Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Local Feature Relevance in Full-Covariance Gaussian Mixture Models via Penalized Classification EM

Wilson Toussile

wilson.toussile@enspy-uy1.cm

Department of Mathematics and Physics, National Advanced School of Engineering (ENSPY), University of Yaoundé 1 https://orcid.org/0009-0002-4330-1148

Siméon Fotso

Department of Mathematics, Higher Teacher Training College (ENS), University of Yaoundé 1

Research Article

Keywords: Gaussian Mixture Models, Full-Covariance, Feature Selection, Variable Selection, Clustering, Classification EM, Down-Projection Operator, Single-Cell Transcriptomics

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11758566/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Local Feature Relevance in Full-Covariance Gaussian Mixture Models via Penalized Classification EM

Wilson Toussile[1*] Siméon Fotso[2†]

2026-11-01

[1]Department of Mathematics and Physics, National Advanced School of Engineering (ENSPY)
University of Yaounde 1, Yaounde, Cameroon

[2]Department of Mathematics, Higher Teacher Training College (ENS)
University of Yaounde 1, Yaounde, Cameroon

In clustering by mixture models, relevant features are often cluster-specific and correlated within each group. Existing methods force an uncomfortable compromise: they either assume conditional independence to select features locally, or retain full covariances at the expense of a single global feature subset. We resolve this dilemma with a Full-Covariance Gaussian Mixture Model with Local Feature Relevance (FC-GMM-LFR). Each observation is governed by its own binary activation profile: active features follow the cluster's full-covariance Gaussian distribution, inactive features follow a shared background distribution, and a null profile naturally isolates non-informative observations. Because feature relevance is not identifiable from the likelihood alone, we select cluster-specific supports through a penalized classification criterion. We maximize this criterion with a monotone block coordinate ascent algorithm that classifies individual profiles, updates full covariances via a missing-data EM step under spectral regularization, and alternates with background estimation. In simulations, our method reliably recovers local active structures and outperforms competitors when clusters differ by their correlations. On sorted single-cell blood data, it accurately separates cell populations (ARI = 0.96), and its estimated activation rates closely reproduce the detection rates of marker genes.

Keywords: Gaussian Mixture Models, Full-Covariance, Feature Selection, Variable Selection, Clustering, Classification EM, Down-Projection Operator, Single-Cell Transcriptomics.

*Corresponding author. Email: `wilson.toussile@enspy-uy1.cm`
†Email: `simeonfotso@yahoo.fr`

1 Introduction

Gaussian Mixture Models (GMM) provide a foundational probabilistic framework for clustering (Bouveyron et al., 2019; Gormley et al., 2023). In high-dimensional applications, clustering structures are rarely supported by all measured features; instead, relevant signals typically lie within lower-dimensional subspaces surrounded by uninformative noise (Toussile and Gassiat, 2009; Fop and Murphy, 2018). Retaining irrelevant features often inflates parameter estimation variance and degrades classification performances, particularly in full-covariance mixtures where the number of parameters scales as $\mathcal{O}(KJ^2)$ (Bouveyron and Brunet-Saumard, 2014; Bouveyron et al., 2019).

To tackle this challenge, existing feature selection methods generally adopt one of two compromises. The first family assumes conditional independence of features within clusters, restricting covariance matrices to be diagonal (Law et al., 2004; Xie et al., 2008). While localized extensions select features on a cluster-by-cluster basis (Li et al., 2009), assuming diagonal covariances discards intra-cluster dependencies and distorts the geometry of continuous subpopulations. The second family preserves full covariances, as pioneered by Tadesse et al. (2005) with latent Bayesian indicators or in stepwise and regularized frequentist formulations (Scrucca and Raftery, 2018; Zhou et al., 2009; Celeux et al., 2019). However, these approaches enforce a single global feature subset across all clusters and scale poorly with feature dimension due to combinatorial searches or slow MCMC mixing. Although cluster-specific relevance of features has also appeared in Bayesian nonparametrics (Hoff, 2006; Quintana et al., 2015), estimating localized full-rank covariance structures remains an important practical challenge.

In this paper, we propose an approach to fill this gap via a Full-Covariance Gaussian Mixture Model with Local Feature Relevance (FC-GMM-LFR). In this model, each observation is associated with its own binary activation profile $\boldsymbol{s}_i \in \{0,1\}^J$: its active coordinates follow a marginal cluster-specific full-covariance Gaussian distribution via a Down-Projection Operator (DPO) $\mathbf{D}(\boldsymbol{s}_i)$, while its inactive coordinates follow a background distribution common to all clusters. Working at the level of individual coordinates rather than the one of entire observations connects naturally with cellwise robust methods (Raymaekers and Rousseeuw, 2024; Zaccaria et al., 2025). In our setting, however, inactive coordinates are not contaminated values to discard, but coordinates at rest that follow an invariant background, with their cluster-specific activation probabilities measuring feature relevance.

A key issue with our model is that feature relevance itself is not identifiable from the observed likelihood alone. Specifically, the activation rate of a feature can vary freely between 0 and 1 without altering the data distribution whenever its cluster parameters align with the background. Because the unpenalized likelihood cannot distinguish a genuinely irrelevant feature from an active one that mimics the background, it tends to overfit by retaining uninformative variables. We show that this is the only obstruction: once such degenerate configurations are excluded, the activation structure and the parameters on it are identified. We resolve the ambiguity through a penalized classification criterion that enforces parsimony, pruning inactive coordinates unless the data justify cluster-specific parameters.

To maximize this objective, we develop a monotone block coordinate ascent algorithm. It updates individual activation profiles coordinate-wise using exact conditional Gaussian distributions, estimates cluster means and full covariance matrices via a missing-data EM step regularized by an eigenvalue floor, and alternates with a shared background estimation. In numerical simulations, our approach reliably uncovers cluster-specific active subspaces and outperforms existing methods when clusters differ by their correlation geometry. Applied to sorted single-cell blood data, it accurately separates cell populations while recovering known marker genes through their estimated activation rates.

The remainder of this article is organized as follows. Section~2 formalizes the model, its finite mixture representation, and the identifiability of the activation structure. Section~3 introduces the penalized classification criterion and the coordinate ascent algorithm. Section~4 details the classification rules and variable relevance metrics. Sections~5 and~6 present the simulation experiments and the application to sorted blood cells using the proposed Julia package `FCGMMLFR.jl`, followed by a concluding discussion in Section~7. Mathematical proofs and computing times are compiled in the appendices.

2 Full-Covariance Gaussian Mixtures with Local Feature Relevance

This section formalizes the Full-Covariance Gaussian Mixture Model with Local Feature Relevance (FC-GMM-LFR). We first define the Down-Projection Operator, an algebraic tool that extracts cluster-specific active subvectors. We then detail the hierarchical generative process, establish the resulting finite mixture representation on $\mathbb{R}^J$, and analyze the structural identifiability of the local activation profiles.

2.1 Notation

To keep notation light, we use lowercase letters for both random variables and their realizations. We also use $p(\cdot)$ generically to denote probability densities, taken with respect to the counting measure for discrete variables (such as cluster indicators and activation profiles) and the Lebesgue measure for continuous variables, with the argument indicating which distribution is referenced.

For a binary selection profile $\boldsymbol{s} = (s_1, \ldots, s_J)^\top \in \{0,1\}^J$, we write $\mathcal{A}(\boldsymbol{s}) = \{j : s_j = 1\}$ for its active coordinates, $d(\boldsymbol{s}) = |\mathcal{A}(\boldsymbol{s})|$ for the active dimension, and $\boldsymbol{s}^c = \mathbf{1}_J - \boldsymbol{s}$ for the complementary profile, whose active coordinates are the inactive coordinates of $\boldsymbol{s}$.

Definition 2.1 (Down-Projection Operator)**.** *For a binary selection profile $\boldsymbol{s} \in \{0,1\}^J$ with active dimension $d(\boldsymbol{s}) \geq 1$, the down-projection operator $\mathbf{D}(\boldsymbol{s}) \in \{0,1\}^{d(\boldsymbol{s}) \times J}$ is the matrix formed by the rows of $\mathbf{I}_J$ indexed by $\mathcal{A}(\boldsymbol{s})$, in increasing order. For the null profile, $\mathbf{D}(\mathbf{0}_J)$ is the empty $0 \times J$ matrix.*

The operator $\mathbf{D}(\boldsymbol{s})$ extracts the active subvector $\mathbf{D}(\boldsymbol{s})\boldsymbol{y} \in \mathbb{R}^{d(\boldsymbol{s})}$ from any observation $\boldsymbol{y}$, while $\mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{s})^\top$ extracts the principal submatrix of $\boldsymbol{\Sigma}$ on $\mathcal{A}(\boldsymbol{s})$. It satisfies $\mathbf{D}(\boldsymbol{s})\mathbf{D}(\boldsymbol{s})^\top = \mathbf{I}_{d(\boldsymbol{s})}$ and induces the diagonal orthogonal projector $\mathbf{P}(\boldsymbol{s}) \equiv \mathbf{D}(\boldsymbol{s})^\top\mathbf{D}(\boldsymbol{s}) = \mathrm{diag}(\boldsymbol{s})$, with $\mathbf{P}(\boldsymbol{s})+\mathbf{P}(\boldsymbol{s}^c) = \mathbf{I}_J$ and $\mathbf{P}(\mathbf{0}_J) = \mathbf{0}_{J\times J}$. For $1 \leq d(\boldsymbol{s}) \leq J-1$, the stacked matrix $\boldsymbol{\Pi}(\boldsymbol{s}) = [\mathbf{D}(\boldsymbol{s})^\top\ \mathbf{D}(\boldsymbol{s}^c)^\top]^\top$ is an orthogonal permutation matrix ($|\det \boldsymbol{\Pi}(\boldsymbol{s})| = 1$) that groups active coordinates first.

2.2 Hierarchical Generative Model and Subspace-Projected Emission

Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top \in \mathbb{R}^{n\times J}$ denote n independent and identically distributed observations of a J-dimensional random vector. Each observation $\boldsymbol{y}_i$ is associated with a latent cluster allocation $z_i \in \{1, \ldots, K\}$ and a latent binary feature relevance profile $\boldsymbol{s}_i = (s_{i,1}, \ldots, s_{i,J})^\top \in \{0,1\}^J$, collected into $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ and $\boldsymbol{s} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$. To reconcile localized feature relevance with unconstrained intra-cluster dependence structures on relevant coordinates, the FC-GMM-LFR framework is founded on the following structural assumption.

Assumption 2.2 (Active-Background Decoupling and Cluster Invariance)**.** *Conditioned on* $(z_i = k, \boldsymbol{s}_i)$*:*

1. *The active subvector* $\mathbf{D}(\boldsymbol{s}_i)\boldsymbol{y}_i$ *and the inactive subvector* $\mathbf{D}(\boldsymbol{s}_i^c)\boldsymbol{y}_i$ *are conditionally independent.*
2. *The conditional distribution of the inactive subvector* $\mathbf{D}(\boldsymbol{s}_i^c)\boldsymbol{y}_i$ *does not depend on the cluster allocation* k*.*

Assumption~2.2 formalizes the principled separation between cluster-specific signals and ambient background noise. Specifically, Item~1 decouples inactive coordinates from the active cluster geometry, ensuring that uninformative features reflect pure background variation rather than distorting intra-cluster correlations. Item~2 guarantees that inactive coordinates carry no discriminative information, ensuring that cluster assignments are driven exclusively by genuinely relevant features.

Under Assumption~2.2, the Full-Covariance Gaussian Mixture Model with Local Feature Relevance (FC-GMM-LFR) generates each observation $\boldsymbol{y}_i$ through the following hierarchical three-level process:

1. **Latent Cluster Allocation**: For observation $i \in \{1, \ldots, n\}$, a latent categorical indicator $z_i \in \{1, \ldots, K\}$ is drawn from a categorical distribution with mixing proportions $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)^\top$:

$$p(z_i = k \mid \boldsymbol{\omega}) = \omega_k, \qquad \sum_{k=1}^{K} \omega_k = 1, \quad \omega_k > 0. \tag{1}$$

2. **Local Subspace Activation Profile**: Conditioned on cluster allocation $z_i = k$, the feature relevance profile $\boldsymbol{s}_i = (s_{i,1}, \ldots, s_{i,J})^\top \in \{0,1\}^J$ is drawn from a product of independent Bernoulli distributions:

$$p(\boldsymbol{s}_i \mid z_i = k, \boldsymbol{\gamma}_k) = \prod_{j=1}^{J} \gamma_{k,j}^{s_{i,j}} (1 - \gamma_{k,j})^{1 - s_{i,j}}, \tag{2}$$

where $\gamma_{k,j} = p(s_{i,j} = 1 \mid z_i = k) \in [0,1]$ represents the cluster-specific prior probability that feature j is active in cluster k, collected in $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)$ with $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \ldots, \gamma_{k,J})^\top \in [0,1]^J$.

3. **Emission**: Conditioned on cluster allocation $z_i = k$ and feature relevance profile $\boldsymbol{s}_i$, the observation $\boldsymbol{y}_i \in \mathbb{R}^J$ is split into active cluster-specific coordinates and inactive background coordinates:

   a) *Active Subspace Coordinates* ($s_{i,j} = 1$): The subvector $\mathbf{D}(\boldsymbol{s}_i)\boldsymbol{y}_i \in \mathbb{R}^{d(\boldsymbol{s}_i)}$ follows a $d(\boldsymbol{s}_i)$-dimensional Gaussian distribution with full-rank subspace covariance:

$$\mathbf{D}(\boldsymbol{s}_i)\boldsymbol{y}_i \mid (z_i = k, \boldsymbol{s}_i) \sim \mathcal{N}_{d(\boldsymbol{s}_i)} \left(\mathbf{D}(\boldsymbol{s}_i)\boldsymbol{\mu}_k,\ \mathbf{D}(\boldsymbol{s}_i)\boldsymbol{\Sigma}_k\mathbf{D}(\boldsymbol{s}_i)^\top \right), \tag{3}$$

   where $\boldsymbol{\mu}_k \in \mathbb{R}^J$ is the cluster center and $\boldsymbol{\Sigma}_k \in \mathcal{S}_{++}^J$ the parent covariance matrix; the active covariance $\mathbf{D}(\boldsymbol{s}_i)\boldsymbol{\Sigma}_k\mathbf{D}(\boldsymbol{s}_i)^\top$ is positive definite by Lemma 3.1. For every profile, the law of the active coordinates is the marginal, on $\mathcal{A}(\boldsymbol{s}_i)$, of the same parent distribution $\mathcal{N}_J(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: the 2^J active laws of a cluster are therefore mutually consistent and share $J + J(J+1)/2$ parameters, instead of a separate mean and covariance for each profile.

   b) *Inactive Background Coordinates* ($s_{i,j} = 0$): The complementary subvector $\mathbf{D}(\boldsymbol{s}_i^c)\boldsymbol{y}_i \in \mathbb{R}^{J - d(\boldsymbol{s}_i)}$ follows a background distribution common to all clusters:

$$\mathbf{D}\left(\boldsymbol{s}_i^c\right)\boldsymbol{y}_i \mid \boldsymbol{s}_i \sim \mathcal{N}_{J-d(\boldsymbol{s}_i)} \left(\mathbf{D}\left(\boldsymbol{s}_i^c\right)\boldsymbol{\mu}_0,\ \mathbf{D}\left(\boldsymbol{s}_i^c\right)\boldsymbol{\Sigma}_0\mathbf{D}\left(\boldsymbol{s}_i^c\right)^\top \right), \tag{4}$$

   where $\boldsymbol{\mu}_0 \in \mathbb{R}^J$ and $\boldsymbol{\Sigma}_0 \in \mathcal{S}_{++}^J$ are the background parameters, which do not depend on the cluster.

The joint conditional emission density of $\boldsymbol{y}_i$ given $(z_i = k, \boldsymbol{s}_i)$ factorizes as the product of the active cluster Gaussian and inactive background Gaussian densities:

$$\begin{aligned} p\left(\boldsymbol{y}_i \mid z_i = k, \boldsymbol{s}_i; \boldsymbol{\theta}\right) = \mathcal{N}_{d(\boldsymbol{s}_i)} &\left(\mathbf{D}(\boldsymbol{s}_i)\boldsymbol{y}_i\ ;\ \mathbf{D}(\boldsymbol{s}_i)\boldsymbol{\mu}_k,\ \mathbf{D}(\boldsymbol{s}_i)\boldsymbol{\Sigma}_k\mathbf{D}(\boldsymbol{s}_i)^\top \right) \\ &\times \mathcal{N}_{J-d(\boldsymbol{s}_i)} \left(\mathbf{D}\left(\boldsymbol{s}_i^c\right)\boldsymbol{y}_i\ ;\ \mathbf{D}\left(\boldsymbol{s}_i^c\right)\boldsymbol{\mu}_0,\ \mathbf{D}\left(\boldsymbol{s}_i^c\right)\boldsymbol{\Sigma}_0\mathbf{D}\left(\boldsymbol{s}_i^c\right)^\top \right). \end{aligned} \tag{5}$$

Remark 2.3 (Boundary Activation Profiles $\boldsymbol{s}_i = \mathbf{0}_J$ and $\boldsymbol{s}_i = \mathbf{1}_J$)**.** *In accordance with Definition 2.1:*

- *When $\boldsymbol{s}_i = \mathbf{1}_J$ (full activation, $d(\boldsymbol{s}_i) = J$), the inactive subspace is empty ($\mathbf{D}(\mathbf{0}_J)\boldsymbol{y}_i \in \mathbb{R}^0$), and $\boldsymbol{y}_i$ is emitted entirely from the substantive cluster Gaussian $\mathcal{N}_J(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.*
- *When $\boldsymbol{s}_i = \mathbf{0}_J$ (null profile, $d(\boldsymbol{s}_i) = 0$), the active subspace is empty ($\mathbf{D}(\mathbf{0}_J)\boldsymbol{y}_i \in \mathbb{R}^0$), and the entire observation $\boldsymbol{y}_i$ is emitted from the background distribution $\mathcal{N}_J(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, independently of the cluster assignment k.*

