IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

# Mixture of Logistic Regression with LASSO Regularization for High-Dimensional Binary Classification with Latent Heterogeneity

Buhari Ishaq[1], Abubakar Usman[2] , Yahaya Zakari[3] and Hillary Ugwu Okwudili[4]

[1,4] Department of Military Science, Nigerian Defence Academy, Kaduna, Nigeria
[2,3] Department of Statistics, Ahmadu Bello University, Zaria, Nigeria
[1]E-mail of First Author : buharii@nda.edu.ng
[2]E-mail of Second Author : abubakarusman28@gmail.com
[3]E-mail of Third Author : yahzaksta@gmail.com
[4]E-mail of Third Author: uh.okwudili@nda.edu.ng

**ARTICLE INFO**

***Article history:***
Received 15 April 2026,
Revised 15 April 2026,
Accepted 27 May 2026,
Available online 10 July 2026

***Keywords:***
Mixture models
LASSO regularisation
high-dimensional classification
EM algorithm
breast cancer diagnosis

**ABSTRACT**

This paper proposes a Mixture of Logistic Regression with LASSO Regularization (MLR-LR) framework for high-dimensional binary classification under latent heterogeneity. The approach integrates finite mixture modelling to capture unobserved subpopulations with $\ell_1$-penalized logistic regression to induce sparsity, enabling simultaneous subgroup identification and variable selection. Parameter estimation is carried out via a penalized Expectation-Maximization algorithm with component-wise coordinate descent. In contrast to the primarily applied framework of Morvan et al. (2021), we develop a rigorous theoretical foundation for penalized mixture logistic models. Under high-dimensional asymptotic regimes, we establish identifiability, estimation consistency, sparsity recovery, and the oracle property of the proposed estimator. Extensive simulations across six heterogeneity regimes (R1-R6) with 200 replications demonstrate that MLR-LR consistently outperforms homogeneous LASSO when latent structure is present, while remaining competitive in homogeneous settings. In particular, the proposed method achieves systematic improvements in classification performance, with gains in AUC ranging from 0.028 to 0.054 under heterogeneous scenarios. Application to the Wisconsin Diagnostic Breast Cancer dataset further illustrates the practical utility of the method in the presence of multicollinearity, where MLR-LR attains superior predictive performance (AUC=0.9903, F1=0.9550, misclassification=0.0292) relative to LASSO, Random Forest, and XGBoost, while yielding interpretable, component-specific coefficient structures that reveal distinct risk profiles. Overall, the proposed framework provides a unified and theoretically grounded approach for high-dimensional heterogeneous binary classification, bridging the gap between mixture modelling and sparse statistical learning.

## 1. Introduction

### 1.1 Background

Binary classification remains a central problem in modern data science, with applications spanning healthcare, bioinformatics, finance, and defence analytics. Classical logistic regression is widely used due to its interpretability and probabilistic foundation; however, it assumes that observations arise from a single homogeneous

Corresponding author E-mail address: buharii@gmail.com
https://doi.org/10.62933/h3051210

population. In many real-world settings, this assumption is unrealistic. Heterogeneous populations frequently arise in clinical diagnosis (e.g., disease subtypes), behavioural modelling, and socio-technical systems, where latent subgroups exhibit distinct predictor-outcome relationships.

Finite mixture models provide a principled framework for modelling such latent heterogeneity by representing the observed distribution as a convex combination of multiple component distributions [1]. Mixture logistic regression, a special case of mixture-of-experts models [24], extends classical logistic regression by allowing multiple latent components, each characterised by its own regression structure. This framework enables subgroup discovery and improves predictive flexibility when the underlying population structure is unknown. However, the benefits of mixture modelling diminish in high-dimensional regimes, where the number of predictors is large relative to the sample size. In such settings, mixture models become prone to overfitting, unstable estimation, and reduced interpretability due to the proliferation of parameters [5].

High-dimensional statistical learning has therefore increasingly relied on regularisation techniques to enforce sparsity and improve generalisation. Among these, the Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Tibshirani [6], has become a cornerstone methodology. By imposing an $\ell_1$ penalty, LASSO simultaneously performs variable selection and shrinkage, yielding parsimonious and interpretable models even when the number of predictors is large. In logistic regression, LASSO has been shown to improve predictive performance and stability under high-dimensional settings, particularly when irrelevant or redundant features are present [5].

