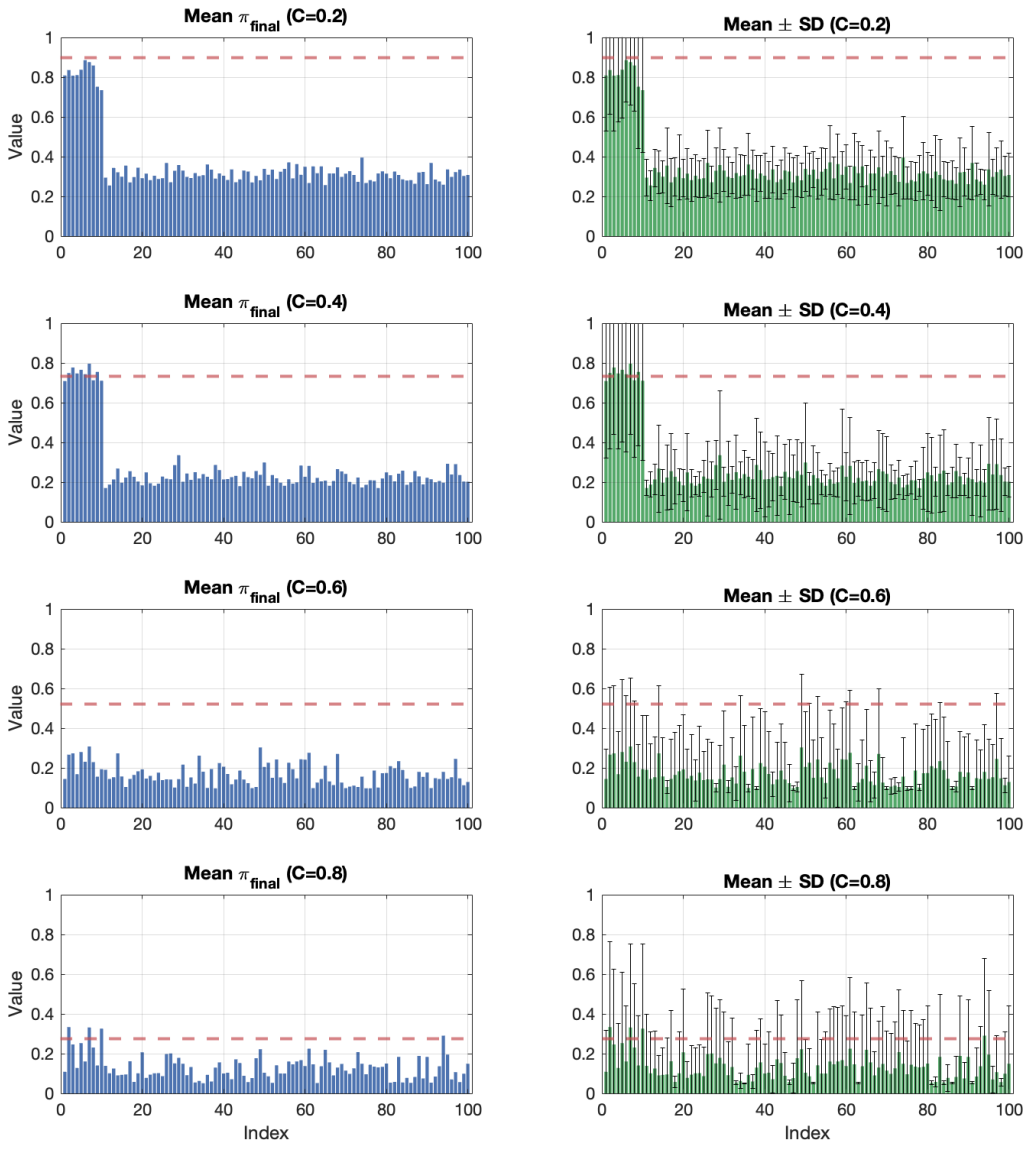


Figure 6: final θ value (beta mean) for different value of C. average over 20 replications.

Table 2: Clustering accuracy on the leukemia gene expression dataset and the MNIST dataset.