2.3 Full-Covariance Gaussian Mixture Representation on $\mathbb{R}^J$

Proposition 2.4 (Full-Covariance Gaussian Mixture Representation and Ambient Regularity)**.** *Under the Down-Projection Operator (DPO) $\mathbf{D}(\boldsymbol{s}_i)$ and the emission structure ((3) and (4)), the joint conditional distribution of the complete vector $\boldsymbol{y}_i \in \mathbb{R}^J$ given $(z_i = k, \boldsymbol{s}_i)$ is a valid non-degenerate J-dimensional multivariate Gaussian distribution:*

$$p(\boldsymbol{y}_i \mid z_i = k, \boldsymbol{s}_i) = \mathcal{N}_J\left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k(\boldsymbol{s}_i),\ \boldsymbol{\Sigma}_k(\boldsymbol{s}_i)\right), \tag{6}$$

with composite mean vector and reconstructed covariance matrix:

$$\boldsymbol{\mu}_k(\boldsymbol{s}_i) = \mathbf{P}(\boldsymbol{s}_i)\boldsymbol{\mu}_k + \mathbf{P}\left(\boldsymbol{s}_i^c\right)\boldsymbol{\mu}_0, \tag{7}$$

$$\boldsymbol{\Sigma}_k(\boldsymbol{s}_i) = \mathbf{P}(\boldsymbol{s}_i)\boldsymbol{\Sigma}_k\mathbf{P}(\boldsymbol{s}_i) + \mathbf{P}\left(\boldsymbol{s}_i^c\right)\boldsymbol{\Sigma}_0\mathbf{P}\left(\boldsymbol{s}_i^c\right). \tag{8}$$

Furthermore:

1. *The reconstructed matrix $\boldsymbol{\Sigma}_k(\boldsymbol{s}_i)$ is strictly symmetric positive-definite on $\mathbb{R}^J$, with condition number bounded by:*

$$\kappa(\boldsymbol{\Sigma}_k(\boldsymbol{s}_i)) \le \frac{\max\left\{\lambda_1(\boldsymbol{\Sigma}_k),\ \lambda_1(\boldsymbol{\Sigma}_0)\right\}}{\min\left\{\lambda_J(\boldsymbol{\Sigma}_k),\ \lambda_J(\boldsymbol{\Sigma}_0)\right\}}. \tag{9}$$

2. *Marginalizing over the latent selection hypercube $\boldsymbol{s}_i \in \{0,1\}^J$, the conditional density of $\boldsymbol{y}_i$ given $z_i = k$ is a finite Gaussian mixture with 2^J components:*

$$p(\boldsymbol{y}_i \mid z_i = k) = \sum_{\boldsymbol{s}\in\{0,1\}^J} \left(\prod_{j=1}^{J} \gamma_{k,j}^{s_j}(1-\gamma_{k,j})^{1-s_j}\right) \mathcal{N}_J\left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k(\boldsymbol{s}),\ \boldsymbol{\Sigma}_k(\boldsymbol{s})\right). \tag{10}$$

3. *The marginal data distribution on $\mathbb{R}^J$ is a finite Gaussian mixture of $K \times 2^J$ latent configurations, with at most $1 + K(2^J - 1)$ distinct components, since the K components of the null profile $\boldsymbol{s} = \mathbf{0}_J$ all equal the background distribution:*

$$p(\boldsymbol{y}_i) = \sum_{k=1}^{K} \omega_k \sum_{\boldsymbol{s}\in\{0,1\}^J} p(\boldsymbol{s} \mid z_i = k)\, \mathcal{N}_J\left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k(\boldsymbol{s}),\ \boldsymbol{\Sigma}_k(\boldsymbol{s})\right). \tag{11}$$

The proof is given in Appendix~B. Proposition~2.4 shows that, for every profile, the conditional law of $\boldsymbol{y}_i$ is a non-degenerate Gaussian law on $\mathbb{R}^J$, so that FC-GMM-LFR is a finite Gaussian mixture with tied parameters.

2.4 Canonical $(K+1)$-Component Partition and Non-Informative Observations

Proposition 2.5 (Canonical $(K+1)$-Component Mixture Partition)**.** *Let $\mathcal{S}^\star = \{0,1\}^J \setminus \{\mathbf{0}_J\}$ denote the set of non-null feature profiles, and let $\widetilde{\omega}_0 = \sum_{k=1}^K \omega_k \prod_{j=1}^J (1-\gamma_{k,j}) \in [0,1)$ denote the marginal probability of the null profile $\boldsymbol{s} = \mathbf{0}_J$, and assume that, for every cluster k, $\gamma_{k,j} > 0$ for at least one j. Then the marginal data density (11) admits a canonical $(K+1)$-component mixture partition:*

$$p(\boldsymbol{y}_i) = \widetilde{\omega}_0 \, \mathcal{N}_J\left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) + \sum_{k=1}^{K} \widetilde{\omega}_k \, f_k^{\text{in}}(\boldsymbol{y}_i), \tag{12}$$

where $\sum_{k=0}^K \widetilde{\omega}_k = \widetilde{\omega}_0 + \sum_{k=1}^K \widetilde{\omega}_k = 1$, with substantive cluster mixture weights:

$$\widetilde{\omega}_k = \omega_k \left(1 - \prod_{j=1}^{J} (1-\gamma_{k,j})\right), \quad k \in \{1,\ldots,K\}, \tag{13}$$

and conditional cluster emission densities supported exclusively on non-empty subspaces:

$$f_k^{\text{in}}(\boldsymbol{y}_i) = \sum_{\boldsymbol{s}\in\mathcal{S}^\star} \frac{\prod_{j=1}^J \gamma_{k,j}^{s_j}(1-\gamma_{k,j})^{1-s_j}}{1-\prod_{j=1}^J(1-\gamma_{k,j})} \mathcal{N}_J\left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s})\right). \tag{14}$$

The proof is detailed in Appendix~B.

Definition 2.6 (Non-Informative Observation)**.** *Observation i is non-informative if its activation profile is null, $\boldsymbol{s}_i = \mathbf{0}_J$. Its distribution is then the background distribution $\mathcal{N}_J(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, whatever its cluster z_i, so that it carries no information on the clusters. It is informative if $\boldsymbol{s}_i \neq \mathbf{0}_J$.*

Following Toussile et al. (2026), who introduced this concept under diagonal conditional independence, the non-informative observations form the component of weight $\widetilde{\omega}_0$ in (12). This weight is not a free parameter: it is the probability of the null profile, a product of inactivation probabilities, and it is small when the activation rates are large. The model therefore represents non-informative observations as cluster members whose coordinates are all inactive, not as a separate population. Data may also contain observations unrelated to the clusters, in a proportion that need not equal $\widetilde{\omega}_0$; Appendix D presents a formulation with a free weight for them, to which the results of this paper carry over.

Under the canonical $(K+1)$-component mixture partition (12), each observation $\boldsymbol{y}_i$ is associated with an extended cluster index $z_i^{\text{ext}} \in \{0,1,\ldots,K\}$, where $z_i^{\text{ext}} = 0$ corresponds to the non-informative background component and $z_i^{\text{ext}} = k \in \{1,\ldots,K\}$ to substantive cluster

k. Applying Bayes' theorem to (12) directly yields the exact $(K+1)$-mixture posterior responsibilities:

$$\tau_{i,0} = p(z_i^{\text{ext}} = 0 \mid \boldsymbol{y}_i) = \frac{\widetilde{\omega}_0 \, \mathcal{N}_J(\boldsymbol{y}_i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{p(\boldsymbol{y}_i)}, \tag{15}$$

$$\tau_{i,k} = p(z_i^{\text{ext}} = k \mid \boldsymbol{y}_i) = \frac{\widetilde{\omega}_k \, f_k^{\text{in}}(\boldsymbol{y}_i)}{p(\boldsymbol{y}_i)}, \qquad k \in \{1, \ldots, K\}, \tag{16}$$

where $\sum_{k=0}^{K} \tau_{i,k} = 1$, and $\widetilde{\omega}_k f_k^{\text{in}}(\boldsymbol{y}_i) = \sum_{\boldsymbol{s} \in \mathcal{S}^\star} \exp\{\eta_{i,k}(\boldsymbol{s})\}$ is evaluated via the configuration log-weights

$$\eta_{i,k}(\boldsymbol{s}) = \log \omega_k + \log p(\boldsymbol{s} \mid z_i = k) + \log \mathcal{N}_J\left(\boldsymbol{y}_i; \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s})\right), \qquad \boldsymbol{s} \in \mathcal{S}^\star. \tag{17}$$

The optimal Bayes classification rule on $\{0, 1, \ldots, K\}$ assigns observation i to:

$$\widehat{z}_i^{\text{ext}} = \arg \max_{k \in \{0,1,\ldots,K\}} \tau_{i,k}. \tag{18}$$

This extended MAP rule admits an intuitive and transparent two-step decision structure:

1. *Step 1: Non-informative vs Informative Screening.* Observation i is isolated as non-informative ($\widehat{z}_i^{\text{ext}} = 0$) if and only if the background responsibility strictly dominates every substantive cluster:

$$\tau_{i,0} > \max_{1 \le k \le K} \tau_{i,k} \iff \widetilde{\omega}_0 \, \mathcal{N}_J(\boldsymbol{y}_i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) > \max_{1 \le k \le K} \widetilde{\omega}_k \, f_k^{\text{in}}(\boldsymbol{y}_i). \tag{19}$$

2. *Step 2: Substantive Cluster Assignment.* Otherwise, observation i is assigned to $\widehat{z}_i^{\text{ext}} = \arg\max_{1 \le k \le K} \tau_{i,k}$. Because the normalization constant $p(\boldsymbol{y}_i)(1 - \tau_{i,0})$ is shared across all substantive clusters, this is identical to maximizing the conditional responsibilities given that the observation is informative ($\boldsymbol{s}_i \neq \mathbf{0}_J$):

$$\tau_{i,k}^{\text{in}} = P(z_i = k \mid \boldsymbol{y}_i, \boldsymbol{s}_i \neq \mathbf{0}_J) = \frac{\tau_{i,k}}{1 - \tau_{i,0}} = \frac{\widetilde{\omega}_k \, f_k^{\text{in}}(\boldsymbol{y}_i)}{\sum_{l=1}^{K} \widetilde{\omega}_l \, f_l^{\text{in}}(\boldsymbol{y}_i)}. \tag{20}$$

Operationally, because $\tau_{i,0}$ incorporates the prior weight $\widetilde{\omega}_0$, an observation is classified as non-informative only when the background distribution is favored by a margin large enough to overcome this small prior probability. Consequently, this classification is naturally conservative: background observations that happen to fall within the high-density core of a cluster cannot be reliably distinguished from genuine cluster members by any rule.

2.5 Identifiability of the Activation Structure

The latent profiles $\boldsymbol{s}_i$ are meant to tell which variables are relevant in each cluster. This raises an identifiability question that is specific to the selection problem: can an inactive variable be distinguished from an active one whose distribution resembles the background?

Proposition 2.7 (Identifiability of the Activation Structure)**.** *Assume that the background is fixed, with* $\boldsymbol{\Sigma}_0 = \mathrm{diag}(\sigma_{0,1}^2, \ldots, \sigma_{0,J}^2)$*, and write* $\mathcal{A}_k = \{j : \gamma_{k,j} > 0\}$ *for the support of cluster* k*.*

(i) Degenerate blocks are not identifiable. *Let* $\boldsymbol{\theta}$ *be a parameter for which variable* j *is inactive in cluster* k *(*$\gamma_{k,j} = 0$*). For any* $\gamma' \in [0,1]$*, let* $\boldsymbol{\theta}'$ *coincide with* $\boldsymbol{\theta}$ *except for* $\gamma'_{k,j} = \gamma'$*,* $\mu'_{k,j} = \mu_{0,j}$*,* $\Sigma'_{k,jj} = \sigma_{0,j}^2$ *and* $\Sigma'_{k,jl} = \Sigma'_{k,lj} = 0$ *for* $l \neq j$ *(so that* $\boldsymbol{\Sigma}'_k$ *is block diagonal and positive definite). Then* $\boldsymbol{\theta}$ *and* $\boldsymbol{\theta}'$ *define the same distribution of* $\boldsymbol{y}_i$*.*

(ii) Otherwise, the activation structure is identifiable. *Let* $\boldsymbol{\theta}$ *and* $\boldsymbol{\theta}'$*, with the same number* K *of clusters, both satisfy: (a)* $\omega_k > 0$ *and* $\mathcal{A}_k \neq \emptyset$ *for every* k*; (b) no block is degenerate,* $(\mu_{k,j}, \Sigma_{k,jj}) \neq (\mu_{0,j}, \sigma_{0,j}^2)$ *for every* $j \in \mathcal{A}_k$*; (c) the clusters are distinct on their common supports,* $(\boldsymbol{\mu}_{k,A}, \boldsymbol{\Sigma}_{k,AA}) \neq (\boldsymbol{\mu}_{l,A}, \boldsymbol{\Sigma}_{l,AA})$ *for* $k \neq l$ *and every non-empty* $A \subseteq \mathcal{A}_k \cap \mathcal{A}_l$*; (d) at most one cluster has a support reduced to one variable. If* $\boldsymbol{\theta}$ *and* $\boldsymbol{\theta}'$ *define the same distribution of* $\boldsymbol{y}_i$*, then, up to a permutation of the labels,* $\mathcal{A}_k = \mathcal{A}'_k$*,* $\omega_k = \omega'_k$*,* $\gamma_{k,j} = \gamma'_{k,j}$*,* $\mu_{k,j} = \mu'_{k,j}$ *and* $\Sigma_{k,jl} = \Sigma'_{k,jl}$ *for every* k *and* $j, l \in \mathcal{A}_k$*.*

The proof is detailed in Appendix~B. Part (i) shows that condition (b) cannot be dropped: a variable whose active distribution in a cluster is the background, with no correlation with the other variables, can be declared active with any rate. Part (ii) shows that this is the only obstruction: once such degenerate blocks are removed, the support of the activation rates, the weights and the parameters on the supports are determined by the distribution of the data. Among the parameters that define the same distribution, the identified support is therefore the smallest one, and it is the target of the selection of Section 3. The parameters outside the supports never enter the distribution of the data and are not identified. Condition (c) holds for almost every parameter. Condition (d) cannot be dropped either: when the support of a cluster is a single variable j, its null profile merges with the background and only $\omega_k \gamma_{k,j}$ is determined, so that the weights of two such clusters cannot be separated. Condition (b) excludes a variable that is relevant in a cluster only through its correlations with the other active variables, with the mean and the variance of the background; the designs of Section 5 keep a mean shift. The identification of the background itself is discussed in Remark 3.4.

Part (i) has a direct consequence for estimation: since the model in which variable j is active in cluster k contains the model in which it is inactive, the likelihood alone cannot select the inactive variables. Deciding that a variable is inactive is a choice between nested models, which Section 3 makes with a penalized criterion.

3 Estimation by Penalized Classification EM

Directly maximizing the mixture likelihood requires summing over 2^J activation profiles per observation. We avoid this combinatorial bottleneck through a penalized classification EM scheme (Celeux and Govaert, 1992), optimizing local profiles and cluster parameters under parsimony penalties. The number of clusters K is assumed known, as is common in work on variable selection for clustering (Witten and Tibshirani, 2010); when it is unknown, it can be chosen by fitting the model for several values of K with a standard model selection criterion. The background parameters $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ are updated in alternation (Section 3.4).

3.1 Bounded Active Densities under an Eigenvalue Floor

Lemma 3.1 (Uniform Control of the Active Covariances)**.** *Let $\boldsymbol{\Sigma} \in \mathcal{S}_{++}^J$ with eigenvalues $\lambda_1(\boldsymbol{\Sigma}) \geq \cdots \geq \lambda_J(\boldsymbol{\Sigma}) > 0$, and let $\boldsymbol{s}$ be a profile with $d(\boldsymbol{s}) \geq 1$.*

(i) The principal submatrix $\mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{s})^\top$ is positive definite, and

$$\lambda_J(\boldsymbol{\Sigma}) \leq \lambda_{\min}\left(\mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{s})^\top\right) \leq \lambda_{\max}\left(\mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{s})^\top\right) \leq \lambda_1(\boldsymbol{\Sigma}), \tag{21}$$

so that its condition number satisfies

$$\kappa\left(\mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{s})^\top\right) \leq \kappa(\boldsymbol{\Sigma}). \tag{22}$$

(ii) If $\lambda_J(\boldsymbol{\Sigma}) \geq \delta > 0$, then for all $\boldsymbol{y}, \boldsymbol{\mu} \in \mathbb{R}^J$,

$$\mathcal{N}_{d(\boldsymbol{s})}\left(\mathbf{D}(\boldsymbol{s})\boldsymbol{y}\ ;\ \mathbf{D}(\boldsymbol{s})\boldsymbol{\mu}, \mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{s})^\top\right) \leq (2\pi\delta)^{-d(\boldsymbol{s})/2}. \tag{23}$$

Item (i) is a consequence of Cauchy's interlacing theorem (Horn and Johnson, 2012, Section 4.3): a single lower bound on the smallest eigenvalue of the parent matrix controls the $2^J - 1$ active covariance matrices at once. It does not prevent the parent matrix itself from degenerating during estimation, since the likelihood of a Gaussian mixture is unbounded when a covariance matrix becomes singular (Fraley and Raftery, 2007). The parent matrices must therefore be estimated under the constraint $\lambda_J(\boldsymbol{\Sigma}_k) \geq \delta$; by item (ii), every active density, and hence the estimation criterion, is then bounded. The proof is given in Appendix A.

3.2 Penalized Classification Criterion

Let $\boldsymbol{r} = (r_{i,k})$ be cluster responsibilities ($r_{i,k} \geq 0$, $\sum_k r_{i,k} = 1$) and, for each observation i and cluster k, let $\boldsymbol{s}_{i,k} \in \{0,1\}^J$ be a classified activation profile. Let $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\gamma}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$ collect all model parameters.

Under the structured variational family $q(\boldsymbol{z}, \boldsymbol{s}) = \prod_{i=1}^{n} q_i(z_i, \boldsymbol{s}_i)$ with $q_i(z_i = k, \boldsymbol{s}_i = \boldsymbol{s}) = r_{i,k}\, \mathbb{I}(\boldsymbol{s} = \boldsymbol{s}_{i,k})$, which pairs soft cluster assignments with cluster-conditional point masses on the classified profiles, the variational Evidence Lower Bound (ELBO) $\mathbb{E}_q\big[\log \frac{p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{s}|\boldsymbol{\theta})}{q(\boldsymbol{z},\boldsymbol{s})}\big]$ defines the classification criterion:

$$\mathcal{F}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{i,k} \Big[\log \omega_k - \log r_{i,k} + g_{i,k}(\boldsymbol{s}_{i,k}; \boldsymbol{\theta}) \Big], \tag{24}$$

where $g_{i,k}(\boldsymbol{s}; \boldsymbol{\theta}) = \log p(\boldsymbol{s} \mid z_i = k; \boldsymbol{\gamma}) + \log p(\boldsymbol{y}_i \mid z_i = k, \boldsymbol{s}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, with profile prior $p(\boldsymbol{s} \mid z_i = k; \boldsymbol{\gamma}) = \prod_{j=1}^{J} \gamma_{k,j}^{s_j} (1 - \gamma_{k,j})^{1-s_j}$ and emission density (5). In (24), the point masses on $\boldsymbol{s}_{i,k}$ contribute zero conditional entropy ($\log 1 = 0$), so that the entropy term $\mathcal{H}(q) = -\sum_{i,k} r_{i,k} \log r_{i,k}$ reflects only cluster uncertainty; hence $\mathcal{F}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}) \leq \log p(\boldsymbol{y}; \boldsymbol{\theta})$ is the exact variational ELBO under q. Like any classification criterion (Celeux and Govaert, 1992), it yields estimates that differ from the maximum likelihood estimates.