Despite these advances, modern datasets increasingly exhibit both high dimensionality and latent heterogeneity. For example, biomedical datasets may contain hundreds of biomarkers alongside unobserved disease subtypes, while behavioural or defence-related datasets often combine complex latent structure with sparse predictive signals. Classical logistic regression fails under heterogeneity, while LASSO-logistic regression assumes population homogeneity. Conversely, traditional mixture models capture heterogeneity but typically lack principled mechanisms for variable selection in high-dimensional settings. This creates a methodological gap at the intersection of sparse learning and heterogeneous modelling.

### 1.2 Research Gap and Motivation

The statistical literature on binary classification has developed along three largely distinct directions: logistic regression for modelling binary outcomes, sparsity-inducing regularisation for high-dimensional inference, and finite mixture models for capturing latent heterogeneity. While each paradigm is well established, its integration remains incomplete, particularly in regimes where the number of predictors is large relative to sample size, and the population exhibits unobserved substructure.

This gap is consequential in modern applications, where high-dimensional covariates and latent heterogeneity arise simultaneously. Standard logistic regression and its $\ell_1$-regularised variants assume population homogeneity and may yield biased or oversimplified models in the presence of latent subgroups. Conversely, mixture logistic regression models account for heterogeneity but are statistically and computationally challenging in high dimensions due to non-convex likelihoods, parameter proliferation, and the absence of intrinsic variable selection mechanisms [9, 10].

To address these limitations, we propose a Mixture of Logistic Regression with LASSO Regularisation (MLR-LR) framework that integrates finite mixture modelling with $\ell_1$-penalised estimation. The proposed approach embeds sparsity-inducing regularisation within a penalised Expectation–Maximisation

framework [16], enabling simultaneous recovery of latent subgroup structure and sparse predictive signals.

In contrast to the primarily applied framework of [13], who developed a penalized mixture logistic regression for NASH prediction, we provide a rigorous theoretical treatment of penalised mixture logistic models. In particular, under high-dimensional asymptotic regimes, we establish identifiability, estimation consistency, sparsity recovery, and the oracle property of the proposed estimator.

### 1.3 Summary of Contributions

The main contributions of this work are fourfold:

- **(i) Modelling contribution:** We propose a finite mixture of logistic regressions with LASSO regularisation (MLR-LR) that simultaneously captures latent population heterogeneity and enforces sparsity in high-dimensional settings. Unlike standard mixture models, our framework performs intrinsic variable selection across components, yielding interpretable subgroup-specific predictive signatures. The model accommodates six distinct regimes of heterogeneity to evaluate performance under varying latent structures systematically.
- **(ii) Algorithmic contribution:** We develop a penalised Expectation-Maximisation (EM) algorithm incorporating a LASSO-penalised M-step via component-wise cyclic coordinate descent. The algorithm efficiently handles high-dimensional predictors ($p>n$) and produces a full regularisation path over $\lambda$, with warm starts for computational efficiency and convergence guarantees to a stationary point.
- **(iii) Theoretical contribution:** Under high-dimensional asymptotic regimes, we establish: (a) identifiability conditions for the penalised mixture logistic model; (b) estimation consistency; (c) sparsity recovery (variable selection consistency); and (d) the oracle property, demonstrating that the estimator performs asymptotically as well as if the true sparse structure and component assignments were known a priori.
- **(iv) Empirical evidence:** We validate the proposed method through extensive simulations across six heterogeneity regimes with 200 replications each, demonstrating that MLR-LR consistently outperforms homogeneous LASSO when latent structure is present (AUC gains of 0.028-0.054) while remaining competitive in homogeneous settings. Application to the Wisconsin Diagnostic Breast Cancer dataset further illustrates practical utility in the presence of multicollinearity, where MLR-LR achieves superior predictive performance (AUC = 0.9903, F1 = 0.9550, misclassification = 0.0292) relative to LASSO, Random Forest, and XGBoost, while yielding interpretable, component-specific coefficient structures that reveal distinct risk profiles.

## 2. Related Work

### 2.1 Mixture Models for Heterogeneous Binary Data

Finite mixture models provide a principled framework for modelling unobserved heterogeneity by representing the observed distribution as a convex combination of component densities [1]. In regression contexts, mixtures allow covariate effects to vary across latent subpopulations, capturing structure that homogeneous models cannot [2]. For binary outcomes, mixture logistic regression has been applied in clinical diagnosis, where disease subtypes exhibit distinct risk factor profiles [3] and in defence analytics, where adversary behaviour may follow latent patterns [4].