	Our method	IFPCA	SKM	SAS
Leukemia	**0.93**	0.84	0.79	0.87
MNIST	**0.94**	0.61	0.57	0.56

the performance of CHIME (Cai et al. 2019) in this real-data analysis, as the original paper does not share the implemented code for initializing cluster centers and covariance, and unlike in simulations, there is no obvious choice of initialization for real datasets. Details of the baseline methods are provided in Appendix **??**. For the k-sparse autoencoder, we used the authors' implementation available on their GitHub repository. The results in Table 2 show that our methods consistently outperform baseline approaches.

4.4.0.1 Additional Simulations Appendix **??** demonstrates that our algorithms maintain strong clustering accuracy under mild violations of the assumptions of Gaussianity, covariance homogeneity, and exact sparsity of model parameters. For scalability, while our focus is on the high-dimensional, low-sample-size regime, real-world datasets may involve much larger samples. For larger n, we recommend approximate solvers (Zhuang et al. 2022, 2024) tailored for the SDP step. Appendix **??** demonstrates that those approximate solvers scale our algorithms up to $n = 100{,}000$. Finally, we compare our methods to subspace clustering algorithms that extracts, rather than selects, features lying in low-dimensional subspaces (Soltanolkotabi et al. 2014, Wang & Xu 2016, Huang & Gu 2025). Specifically, Appendix **??** shows that our method excels the Johnson–Lindenstrauss transform-based K-means (Boutsidis et al. 2010) in sparse settings.

5 DISCUSSION

Under a known isotropic covariance, we derived the minimax separation required for exact cluster recovery across varying signal feature subset estimates, achieved by a modified SDP K-means. This result highlights the critical role of feature selection in sparse clustering while showing that slight overselection of features is not detrimental. Building on this insight, we proposed iterative algorithms that integrate feature selection into SDP K-means, avoiding the need to precisely estimate nuisance parameters such as cluster centers and the covariance matrix (which may not be the identity). There are two potential directions for future research: i) establish theoretical guarantees that our algorithm can identify a subset with sufficient separation, and ii) extend our minimax framework to the general case of an unknown covariance matrix.

6 Disclosure statement

The authors have the following conflicts of interest to declare (or replace with a statement that no conflicts of interest exist).

7 Data Availability Statement

Deidentified data have been made available at the following URL: XX.

SUPPLEMENTARY MATERIAL

Title: Brief description. (file type)

R-package for MYNEW routine: R-package MYNEW containing code to perform the diagnostic methods described in the article. The package also contains all datasets used as examples in the article. (GNU zipped tar file)

HIV data set: Data set used in the illustration of MYNEW method in Section **??** (.txt file).

References

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

Azizyan, M., Singh, A. & Wasserman, L. (2013), Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation, *in* 'Advances in Neural Information Processing Systems (NeurIPS)'.

Azizyan, M., Singh, A. & Wasserman, L. (2015), Efficient sparse clustering of high-dimensional non-spherical gaussian mixtures, *in* 'International Conference on Artificial Intelligence and Statistics (AISTATS)'.

Banerjee, T., Mukherjee, G. & Radchenko, P. (2017), 'Feature screening in large scale cluster analysis', *Journal of Multivariate Analysis* **161**, 191–212.

Bing, X., Bunea, F. & Wegkamp, M. (2020), 'Optimal estimation of sparse topic models', *Journal of Machine Learning Research* **21**(177), 1–45.

Boutsidis, C., Zouzias, A. & Drineas, P. (2010), Random projections for k-means clustering, *in* 'Advances in Neural Information Processing Systems (NeurIPS)'.

Bouveyron, C. & Brunet-Saumard, C. (2014), 'Discriminative variable selection for clustering with the sparse Fisher-EM algorithm', *Computational Statistics* **29**(3), 489–513.

Bouveyron, C., Girard, S. & Schmid, C. (2007), 'High-dimensional data clustering', *Computational Statistics & Data Analysis* **52**(1), 502–519.

Brodinová, Š., Filzmoser, P., Ortner, T., Breiteneder, C. & Rohm, M. (2019), 'Robust and

sparse k-means clustering for high-dimensional data', *Advances in Data Analysis and Classification* **13**(4), 905–932.