Maximizing $\mathcal{F}$ alone cannot select the relevant variables. By Proposition~2.7(i), the likelihood does not distinguish a degenerate block from an inactive one, and cluster-specific parameters fitted to the data describe an irrelevant variable at least as well as the background, so that maximizing $\mathcal{F}$ has no systematic reason to make a variable inactive. (Along the iterations, profile classification may still leave irrelevant variables inactive; Section 5.1 shows that this occurs in large samples but not in small ones.)

To resolve this non-identifiability, we decouple structural variable selection from the estimation of activation rates by introducing a binary matrix $\mathbf{B} \in \{0,1\}^{K \times J}$ of active blocks. Block (k, j) is active ($B_{k,j} = 1$) if variable j is permitted to be relevant in cluster k, in which case its activation rate satisfies $\gamma_{k,j} \in [\epsilon, 1 - \epsilon]$; otherwise $B_{k,j} = 0$, which strictly enforces $\gamma_{k,j} = 0$ and $s_{i,k,j} = 0$ for all i. We then arbitrate between nested activation structures by penalizing the model complexity associated with $\mathbf{B}$, in the spirit of BIC (Schwarz, 1978) and ICL (Biernacki et al., 2000):

$$\mathcal{G}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}, \mathbf{B}) = \mathcal{F}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}) - \frac{\log n}{2}\, \nu(\mathbf{B}), \qquad \nu(\mathbf{B}) = \sum_{k=1}^{K} \left[3a_k + \frac{a_k(a_k - 1)}{2} \right], \quad a_k = \sum_{j=1}^{J} B_{k,j}, \tag{25}$$

where $\nu(\mathbf{B})$ counts three parameters for each active block (mean, variance, and activation rate), along with the $a_k(a_k - 1)/2$ pairwise covariances between active variables in each cluster. Parameters associated with inactive blocks do not enter $\mathcal{G}$. Because the number of clusters K is fixed, the mixing proportions contribute a constant penalty $(K - 1)\frac{\log n}{2}$ and are omitted. The criterion is maximized under the spectral constraints $\lambda_J(\boldsymbol{\Sigma}_k) \geq \delta$ for all k.

Proposition 3.2 (Boundedness and Parsimony of the Penalized Criterion)**.** *Let the criterion $\mathcal{G}$ be defined by (25).*

(i) Uniform Boundedness. *Under the spectral constraints $\lambda_J(\boldsymbol{\Sigma}_k) \geq \delta > 0$, $\mathcal{G}$ is bounded above across all configurations of $(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}, \mathbf{B})$.*

(ii) Deterministic Resolution of Non-Identifiability. *Let* $\mathbf{B}$ *and* $\mathbf{B}'$ *be two activation structures such that* $\mathbf{B}'$ *strictly contains* $\mathbf{B}$ *($B_{k,j} \leq B'_{k,j}$ for all (k,j), with at least one strict inequality). Let $(\boldsymbol{s}', \boldsymbol{\theta}')$ coincide with $(\boldsymbol{s}, \boldsymbol{\theta})$ except on the extra blocks of $\mathbf{B}'$, whose parameters are set to those of the background (as in Proposition 2.7(i), with $\sigma^2_{0,j} \geq \delta$). Then*

$$\mathcal{G}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}, \mathbf{B}) - \mathcal{G}(\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{\theta}', \mathbf{B}') \geq \frac{\log n}{2} \left[\nu(\mathbf{B}') - \nu(\mathbf{B})\right] > 0, \quad \forall n > 1. \tag{26}$$

Among parameters that define the same distribution, the penalty thus strictly prefers the one without degenerate blocks, whose support is identified by Proposition 2.7(ii).

The proof is detailed in Appendix~C. Proposition~3.2 has two consequences: item~(i) rules out the singularities of unconstrained Gaussian mixtures and ensures that the values of $\mathcal{G}$ along the algorithm converge (Section 3.3), while item~(ii) shows that degenerate blocks, which leave the distribution unchanged by Proposition~2.7(i), are strictly penalized. It does not by itself guarantee that the maximizer of $\mathcal{G}$ recovers the minimal support, which is studied empirically in Section 5.1.

3.3 Monotone Block Coordinate Ascent

The criterion $\mathcal{G}$ is maximized by block coordinate ascent. Each step below maximizes $\mathcal{G}$, or does not decrease it, with the other blocks held fixed.

Proposition 3.3 (Monotone Updates)**.** *Each of the following steps does not decrease $\mathcal{G}$.*

1. ***Responsibilities.*** $r_{i,k} \propto \omega_k \exp\{g_{i,k}(\boldsymbol{s}_{i,k}; \boldsymbol{\theta})\}$ *(exact maximizer).*
2. ***Profiles.*** *For each (i,k), $\boldsymbol{s}_{i,k}$ is changed one coordinate at a time, among active blocks, and a change is kept only if it increases $g_{i,k}$. For a coordinate $j \notin \mathcal{A} = \mathcal{A}(\boldsymbol{s}_{i,k})$, adding j changes $g_{i,k}$ by*

$$\begin{aligned} &\log \frac{\gamma_{k,j}}{1-\gamma_{k,j}} \\ &\quad + \log \mathcal{N}\Big(y_{i,j};\, \mu_{k,j} + \boldsymbol{\Sigma}_{k,j\mathcal{A}} \boldsymbol{\Sigma}^{-1}_{k,\mathcal{A}\mathcal{A}} (\boldsymbol{y}_{i,\mathcal{A}} - \boldsymbol{\mu}_{k,\mathcal{A}}),\, \Sigma_{k,jj} - \boldsymbol{\Sigma}_{k,j\mathcal{A}} \boldsymbol{\Sigma}^{-1}_{k,\mathcal{A}\mathcal{A}} \boldsymbol{\Sigma}_{k,\mathcal{A}j}\Big) \\ &\quad - \log \mathcal{N}(y_{i,j};\, \mu_{0,j}, \sigma^2_{0,j}), \end{aligned} \tag{27}$$

   the conditional law of $y_{i,j}$ given the active coordinates; removing $j \in \mathcal{A}$ changes it by the opposite of the same quantity computed with $\mathcal{A} \setminus \{j\}$. The profile obtained, on the active blocks, from the univariate log-odds (with $\mathcal{A} = \emptyset$ in (27)) is used as a starting point when it gives a larger $g_{i,k}$ than the current profile.
3. ***Mixing proportions and activation rates.*** $\omega_k = n^{-1} \sum_i r_{i,k}$ *and, for active blocks,* $\gamma_{k,j} = \min\{\max(\sum_i r_{i,k} s_{i,k,j} / \sum_i r_{i,k}, \epsilon), 1-\epsilon\}$ *(exact maximizers).*

4. ***Cluster means and covariances.*** *For each cluster, the coordinates that are inactive for observation i are treated as missing data, and one iteration of the EM algorithm for a Gaussian sample with missing values (Dempster et al., 1977; Little and Rubin, 2019) is applied: with $\mathcal{A}_i = \mathcal{A}(\boldsymbol{s}_{i,k})$ and $\mathcal{M}_i$ its complement, the completed vector $\widehat{\boldsymbol{x}}_{i,k}$ equals $\boldsymbol{y}_i$ on $\mathcal{A}_i$ and $\boldsymbol{\mu}_{k,\mathcal{M}_i} + \boldsymbol{\Sigma}_{k,\mathcal{M}_i\mathcal{A}_i}\boldsymbol{\Sigma}_{k,\mathcal{A}_i\mathcal{A}_i}^{-1}(\boldsymbol{y}_{i,\mathcal{A}_i} - \boldsymbol{\mu}_{k,\mathcal{A}_i})$ on $\mathcal{M}_i$, with conditional covariance $\mathbf{C}_{i,k}$ on $\mathcal{M}_i \times \mathcal{M}_i$, and*

$$\boldsymbol{\mu}_k = \frac{\sum_i r_{i,k}\,\widehat{\boldsymbol{x}}_{i,k}}{\sum_i r_{i,k}}, \qquad \mathbf{S}_k = \frac{\sum_i r_{i,k}\left[(\widehat{\boldsymbol{x}}_{i,k} - \boldsymbol{\mu}_k)(\widehat{\boldsymbol{x}}_{i,k} - \boldsymbol{\mu}_k)^\top + \mathbf{C}_{i,k}\right]}{\sum_i r_{i,k}},$$
$$\boldsymbol{\Sigma}_k = \mathbf{U}_k \max(\boldsymbol{\Lambda}_k, \delta)\mathbf{U}_k^\top, \tag{28}$$

where $\mathbf{S}_k = \mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{U}_k^\top$ and the maximum is taken entrywise on the eigenvalues; the sums run over observations with at least one active coordinate in cluster k.

5. ***Active blocks.*** *For each (k, j), the block is switched off (setting $s_{i,k,j} = 0$ for all i) if it is active, or switched on if it is inactive, and the change is kept only if it increases $\mathcal{G}$. When a block is switched on, $\mu_{k,j}$ and $\Sigma_{k,jj}$ are initialized by the weighted mean and variance of $y_{\bullet,j}$ in cluster k (the variance being at least δ), $\Sigma_{k,jl} = 0$ for $l \neq j$, $s_{i,k,j} = 1$ where $\mathcal{N}(y_{i,j}; \mu_{k,j}, \Sigma_{k,jj}) > \mathcal{N}(y_{i,j}; \mu_{0,j}, \sigma_{0,j}^2)$, and $\gamma_{k,j}$ by step 3.*

The proof and pseudo-code (Algorithm 1) are in Appendix C. By Propositions 3.2 and 3.3, the sequence of criterion values $(\mathcal{G}^{(t)})$ is non-decreasing and bounded, hence convergent; the convergence of the parameters is not guaranteed, and the limit may correspond to a local maximum. The algorithm operates as a generalized classification EM: unlike classical CEM (Celeux and Govaert, 1992), it retains soft cluster responsibilities $r_{i,k}$, and its regularized covariance update guarantees monotonicity without closed-form global maximization. Alternating discrete moves on the model with the other updates, within the maximization of a single criterion, echoes Marbac and Sedki (2017), whose algorithm alternates a partition step and a model step, and the structural EM of Fop et al. (2019), which alternates moves on cluster-specific covariance graphs with parameter estimation. Furthermore, treating inactive coordinates as missing data in step 4 relates to the cellwise robust clustering of Zaccaria et al. (2025), though here the sparsity pattern is cluster-specific and regularized by the discrete block penalty $\nu(\mathbf{B})$.

In practice, the algorithm starts from soft responsibilities supplied by k-means++, with every block active ($\mathbf{B} = \mathbf{1}_{K\times J}$), every profile fully active ($\boldsymbol{s}_{i,k} = \mathbf{1}_J$), and intermediate activation rates $\gamma_{k,j} = 1/2$. Initializing rates at the upper bound $\gamma_{k,j} = 1 - \epsilon$ would make fully active profiles an artificial fixed point of step 2, whereas $\gamma_{k,j} = 1/2$ gives uninformative initial prior odds and lets the conditional likelihood drive early profile updates. Several starting points are evaluated, and the solution achieving the highest penalized criterion $\mathcal{G}$ is retained.

Computational cost.

In the profile step, the inverse of the active covariance matrix is updated exactly when a coordinate is added or removed, so that each candidate change costs $\mathcal{O}(d^2)$, with d the number

of active coordinates. The mean and covariance step costs $\mathcal{O}(nKJ^3)$. In the block step, the change of $\mathcal{G}$ due to switching a block on or off has a closed form, which costs $\mathcal{O}(n)$ for switching on and $\mathcal{O}(nd^3)$ for switching off. The estimation of the background (Section 3.4) fits 2^K univariate mixtures per feature, which restricts it to moderate numbers of clusters. Computing times are reported in Appendix E.

3.4 Estimation of the Background

We assume a diagonal background covariance $\mathbf{\Sigma}_0 = \mathrm{diag}(\sigma_{0,1}^2, \ldots, \sigma_{0,J}^2)$, reflecting the premise that inactive coordinates carry no dependence structure, and making the background term separable across coordinates, unlike a full-covariance background (Tadesse et al., 2005). The background is estimated separately from $\mathcal{G}$. Under the model, feature j within cluster k follows a marginal two-component mixture $\gamma_{k,j}\,\mathcal{N}(\mu_{k,j}, \Sigma_{k,jj}) + (1-\gamma_{k,j})\,\mathcal{N}(\mu_{0,j}, \sigma_{0,j}^2)$ with a shared background. Given responsibilities $r_{i,k}$, the background parameters $(\mu_{0,j}, \sigma_{0,j}^2)$ are estimated feature by feature via an auxiliary EM algorithm maximizing the weighted log-likelihood:

$$\sum_{i=1}^{n}\sum_{k=1}^{K} r_{i,k} \log\left\{\alpha_{k,j}\,\mathcal{N}(y_{i,j}; m_{k,j}, v_{k,j}) + (1-\alpha_{k,j})\,\mathcal{N}(y_{i,j}; \mu_{0,j}, \sigma_{0,j}^2)\right\}, \tag{29}$$

where $(\alpha_{k,j}, m_{k,j}, v_{k,j})$ are auxiliary cluster parameters, used only to estimate the background. To prevent these components from absorbing background tails on irrelevant features, the active subset $\{k : \alpha_{k,j} > 0\}$ is selected by BIC among the 2^K candidate configurations; this selection is not the support of the activation rates. In practice, background estimation alternates with the algorithm of Section 3: initialized from k-means++ responsibilities, the background is updated at each outer iteration until stabilization. If cycling occurs across outer iterations, the candidate maximizing $\mathcal{G}$ is retained. Each fit increases $\mathcal{G}$ for its background; the alternation itself is not covered by Proposition 3.3.

Remark 3.4 (Identifiability of the Background)**.** *The background distribution $\mathcal{N}(\mu_{0,j}, \sigma_{0,j}^2)$ is identified by (29) whenever feature j is inactive in at least two clusters (or partially inactive, $0 < \alpha_{k,j} < 1$, in an active cluster), as the shared component is then uniquely determined across groups. In contrast, when the background appears in only one cluster ($\alpha_{k,j} = 1$ in all others, such as when $K = 2$), assigning the background role to one cluster rather than another yields identical mixture likelihoods. In this ambiguous case, the data support the separation between groups, but the active or background status is a matter of convention: among models within 10 BIC units of the minimum, we select the background whose mean is closest to the sample median of feature j, and report the feature as ambiguous. Finally, if feature j is active in all clusters ($\alpha_{k,j} = 1$ everywhere), the background does not enter the data distribution.*

4 Post-Estimation Diagnostics

4.1 Classification of the Observations

Given estimated parameters $\widehat{\boldsymbol{\theta}}$, the two-step rule (19)–(20) evaluates $\eta_{i,k}(\boldsymbol{s}) = \log \omega_k + g_{i,k}(\boldsymbol{s}; \widehat{\boldsymbol{\theta}})$. Because only profiles supported by the estimated active set $\mathcal{A}_k = \{j : \gamma_{k,j} > 0\}$ have positive probability, the sum and maximum over $\mathcal{S}^\star$ involve only $2^{|\mathcal{A}_k|} - 1$ configurations, which are computed exactly for moderate supports. For larger supports, following the classification principle, both are approximated by the converged profile $\widehat{\boldsymbol{s}}_{i,k}$ (or $-\infty$ if null). Standard K-class MAP assignment without outlier screening is given by $\widehat{z}_i = \arg\max_k r_{i,k}$.

Inactivation Rate.

Alternatively, building on Toussile et al. (2026), a heuristic coordinate inactivation rate screens observations without requiring the prior weight $\widetilde{\omega}_0$:

$$\bar{\rho}_i = 1 - \frac{1}{|\mathcal{A}_{\widehat{z}_i}|} \sum_{j \in \mathcal{A}_{\widehat{z}_i}} \sigma(\Delta_{i,\widehat{z}_i,j}), \tag{30}$$

where $\sigma(\Delta_{i,k,j})$ is the conditional activation probability of cell (i, j) in cluster k derived from increment (27). Observation i is screened as non-informative when $\bar{\rho}_i \geq 1/2$, meaning that most features in its cluster support are inactive. While computationally trivial, this rate tends to over-screen informative observations whenever true activation rates are moderate (Section 5.4).

4.2 Relevance of the Variables

The model provides two complementary levels of variable importance. At the cluster level, the support $\mathcal{A}_k = \{j : \gamma_{k,j} > 0\}$, selected by the penalized criterion, identifies the variables relevant to cluster k; variables outside all supports are globally irrelevant. Within each support, the activation rate $\hat{\gamma}_{k,j}$ quantifies how frequently feature j follows the cluster distribution rather than the background, while cell-level conditional probabilities $\sigma(\Delta_{i,k,j})$ measure activation for individual observations. The magnitude of departure from the background is directly captured by the standardized mean shift $(\hat{\mu}_{k,j} - \hat{\mu}_{0,j})/\hat{\sigma}_{0,j}$ and the variance ratio $\hat{\Sigma}_{k,jj}/\hat{\sigma}^2_{0,j}$, requiring no ad hoc importance index.

5 Simulation Study

All simulations are implemented in the `FCGMMLFR.jl` package, with $R = 20$ replications per configuration. Tables report Monte Carlo means with standard errors in parentheses.

The base design comprises $K = 3$ equiprobable clusters and $J = 20$ features with $n = 1\,000$. Each cluster k is active on four features $\{4k-3, \ldots, 4k\}$ with activation rate γ (and zero elsewhere), means $\mu_{k,j} = \pm\Delta$ (signs $+, +, -, -$ within the block), unit variances, and autoregressive correlations $\rho^{|a-b|}$; inactive cells follow the background $\mathcal{N}(0,1)$. Thus, 12 of the 60 blocks are active. Unless stated otherwise, $\Delta = 2.5$, $\gamma = 0.9$, and $\rho = 0.5$.

Models are fitted with the default penalty (25), eigenvalue floor $\delta = 10^{-3} \min_j \sigma_{0,j}^2$, $\epsilon = 10^{-4}$, and five k-means++ restarts; the background is updated via (29) until stabilization. Performance is assessed by: (i) partition quality via the adjusted Rand index (ARI); (ii) block support recovery via True and False Positive Rates (TPR, FPR); (iii) cell-level profile accuracy via active and inactive cell classification rates along with Expected Calibration Error (ECE) of $\sigma(\Delta_{i,k,j})$ over ten bins; and (iv) active covariance estimation on $\mathcal{A}_k$ via relative Frobenius error $\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}_k\mathcal{A}_k} - \boldsymbol{\Sigma}_{k,\mathcal{A}_k\mathcal{A}_k}\|_F / \|\boldsymbol{\Sigma}_{k,\mathcal{A}_k\mathcal{A}_k}\|_F$ and maximum absolute correlation error.