However, standard estimation using the EM algorithm becomes unstable in high-dimensional settings. With p predictors and K components, the model requires K×p regression coefficients, leading to high variance, poor generalisation, and computational inefficiency [5]. This overparameterisation not only degrades predictive performance but also obscures interpretability, as irrelevant predictors dilute subgroup-specific signals. Regularisation offers a natural remedy by shrinking redundant coefficients toward zero while preserving mixture structure (a motivation that directly connects to the penalisation framework discussed below).

## 2.2 Sparsity-Inducing Regularisation in High Dimensions

Penalised likelihood methods with $\ell_1$ regularisation (LASSO) have become standard for high-dimensional regression [6]. LASSO simultaneously performs variable selection and shrinkage, producing sparse and interpretable models. For logistic regression, coordinate descent algorithms efficiently handle cases where $p > n$ [7], and consistent variable selection is guaranteed under irrepresentable conditions [8].

A parallel line of work explored penalised logistic regression for high-dimensional classification. [14] proposed $\ell 2$-penalised (ridge) logistic regression for gene microarray classification, demonstrating improved predictive performance over unregularised methods. However, their approach uses ridge regularisation, which shrinks coefficients but does not enforce sparsity, limiting interpretability when many irrelevant predictors are present. The LASSO-based framework we adopt addresses this limitation by enabling simultaneous variable selection and coefficient estimation.

Despite these advances, conventional logistic LASSO assumes a homogeneous population, limiting its applicability when latent subgroups are present. Ignoring heterogeneity can bias estimates, reduce classification accuracy, and obscure subgroup-specific predictive relationships. This limitation motivates the need for frameworks that integrate sparsity induction with mixture modelling.

## 2.3 Penalised Mixture Models: Unifying Sparsity and Heterogeneity

The integration of sparsity-inducing penalties into mixture models has been explored primarily for continuous outcomes. [9] proposed LASSO-based variable selection in finite mixtures of linear regression, establishing consistency and oracle properties under fixed-p asymptotics. [10,11] extended these results to high-dimensional Gaussian mixtures, deriving non-asymptotic $\ell_1$-oracle inequalities and demonstrating that penalised EM algorithms can recover sparse structures even when the number of predictors grows exponentially with sample size. These frameworks, however, assume sub-Gaussian errors and require careful tuning to balance sparsity and mixture complexity.

For binary outcomes, penalised mixture models remain underdeveloped. The logistic likelihood introduces non-convexity that complicates both optimisation and theoretical analysis, unlike the Gaussian case, where closed-form updates exist within the EM algorithm. Furthermore, identifiability becomes more delicate when mixing binary regression components, requiring careful constraints to separate component-specific effects from overall intercept shifts. Existing attempts to address these challenges remain limited: [11] note that extending their theoretical guarantees to non-Gaussian settings is non-trivial, while [12] explicitly leaves binary outcomes as an open problem.

Related approaches in mixture-of-experts (MoE) models have addressed some of these challenges, but with critical limitations. [15] provided computational frameworks for fitting finite mixtures of generalised linear regressions in R, yet their implementation does not incorporate sparsity-inducing penalties for high-dimensional settings. [27] developed estimation and feature selection methods for

high-dimensional MoE models, but the theoretical guarantees assume specific regularity conditions that may not hold for binary responses with weak signal-to-noise ratios. More recently, [28] proposed prediction sets for high-dimensional mixture of experts models, addressing uncertainty quantification rather than variable selection or oracle properties. Moreover, these existing MoE methods either assume continuous experts or require the gating function to be partially known, restricting applicability when subgroup membership is entirely latent and responses are binary.

The most directly related work is that of [13], who developed a penalised mixture of logistic regressions for predicting non-alcoholic steatohepatitis (NASH). While their applied framework demonstrates the practical promise of penalised mixture logistic models, it does not provide theoretical guarantees on sparsity recovery, estimation consistency, or oracle properties under high-dimensional asymptotics.