Cai, T. T., Ma, J. & Zhang, L. (2019), 'CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality', *The Annals of Statistics* **47**(3), 1234–1267.

Chakraborty, S. & and Xu, J. (2023), 'Biconvex clustering', *Journal of Computational and Graphical Statistics* **32**(4), 1524–1536.

Chakraborty, S., Paul, D. & Das, S. (2023), 'On consistent entropy-regularized k-means clustering with feature weight learning: Algorithm and statistical analyses', *IEEE Transactions on Cybernetics* **53**(8), 4779–4790.

Chang, W.-C. (1983), 'On using principal components before separating a mixture of two multivariate normal distributions', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **32**(3), 267–275.

Chen, X. & Yang, Y. (2021), 'Cutoff for exact recovery of gaussian mixture models', *IEEE Transactions on Information Theory* **67**(6), 4223–4238.

Chen, X. & Zhang, A. Y. (2024), 'Achieving optimal clustering in Gaussian mixture models with anisotropic covariance structures', *Advances in Neural Information Processing Systems (NeurIPS)* .

Cunningham, J. P. & Yu, B. M. (2014), 'Dimensionality reduction for large-scale neural recordings', *Nature Neuroscience* **17**(11), 1500–1509.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Leigh Murphy, Ellis, I., Purushotham, A., Børresen-Dale,

A.-L., Brenton, J. D., Tavaré, S., Caldas, C. & Aparicio, S. (2012), 'The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups', *Nature* **486**(7403), 346–352.

Dharamshi, A., Neufeld, A., Motwani, K., Gao, L. L., Witten, D. & Bien, J. (2025), 'Generalized data thinning using sufficient statistics', *Journal of the American Statistical Association* **120**(549), 511–523.

Dong, W., Xu, C., Xie, J. & Tang, N. (2024), 'Tuning-free sparse clustering via alternating hard-thresholding', *Journal of Multivariate Analysis* **203**, 105330.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0047259X2400037X*

Even, B., Giraud, C. & Verzelen, N. (2025), 'Computational lower bounds in latent models: Clustering, sparse-clustering, biclustering', *arXiv preprint arXiv:2506.13647* .

Fan, Y., Jin, J. & Yao, Z. (2013), 'Optimal classification in sparse Gaussian graphic model', *The Annals of Statistics* **41**(5), 2537–2571.

Fan, Y., Kong, Y., Li, D. & Zheng, Z. (2015), 'Innovated interaction screening for high-dimensional nonlinear classification', *The Annals of Statistics* **43**(3).
URL: *http://dx.doi.org/10.1214/14-AOS1308*

Fan, Y. & Lv, J. (2016), 'Innovated scalable efficient estimation in ultra-large Gaussian graphical models', *The Annals of Statistics* **44**(5), 2098–2126.

Fraley, C. & Raftery, A. E. (2007), 'Bayesian regularization for normal mixture estimation and model-based clustering', *Journal of Classification* **24**(2), 155–181.

Fu, Y., Liu, X., Sarkar, S. & Wu, T. (2021), 'Gaussian mixture model with feature selection: An embedded approach', *Computers & Industrial Engineering* **152**, 107000.

Ghosh, D. & Chinnaiyan, A. M. (2002), 'Mixture modelling of gene expression data from

microarray experiments', *Bioinformatics* **18**(2), 275–286.

URL: *https://doi.org/10.1093/bioinformatics/18.2.275*

Giraud, C. & Verzelen, N. (2019), 'Partial recovery bounds for clustering with the relaxed K-means', *Mathematical Statistics and Learning* **1**(3), 317–374.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**(5439), 531–537.

Guo, J., Levina, E., Michailidis, G. & Zhu, J. (2010), 'Pairwise variable selection for high-dimensional model-based clustering', *Biometrics* **66**(3), 793–804.

Han, X., Tong, X. & Fan, Y. (2023), 'Eigen selection in spectral clustering: A theory-guided practice', *Journal of the American Statistical Association* **118**(541), 109–121.