5.1 Recovery of the Active Structure

Table 1 reports structure recovery across sample sizes $n \in \{200, 500, 1\,000, 2\,000\}$ and separations $\Delta \in \{1.5, 2.5, 3.5\}$. Across all configurations, the default penalty prevents spurious activations (FPR $= 0$). Active blocks are fully recovered (TPR $= 1$) except under low separation with small samples ($\Delta = 1.5$: TPR $= 0.713$ at $n = 200$, 0.996 at $n = 500$). Partition accuracy (ARI) increases with separation and stabilizes with n, while covariance errors steadily decrease (e.g., relative Frobenius error drops from 0.262 to 0.102 at $\Delta = 2.5$). The estimation error on activation rates $|\hat{\gamma} - \gamma|$ decreases with n at $\Delta = 3.5$ (0.030 to 0.012) and stabilizes near 0.03 at $\Delta = 2.5$; at $\Delta = 1.5$ it stays near 0.10 up to $n = 2\,000$ ($\hat{\gamma} \approx 1$ vs true $\gamma = 0.9$): under distribution overlap, cell classification induces an upward bias that does not decrease with n in this range.

Table 2 examines sensitivity to the penalty charge per parameter at $\Delta = 2.5$. Without penalty, 26.7% of inactive blocks are falsely activated at $n = 200$; for $n \geq 500$, however, unpenalized and penalized fits strictly coincide (FPR $= 0$, TPR $= 1$). Profile classification thus naturally suppresses irrelevant features in large samples, whereas the penalty is indispensable in small samples.

5.2 Estimation of the Background

Table 3 evaluates four treatments of the background: the alternating shared-component estimator (29); the same estimator with normal rank scores and an added 10-point BIC cost per cluster component; the pooled sample median and $1.4826 \times$ MAD; and the oracle background. Four Gaussian configurations vary the proportion of active cells per relevant feature (30%, 60%, 90%), alongside two non-Gaussian designs with unit variance (Student t_3 and centred exponential).

Table 1: Recovery of the active structure with the default penalty: partition, active blocks, mean absolute error of the activation rates on the true active blocks and covariance errors ($\gamma = 0.9$, $\rho = 0.5$, $R = 20$).

Δ	n	ARI	TPR (blocks)	FPR (blocks)	$\lvert\hat{\gamma} - \gamma\rvert$	Frobenius	max corr.
1.5	200	0.646 (0.033)	0.713 (0.039)	0.000 (0.000)	0.317 (0.033)	0.338 (0.016)	0.285 (0.014)
1.5	500	0.811 (0.007)	0.996 (0.004)	0.000 (0.000)	0.091 (0.004)	0.223 (0.006)	0.177 (0.009)
1.5	1000	0.819 (0.005)	1.000 (0.000)	0.000 (0.000)	0.097 (0.001)	0.209 (0.006)	0.162 (0.006)
1.5	2000	0.819 (0.004)	1.000 (0.000)	0.000 (0.000)	0.099 (0.000)	0.194 (0.004)	0.160 (0.006)
2.5	200	0.983 (0.004)	1.000 (0.000)	0.000 (0.000)	0.042 (0.001)	0.262 (0.011)	0.198 (0.009)
2.5	500	0.988 (0.002)	1.000 (0.000)	0.000 (0.000)	0.032 (0.002)	0.190 (0.006)	0.142 (0.008)
2.5	1000	0.984 (0.001)	1.000 (0.000)	0.000 (0.000)	0.030 (0.001)	0.123 (0.005)	0.096 (0.005)
2.5	2000	0.988 (0.001)	1.000 (0.000)	0.000 (0.000)	0.029 (0.001)	0.102 (0.005)	0.066 (0.003)
3.5	200	0.999 (0.001)	1.000 (0.000)	0.000 (0.000)	0.030 (0.001)	0.245 (0.011)	0.194 (0.010)
3.5	500	0.999 (0.001)	1.000 (0.000)	0.000 (0.000)	0.021 (0.001)	0.154 (0.005)	0.129 (0.006)
3.5	1000	0.999 (0.000)	1.000 (0.000)	0.000 (0.000)	0.016 (0.001)	0.119 (0.005)	0.089 (0.004)
3.5	2000	0.999 (0.000)	1.000 (0.000)	0.000 (0.000)	0.012 (0.001)	0.089 (0.004)	0.065 (0.004)

Table 2: Sensitivity to the penalty ($\Delta = 2.5$, $R = 20$): charge per parameter of the active structure.

n	penalty per parameter	ARI	TPR (blocks)	FPR (blocks)
200	none	0.790 (0.060)	0.983 (0.010)	0.267 (0.065)
200	$\frac{1}{4}\log n$	0.983 (0.004)	1.000 (0.000)	0.000 (0.000)
200	$\frac{1}{2}\log n$ (default)	0.983 (0.004)	1.000 (0.000)	0.000 (0.000)
200	$\log n$	0.984 (0.003)	1.000 (0.000)	0.000 (0.000)
500	none	0.988 (0.002)	1.000 (0.000)	0.000 (0.000)
500	$\frac{1}{4}\log n$	0.988 (0.002)	1.000 (0.000)	0.000 (0.000)
500	$\frac{1}{2}\log n$ (default)	0.988 (0.002)	1.000 (0.000)	0.000 (0.000)
500	$\log n$	0.988 (0.002)	1.000 (0.000)	0.000 (0.000)
1000	none	0.984 (0.001)	1.000 (0.000)	0.000 (0.000)
1000	$\frac{1}{4}\log n$	0.984 (0.001)	1.000 (0.000)	0.000 (0.000)
1000	$\frac{1}{2}\log n$ (default)	0.984 (0.001)	1.000 (0.000)	0.000 (0.000)
1000	$\log n$	0.984 (0.001)	1.000 (0.000)	0.000 (0.000)
2000	none	0.988 (0.001)	1.000 (0.000)	0.000 (0.000)
2000	$\frac{1}{4}\log n$	0.988 (0.001)	1.000 (0.000)	0.000 (0.000)
2000	$\frac{1}{2}\log n$ (default)	0.988 (0.001)	1.000 (0.000)	0.000 (0.000)
2000	$\log n$	0.988 (0.001)	1.000 (0.000)	0.000 (0.000)

Table 3: Estimation of the background: partition, recovery of the active blocks and estimated background standard deviation on relevant and irrelevant features (true value 1; on the transformed scale for the normal-scores variant), for four treatments of the background ($n = 1\,000$, $R = 20$). The condition gives the proportion of active cells among the cells of a relevant feature ($\Delta = 2.5$ unless stated otherwise) or the distribution of the background.

condition	background	ARI	TPR (blocks)	FPR (blocks)	$\hat{\sigma}_0$ relevant	$\hat{\sigma}_0$ irrelevant
30%, $\Delta = 1.5$	shared component	0.820 (0.005)	1.000 (0.000)	0.000 (0.000)	1.004 (0.002)	1.001 (0.002)
30%, $\Delta = 1.5$	normal scores + shared	0.806 (0.005)	1.000 (0.000)	0.000 (0.000)	0.856 (0.001)	0.999 (0.000)
30%, $\Delta = 1.5$	median/MAD	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.230 (0.002)	1.001 (0.002)
30%, $\Delta = 1.5$	true background	0.824 (0.005)	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)
30%	shared component	0.989 (0.001)	1.000 (0.000)	0.000 (0.000)	0.999 (0.002)	1.001 (0.002)
30%	normal scores + shared	0.983 (0.002)	1.000 (0.000)	0.000 (0.000)	0.748 (0.001)	0.999 (0.000)
30%	median/MAD	0.873 (0.049)	0.950 (0.027)	0.022 (0.012)	1.547 (0.004)	0.999 (0.003)
30%	true background	0.989 (0.001)	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)
60%	shared component	0.993 (0.001)	1.000 (0.000)	0.000 (0.000)	1.001 (0.003)	0.999 (0.002)
60%	normal scores + shared	0.990 (0.001)	0.988 (0.007)	0.003 (0.002)	0.407 (0.003)	0.999 (0.000)
60%	median/MAD	0.987 (0.001)	1.000 (0.000)	0.125 (0.000)	2.570 (0.008)	1.000 (0.003)
60%	true background	0.993 (0.001)	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)
90%	shared component	0.944 (0.006)	0.700 (0.044)	0.000 (0.000)	1.178 (0.028)	0.999 (0.001)
90%	normal scores + shared	0.893 (0.033)	0.458 (0.025)	0.000 (0.000)	0.644 (0.012)	0.999 (0.000)
90%	median/MAD	0.848 (0.029)	0.729 (0.020)	0.000 (0.000)	2.386 (0.009)	0.998 (0.002)
90%	true background	0.967 (0.002)	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Student t_3	shared component	0.848 (0.035)	0.950 (0.025)	0.503 (0.022)	1.918 (0.064)	0.602 (0.005)
Student t_3	normal scores + shared	0.984 (0.002)	1.000 (0.000)	0.000 (0.000)	0.757 (0.001)	0.999 (0.000)
Student t_3	median/MAD	0.851 (0.034)	0.958 (0.023)	0.131 (0.012)	1.175 (0.004)	0.654 (0.002)
Student t_3	true background	0.966 (0.003)	1.000 (0.000)	0.018 (0.005)	1.000 (0.000)	1.000 (0.000)
exponential	shared component	0.982 (0.002)	0.992 (0.008)	0.842 (0.012)	0.903 (0.015)	0.331 (0.003)
exponential	normal scores + shared	0.982 (0.001)	1.000 (0.000)	0.008 (0.004)	0.735 (0.004)	0.999 (0.000)
exponential	median/MAD	0.819 (0.023)	0.983 (0.017)	0.193 (0.010)	1.266 (0.003)	0.715 (0.003)
exponential	true background	0.968 (0.002)	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Under Gaussian backgrounds with up to 60% active cells, the shared-component estimator matches the oracle (TPR = 1, FPR = 0, $\hat{\sigma}_0 \in [0.999, 1.004]$). Conversely, the pooled median/MAD is contaminated by active cells, widening the estimated background on relevant features ($\hat{\sigma}_0$ from 1.23 to 2.57); this induces total support collapse at $\Delta = 1.5$ (ARI = 0) and spurious activations at 60% (FPR = 0.125). At 90% active cells, the background is weakly identified (Remark 3.4), causing the shared estimator to widen ($\hat{\sigma}_0 = 1.18$) and miss 30% of active blocks; the alternation takes nine outer iterations on average and often reaches the limit of ten.

With non-Gaussian backgrounds, Gaussian components absorb background tails on irrelevant features, causing severe false activations (FPR = 0.503 on t_3, 0.842 on exponential). The normal-scores transformation resolves this failure, eliminating false positives on t_3 and reducing them to 0.8% on exponential, achieving top ARI (0.984 and 0.982). On Gaussian data, the transformation costs little with 30% of active cells (ARI = 0.983 vs 0.989) but much when most cells of the relevant features are active (TPR = 0.458 vs 0.700 with 90%); it is therefore recommended when the features exhibit heavy tails or asymmetry.

5.3 Activation Profiles and Covariances under Partial Activation

We evaluate cell-level classification under $\gamma \in \{0.6, 0.8, 0.95\}$ and $\rho \in \{0, 0.8\}$ across three rules: the full model rule (27) conditioning on currently active coordinates $\mathcal{A}$, a univariate benchmark setting $\mathcal{A} = \emptyset$ with the same estimated parameters, and an oracle Bayes rule evaluating (27) with true parameters and true active sets. Table 4 reports the results.

Table 4: Partial activation: partition, classification of the cells on the true active blocks, calibration of the activation probabilities and estimation of the covariances ($n = 1\,000$, $R = 20$). In every configuration, the active structure is recovered exactly (TPR = 1, FPR = 0 in all replications).

			inactive cells classified active			active cells classified active			covariance error	
γ	ρ	ARI	model	univariate	Bayes	model	univariate	ECE	Frobenius	max corr.
0.60	0.0	0.812 (0.005)	0.113 (0.003)	0.110 (0.002)	0.140 (0.002)	0.897 (0.002)	0.896 (0.002)	0.043 (0.002)	0.379 (0.005)	0.304 (0.012)
0.60	0.8	0.838 (0.005)	0.142 (0.003)	0.152 (0.003)	0.126 (0.002)	0.961 (0.001)	0.928 (0.001)	0.022 (0.001)	0.144 (0.008)	0.091 (0.004)
0.80	0.0	0.959 (0.002)	0.264 (0.008)	0.264 (0.008)	0.243 (0.003)	0.964 (0.001)	0.964 (0.001)	0.019 (0.001)	0.232 (0.008)	0.166 (0.005)
0.80	0.8	0.970 (0.002)	0.194 (0.004)	0.276 (0.005)	0.165 (0.004)	0.989 (0.001)	0.972 (0.001)	0.013 (0.001)	0.086 (0.007)	0.059 (0.003)
0.95	0.0	0.993 (0.001)	0.670 (0.019)	0.683 (0.018)	0.492 (0.008)	0.996 (0.001)	0.997 (0.001)	0.025 (0.001)	0.197 (0.008)	0.106 (0.004)
0.95	0.8	0.997 (0.001)	0.249 (0.005)	0.522 (0.011)	0.228 (0.004)	0.998 (0.000)	0.995 (0.000)	0.005 (0.000)	0.083 (0.006)	0.051 (0.002)

The block-level active structure is recovered exactly across all replications (TPR = 1, FPR = 0). For uncorrelated features ($\rho = 0$), off-diagonal covariances vanish, making conditional and univariate updates virtually identical. In contrast, under strong correlation ($\rho = 0.8$), conditioning on active coordinates substantially reduces false active cell classifications: from 0.276 to 0.194 at $\gamma = 0.8$, and from 0.522 to 0.249 at $\gamma = 0.95$, while increasing true active detection from 0.928 to 0.961 at $\gamma = 0.6$, closely tracking the Bayes oracle (0.228 false active at $\gamma = 0.95$). The largest gap with the oracle occurs at $\gamma = 0.95$ and $\rho = 0$ (0.670 vs 0.492), where the large prior odds $\hat{\gamma}_k/(1 - \hat{\gamma}_k) \approx 19$ push marginal cells toward activation. Across

all settings, posterior probabilities $\sigma(\Delta_{i,k,j})$ are well calibrated (ECE ≤ 0.043). Covariance errors are lowest when activation is dense and correlations are strong, reaching their maximum correlation error (0.304) when 40% of cells are background noise ($\gamma = 0.6$, $\rho = 0$).

5.4 Detection of Non-Informative Observations

Non-informative observations arise naturally under our generative model when all coordinates on the cluster support are inactive. In the base design ($\Delta = 2.5$, $\gamma = 0.5$), an observation is fully inactive with probability $(1-\gamma)^4 = 0.5^4 = 6.25\%$. We compare the two-step diagnostic (19) and (20) under estimated parameters against the same rule with true parameters, and against the heuristic inactivation rate (30) (Table 5).

Table 5: Detection of non-informative observations generated by the model ($\Delta = 2.5$, $\gamma = 0.5$, $n = 1\,000$, $R = 20$; 6.2% null profiles on average): proportion of non-informative observations detected, proportion of informative observations wrongly flagged (false alarms), and ARI of the retained informative observations.

rule	detection	false alarms	ARI (kept)
two-step rule, estimated parameters	0.654 (0.020)	0.065 (0.003)	0.859 (0.006)
two-step rule, true parameters	0.560 (0.017)	0.044 (0.001)	–
inactivation rate	0.996 (0.002)	0.610 (0.006)	0.989 (0.002)

Under estimated parameters, the two-step rule detects 65.4% of non-informative observations with only 6.5% false alarms, close to the same rule with true parameters (56.0% detection and 4.4% false alarms) and preserving clustering accuracy on retained observations (ARI = 0.859). Undetected null profiles correspond to background draws falling within the high-density support of the cluster, which are intrinsically indistinguishable from genuine points. In contrast, thresholding the average inactivation rate (30) detects 99.6% of null profiles but discards 61.0% of informative observations. Because $\gamma = 0.5$ leaves two cells inactive on average per informative point, marginal binomial variations frequently exceed the threshold. The two-step rule avoids this pitfall by comparing the null profile with the informative configurations instead of counting inactive cells.

5.5 Comparison with Competing Methods

We benchmark FC-GMM-LFR on the base design ($\Delta \in \{1.5, 2.5\}$, $\rho \in \{0, 0.5, 0.8\}$, $K = 3$) against four paradigms: the local diagonal model of Toussile et al. (2026) (Gaussian margins, selecting features with posterior activation rate exceeding 1/2 in at least one cluster), VarSelLCM (Marbac and Sedki, 2017, 2019) (global selection under conditional independence), clustvarsel (Raftery and Dean, 2006; Scrucca and Raftery, 2018) (global selection under full covariance), and standard full-covariance Gaussian mixtures fitted by mclust (Scrucca et al., 2016) on all

features and on the true relevant features (oracle). Table 6 reports clustering accuracy (ARI), and Table 7 details feature selection and covariance errors.

Table 6: Comparison with competing methods: ARI ($n = 1\,000$, $K = 3$ given to every method, $R = 20$).

Δ	ρ	FC-GMM-LFR	Local diag.	VarSelLCM	clustvarsel	Full GMM	Full GMM (oracle)
2.5	0.0	0.987 (0.002)	0.985 (0.002)	0.985 (0.002)	0.985 (0.002)	0.981 (0.002)	0.984 (0.002)
2.5	0.5	0.989 (0.002)	0.987 (0.002)	0.987 (0.002)	0.987 (0.002)	0.975 (0.002)	0.978 (0.002)
2.5	0.8	0.993 (0.002)	0.989 (0.002)	0.989 (0.002)	0.961 (0.012)	0.951 (0.011)	0.884 (0.023)
1.5	0.0	0.822 (0.004)	0.821 (0.004)	0.823 (0.004)	0.820 (0.004)	0.720 (0.014)	0.784 (0.006)
1.5	0.5	0.814 (0.004)	0.824 (0.003)	0.824 (0.004)	0.620 (0.063)	0.727 (0.011)	0.783 (0.006)
1.5	0.8	0.901 (0.004)	0.866 (0.004)	0.866 (0.004)	0.657 (0.033)	0.721 (0.013)	0.710 (0.016)

Table 7: Comparison with competing methods: proportion of irrelevant features selected by FC-GMM-LFR and local diagonal models, proportion of relevant features selected by clustvarsel, and relative Frobenius error of covariance matrices on true active blocks ($R = 20$). FC-GMM-LFR selects every relevant feature in all configurations.