### 2.4 Gaps and Positioning

The literature thus reveals three key limitations that motivate our proposed framework:

i. **Outcome specificity:** Theoretical guarantees for penalised mixtures focus on Gaussian responses, leaving binary outcomes without comparable results. The non-convexity of the logistic likelihood and the identifiability challenges it introduces remain largely unaddressed.

ii. **Methodological integration:** No existing framework simultaneously addresses latent heterogeneity, high-dimensional sparsity, and binary likelihoods within a tractable estimation procedure accompanied by formal theoretical guarantees. Current methods either sacrifice heterogeneity for sparsity (logistic LASSO [6, 7]), sacrifice sparsity for heterogeneity (unpenalised mixture models), or lack rigorous asymptotic properties (applied frameworks such as [13]).

iii. **Computational scalability:** Available software implementations for mixture of GLMs [15] do not incorporate regularisation for high-dimensional logistic components with theoretical guarantees, limiting their practical utility in modern settings where both p and latent structure pose challenges.

Our proposed Mixture of Logistic Regression with LASSO Regularisation (MLR-LR) framework directly addresses these gaps. Unlike standard penalised logistic regression [6,14], MLR-LR accommodates latent subgroups; unlike penalised Gaussian mixtures [10,11], it targets binary outcomes with non-asymptotic theoretical guarantees; unlike [27] and [28], it focuses on sparsity recovery and oracle properties rather than feature selection or prediction sets; and unlike the applied framework of [13], it provides rigorous theoretical foundations including identifiability, consistency, sparsity recovery, and the oracle property under high-dimensional regimes. By explicitly handling the non-convexity and identifiability challenges inherent in binary mixture models, our approach advances the methodological frontier at the intersection of penalised regression, mixture modelling, and high-dimensional classification.

## 3. Methodology

### 3.1 Introduction

This chapter develops a Mixture of Logistic Regression with LASSO Regularisation (MLR-LR) framework for high-dimensional binary classification under latent heterogeneity. The proposed method simultaneously identifies latent subpopulations, selects relevant predictors, and ensures parsimonious inference. Parameter estimation proceeds via a penalised Expectation Maximisation (EM) algorithm with component-wise coordinate descent. We present model selection criteria, theoretical guarantees, simulation protocols, and real-data evaluation strategies. The framework balances model complexity with interpretability while explicitly

addressing sparsity and unobserved heterogeneity.

### 3.2 Model Formulation

#### 3.2.1 Mixture Logistic Regression

Let $Y_i \in \{0,1\}$ denote a binary response for observation *i*=1,…, n, and let $X_i = (X_{i1}, X_{i2}, ..., X_{ip})^T \in \mathbb{R}_p$ be a *p*-dimensional covariate vector. Standard logistic regression assumes population homogeneity, an assumption violated when latent subgroups exhibit distinct predictor outcome relationships. We therefore model the population as a mixture of *K* latent components, each governed by its own logistic regression model.

For component $k \in \{1,\ldots,K\}$, the conditional probability of success is:

$$p_k(\mathbf{X}_i) = \Pr(Y_i = 1 | \mathbf{X}_i, Z_i = k) = \frac{\exp(\mathbf{X}_i^{\cdot} \beta_k)}{1 + \exp(\mathbf{X}_i^{\cdot} \beta_k)}, \quad (3.1)$$

where $\beta_k \in \mathbb{R}_p$ is the component-specific coefficient vector and $Z_i \in \{1,\ldots,K\}$ is a latent class indicator. Conditional on component membership, the response follows a Bernoulli distribution with probability mass function:

$$f_k(y_i | \mathbf{X}_i; \beta_k) = \left(p_k(\mathbf{X}_i)\right)^{y_i} \left(1 - p_k(\mathbf{X}_i)\right)^{1-y_i}, \quad (3.2)$$

Let $\pi_k = \Pr(Z_i = k)$ denote the mixing proportions, satisfying $\pi_k > 0$ for all *k*=1,…,*K* and $\sum_{k=1}^{K} \pi_k = 1$. The marginal probability of success is:

$$f(y_i | \mathbf{X}_i; \Theta) = \sum_{k=1}^{K} \pi_k f_k(y_i | \mathbf{X}_i; \beta_k), \quad (3.3)$$