He, Y., Tang, X., Huang, J., Ren, J., Zhou, H., Chen, K., Liu, A., Shi, H., Lin, Z., Li, Q., Aditham, A., Ounadjela, J., Grody, E. I., Shu, J., Liu, J. & Wang, X. (2021), 'ClusterMap for multi-scale clustering analysis of spatial gene expression', *Nature Communications* **12**(1), 5909.

Huang, C. & Gu, Y. (2025), 'Minimax-optimal dimension-reduced clustering for high-dimensional nonspherical mixtures', *arXiv preprint arXiv:2502.02580* .

Huo, Z. & Tseng, G. (2017), 'Integrative sparse K-means with overlapping group lasso in genomic applications for disease subtype discovery', *The Annals of Applied Statistics* **11**(2), 1011–1039.

Javanmard, A. & Mirrokni, V. (2023), Anonymous learning via look-alike clustering: A precise analysis of model generalization, *in* 'Advances in Neural Information Processing Systems (NeurIPS)'.

Jin, J., Ke, Z. T. & Wang, W. (2017), 'Phase transitions for high dimensional clustering and related problems', *The Annals of Statistics* **45**(5), 2151–2189.

Jin, J. & Wang, W. (2016), 'Influential features pca for high dimensional clustering', *The Annals of Statistics* **44**(6), 2323–2359.

Kadir, S. N., Goodman, D. F. M. & Harris, K. D. (2014), 'High-dimensional cluster analysis with the masked EM algorithm', *Neural Computation* **26**(11), 2379–2394.

Kim, H., Kim, H. & De Veciana, G. (2024), Clustered federated learning via gradient-based partitioning, *in* 'International Conference on Machine Learning (ICML)'.

Leiner, J., Duan, B., Wasserman, L. & Ramdas, A. (2025), 'Data Fission: Splitting a Single Data Point', *Journal of the American Statistical Association* **120**(549), 135–146.

Li, R., Chang, X., Wang, Y. & Xu, Z. (2018), 'Sparse k-means with ℓ_∞/ℓ_0 penalty for high-dimensional data clustering', *Statistica Sinica* .

Liu, J., Zhang, J., Palumbo, M. & Lawrence, C. (2003), 'Bayesian clustering with variable and transformation selections', *Bayesian Statistics* **7**, 249–275.

Liu, R. G. & Frank, M. J. (2022), 'Hierarchical clustering optimizes the tradeoff between compositionality and expressivity of task structures for flexible reinforcement learning', *Artificial Intelligence* **312**, 103770.

Liu, T., Lu, Y., Zhu, B. & Zhao, H. (2023), 'Clustering high-dimensional data via feature selection', *Biometrics* **79**(2), 940–950.

Liu, Y. & Ročková, V. (2023), 'Variable selection via thompson sampling', *Journal of the American Statistical Association* **118**(541), 287–304.

Löffler, M., Wein, A. S. & Bandeira, A. S. (2022), 'Computationally efficient sparse clustering', *Information and Inference: A Journal of the IMA* **11**(4), 1255–1286.

Löffler, M., Zhang, A. Y. & Zhou, H. H. (2021), 'Optimality of spectral clustering in the Gaussian mixture model', *The Annals of Statistics* **49**(5), 2506–2530.

Lu, Y. Y., Lv, J., Fuhrman, J. A. & Sun, F. (2017), 'Towards enhanced and interpretable clustering/classification in integrative genomics', *Nucleic Acids Research* **45**(20), e169.

Makhzani, A. & Frey, B. (2013), K-Sparse autoencoders, *in* 'International Conference on Learning Representations (ICLR)'.

Maugis, C., Celeux, G. & Martin-Magniette, M.-L. (2009), 'Variable selection for clustering with gaussian mixture models', *Biometrics* **65**(3), 701–709.

Mishra, S., Gordon, B. A., Su, Y., Christensen, J., Friedrichsen, K., Jackson, K., Hornbeck, R., Balota, D. A., Cairns, N. J., Morris, J. C., Ances, B. M. & Benzinger, T. L. S. (2017), 'AV-1451 PET imaging of tau pathology in preclinical Alzheimer disease: Defining a summary measure', *NeuroImage* **161**, 171–178.