Δ	ρ	irrelevant sel. (FC-GMM-LFR)	irrelevant sel. (local diag.)	relevant sel. (clustvarsel)	Frobenius (FC-GMM-LFR)	Frobenius (full GMM)
2.5	0.0	0.000 (0.000)	0.056 (0.021)	1.000 (0.000)	0.252 (0.016)	0.588 (0.008)
2.5	0.5	0.000 (0.000)	0.056 (0.021)	1.000 (0.000)	0.137 (0.005)	0.494 (0.006)
2.5	0.8	0.000 (0.000)	0.056 (0.021)	1.000 (0.000)	0.081 (0.004)	0.409 (0.011)
1.5	0.0	0.000 (0.000)	0.006 (0.006)	1.000 (0.000)	0.245 (0.007)	0.281 (0.013)
1.5	0.5	0.000 (0.000)	0.006 (0.006)	0.787 (0.067)	0.204 (0.004)	0.241 (0.008)
1.5	0.8	0.000 (0.000)	0.006 (0.006)	0.900 (0.045)	0.095 (0.004)	0.235 (0.009)

When clusters are well separated ($\Delta = 2.5$), all methods reach ARI ≥ 0.95 except the full mixture on the true features under strong correlation (ARI $= 0.884$ at $\rho = 0.8$). FC-GMM-LFR selects zero irrelevant features across all runs, whereas the local diagonal model includes 5.6% false positives.

At moderate separation ($\Delta = 1.5$), diagonal methods remain competitive when correlations are absent or moderate (ARI ≈ 0.82; at $\rho = 0.5$ they slightly outperform FC-GMM-LFR, 0.824 vs 0.814), but FC-GMM-LFR gains a decisive advantage under strong correlation (ARI $= 0.901$ vs 0.866 at $\rho = 0.8$). clustvarsel exhibits severe instability under correlation, dropping to ARI $\in [0.620, 0.657]$ and missing up to 21.3% of relevant features. Unconstrained full mixtures lag further behind (ARI ≤ 0.727). Even the oracle GMM is outperformed (ARI $= 0.710$ at $\rho = 0.8$): because standard mixtures lack a cell-level background mechanism, inactive cells within relevant features inflate cluster variances, whereas FC-GMM-LFR explicitly pools them into the background. Consequently, relative Frobenius errors for FC-GMM-LFR are two to five times lower than those of full GMMs at $\Delta = 2.5$ (0.081 vs 0.409 at $\rho = 0.8$) and 1.2 to 2.5 times lower at $\Delta = 1.5$.

Clusters Differing by their Correlations.

In a correlation-driven design ($\rho = 0.9$, $\gamma = 0.9$, Table 8), clusters differ primarily through internal dependence rather than mean separation. At $\Delta = 1.0$, diagonal methods break down entirely (ARI $= 0.045$ for local diagonal, 0.133 for VarSelLCM). clustvarsel selects only a third of relevant features (33.7%, ARI $= 0.240$), and full mixtures plateau at ARI ≈ 0.580 (0.601 for the oracle). In contrast, FC-GMM-LFR achieves ARI $= 0.772$, retaining 96.7% of relevant features with zero false positives and a 2.5-fold reduction in covariance error (0.086 vs 0.217). At $\Delta = 1.5$, FC-GMM-LFR reaches ARI $= 0.936$ (vs 0.890 for VarSelLCM, 0.879 for local diagonal and 0.646 for full GMM). When mean shifts vanish completely, mean-based initializations like k-means++ provide no discriminatory signal; extending the initialization to covariance-sensitive spectral anchors remains a promising direction.

Table 8: Clusters differing mainly by their correlations ($\rho = 0.9$): ARI ($n = 1\,000$, $K = 3$ given to every method, $R = 20$).

Δ	ρ	FC-GMM-LFR	Local diag.	VarSelLCM	clustvarsel	Full GMM	Full GMM (oracle)
1.5	0.9	0.936 (0.003)	0.879 (0.004)	0.890 (0.004)	0.509 (0.034)	0.646 (0.016)	0.624 (0.012)
1.0	0.9	0.772 (0.019)	0.045 (0.011)	0.133 (0.011)	0.240 (0.008)	0.580 (0.028)	0.601 (0.019)

6 Application to Sorted Blood Cells

Our framework posits that features follow a shared background distribution in clusters where they are uninformative. This premise is particularly suited to molecular profiling where marker genes depart from basal levels only in specific cell types; conversely, when all groups differ across all features, relevance reduces to the symmetry convention of Remark 3.4. We evaluate FC-GMM-LFR on sorted peripheral blood mononuclear cells, where this local structure arises by biological design. Competing methods are evaluated under the true number of clusters, including full Gaussian mixtures with the conjugate prior of Fraley and Raftery (2007).

Cell-type discovery from single-cell RNA sequencing poses severe statistical challenges due to high dimensionality, sparsity, and transcriptional drop-out (Kiselev et al., 2019). We analyze $n = 2\,000$ cells from the bead-sorted dataset of Zheng et al. (2017), drawing 500 cells from each of four known populations: B cells, CD14 monocytes, Natural Killer (NK) cells, and CD8 cytotoxic T cells. Cell identities derive from antibody sorting rather than transcriptomic clustering, providing an objective ground truth. We consider nineteen curated marker genes: four for B cells (MS4A1, CD79A, CD79B, CD19), five for monocytes (CD14, LYZ, S100A8, S100A9, FCN1), two for NK cells (NCAM1, KLRF1), four for T cells (CD8A, CD8B, CD3E, CD3D), and four cytotoxic effector genes shared by NK and CD8 T cells (NKG7, GZMB, PRF1, GNLY). Expression values are normalized to 10^4 library size and $\log(1+x)$ transformed. Because undetected genes produce an exact point mass at zero, cluster and background variances share a protective lower bound of 0.25 on the log scale (corresponding to a measurement resolution of 0.5) to prevent spurious variance collapse onto the zero mass. The model is fitted with $K = 4$ across five random restarts.

Table 9: Sorted blood cells: ARI with the sorted populations ($n = 2\,000$, $J = 19$, $K = 4$). Unregularized full GMM and VarSelLCM fail due to covariance degeneracies caused by zeros.

method	ARI
FC-GMM-LFR	0.958
Local diag.	0.679
Full GMM, prior	0.648
clustvarsel	0.357
Full GMM	failed
VarSelLCM	failed

FC-GMM-LFR recovers the four cell populations with high fidelity (ARI = 0.958, Table 9), substantially outperforming the local diagonal model (ARI = 0.679), the Bayesian full GMM (ARI = 0.648), and clustvarsel (ARI = 0.357). Unregularized full mixtures and VarSelLCM fail entirely, as excessive zero counts produce degenerate sample covariance matrices.

Beyond clustering accuracy, the estimated activation parameters $\hat{\gamma}_{k,j}$ have a direct physical interpretation: they estimate the probability that a marker is actively expressed in a single cell of that population (Table 10). Each marker is correctly selected for its target lineage, with estimated activation rates closely matching empirical detection frequencies: 0.69 vs 0.70 for MS4A1 in B cells, 0.50 vs 0.50 for CD14 in monocytes, 0.34 vs 0.34 for CD8A in CD8 T cells, and 0.84 vs 0.82 for GZMB in NK cells.

Two genes exhibit background inversion (Table 10): for CD79A and GNLY, the shared background mean is placed on the expressed state, causing both genes to appear active with rates near 1 in unexpressing populations. Because a unimodal Gaussian background cannot perfectly represent a sharp zero point mass alongside sporadic background leakage, the estimator occasionally anchors $\hat{\mu}_{0,j}$ to the nonzero cluster. Varying the variance floor (Table 11) shifts these boundary choices without altering the overall partition quality (ARI = 0.946 at floor 0.01, and 0.958 at 0.10 and 0.25).

Finally, the estimated intra-cluster correlation structures (Figure 2) illuminate why full covariances are critical. In monocytes, S100A8 and S100A9 are strongly correlated (0.74), consistent with their joint expression as the calprotectin complex. The cytotoxic effectors NKG7, GZMB and GNLY are moderately correlated in CD8 T cells (0.30 to 0.50) and nearly uncorrelated in NK cells (at most 0.11); since the background of GNLY lies at its expressed level, its correlations in CD8 T cells describe the variation of its low expression rather than a joint activation. These group-specific dependencies are retained by the model, whereas a diagonal model would discard them.

Table 10: Sorted blood cells: estimated activation rate $\hat{\gamma}_{k,j}$ of each active block alongside observed detection rate (fraction of cells with non-zero expression) in parentheses; a dash indicates an inactive block.

gene	B	CD14 monocytes	NK	CD8 T
MS4A1	0.69 (0.70)	–	–	–
CD79A	0.94 (0.95)	0.99 (0.01)	0.99 (0.01)	0.95 (0.06)
CD79B	0.89 (0.90)	–	–	–
CD19	0.20 (0.20)	–	–	–
CD14	–	0.50 (0.50)	–	–
LYZ	–	0.88 (0.88)	–	–
S100A8	–	0.98 (0.97)	–	–
S100A9	–	0.97 (0.97)	–	–
FCN1	–	0.73 (0.74)	–	–
NCAM1	–	–	0.03 (0.04)	–
KLRF1	–	–	0.59 (0.58)	–
CD8A	–	–	–	0.34 (0.34)
CD8B	–	–	–	0.61 (0.61)
CD3E	–	–	0.08 (0.07)	0.75 (0.77)
CD3D	–	–	0.04 (0.03)	0.78 (0.80)
NKG7	–	0.03 (0.03)	1.00 (0.98)	0.38 (0.39)
GZMB	–	–	0.84 (0.82)	0.08 (0.10)
PRF1	–	–	0.74 (0.72)	0.07 (0.09)
GNLY	1.00 (0.00)	0.99 (0.01)	0.99 (0.99)	1.00 (0.22)

Table 11: Sorted blood cells: sensitivity to the variance floor across background and cluster covariances (five restarts per setting). The last column reports genes whose estimated background mean exceeds 0.5.

variance floor	ARI	active blocks	background at the expressed level
0.01	0.946	46	CD79A CD79B NKG7 GNLY
0.10	0.958	35	CD79A CD79B GNLY
0.25	0.958	31	CD79A GNLY

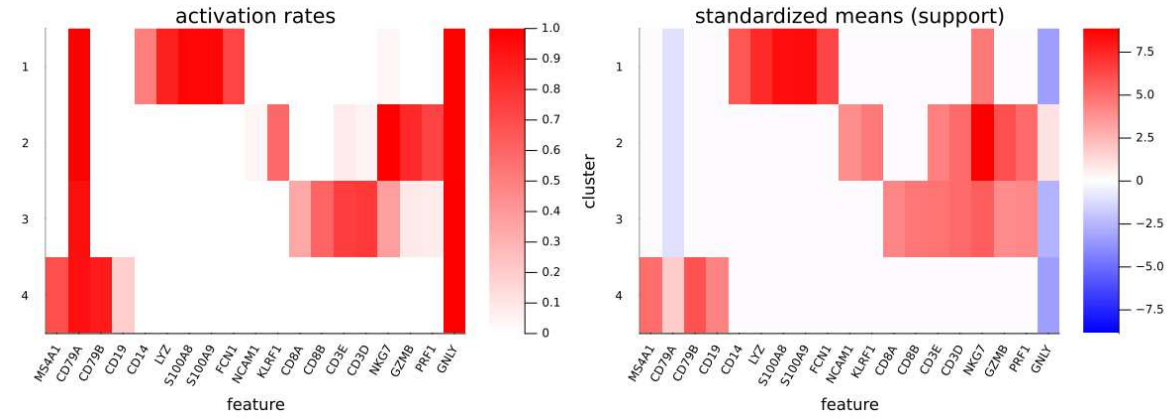


Figure 1: Sorted blood cells: estimated activation rates (left) and standardized mean shifts $(\hat{\mu}_{k,j} - \hat{\mu}_{0,j})/\hat{\sigma}_{0,j}$ of active blocks (right).

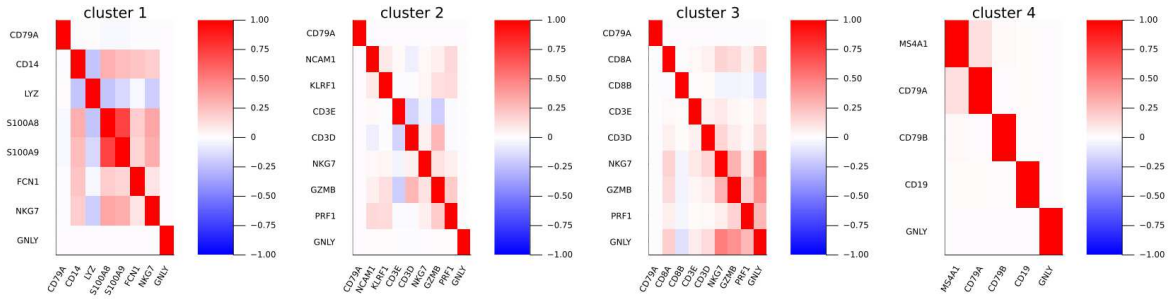


Figure 2: Sorted blood cells: estimated correlations between active genes in each cluster (blue: negative, white: zero, red: positive). Clusters 1 to 4 correspond to monocytes, NK cells, CD8 T cells, and B cells.

7 Discussion and Concluding Remarks

We have introduced the Full-Covariance Gaussian Mixture Model with Local Feature Relevance (FC-GMM-LFR). By allowing feature relevance to be determined locally per observation through binary activation profiles while retaining unconstrained cluster covariances, our framework resolves the longstanding conflict between local feature selection and intra-cluster correlation modeling. Cluster-specific supports are selected through a penalized classification criterion maximized via a monotone block coordinate ascent algorithm, where individual cell profiles are classified, covariance matrices are updated via missing-data EM with spectral regularization, and background parameters are pooled across uninformative cells.

Empirical Findings.

Across diverse simulations, FC-GMM-LFR consistently isolated the true active structures without selecting irrelevant features, achieving covariance estimation errors two to five times lower than standard full GMMs at $\Delta = 2.5$. Estimating the background as a shared component proved essential: naive empirical estimators (such as pooled medians and MADs) suffer severe bias whenever active cells are prevalent, inducing support collapse or spurious activations. Furthermore, conditioning on active coordinates substantially reduced cell misclassification rates under strong correlation, and posterior probabilities remained well calibrated (ECE ≤ 0.043). In real-world single-cell transcriptomics, the method accurately separated sorted blood cell populations (ARI $= 0.958$), and estimated activation rates faithfully mirrored empirical marker gene detection frequencies under drop-out.

Limitations.

We highlight several structural limitations. First, uninformative observations can only be distinguished if cluster separation is sufficient to prevent background points from landing in high-density cluster cores. In addition, the proportion of null profiles, $\widetilde{\omega}_0 = \sum_{k=1}^{K} \omega_k \prod_{j=1}^{J} (1 - \gamma_{k,j})$ (Proposition~2.5), is determined by the activation rates rather than parameterized by an independent contamination weight (see Appendix~D for a free weight). Second, hard cell classification exhibits an upward bias at moderate separation ($\Delta = 1.5$) that does not decrease with sample size in the tested range. Third, background identifiability requires that uninformative states appear in at least two clusters or partially within an active group; for $K = 2$, distinguishing which group holds the active state rests on the median convention (Remark 3.4). Fourth, heavy tails, skewness, or zero inflation can distort a Gaussian background, necessitating rank-scores preprocessing or variance floors. More generally, the model is designed for data in which variables are relevant for some classes only: when the classes differ on every variable, as in many morphometric data sets, it has no advantage over full mixtures with global selection, and a class lying at the common level of every feature is represented by the background, that is, as non-informative. Fifth, initializations relying purely on cluster

centroids (such as k-means++) fail when groups differ exclusively through correlations. Finally, coordinate-wise cell updates scale superlinearly with dimension J.

Perspectives.

Several promising avenues emerge from this work. Theoretically, establishing the asymptotic selection consistency $\mathbb{P}(\widehat{\mathbf{B}} = \mathbf{B}^\star) \to 1$ of the penalized classification criterion under latent profile uncertainty remains an open challenge. Methodologically, natural extensions include designing covariance-aware initializations, modeling full background covariances, incorporating explicit zero-inflated emission models for count data, and selecting the number of clusters K jointly with feature supports. Extending this framework to heterogeneous mixed-type tables via vine copulas or variational formulations constitutes another compelling direction.

Reproducibility and Code Availability

The model and the estimation algorithm are implemented in the open-source Julia package `FCGMMLFR.jl`, available at https://github.com/toussile/FCGMMLFR.jl. The directory `experiments` of the package contains one script per simulation experiment, the script for the application, the scripts that generate the tables, and the preparation of the blood cell data; the random seeds depend only on the experiment and the replication.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used LLM to translate the manuscript text into English, improve its readability, and assist with the modular structure and organization of the accompanying Julia code. Following the use of this tool, the author thoroughly reviewed, verified, and edited all material, and takes full intellectual and scientific responsibility for the contents of the publication.

References

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000. doi: 10.1109/34.865189.

Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014. doi: 10.1016/j.csda.2012.12.008.

Charles Bouveyron, Gilles Celeux, Thomas Brendan Murphy, and Adrian E. Raftery. *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge University Press, Cambridge, 2019. doi: 10.1017/9781108644181.

Gilles Celeux and Gérard Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992. doi: 10.1016/0167-9473(92)90042-E.

Gilles Celeux, Cathy Maugis-Rabusseau, and Mohammed Sedki. Variable selection in model-based clustering and discriminant analysis with a regularization approach. *Advances in Data Analysis and Classification*, 13(1):259–278, 2019. doi: 10.1007/s11634-018-0322-5.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–38, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Michael Fop and Thomas Brendan Murphy. Variable selection methods for model-based clustering. *Statistics Surveys*, 12:18–65, 2018. doi: 10.1214/18-SS119.

Michael Fop, Thomas Brendan Murphy, and Luca Scrucca. Model-based clustering with sparse covariance matrices. *Statistics and Computing*, 29(4):791–819, 2019. doi: 10.1007/s11222-018-9838-y.

Chris Fraley and Adrian E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, 2007. doi: 10.1007/s00357-007-0004-5.

Isobel Claire Gormley, Thomas Brendan Murphy, and Adrian E. Raftery. Model-based clustering. *Annual Review of Statistics and Its Application*, 10:573–595, 2023. doi: 10.1146/annurev-statistics-033121-115326.

Peter D. Hoff. Model-based subspace clustering. *Bayesian Analysis*, 1(2):321–344, 2006. doi: 10.1214/06-BA111.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2nd edition, 2012. doi: 10.1017/CBO9781139020411.

Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019. doi: 10.1038/s41576-018-0088-9.

Martin H. C. Law, Mário A. T. Figueiredo, and Anil K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, 2004. doi: 10.1109/TPAMI.2004.71.

Yuanhong Li, Ming Dong, and Jing Hua. Simultaneous localized feature selection and model detection for Gaussian mixtures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):953–960, 2009. doi: 10.1109/TPAMI.2008.261.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ, 3rd edition, 2019. doi: 10.1002/9781119482260.