The parameter set is $\Theta = \{\pi_1,\ldots,\pi_K, \beta_1,\ldots,\beta_K\}$. Substituting (3.2) into (3.3) and expanding $p_k(\mathbf{X}_i)$ yields the explicit form:

$$f(y_i | \mathbf{X}_i; \Theta) = \sum_{k=1}^{K} \pi_k \left(\frac{\exp(\mathbf{X}_i^{\cdot} \beta_k)}{1 + \exp(\mathbf{X}_i^{\cdot} \beta_k)}\right)^{y_i} \left(1 - \frac{\exp(\mathbf{X}_i^{\cdot} \beta_k)}{1 + \exp(\mathbf{X}_i^{\cdot} \beta_k)}\right)^{1-y_i} \quad (3.4)$$

Equation (3.4) defines the Mixture of Logistic Regression model. For a dataset $D = \{(X_i, Y_i)\}_{i=1}^{n}$, the observed-data log-likelihood is:

$$\ell(\Theta) = \sum_{i=1}^{n} \log f(y_i | \mathbf{X}_i; \Theta) = \sum_{i=1}^{n} \log \left[ \sum_{k=1}^{K} \pi_k \left(\frac{e^{(\mathbf{X}_i^{\cdot} \beta_k)}}{1 + e^{(\mathbf{X}_i^{\cdot} \beta_k)}}\right)^{y_i} \left(1 - \frac{e^{(\mathbf{X}_i^{\cdot} \beta_k)}}{1 + e^{(\mathbf{X}_i^{\cdot} \beta_k)}}\right)^{1-y_i} \right] \quad (3.5)$$

#### 3.2.2 LASSO-Regularised Extension

In high-dimensional settings where *p* may exceed *n*, unpenalised estimation becomes unstable and yields poor generalisation. We incorporate an $\ell_1$ (LASSO) penalty to enforce sparsity within each component, shrinking irrelevant coefficients to zero. The penalised log-likelihood is:

$$\ell_p(\Theta) = \sum_{i=1}^{n} \log \left[ \sum_{k=1}^{K} \pi_k \left(\frac{e^{(\mathbf{X}_i^{\cdot} \beta_k)}}{1 + e^{(\mathbf{X}_i^{\cdot} \beta_k)}}\right)^{y_i} \left(1 - \frac{e^{(\mathbf{X}_i^{\cdot} \beta_k)}}{1 + e^{(\mathbf{X}_i^{\cdot} \beta_k)}}\right)^{1-y_i} \right] - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} |\beta_{kj}|, \quad (3.6)$$

where $\lambda \geq 0$ is a tuning parameter controlling the regularisation strength, $\lambda \sum_{k=1}^{K} \sum_{j=1}^{p} |\beta_{kj}|$ induces sparsity by driving coefficients of irrelevant predictors to exactly zero, producing sparse and

interpretable component-specific models. The choice of $\lambda$ balances model fit against complexity; larger values yield sparser solutions, while λ=0 reduces to the unpenalised mixture in (3.5). This formulation unifies two key methodological strands: mixture modelling for latent heterogeneity and LASSO regularisation for high-dimensional sparsity. By simultaneously addressing both challenges, the MLR-LR framework provides a flexible yet parsimonious approach to binary classification in complex, heterogeneous datasets.

A single global regularisation parameter λ is used across all *K* components. This choice enforces a uniform level of sparsity across subgroups, which is appropriate when the underlying sparsity structure is similar across components. While component-specific $\lambda_k$ could be considered (e.g., adaptive LASSO), this would introduce an additional *K*-dimensional tuning problem, complicating model selection and theoretical analysis.

### 3.3 Theoretical Properties of the Penalised Mixture of Logistic Regression

We establish theoretical guarantees for the proposed MLR–LR model, focusing on three fundamental aspects: identifiability, convergence of the penalised EM algorithm, and sparsity recovery. These properties ensure that the model is well-defined, that the estimation procedure is reliable, and that the resulting estimators possess desirable statistical behaviour in high-dimensional settings.

### 3.3.1 Identifiability

A model is identifiable if distinct parameter sets yield distinct distributions of the observed data, up to label switching in mixture models. The presence of clustering structure can be assessed using the Hopkins statistic [25]. A value exceeding 0.7 indicates strong clustering tendency, providing empirical support for the mixture modelling approach.

**Proposition 3.1 (Identifiability of MLR–LR)**

Let $\Theta = \{\pi_1,...,\pi_K,\beta_1,...,\beta_K\}$ denote the parameter set of a K-component mixture of logistic regression models. Assume the following conditions hold: Assume the following conditions hold:

(A1) Distinct components: $\beta_k \neq \beta_l \forall k \neq l$.