Namgung, J. Y., Mun, J., Park, Y., Kim, J. & Park, B.-y. (2024), 'Sex differences in autism spectrum disorder using class imbalance adjusted functional connectivity', *NeuroImage* **304**, 120956.

Nystrup, P., Kolm, P. N. & Lindström, E. (2021), 'Feature selection in jump models', *Expert Systems with Applications* **184**, 115558.

Pan, W. & Shen, X. (2007), 'Penalized model-based clustering with application to variable selection', *Journal of Machine Learning Research* **8**(41), 1145–1164.

Park, Y.-G., Kwon, Y. W., Koh, C. S., Kim, E., Lee, D. H., Kim, S., Mun, J., Hong, Y.-M., Lee, S., Kim, J.-Y., Lee, J.-H., Jung, H. H., Cheon, J., Chang, J. W. & Park, J.-U. (2024), 'In-vivo integration of soft neural probes through high-resolution printing of liquid electronics on the cranium', *Nature Communications* **15**(1), 1772.

Peng, J. & Wei, Y. (2007), 'Approximating K-means-type clustering via semidefinite programming', *SIAM Journal on Optimization* **18**(1), 186–205.

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

Royer, M. (2017), Adaptive clustering through semidefinite programming, *in* 'Advances in Neural Information Processing Systems (NeurIPS)'.

Soltanolkotabi, M., Elhamifar, E. & Candès, E. J. (2014), 'Robust subspace clustering', *The Annals of Statistics* **42**(2), 669–699.

Sun, W., Wang, J. & Fang, Y. (2012), 'Regularized k-means clustering of high-dimensional data and its asymptotic consistency', *Electronic Journal of Statistics* **6**, 148–167.

Tadesse, M. G., , Naijun, S. & and Vannucci, M. (2005), 'Bayesian Variable Selection in Clustering High-Dimensional Data', *Journal of the American Statistical Association* **100**(470), 602–617.

Tamayo, P., Scanfeld, D., Ebert, B. L., Gillette, M. A., Roberts, C. W. M. & Mesirov, J. P. (2007), 'Metagene projection for cross-platform, cross-species characterization of global transcriptional states', *Proceedings of the National Academy of Sciences* **104**(14), 5959–5964.

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

Tsai, C.-Y. & Chiu, C.-C. (2008), 'Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm', *Computational Statistics & Data Analysis* **52**(10), 4658–4672.

Verzelen, N. & Arias-Castro, E. (2017), 'Detection and feature selection in sparse mixture models', *The Annals of Statistics* **45**(5), 1920–1950.

Wang, B., Zhang, Y., Sun, W. W. & Fang, Y. (2018), 'Sparse convex clustering', *Journal of Computational and Graphical Statistics* **27**(2), 393–403.

Wang, S. & Zhu, J. (2008), 'Variable selection for model-based high-dimensional clustering and its application to microarray data', *Biometrics* **64**(2), 440–448.

Wang, Y.-X. & Xu, H. (2016), 'Noisy sparse subspace clustering', *Journal of Machine Learning Research* **17**(12), 1–41.

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

Wu, R., , Linjun, Z. & and Tony Cai, T. (2023), 'Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference', *Journal of the American Statistical Association* **118**(543), 1849–1861.

Xie, B., Pan, W. & Shen, X. (2008), 'Variable selection in penalized model-based clustering via regularization on grouped parameters', *Biometrics* **64**(3), 921–930.

Yao, D., Xie, F. & Xu, Y. (2025), 'Bayesian sparse gaussian mixture model for clustering in high dimensions', *Journal of Machine Learning Research* **26**(21), 1–50.

Yi, J., Zhang, L., Wang, J., Jin, R. & Jain, A. (2014), A single-pass algorithm for efficiently recovering sparse cluster centers of high-dimensional data, *in* 'International Conference on Machine Learning (ICML)'.

Yuan, Z., Chen, J., Qiu, H., Wang, H. & Huang, Y. (2024), 'Adaptive sufficient sparse clustering by controlling false discovery', *Statistics and Computing* **34**(6), 1–36.