Matthieu Marbac and Mohammed Sedki. Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, 27(4):1049–1063, 2017. doi: 10.1007/s11222-016-9670-1.

Matthieu Marbac and Mohammed Sedki. VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values. *Bioinformatics*, 35(7):1255–1257, 2019. doi: 10.1093/bioinformatics/bty786.

Fernando A. Quintana, Peter Müller, and Ana Luisa Papoila. Cluster-specific variable selection for product partition models. *Scandinavian Journal of Statistics*, 42(4):1065–1077, 2015. doi: 10.1111/sjos.12151.

Adrian E. Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006. doi: 10.1198/016214506000000113.

Jakob Raymaekers and Peter J. Rousseeuw. The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association*, 119(548):2610–2621, 2024. doi: 10.1080/01621459.2023.2267777.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136.

Luca Scrucca and Adrian E. Raftery. clustvarsel: A package implementing variable selection for Gaussian model-based clustering in R. *Journal of Statistical Software*, 84(1):1–28, 2018. doi: 10.18637/jss.v084.i01.

Luca Scrucca, Michael Fop, Thomas Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317, 2016. doi: 10.32614/RJ-2016-021.

Mahlet G. Tadesse, Naijun Sha, and Marina Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617, 2005. doi: 10.1198/016214504000001565.

Wilson Toussile and Elisabeth Gassiat. Variable selection in model-based clustering using multilocus genotype data. *Advances in Data Analysis and Classification*, 3(2):109–134, 2009. doi: 10.1007/s11634-009-0043-x.

Wilson Toussile, Simeon Fotso, and Patrice Takam Soh. Local feature relevance in model-based clustering of mixed data via variational inference. Submitted, 2026.

Daniela M. Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010. doi: 10.1198/jasa.2010.tm09415.

Benhuai Xie, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics*, 2:168–212, 2008. doi: 10.1214/08-EJS194.

Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968. doi: 10.1214/aoms/1177698520.

Giorgia Zaccaria, Luis A. García-Escudero, Francesca Greselin, and Agustín Mayo-Íscar. Cellwise outlier detection in heterogeneous populations. *Technometrics*, 67(4):643–654, 2025. doi: 10.1080/00401706.2025.2497822.

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8: 14049, 2017. doi: 10.1038/ncomms14049.

Hui Zhou, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3:1473–1496, 2009. doi: 10.1214/09-EJS487.

A Auxiliary Lemma and Proof of Lemma 3.1

Lemma A.1 (Gaussian Density with Independent Active and Background Blocks)**.** *Let* $1 \le d(\boldsymbol{s}) \le J-1$, $\boldsymbol{\mu}_{\text{act}} \in \mathbb{R}^{d(\boldsymbol{s})}$, $\boldsymbol{\Sigma}_{\text{act}} \in \mathcal{S}_{++}^{d(\boldsymbol{s})}$, $\boldsymbol{\mu}_{\text{bg}} \in \mathbb{R}^{J-d(\boldsymbol{s})}$ *and* $\boldsymbol{\Sigma}_{\text{bg}} \in \mathcal{S}_{++}^{J-d(\boldsymbol{s})}$. *If* $\mathbf{D}(\boldsymbol{s})\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{act}}, \boldsymbol{\Sigma}_{\text{act}})$ *and* $\mathbf{D}(\boldsymbol{s}^c)\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{bg}}, \boldsymbol{\Sigma}_{\text{bg}})$ *are independent, then* $\boldsymbol{y} \sim \mathcal{N}_J(\boldsymbol{\mu}(\boldsymbol{s}), \boldsymbol{\Sigma}(\boldsymbol{s}))$ *on* $\mathbb{R}^J$, *with*

$$\boldsymbol{\mu}(\boldsymbol{s}) = \mathbf{D}(\boldsymbol{s})^\top \boldsymbol{\mu}_{\text{act}} + \mathbf{D}(\boldsymbol{s}^c)^\top \boldsymbol{\mu}_{\text{bg}}, \qquad \boldsymbol{\Sigma}(\boldsymbol{s}) = \mathbf{D}(\boldsymbol{s})^\top \boldsymbol{\Sigma}_{\text{act}} \mathbf{D}(\boldsymbol{s}) + \mathbf{D}(\boldsymbol{s}^c)^\top \boldsymbol{\Sigma}_{\text{bg}} \mathbf{D}(\boldsymbol{s}^c) \in \mathcal{S}_{++}^J, \tag{31}$$

$$\boldsymbol{\Sigma}(\boldsymbol{s})^{-1} = \mathbf{D}(\boldsymbol{s})^\top \boldsymbol{\Sigma}_{\text{act}}^{-1} \mathbf{D}(\boldsymbol{s}) + \mathbf{D}(\boldsymbol{s}^c)^\top \boldsymbol{\Sigma}_{\text{bg}}^{-1} \mathbf{D}(\boldsymbol{s}^c). \tag{32}$$

The density of $\boldsymbol{y}$ *is the product of the two block densities, its Mahalanobis form is the sum of the two block forms, and* $\boldsymbol{\Pi}(\boldsymbol{s})\boldsymbol{\Sigma}(\boldsymbol{s})\boldsymbol{\Pi}(\boldsymbol{s})^\top = \operatorname{diag}(\boldsymbol{\Sigma}_{\text{act}}, \boldsymbol{\Sigma}_{\text{bg}})$, *so that the spectrum of* $\boldsymbol{\Sigma}(\boldsymbol{s})$ *is the union of the spectra of* $\boldsymbol{\Sigma}_{\text{act}}$ *and* $\boldsymbol{\Sigma}_{\text{bg}}$.

In the model, the independence of the two blocks is Assumption 2.2; the lemma gives the explicit form of the resulting density on $\mathbb{R}^J$ and is used in the proof of Proposition 2.4.

A.1 Proof of Lemma A.1

Proof. Write $\boldsymbol{u} = \boldsymbol{\Pi}(\boldsymbol{s})\boldsymbol{y} = (\mathbf{D}(\boldsymbol{s})\boldsymbol{y}, \mathbf{D}(\boldsymbol{s}^c)\boldsymbol{y})$. By independence, $\boldsymbol{u}$ has the product density $\mathcal{N}(\boldsymbol{u}_{\text{act}}; \boldsymbol{\mu}_{\text{act}}, \boldsymbol{\Sigma}_{\text{act}})\,\mathcal{N}(\boldsymbol{u}_{\text{bg}}; \boldsymbol{\mu}_{\text{bg}}, \boldsymbol{\Sigma}_{\text{bg}})$, that is, $\boldsymbol{u} \sim \mathcal{N}_J((\boldsymbol{\mu}_{\text{act}}, \boldsymbol{\mu}_{\text{bg}}), \operatorname{diag}(\boldsymbol{\Sigma}_{\text{act}}, \boldsymbol{\Sigma}_{\text{bg}}))$. Since $\boldsymbol{\Pi}(\boldsymbol{s})$ is a permutation matrix ($|\det \boldsymbol{\Pi}(\boldsymbol{s})| = 1$), $\boldsymbol{y} = \boldsymbol{\Pi}(\boldsymbol{s})^\top \boldsymbol{u}$ has density equal to that of $\boldsymbol{u}$ evaluated at $\boldsymbol{\Pi}(\boldsymbol{s})\boldsymbol{y}$, and $\boldsymbol{y} \sim \mathcal{N}_J(\boldsymbol{\Pi}(\boldsymbol{s})^\top(\boldsymbol{\mu}_{\text{act}}, \boldsymbol{\mu}_{\text{bg}}), \boldsymbol{\Pi}(\boldsymbol{s})^\top \operatorname{diag}(\boldsymbol{\Sigma}_{\text{act}}, \boldsymbol{\Sigma}_{\text{bg}})\boldsymbol{\Pi}(\boldsymbol{s}))$, which is (31) because $\boldsymbol{\Pi}(\boldsymbol{s})^\top = [\mathbf{D}(\boldsymbol{s})^\top\ \mathbf{D}(\boldsymbol{s}^c)^\top]$. Inverting the block-diagonal matrix and conjugating by $\boldsymbol{\Pi}(\boldsymbol{s})$ gives (32); substituting (32) in the quadratic form of $\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{s})$ gives the additive Mahalanobis form. Conjugation by an orthogonal matrix preserves the spectrum. □

A.2 Proof of Lemma 3.1

Proof. **Proof of item (i): Eigenvalue Interlacing and Condition Number.** Let $\mathbf{S} = \mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}\mathbf{D}(\boldsymbol{s})^\top \in \mathcal{S}_{++}^{d(\boldsymbol{s})}$. For any unit vector $\boldsymbol{u} \in \mathbb{R}^{d(\boldsymbol{s})}$ with $\|\boldsymbol{u}\|_2 = 1$, the lifted vector $\boldsymbol{v} = \mathbf{D}(\boldsymbol{s})^\top \boldsymbol{u} \in \mathbb{R}^J$ satisfies:

$$\|\boldsymbol{v}\|_2^2 = \boldsymbol{u}^\top \mathbf{D}(\boldsymbol{s})\mathbf{D}(\boldsymbol{s})^\top \boldsymbol{u} = \boldsymbol{u}^\top \mathbf{I}_{d(\boldsymbol{s})} \boldsymbol{u} = 1, \tag{33}$$

and the Rayleigh quotient of $\mathbf{S}$ coincides with that of $\boldsymbol{\Sigma}$:

$$\boldsymbol{u}^\top \mathbf{S} \boldsymbol{u} = \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v} \in [\lambda_J(\boldsymbol{\Sigma}),\ \lambda_1(\boldsymbol{\Sigma})]\,. \tag{34}$$

Taking the minimum and maximum over $\|\boldsymbol{u}\|_2 = 1$ yields the extreme eigenvalue bounds:

$$\lambda_1(\mathbf{S}) \le \lambda_1(\boldsymbol{\Sigma}), \qquad \lambda_{d(\boldsymbol{s})}(\mathbf{S}) \ge \lambda_J(\boldsymbol{\Sigma}). \tag{35}$$

More generally, Cauchy's interlacing theorem (Horn and Johnson, 2012, Section 4.3) guarantees that $\lambda_{J-d(\boldsymbol{s})+m}(\boldsymbol{\Sigma}) \le \lambda_m(\mathbf{S}) \le \lambda_m(\boldsymbol{\Sigma})$ for all $m \in \{1, \dots, d(\boldsymbol{s})\}$. Consequently, $\lambda_{\min}(\mathbf{S}) \ge \lambda_J(\boldsymbol{\Sigma}) > 0$, ensuring $\mathbf{S} \succ 0$, and the condition number satisfies:

$$\kappa(\mathbf{S}) = \frac{\lambda_1(\mathbf{S})}{\lambda_{d(\boldsymbol{s})}(\mathbf{S})} \le \frac{\lambda_1(\boldsymbol{\Sigma})}{\lambda_J(\boldsymbol{\Sigma})} = \kappa(\boldsymbol{\Sigma}). \tag{36}$$

Proof of item (ii): Uniform Density Bound. Because $\mathbf{S} \succ 0$, we have $\det \mathbf{S} = \prod_{m=1}^{d(\boldsymbol{s})} \lambda_m(\mathbf{S}) \ge [\lambda_{d(\boldsymbol{s})}(\mathbf{S})]^{d(\boldsymbol{s})} \ge \delta^{d(\boldsymbol{s})}$. Since the exponential factor of any multivariate Gaussian density is bounded above by 1, the active density satisfies:

$$\mathcal{N}_{d(\boldsymbol{s})}(\mathbf{D}(\boldsymbol{s})\boldsymbol{y}; \boldsymbol{\mu}_{\text{act}}, \mathbf{S}) \le (2\pi)^{-d(\boldsymbol{s})/2} (\det \mathbf{S})^{-1/2} \le (2\pi\delta)^{-d(\boldsymbol{s})/2}, \tag{37}$$

which establishes the uniform bound (23). □

B Subspace Mixture Expansion, Canonical Partition, and Identifiability

B.1 Proof of Proposition 2.4 (Full-Covariance Gaussian Mixture Representation and Ambient Spectral Regularity)

Proof. The proof proceeds in two steps: first establishing ambient positive-definiteness and the uniform condition number bound, then deriving the subspace product density and the resulting marginal finite mixture expansion.

For the boundary profiles $\boldsymbol{s} = \mathbf{1}_J$ and $\boldsymbol{s} = \mathbf{0}_J$, the conditional law of $\boldsymbol{y}_i$ is $\mathcal{N}_J(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $\mathcal{N}_J(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ respectively (Remark 2.3), which trivially agrees with (7)–(8) and satisfies (9). We thus assume $1 \le d(\boldsymbol{s}) \le J-1$ in what follows.

Step 1: Ambient Positive-Definiteness and Uniform Conditioning. By Lemma A.1, conjugating $\boldsymbol{\Sigma}_k(\boldsymbol{s})$ by the coordinate permutation matrix $\boldsymbol{\Pi}(\boldsymbol{s})$ yields a block-diagonal matrix:

$$\boldsymbol{\Pi}(\boldsymbol{s})\boldsymbol{\Sigma}_k(\boldsymbol{s})\boldsymbol{\Pi}(\boldsymbol{s})^\top = \operatorname{diag}\left(\mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}_k\mathbf{D}(\boldsymbol{s})^\top,\ \mathbf{D}(\boldsymbol{s}^c)\boldsymbol{\Sigma}_0\mathbf{D}(\boldsymbol{s}^c)^\top\right). \tag{38}$$

Because conjugation by an orthogonal matrix preserves eigenvalues, the spectrum of $\boldsymbol{\Sigma}_k(\boldsymbol{s})$ is the exact union of the spectra of the active subspace covariance $\mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}_k\mathbf{D}(\boldsymbol{s})^\top$ and the background subspace covariance $\mathbf{D}(\boldsymbol{s}^c)\boldsymbol{\Sigma}_0\mathbf{D}(\boldsymbol{s}^c)^\top$. Applying Lemma 3.1 to each block, we obtain:

$$\lambda_{\max}(\boldsymbol{\Sigma}_k(\boldsymbol{s})) \le \max\left\{\lambda_1(\boldsymbol{\Sigma}_k),\ \lambda_1(\boldsymbol{\Sigma}_0)\right\}, \tag{39}$$

$$\lambda_{\min}(\boldsymbol{\Sigma}_k(\boldsymbol{s})) \ge \min\left\{\lambda_J(\boldsymbol{\Sigma}_k),\ \lambda_J(\boldsymbol{\Sigma}_0)\right\} > 0. \tag{40}$$

Hence $\boldsymbol{\Sigma}_k(\boldsymbol{s}) \in \mathcal{S}^J_{++}$, and taking the ratio of these extreme eigenvalues establishes the condition number bound (9).

Step 2: Subspace Product Density and Marginal Mixture Expansion. Because $\boldsymbol{\Pi}(\boldsymbol{s})$ is an orthogonal permutation matrix ($|\det \boldsymbol{\Pi}(\boldsymbol{s})| = 1$), the Lebesgue measure on $\mathbb{R}^J$ decomposes as the product of the Lebesgue measures on the active and inactive coordinate subspaces. Lemma A.1 yields:

$$\begin{aligned} p(\boldsymbol{y}_i \mid z_i = k, \boldsymbol{s}_i = \boldsymbol{s}) &= \mathcal{N}_{d(\boldsymbol{s})}\left(\mathbf{D}(\boldsymbol{s})\boldsymbol{y}_i \,;\, \mathbf{D}(\boldsymbol{s})\boldsymbol{\mu}_k, \mathbf{D}(\boldsymbol{s})\boldsymbol{\Sigma}_k\mathbf{D}(\boldsymbol{s})^\top\right) \\ &\quad \times \mathcal{N}_{J-d(\boldsymbol{s})}\left(\mathbf{D}\left(\boldsymbol{s}^c\right)\boldsymbol{y}_i \,;\, \mathbf{D}\left(\boldsymbol{s}^c\right)\boldsymbol{\mu}_0, \mathbf{D}\left(\boldsymbol{s}^c\right)\boldsymbol{\Sigma}_0\mathbf{D}\left(\boldsymbol{s}^c\right)^\top\right) \\ &= (2\pi)^{-J/2}|\det \boldsymbol{\Sigma}_k(\boldsymbol{s})|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_k(\boldsymbol{s}))^\top \boldsymbol{\Sigma}_k(\boldsymbol{s})^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_k(\boldsymbol{s}))\right) \\ &= \mathcal{N}_J\left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s})\right). \end{aligned} \tag{41}$$

Marginalizing over $\boldsymbol{s} \in \{0,1\}^J$ with prior weights $p(\boldsymbol{s} \mid z_i = k) = \prod_{j=1}^J \gamma_{k,j}^{s_j}(1-\gamma_{k,j})^{1-s_j}$ produces the 2^J-component Gaussian mixture (10) for each cluster k. Weighting by ω_k and summing over $k \in \{1,\ldots,K\}$ yields the extended mixture (11) over the $K \times 2^J$ latent configurations. Because $\boldsymbol{\mu}_k(\mathbf{0}_J) = \boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_k(\mathbf{0}_J) = \boldsymbol{\Sigma}_0$ identically for all $k \in \{1,\ldots,K\}$, the K null terms merge into the single background Gaussian distribution, leaving at most $1 + K(2^J - 1)$ distinct Gaussian components. □

B.2 Proof of Proposition 2.5 (Canonical $(K+1)$-Component Mixture Partition)

Proof. The proof proceeds in three steps: isolating the null profile $\boldsymbol{s} = \mathbf{0}_J$, normalizing the substantive cluster components, and verifying that the total mixture weights sum to one.