(A2) Full rank design: The design matrix $X \in \mathbb{R}^{n\times p}$ has full column rank.

(A3) Positive mixing: $\pi_k > 0$ for all *k*=1,…, *K*.

Then Θ is identifiable up to label switching. That is, if two parameter sets Θ and $\tilde{\Theta}$ induce the same mixture distribution

$$\sum_{k=1}^{K}\pi_k\,\sigma(\mathbf{X}^{\cdot}\,\beta_k)^y\left(1-\sigma(\mathbf{X}^{\cdot}\,\beta_k)\right)^{1-y}$$
$$=\sum_{k=1}^{K}\tilde{\pi}_k\,\sigma(\mathbf{X}^{\cdot}\,\beta_k)^y\left(1-\sigma(\mathbf{X}^{\cdot}\,\beta_k)\right)^{1-y}\quad \forall \mathbf{X}, y\in\{0,1\},$$

then there exists a permutation $\tau$ of {1,…,*K*} such that $\tilde{\pi}_k = \pi_{\tau(k)}$, and $\beta_k = \beta_{\tau(k)}$ for all *k*.

*Proof sketch.* The result follows from [17] and [18]. Logistic regression components are identifiable, and the distinctness condition ensures separation of component-specific regression functions. Label switching is the only source of non-identifiability.

### 3.3.2 Convergence of the Penalised EM Algorithm

The penalised EM algorithm introduced in Section 3.4 generates a sequence of parameter estimates $\{\Theta^{(t)}\}_{t\geq 0}$. We establish its monotonicity and convergence to the stationary points of the penalised log-likelihood.

**Theorem 3.1 (Monotone Ascent).** Let $\ell_p(\Theta)$ be the penalised log-likelihood

defined in (3.6). For any sequence $\{\Theta^{(t)}\}_{t \geq 0}$. generated by the penalised EM algorithm,

$\ell_p(\Theta^{(t+1)}) \geq \ell_p(\Theta^{(t)})$ for all $t \geq 0$. Thus, the sequence $\{\ell_p(\Theta^{(t)})\}$ is monotone non-decreasing.

*Proof.* The standard EM ascent property extends to the penalised setting because the penalty term $\lambda \sum_{j=1}^{p} \sum_{k=1}^{K} |\beta_{jk}|$ is concave in $\Theta$ (it is linear in the absolute values, which are convex; however, the negative penalty is concave). The Q-function satisfies $Q(\Theta^{(t+1)} | \Theta^{(t)}) \geq Q(\Theta^{(t)} | \Theta^{(t)})$ by construction of the M-step, and the penalised log-likelihood can be expressed as $\ell_p(\Theta) = Q(\Theta | \Theta^{(t)}) + H(\Theta | \Theta^{(t)})$ where H satisfies $H(\Theta | \Theta^{(t)}) \leq H(\Theta^{(t)} | \Theta^{(t)})$. Combining these inequalities yields the result.

**Proposition 3.2 (Convergence to Stationary Point).**
Assume the following regularity conditions:

(B1) $\ell_p(\Theta)$ is continuous and bounded above on the parameter space.

(B2) The parameter space $\Omega$ is compact or $\ell_p$ is coercive.

(B3) The Q-function $Q(\Theta | \Theta')$ in the EM algorithm is concave in $\Theta$ for each fixed $\Theta'$.

(B4) The LASSO penalty is convex.

Then every limit point of $\{\Theta^{(t)}\}$ is a stationary point of $\ell_p$. If $\ell_p$ has isolated stationary points, the sequence $\{\Theta^{(t)}\}$ converges to a stationary point.

*Proof.* The monotone ascent property (Theorem 3.1) together with boundedness (B2) ensures that $\{\ell_p(\Theta^{(t)})\}$ converges to a finite limit $\ell^*$. Continuity (B1) and compactness imply that the sequence $\{\Theta^{(t)}\}$ has at least one limit point. The concave-convex structure of the Q-function and penalty ensures that any such limit point satisfies the Karush Kuhn Tucker conditions for $\ell_p$, hence is stationary.

### 3.3.3 Consistency of the Penalised Maximum Likelihood Estimator

**Theorem 3.2 (Consistency)**

Let $\hat{\Theta}_n$ denote the penalised maximum likelihood estimator obtained by maximising (3.6), and let $\Theta_0$ denote the true parameter value. Assume:

(C1) Identifiability holds (Proposition 3.1).