Zhang, Y., Wu, W., Toll, R. T., Naparstek, S., Maron-Katz, A., Watts, M., Gordon, J., Jeong,

J., Astolfi, L., Shpigel, E., Longwell, P., Sarhadi, K., El-Said, D., Li, Y., Cooper, C., Chin-Fatt, C., Arns, M., Goodkind, M. S., Trivedi, M. H., Marmar, C. R. & Etkin, A. (2021), 'Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography', *Nature Biomedical Engineering* **5**(4), 309–323.

Zhang, Z., Lange, K. & Xu, J. (2020), Simple and scalable sparse k-means clustering via feature ranking, *in* 'Advances in Neural Information Processing Systems (NeurIPS)'.

Zhou, H., Pan, W. & Shen, X. (2009), 'Penalized model-based clustering with unconstrained covariance matrices', *Electronic Journal of Statistics* **3**, 1473–1496.

Zhuang, Y., Chen, X. & Yang, Y. (2022), Sketch-and-lift: Scalable subsampled semidefinite program for k-means clustering, *in* 'International Conference on Artificial Intelligence and Statistics (AISTATS)'.

Zhuang, Y., Chen, X. & Yang, Y. (2023), Likelihood adjusted semidefinite programs for clustering heterogeneous data, *in* 'International Conference on Machine Learning (ICML)'.

Zhuang, Y., Chen, X., Yang, Y. & Zhang, R. Y. (2024), Statistically optimal K-means clustering via nonnegative low-rank semidefinite programming, *in* 'International Conference on Learning Representations (ICLR)'.

Algorithm 3 Block Coordinate Ascent with Thompson Sampling

Require: Data $\mathbf{X}$, Covariance $\mathbf{\Sigma}$, Iterations T, Regularization C, Permutations B, FDR q

Initialize priors: $a_{j,1} \leftarrow 1, b_{j,1} \leftarrow 1$ for all $j \in [p]$

Calculate threshold: $\tau_C \leftarrow \frac{\log(1/C)}{\log((C+1)/C)}$

for $t = 1, 2, \ldots, T$ **do**

  Step 1: Choose (Sampling)

  Sample $\hat{\theta}_j^t \sim \text{Beta}(a_j^t, b_j^t)$ for $j = 1, \ldots, p$

  $\hat{S}^t \leftarrow \{j : \hat{\theta}_j^t > \tau_C\}$

  if $\hat{S}^t = \emptyset$ **then** $\hat{S}^t \leftarrow$ random feature ▷ Avoid empty set

  Step 2: Clustering

  $\tilde{\mathbf{X}} \leftarrow \mathbf{\Sigma}^{-1}\mathbf{X}$

  $\hat{G}_1^t, \hat{G}_2^t \leftarrow \texttt{SDPclust}(\tilde{\mathbf{X}}_{\hat{S}^t,\cdot}, \mathbf{\Sigma}_{\hat{S}^t,\hat{S}^t})$ ▷ Algorithm 1

  Step 3: Reward & Update

  Compute observed stats: $\mathbf{T}^{obs} \leftarrow |\bar{\tilde{\mathbf{X}}}_{\hat{G}_1^t} - \bar{\tilde{\mathbf{X}}}_{\hat{G}_2^t}|$

  for $b = 1, \ldots, B$ **do**

    Generate permuted labels $\pi_b(\hat{G}_1^t), \pi_b(\hat{G}_2^t)$ and null stats $\mathbf{T}^{(b)}$

  for $j \in \hat{S}^t$ **do** ▷ Evaluate only selected features

    $p_j \leftarrow \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(T_j^{(b)} \geq T_j^{obs})$

  Determine significant features via $\{p_j\}_{j \in \hat{S}^t}$

  for $j \in \hat{S}^t$ **do**

    if j is rejected (significant) **then**

      $Y_j^t \leftarrow 1$

    else

      $Y_j^t \leftarrow 0$

    $a_{j,t+1} \leftarrow a_j^t + Y_j^t$

    $b_{j,t+1} \leftarrow b_j^t + (1 - Y_j^t)$

  for $j \notin \hat{S}^t$ **do**

    $a_{j,t+1} \leftarrow a_j^t, \quad b_{j,t+1} \leftarrow b_j^t$