Step 1: Isolation and Merging of Null Profiles. Partition the binary hypercube $\{0,1\}^J$ into the null profile $\{\mathbf{0}_J\}$ and the non-null configurations $\mathcal{S}^\star = \{0,1\}^J \setminus \{\mathbf{0}_J\}$. For $\boldsymbol{s} = \mathbf{0}_J$, the down-projection operators satisfy $\mathbf{P}(\mathbf{0}_J) = \mathbf{0}_{J\times J}$ and $\mathbf{P}(\mathbf{0}_J^c) = \mathbf{I}_J$. By equations (7) and (8):

$$\boldsymbol{\mu}_k(\mathbf{0}_J) = \mathbf{0} + \mathbf{I}_J\boldsymbol{\mu}_0 = \boldsymbol{\mu}_0, \qquad \boldsymbol{\Sigma}_k(\mathbf{0}_J) = \mathbf{0} + \mathbf{I}_J\boldsymbol{\Sigma}_0\mathbf{I}_J = \boldsymbol{\Sigma}_0, \tag{42}$$

which is independent of k. The prior probability of this null profile given $z_i = k$ is:

$$p(\boldsymbol{s} = \mathbf{0}_J \mid z_i = k) = \prod_{j=1}^J (1 - \gamma_{k,j}). \tag{43}$$

Separating the null profile $\boldsymbol{s} = \mathbf{0}_J$ in the marginal mixture density (11) yields:

$$p(\boldsymbol{y}_i) = \sum_{k=1}^{K} \omega_k \left[p(\boldsymbol{s} = \mathbf{0}_J \mid z_i = k) \mathcal{N}_J \left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 \right) + \sum_{\boldsymbol{s} \in \mathcal{S}^\star} p(\boldsymbol{s} \mid z_i = k) \mathcal{N}_J \left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s}) \right) \right]$$
$$= \underbrace{\left(\sum_{k=1}^{K} \omega_k \prod_{j=1}^{J} (1 - \gamma_{k,j}) \right)}_{\equiv \widetilde{\omega}_0} \mathcal{N}_J \left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 \right) + \sum_{k=1}^{K} \omega_k \sum_{\boldsymbol{s} \in \mathcal{S}^\star} p(\boldsymbol{s} \mid z_i = k) \mathcal{N}_J \left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s}) \right). \tag{44}$$

Step 2: Normalization of Substantive Cluster Components. For each cluster k, the probability of sampling a non-null profile is:

$$p(\mathcal{S}^\star \mid z_i = k) = 1 - \prod_{j=1}^{J} (1 - \gamma_{k,j}) > 0, \tag{45}$$

which is strictly positive because $\gamma_{k,j} > 0$ for at least one $j \in \{1, \ldots, J\}$. Defining the substantive mixing proportions and normalized component densities by:

$$\widetilde{\omega}_k = \omega_k \left(1 - \prod_{j=1}^{J} (1 - \gamma_{k,j}) \right), \tag{46}$$

$$f_k^{\text{in}}(\boldsymbol{y}_i) = \sum_{\boldsymbol{s} \in \mathcal{S}^\star} \frac{p(\boldsymbol{s} \mid z_i = k)}{1 - \prod_{j=1}^{J} (1 - \gamma_{k,j})} \mathcal{N}_J \left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s}) \right), \tag{47}$$

we obtain the canonical representation $p(\boldsymbol{y}_i) = \widetilde{\omega}_0 \mathcal{N}_J(\boldsymbol{y}_i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \sum_{k=1}^{K} \widetilde{\omega}_k f_k^{\text{in}}(\boldsymbol{y}_i)$.

Step 3: Verification of the Extended Mixture Proportions. Summing the $(K+1)$ mixture weights gives:

$$\widetilde{\omega}_0 + \sum_{k=1}^{K} \widetilde{\omega}_k = \sum_{k=1}^{K} \omega_k \prod_{j=1}^{J} (1 - \gamma_{k,j}) + \sum_{k=1}^{K} \omega_k \left(1 - \prod_{j=1}^{J} (1 - \gamma_{k,j}) \right) = \sum_{k=1}^{K} \omega_k = 1, \tag{48}$$

which completes the proof. □

B.3 Proof of Proposition 2.7 (Identifiability of the Activation Structure)

B.3.1 Proof of Proposition 2.7(i): Non-Identifiability of Uncorrelated Null Coordinates

Proof. The proof establishes that the marginal data density $p(\boldsymbol{y}_i)$ is strictly invariant to $\gamma' \in [0, 1]$ in three steps: coordinate-wise factorization on profile $\boldsymbol{s}^{(j)}$, equivalence with the inactive background emission, and pairwise cancellation across the binary hypercube.

Step 1: Coordinate-wise Factorization of Active Emission Density. Under parameter $\boldsymbol{\theta}'$, coordinate j is uncorrelated with all other coordinates in cluster k ($\Sigma'_{k,jl} = \Sigma'_{k,lj} = 0$ for all $l \neq j$). Consequently, for any profile $\boldsymbol{s}^{(j)}$ with $s_j^{(j)} = 1$, the active parent covariance matrix on $\mathcal{A}(\boldsymbol{s}^{(j)})$ is block-diagonal, with coordinate j forming an isolated scalar block of mean $\mu'_{k,j} = \mu_{0,j}$ and variance $\Sigma'_{k,jj} = \sigma^2_{0,j}$. The active Gaussian distribution therefore factorizes as:

$$\begin{aligned}
&\mathcal{N}_{d(\boldsymbol{s}^{(j)})}\left(\mathbf{D}(\boldsymbol{s}^{(j)})\boldsymbol{y}_i \,;\, \mathbf{D}(\boldsymbol{s}^{(j)})\boldsymbol{\mu}'_k,\, \mathbf{D}(\boldsymbol{s}^{(j)})\boldsymbol{\Sigma}'_k\mathbf{D}(\boldsymbol{s}^{(j)})^\top\right) \\
&\quad = \mathcal{N}\left(y_{i,j} \,;\, \mu_{0,j}, \sigma^2_{0,j}\right) \times \mathcal{N}_{d(\boldsymbol{s})}\left(\mathbf{D}\left(\boldsymbol{s}\right)\boldsymbol{y}_i \,;\, \mathbf{D}\left(\boldsymbol{s}\right)\boldsymbol{\mu}_k,\, \mathbf{D}\left(\boldsymbol{s}\right)\boldsymbol{\Sigma}_k\mathbf{D}\left(\boldsymbol{s}\right)^\top\right),
\end{aligned} \tag{49}$$

where $\boldsymbol{s}$ denotes the profile obtained by setting coordinate j to 0 ($s_j = 0$), and the second factor is identical under $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ since the two parameter sets coincide outside coordinate j.

Step 2: Equivalence with Inactive Background Emission. Because the background covariance matrix $\boldsymbol{\Sigma}_0 = \mathrm{diag}(\sigma^2_{0,1}, \ldots, \sigma^2_{0,J})$ is diagonal, inactive coordinates also factorize coordinate-wise. When coordinate j is inactive ($s_j = 0$), its contribution to the background emission density is precisely $\mathcal{N}(y_{i,j}; \mu_{0,j}, \sigma^2_{0,j})$. Multiplying the active and inactive blocks together, we find that for any profile $\boldsymbol{s}$ with $s_j = 0$ and its activated counterpart $\boldsymbol{s}^{(j)}$ with $s_j^{(j)} = 1$, the joint emission densities on the entire observation $\boldsymbol{y}_i \in \mathbb{R}^J$ are identical:

$$\mathcal{N}_J\left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}'_k\left(\boldsymbol{s}^{(j)}\right), \boldsymbol{\Sigma}'_k\left(\boldsymbol{s}^{(j)}\right)\right) = \mathcal{N}_J\left(\boldsymbol{y}_i \,;\, \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s})\right). \tag{50}$$

Step 3: Pairwise Cancellation Across the Binary Hypercube. Partition the 2^J binary profiles into 2^{J-1} disjoint pairs $\{\boldsymbol{s}, \boldsymbol{s}^{(j)}\}$ where $s_j = 0$ and $s_j^{(j)} = 1$. For each pair, write the product of Bernoulli probabilities over the remaining $J-1$ coordinates as:

$$w_{-j}(\boldsymbol{s}) = \prod_{l \neq j} \gamma_{k,l}^{s_l}(1-\gamma_{k,l})^{1-s_l}. \tag{51}$$

In the marginal cluster density (10), the sum of the two corresponding terms factors out the common joint emission density:

$$\begin{aligned}
&p\left(\boldsymbol{s} \mid z_i = k; \boldsymbol{\theta}'\right)\mathcal{N}_J\left(\boldsymbol{y}_i; \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s})\right) + p\left(\boldsymbol{s}^{(j)} \mid z_i = k; \boldsymbol{\theta}'\right)\mathcal{N}_J\left(\boldsymbol{y}_i; \boldsymbol{\mu}'_k\left(\boldsymbol{s}^{(j)}\right), \boldsymbol{\Sigma}'_k\left(\boldsymbol{s}^{(j)}\right)\right) \\
&\quad = \left[(1-\gamma')w_{-j}(\boldsymbol{s}) + \gamma' w_{-j}(\boldsymbol{s})\right]\mathcal{N}_J\left(\boldsymbol{y}_i; \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s})\right) \\
&\quad = w_{-j}(\boldsymbol{s})\mathcal{N}_J\left(\boldsymbol{y}_i; \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s})\right).
\end{aligned} \tag{52}$$

Because $\gamma_{k,j} = 0$ under $\boldsymbol{\theta}$, this combined term is identically the contribution of profile $\boldsymbol{s}$ under the original parameter $\boldsymbol{\theta}$. Summing over all 2^{J-1} disjoint pairs reveals that the conditional cluster density $p(\boldsymbol{y}_i \mid z_i = k)$ is completely invariant to $\gamma' \in [0,1]$. Because the cluster mixing proportions ω_k and the distributions of all other clusters $l \neq k$ remain unchanged, the marginal data density $p(\boldsymbol{y}_i) = \sum_k \omega_k p(\boldsymbol{y}_i \mid z_i = k)$ under $\boldsymbol{\theta}'$ equals that under $\boldsymbol{\theta}$ for all $\boldsymbol{y}_i \in \mathbb{R}^J$, which completes the proof. □

B.3.2 Proof of Proposition 2.7(ii): Generic Identifiability under Non-Degeneracy

Proof. Identifiability up to label permutation is established in three sequential steps: uniquely identifying the informative Gaussian components and their mixture weights from the data distribution; reconstructing the clusters, their variable supports, and continuous parameters via marginal inclusion relations; and determining the mixing proportions and activation rates via strictly monotone inversion.

Step 1: Identification of Informative Gaussian Components and Weights. By equation (11), the marginal data density expands as:

$$p(\boldsymbol{y}_i) = \widetilde{\omega}_0 \, \mathcal{N}_J(\boldsymbol{y}_i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \sum_{k=1}^{K} \sum_{\boldsymbol{s} \in \mathcal{S}^\star} \omega_k \, p(\boldsymbol{s} \mid z_i = k) \, \mathcal{N}_J(\boldsymbol{y}_i; \boldsymbol{\mu}_k(\boldsymbol{s}), \boldsymbol{\Sigma}_k(\boldsymbol{s})), \tag{53}$$

where the inner sum runs over non-null profiles with $\mathcal{A}(\boldsymbol{s}) \subseteq \mathcal{A}_k$ and $p(\boldsymbol{s} \mid z_i = k) > 0$; we refer to these terms as the *informative components* of cluster k. For any pair $(k, \boldsymbol{s})$, the j-th marginal distribution of the corresponding component is $\mathcal{N}(\mu_{k,j}, \Sigma_{k,jj})$ if $s_j = 1$ and $\mathcal{N}(\mu_{0,j}, \sigma^2_{0,j})$ if $s_j = 0$. By condition (b), the active set $\mathcal{A}(\boldsymbol{s})$ is therefore uniquely characterized as the set of coordinates whose marginal distribution differs from that of the background:

$$\mathcal{A}(\boldsymbol{s}) = \left\{ j \in \{1, \ldots, J\} : (\mu_{k,j}, \Sigma_{k,jj}) \neq (\mu_{0,j}, \sigma^2_{0,j}) \right\}. \tag{54}$$

Consequently, every informative component differs from the background distribution. Furthermore:

- Two informative components with different active sets $\mathcal{A}(\boldsymbol{s}) \neq \mathcal{A}(\boldsymbol{s}')$ are necessarily distinct.
- Two informative components from distinct clusters $k \neq l$ sharing the same active set $A \subseteq \mathcal{A}_k \cap \mathcal{A}_l$ differ by their joint marginal distribution on A, by condition (c).

The informative components are therefore pairwise distinct and distinct from the background distribution, so that $p(\boldsymbol{y}_i)$ is a finite mixture of mutually distinct Gaussian distributions with strictly positive weights. The same representation holds under any observationally equivalent parameter $\boldsymbol{\theta}'$. Because finite mixtures of multivariate Gaussian distributions are identifiable (Yakowitz and Spragins, 1968), the set of informative Gaussian components and their corresponding weights $\omega_k \, p(\boldsymbol{s} \mid z_i = k)$ are uniquely identified.

Step 2: Reconstruction of Clusters, Supports, and Parameters via Marginal Projections. We say that an informative component with active set A' is a *marginal* of another component with active set $A \supsetneq A'$ if its mean vector and covariance matrix restricted to A' equal those of the latter. For each cluster k, the full profile $\boldsymbol{s}_k^{\text{full}}$ (equal to 1 on $\mathcal{A}_k$ and 0 elsewhere) has strictly positive prior probability $\prod_{j \in \mathcal{A}_k} \gamma_{k,j} > 0$. Every other informative component of cluster k has active set $A \subsetneq \mathcal{A}_k$ and is a marginal of $\mathcal{N}_J(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ on A, and hence a marginal of the component associated with $\boldsymbol{s}_k^{\text{full}}$. Conversely, an informative component of

cluster $l \neq k$ cannot be a marginal of a component of cluster k: this would require clusters k and l to coincide on a non-empty subset $A \subseteq \mathcal{A}_k \cap \mathcal{A}_l$, which is ruled out by condition (c).

Consequently, the informative components that are maximal (i.e., not marginals of any other component) are exactly the K full-profile components. Every other informative component is a marginal of exactly one maximal component. The identified set of informative components therefore uniquely determines:

1. the partition of components into the K clusters;
2. the variable supports $\mathcal{A}_k$ (as the active sets of the K maximal components);
3. the continuous cluster means $\mu_{k,j}$ and covariances $\Sigma_{k,jl}$ for all $j, l \in \mathcal{A}_k$.

Step 3: Strict Monotonicity and Inversion of Mixing Proportions and Activation Rates. For each cluster k and each active coordinate $j \in \mathcal{A}_k$, summing the weights of all informative components of cluster k that activate coordinate j yields the identified quantities:

$$a_{k,j} \equiv \sum_{\boldsymbol{s}: s_j = 1} \omega_k \, p(\boldsymbol{s} \mid z_i = k) = \omega_k \gamma_{k,j}, \tag{55}$$

while summing the weights of all non-null components of cluster k yields:

$$W_k \equiv \sum_{\boldsymbol{s} \neq \mathbf{0}_J} \omega_k \, p(\boldsymbol{s} \mid z_i = k) = \omega_k \left[1 - \prod_{j \in \mathcal{A}_k} (1 - \gamma_{k,j}) \right]. \tag{56}$$

For any $\omega \geq \max_{j \in \mathcal{A}_k} a_{k,j}$, define the scalar function:

$$f_k(\omega) = \omega \left[1 - \prod_{j \in \mathcal{A}_k} \left(1 - \frac{a_{k,j}}{\omega} \right) \right], \tag{57}$$

so that $f_k(\omega_k) = W_k$. Setting $p_j(\omega) = a_{k,j}/\omega \in (0, 1]$ and letting N denote the sum of independent Bernoulli random variables with success probabilities $p_j(\omega)$ for $j \in \mathcal{A}_k$, differentiating f_k with respect to ω gives:

$$\begin{aligned} f_k'(\omega) &= 1 - \prod_{j \in \mathcal{A}_k} (1 - p_j) - \sum_{j \in \mathcal{A}_k} p_j \prod_{l \neq j} (1 - p_l) \\ &= 1 - \mathbb{P}(N = 0) - \mathbb{P}(N = 1) = \mathbb{P}(N \geq 2). \end{aligned} \tag{58}$$

Whenever $|\mathcal{A}_k| \geq 2$, the random variable N can take values ≥ 2 with non-zero probability, so $f_k'(\omega) > 0$. The function f_k is therefore strictly increasing on $[\max_j a_{k,j}, 1]$, and ω_k is the unique solution to $f_k(\omega) = W_k$. By condition (d), at most one cluster satisfies $|\mathcal{A}_k| = 1$; its mixing proportion is then uniquely determined by the simplex constraint $\omega_k = 1 - \sum_{l \neq k} \omega_l$. Finally, having uniquely recovered ω_k, each activation rate is determined by:

$$\gamma_{k,j} = \frac{a_{k,j}}{\omega_k} \qquad \text{for all } j \in \mathcal{A}_k. \tag{59}$$

This completes the proof of identifiability up to label permutation. □

C Proofs of Section 3 and Algorithm

C.1 Proof of Proposition 3.2 (Boundedness and Parsimony of the Penalized Criterion)

Proof. The proof is divided into two parts: establishing the uniform upper bound of $\mathcal{G}$ across all configurations, and demonstrating that redundant active blocks strictly decrease the penalized criterion.

Part (i): Uniform Boundedness of the Objective Function. The cluster mixing log-weights $\log \omega_k \leq 0$ and the profile prior log-probabilities $\log p(\boldsymbol{s} \mid z_i = k) \leq 0$ are non-positive. The discrete Shannon entropy of responsibilities satisfies $-\sum_{k=1}^{K} r_{i,k} \log r_{i,k} \leq \log K$, and the dimension penalty satisfies $\nu(\mathbf{B}) \geq 0$. Furthermore, by Lemma 3.1(ii), every active Gaussian density is uniformly bounded by $(2\pi\delta)^{-d/2} \leq \max\{1, (2\pi\delta)^{-J/2}\}$, and every background density is bounded above by $(2\pi)^{-(J-d)/2} \prod_{j \in \mathcal{A}^c} \sigma_{0,j}^{-1}$ because the background covariance $\boldsymbol{\Sigma}_0$ is fixed and positive definite. Each individual observation's contribution in $\mathcal{F}$ is therefore bounded above by a finite constant independent of the parameters. Because $\mathcal{G} \leq \mathcal{F}$, the objective $\mathcal{G}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}, \mathbf{B})$ is uniformly bounded above across all parameter configurations and latent profiles.

Part (ii): Deterministic Resolution of Non-Identifiability via Parsimony. Let coordinate j be an extra active block in $\mathbf{B}'$ that is inactive in $\mathbf{B}$ ($B'_{k,j} = 1$ and $B_{k,j} = 0$). By construction of $\boldsymbol{\theta}'$, coordinate j has mean $\mu'_{k,j} = \mu_{0,j}$, variance $\Sigma'_{k,jj} = \sigma^2_{0,j}$, and zero covariance with all other coordinates. Hence, whether cell (i, j) is classified as active ($s'_{i,k,j} = 1$) or inactive ($s'_{i,k,j} = 0$), its emission density under $\boldsymbol{\theta}'$ equals the background density $\mathcal{N}(y_{i,j}; \mu_{0,j}, \sigma^2_{0,j})$, which is identical to its contribution under $\boldsymbol{\theta}$; since coordinate j is uncorrelated with the other coordinates, the active density of the other coordinates is unchanged as well. The Bernoulli profile prior contributes $\log \gamma'_{k,j} \leq 0$ if $s'_{i,k,j} = 1$ and $\log(1 - \gamma'_{k,j}) \leq 0$ if $s'_{i,k,j} = 0$; hence $\mathcal{F}(\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{\theta}') \leq \mathcal{F}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta})$. Subtracting the dimension penalties yields:

$$\mathcal{G}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}, \mathbf{B}) - \mathcal{G}(\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{\theta}', \mathbf{B}') \geq \frac{\log n}{2} \left[\nu(\mathbf{B}') - \nu(\mathbf{B})\right]. \tag{60}$$

Because $\mathbf{B}'$ strictly contains $\mathbf{B}$, activating coordinate j in cluster k adds at least one mean, one variance, and one activation rate, so that $\nu(\mathbf{B}') - \nu(\mathbf{B}) \geq 3$. This ensures that the penalty gap is strictly positive for all sample sizes $n > 1$, which strictly favors the parsimonious support $\mathbf{B}$. □

C.2 Proof of Proposition 3.3 (Monotone Updates)

Proof of Proposition 3.3. The coordinate ascent algorithm updates the parameters and latent variables cyclically. We show that each of the five conditional steps either strictly increases $\mathcal{G}$ or leaves it unchanged while maintaining spectral feasibility $\boldsymbol{\Sigma}_k \succeq \delta \mathbf{I}_J$.

Step 1: Responsibilities Update ($\boldsymbol{r}$). For fixed $(\boldsymbol{s}, \boldsymbol{\theta}, \mathbf{B})$, the objective $\mathcal{G}$ depends on $\boldsymbol{r}_i = (r_{i,1}, \ldots, r_{i,K})$ through:

$$\sum_{k=1}^{K} r_{i,k} \left[\log \omega_k + g_{i,k}(\boldsymbol{s}_{i,k}) - \log r_{i,k}\right]. \tag{61}$$

This is a strictly concave function of $\boldsymbol{r}_i$ on the probability simplex $\sum_{k=1}^{K} r_{i,k} = 1$. By Gibbs' inequality (or relative entropy minimization), the unconstrained maximizer on the simplex is unique and given by the normalized softmax:

$$r_{i,k} = \frac{\omega_k \exp(g_{i,k}(\boldsymbol{s}_{i,k}))}{\sum_{l=1}^{K} \omega_l \exp(g_{i,l}(\boldsymbol{s}_{i,l}))}, \tag{62}$$

which guarantees a non-decreasing step in $\mathcal{G}$.

Step 2: Latent Binary Profile Classification ($\boldsymbol{s}$). For fixed $(\boldsymbol{r}, \boldsymbol{\theta}, \mathbf{B})$, $\mathcal{G}$ depends on the profile $\boldsymbol{s}_{i,k}$ only through the term $r_{i,k} g_{i,k}(\boldsymbol{s}_{i,k})$. Because $r_{i,k} \geq 0$, any modification of $\boldsymbol{s}_{i,k}$ that increases $g_{i,k}$ increases $\mathcal{G}$. Changes are restricted to active blocks ($B_{k,j} = 1$), on which $g_{i,k}$ is finite; activating a coordinate belonging to an inactive block would set $g_{i,k} = -\infty$. Evaluating the coordinate-wise candidate change $s_{i,k,j} \leftarrow 1 - s_{i,k,j}$, the increment $\Delta\mathcal{G}_{i,k,j}$ follows from the factorization of the Gaussian density on $\mathcal{A} \cup \{j\}$ into the marginal density on $\mathcal{A}$ and the conditional density of $y_{i,j}$ given $\boldsymbol{y}_{i,\mathcal{A}}$, the background factor of coordinate j being removed and the profile prior ratio being multiplied by $\gamma_{k,j}/(1 - \gamma_{k,j})$. Since the update rule accepts a candidate flip if and only if $\Delta\mathcal{G}_{i,k,j} > 0$, the objective $\mathcal{G}$ is non-decreasing.

Step 3: Mixing Proportions and Activation Rates ($\boldsymbol{\omega}, \boldsymbol{\gamma}$). Maximizing $\sum_{i=1}^{n} \sum_{k=1}^{K} r_{i,k} \log \omega_k$ subject to $\sum_{k=1}^{K} \omega_k = 1$ is the canonical multinomial problem, maximized by the empirical cluster weights $\omega_k = n^{-1} \sum_{i=1}^{n} r_{i,k}$. For each active block (k, j), the criterion depends on $\gamma_{k,j}$ through:

$$\sum_{i=1}^{n} r_{i,k} \left[s_{i,k,j} \log \gamma_{k,j} + (1 - s_{i,k,j}) \log(1 - \gamma_{k,j})\right]. \tag{63}$$

This function is strictly concave in $\gamma_{k,j}$. Its unconstrained global maximizer is the weighted empirical frequency $\bar{s}_{k,j} = \sum_i r_{i,k} s_{i,k,j} / \sum_i r_{i,k}$. Since the objective is concave, its unique maximizer on the compact interval $[\epsilon, 1 - \epsilon]$ is obtained by projecting $\bar{s}_{k,j}$ onto $[\epsilon, 1 - \epsilon]$, that is, $\gamma_{k,j} = \mathrm{clip}(\bar{s}_{k,j}, \epsilon, 1 - \epsilon)$.

Step 4: Continuous Gaussian Parameters under Spectral Regularization ($\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$). For fixed $(\boldsymbol{r}, \boldsymbol{s})$, $\mathcal{G}$ depends on $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ through:

$$\sum_{i=1}^{n} r_{i,k} \log \mathcal{N}\left(\mathbf{D}(\boldsymbol{s}_{i,k})\boldsymbol{y}_i \,;\, \mathbf{D}(\boldsymbol{s}_{i,k})\boldsymbol{\mu}_k,\, \mathbf{D}(\boldsymbol{s}_{i,k})\boldsymbol{\Sigma}_k \mathbf{D}(\boldsymbol{s}_{i,k})^\top\right). \tag{64}$$

This objective is the weighted log-likelihood of a multivariate Gaussian sample with missing coordinates $\mathcal{M}_{i,k} = \mathcal{A}(\boldsymbol{s}_{i,k})^c$.

- *Missing-data representation and mean update:* Following the EM framework for missing data (Dempster et al., 1977), the conditional expectations of the complete-data sufficient statistics yield the completed observations $\widehat{\boldsymbol{y}}_{i,k} = \mathbb{E}[\boldsymbol{y} \mid \mathbf{D}(\boldsymbol{s}_{i,k})\boldsymbol{y}_i, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}]$ and conditional cross-moments $\mathbf{C}_{i,k} = \mathrm{Var}(\boldsymbol{y} \mid \mathbf{D}(\boldsymbol{s}_{i,k})\boldsymbol{y}_i, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})$. The expected complete-data log-likelihood equals, up to an additive constant:

$$Q_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N_k}{2}\left[\log\det\boldsymbol{\Sigma} + \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}_k(\boldsymbol{\mu})\right)\right], \qquad N_k = \sum_{i=1}^{n} r_{i,k}, \tag{65}$$

where $\mathbf{S}_k(\boldsymbol{\mu}) = N_k^{-1}\sum_i r_{i,k}[(\widehat{\boldsymbol{y}}_{i,k} - \boldsymbol{\mu})(\widehat{\boldsymbol{y}}_{i,k} - \boldsymbol{\mu})^\top + \mathbf{C}_{i,k}]$. Differentiating with respect to $\boldsymbol{\mu}$ shows that Q_k is uniquely maximized by the completed empirical mean $\boldsymbol{\mu}_k = N_k^{-1}\sum_i r_{i,k}\widehat{\boldsymbol{y}}_{i,k}$ for any positive definite $\boldsymbol{\Sigma}$.

- *Constrained covariance update via von Neumann's inequality:* Let $\mathbf{S}_k = \mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{U}_k^\top$ be the eigendecomposition of the completed scatter matrix $\mathbf{S}_k(\boldsymbol{\mu}_k)$, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_J \geq 0$. For any candidate covariance $\boldsymbol{\Sigma}$ with ordered eigenvalues $\sigma_1 \geq \cdots \geq \sigma_J$, von Neumann's trace inequality yields:

$$\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_k) \geq \sum_{m=1}^{J}\frac{\lambda_m}{\sigma_m}, \tag{66}$$

with equality if and only if $\boldsymbol{\Sigma}$ has the same eigenvectors as $\mathbf{S}_k$ arranged in the same order. Under the spectral floor $\lambda_J(\boldsymbol{\Sigma}) \geq \delta$, maximizing Q_k over $\boldsymbol{\Sigma} \succeq \delta\mathbf{I}_J$ thus separates into J independent scalar convex minimization problems:

$$\min_{\sigma_m \geq \delta}\left\{\log\sigma_m + \frac{\lambda_m}{\sigma_m}\right\}, \qquad m = 1, \ldots, J. \tag{67}$$

The unique minimizer on $[\delta, +\infty)$ is $\sigma_m^* = \max(\lambda_m, \delta)$, so that the optimal matrix is $\boldsymbol{\Sigma}_k^* = \mathbf{U}_k\max(\boldsymbol{\Lambda}_k, \delta)\mathbf{U}_k^\top$.

- *Feasibility and ascent property:* Because $\boldsymbol{\Sigma}_k^*$ maximizes the surrogate Q_k over the feasible cone $\mathcal{S}_{++}^J(\delta)$, Dempster's minorization theorem guarantees that the observed incomplete log-likelihood does not decrease. Moreover, the updated matrix satisfies $\lambda_{\min}(\boldsymbol{\Sigma}_k^*) \geq \delta$ by construction, maintaining strict spectral feasibility.

Step 5: Support Matrix Sparsification and Re-activation (B). Switching an active block off ($B_{k,j} \leftarrow 0$) sets $s_{i,k,j} = 0$ for all i and $\gamma_{k,j} = 0$, leaving the continuous parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ unchanged; switching an inactive block on ($B_{k,j} \leftarrow 1$) initializes coordinate j with background statistics as described. Because the parameters of an inactive block do not enter $\mathcal{G}$, the value of the criterion before the proposed move matches the current state. In both directions, the move is accepted if and only if the change in penalized criterion $\Delta\mathcal{G}$ (including the modification of $\nu(\mathbf{B})$) satisfies $\Delta\mathcal{G} \geq 0$. The criterion $\mathcal{G}$ therefore never decreases. When a block is switched on, the augmented covariance matrix is block-diagonal with blocks $\boldsymbol{\Sigma}_{k,-j,-j}$ and $\Sigma_{k,jj} \geq \delta$. By Lemma 3.1, its eigenvalues satisfy $\lambda_{\min} \geq \min\{\lambda_J(\boldsymbol{\Sigma}_k), \Sigma_{k,jj}\} \geq \delta$, so spectral feasibility is preserved across support updates. □

Algorithm 1 Penalized Classification EM for FC-GMM-LFR (K known)

Require: Data $\boldsymbol{y} \in \mathbb{R}^{n \times J}$, number of clusters K, floor $\delta > 0$, $\epsilon > 0$, tolerance ε_{tol}, number of starting points S.

Draw S initial responsibilities $\boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(S)}$ by k-means++ and estimate the background $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ from $\boldsymbol{r}^{(1)}$ by (29).
repeat ▷ alternation with the background
  for $m = 1, \ldots, S$ **do**
    Initialize $\boldsymbol{r} = \boldsymbol{r}^{(m)}$, $\mathbf{B} = \mathbf{1}_{K \times J}$, $\boldsymbol{s}_{i,k} = \mathbf{1}_J$, $\gamma_{k,j} = 1/2$, $\boldsymbol{\omega}$ by step 3 and $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ by the weighted moments of $\boldsymbol{y}$ with eigenvalues clipped at δ.
    repeat
      Profiles $\boldsymbol{s}$ (step 2); responsibilities $\boldsymbol{r}$ (step 1); $\boldsymbol{\omega}, \boldsymbol{\gamma}$ (step 3); $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ (step 4); active blocks $\mathbf{B}$ (step 5).
    until the relative change of $\mathcal{G}$ over an iteration is below ε_{tol}, or 200 iterations
  end for
  Keep the fit with the largest $\mathcal{G}$ among the S starting points, and re-estimate the background from its responsibilities by (29).
until the background no longer changes, returns to a background already visited (then the fit with the largest $\mathcal{G}$ in the cycle is kept), or 10 passes
return $(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}, \mathbf{B}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ of the retained fit.

D A Model with a Free Weight for Non-Informative Observations

In the model of Section 2, the weight of the non-informative observations is the probability of the null profile, $\widetilde{\omega}_0 = \sum_k \omega_k \prod_j (1 - \gamma_{k,j})$, which is determined by the activation rates. Data may also contain observations that are unrelated to the clusters, in a proportion that has no reason to equal $\widetilde{\omega}_0$. A second formulation adds a class 0 that follows the background, with a free weight ω_0:

$$p(\boldsymbol{y}_i) = \omega_0 \, \mathcal{N}_J(\boldsymbol{y}_i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \sum_{k=1}^{K} \omega_k \, p(\boldsymbol{y}_i \mid z_i = k), \qquad \omega_0 + \sum_{k=1}^{K} \omega_k = 1, \tag{68}$$

where $p(\boldsymbol{y}_i \mid z_i = k)$ is the cluster density (10). The model of Section 2 is the case $\omega_0 = 0$.

Theory.

The results of Sections 2 and 3 carry over with one more class. In Proposition 2.5, the weight of the non-informative component becomes $\omega_0 + \sum_k \omega_k \prod_j (1 - \gamma_{k,j})$. Part (i) of Proposition 2.7 is unchanged; in part (ii), the proof only uses the informative components and holds without change, except that the weight of a cluster whose support is a single variable can no longer be obtained from the sum of the weights, so that condition (d) becomes: no support is reduced to one variable. The classification criterion receives the term $\sum_i r_{i,0} [\log \omega_0 - \log r_{i,0} +$

$\log \mathcal{N}_J(\boldsymbol{y}_i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)]$, where $r_{i,0}$ is the responsibility of class 0; it remains a lower bound of the log-likelihood and, the background density being bounded, Proposition 3.2 holds. In Proposition 3.3, step 1 becomes the softmax over $K+1$ classes and step 3 adds $\omega_0 = n^{-1} \sum_i r_{i,0}$, both exact maximizers; steps 2, 4 and 5 are unchanged. In the estimation of the background (29), the observations of class 0 enter the background component with weight $r_{i,0}$. In the two-step rule (19) and (20), $\widetilde{\omega}_0$ is replaced by $\omega_0 + \widetilde{\omega}_0$.

Identifiability.

An observation of class 0 and a cluster member whose coordinates are all inactive have the same distribution, the background, so that the two cannot be told apart individually. Their proportions are nevertheless identified under the conditions of Proposition 2.7(ii) with no support reduced to one variable: the cluster weights and activation rates are identified from the informative components, hence the weight $\sum_k \omega_k \prod_j (1 - \gamma_{k,j})$ of the null profiles, and $\omega_0 = 1 - \sum_k \omega_k$. This identification rests entirely on the activation rates, which explains why the free weight is hard to estimate when the clusters are weakly separated (see below). It does not affect the classification, since both kinds of observations are non-informative.

Illustration.

A proportion $\pi_0 = 0.10$ of the $n = 1\,000$ observations is drawn from the background, the others from the base design with $\Delta = 2.5$ and $\gamma = 0.9$, so that the data follow (68) with $\omega_0 = 0.10$. The model of Section 2 and the extended model are fitted on the same simulated data (Table 12). The model of Section 2 detects 6% of the observations drawn from the background with its two-step rule and 49% with the inactivation rate: its weight of the null profile, determined by activation rates close to 0.9, is much smaller than 0.10. The extended model estimates the weight of the non-informative observations at 0.115 and detects 94% of them with 2.1% of false alarms, close to its two-step rule with the true parameters (92.5% and 1.7%). On trial data sets with smaller separations ($\Delta = 2.0$ and $\Delta = 1.5$), the extended model did not approach the rule with the true parameters: the non-informative class either vanished or absorbed most observations.

Table 12: Observations drawn from the background in proportion $\pi_0 = 0.10$ ($\Delta = 2.5$, $\gamma = 0.9$, $n = 1\,000$, $R = 20$): detection, false alarms, ARI of the observations kept and estimated weight of the non-informative observations, for the model of Section 2 and the extended model.

rule	detection	false alarms	ARI (kept)	estimated weight of non-informative obs.
main model, two-step rule	0.057 (0.016)	0.000 (0.000)	0.977 (0.003)	0.000 (0.000)
main model, inactivation rate	0.491 (0.066)	0.010 (0.001)	0.982 (0.004)	–
extended model, two-step rule	0.943 (0.004)	0.021 (0.001)	0.994 (0.001)	0.115 (0.001)
extended model, true parameters	0.925 (0.004)	0.017 (0.001)	–	–

E Computing Time

The base design ($\Delta = 2.5$) is generated with $n \in \{500, 1\,000, 2\,000, 5\,000\}$ and $J \in \{20, 40, 80\}$, the four relevant features of each cluster being complemented by irrelevant ones, and fitted with the default settings (five starting points, background estimated in alternation with the fit), one thread per fit and four fits running concurrently on a laptop computer. Table 13 reports the elapsed time of a complete fit, the number of background passes, the number of iterations of the retained run, the proportion of starting points that reach the best criterion and the ARI ($R = 10$). The active structure is recovered without error in every configuration. The time grows roughly linearly with n and faster than linearly with J, from about 26 seconds ($n = 500$, $J = 20$) to 27 minutes ($n = 5\,000$, $J = 80$); between 46% and 98% of the starting points reach the best criterion.

Table 13: Computing time of a complete fit (five starting points, background alternated with the fit) and convergence diagnostics ($\Delta = 2.5$, $R = 10$).

J	n	seconds	background passes	iterations	starts at the best criterion	ARI
20	500	26 (5)	2.4 (0.2)	9.5 (0.4)	0.66 (0.10)	0.988 (0.003)
20	1000	46 (6)	2.3 (0.2)	10.3 (0.6)	0.90 (0.04)	0.988 (0.002)
20	2000	86 (10)	2.2 (0.1)	12.6 (0.6)	0.82 (0.08)	0.987 (0.002)
20	5000	283 (43)	2.8 (0.1)	13.1 (0.6)	0.86 (0.05)	0.988 (0.001)
40	500	68 (7)	2.6 (0.2)	10.4 (1.1)	0.64 (0.08)	0.990 (0.003)
40	1000	110 (16)	2.5 (0.2)	10.7 (0.5)	0.90 (0.04)	0.986 (0.002)
40	2000	194 (19)	2.3 (0.2)	15.4 (1.2)	0.88 (0.07)	0.990 (0.001)
40	5000	469 (27)	2.9 (0.1)	14.2 (0.6)	0.98 (0.02)	0.988 (0.001)
80	500	201 (12)	3.0 (0.1)	10.9 (1.1)	0.46 (0.10)	0.989 (0.003)
80	1000	306 (36)	2.2 (0.1)	10.9 (0.8)	0.60 (0.10)	0.986 (0.002)
80	2000	526 (25)	2.1 (0.1)	13.0 (1.0)	0.82 (0.06)	0.988 (0.001)
80	5000	1628 (111)	2.9 (0.1)	16.0 (1.0)	0.96 (0.03)	0.987 (0.001)