(C2) The parameter space $\Omega$ is compact.

(C3) Mixing proportions: $\delta > 0$ such that $\delta \leq \pi_k \leq 1-\delta$ for all $k$.

(C4) The design matrix has full rank, and the Fisher information matrix is positive definite.

(C5) The penalty parameter satisfies $\lambda_n \to 0$ as $n \to \infty$.

Then $\hat{\Theta}_n \xrightarrow{p} \Theta_0$ as $n \to \infty$.

*Proof sketch.* The proof follows standard arguments for penalised M-estimators [19]. Compactness (C2) ensures the existence of the estimator. Identifiability (C1) guarantees that the limiting objective function is uniquely maximised at $\Theta_0$. Uniform convergence of the normalised penalised log-likelihood to its expectation follows from the uniform law of large numbers under (C4). The penalty term vanishes asymptotically by (C5), so it does not shift the limit. Consistency then follows from standard results on extremum estimators.

### 3.3.4 Sparsity Recovery

A key motivation for the LASSO penalty is its ability to perform automatic variable selection by shrinking irrelevant coefficients exactly to zero. We establish that this property

holds with probability tending to one in the MLR–LR framework.

**Theorem 3.3 (Support Recovery).**

For each component k=1, …, K, define the set of inactive predictors:

$$S_k = \{j \in \{1,\ldots,p\} : \beta_{kj}^0 = 0\},$$

where $\beta_{kj}^0 = 0$ denotes the $j$-th coefficient of the true parameter $\beta_k^0 = 0$. Assume conditions (C1)-(C5) of Theorem 3.2 hold, and additionally that the penalty parameter $\lambda_n$ satisfies:

$$\lambda_n \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \lambda_n = o(1).$$

Then the penalised MLE $\hat{\Theta}_n$ achieves consistent variable selection:

$$\Pr\left(\hat{\beta}_{kj} = 0 \ \forall j \in S_k\right) \to 1 \quad \text{as } n \to \infty, \quad \text{for each } k = 1,\ldots,K.$$

*Proof sketch.* The result follows from adapting the sparsity recovery arguments of Städler et al. (2010) to the logistic mixture setting. The $\lambda_n \sqrt{\frac{\log p}{n}}$ condition ensures that the penalty dominates the estimation error for truly zero coefficients, forcing them to be estimated as exactly zero. The $\lambda_n = o(1)$ condition ensures that the penalty does not overshrink non-zero coefficients. Component-wise application of these arguments, combined with the posterior weights $\tau_{ik}$ concentrating on the correct components, yields the result.

### 3.3.5 Unified Recovery of Latent Structure and Sparsity

We now present a stronger result: under suitable conditions, the MLR–LR estimator simultaneously recovers the true number of components, the correct sparsity pattern within each component, and achieves oracle efficiency for the non-zero coefficients.

**Theorem 3.4 (Simultaneous Recovery).** Let $\{(X_i, Y_i)\}_{i=1}^n$ be generated from a $K_0$ component mixture of logistic regression models with true parameters $\Theta^0 = \{\pi_k^0, \beta_k^0\}_{k=1}^{K_0}$, where each $\beta_k^0$ is sparse with support $S_k^0 = \{j : \beta_{kj}^0 \neq 0\}$. Assume:

(D1) Identifiability: Conditions (A1)-(A3) of Proposition 3.1 hold for K=K$_0$.

(D2) High-dimensional regime: $p \geq n^\gamma$ for some $\gamma > 0$.

(D3) Component separation: $\min_{k \neq l} \left\|\beta_k^0 - \beta_l^0\right\|_2 \geq \delta_n$ with $\delta_n \to 0$ at rate $\delta_n \sqrt{\frac{\log p}{n}}$.

(D4) Restricted eigenvalue condition holds uniformly for component-weighted design matrices.

(D5) Signal strength: $\min_{(k,j) \in S_k^0} \left|\beta_{kj}^0\right| \geq \tau_n$ with $\tau_n \sqrt{\frac{\log p}{n}}$.

Let $\hat{\Theta}_n$ be the penalised MLE obtained from (3.6). Then as n→∞

i. **Mixture order consistency:** $\Pr(\hat{K} = K_0) \to 1$, where $\hat{K}$ is selected by the BIC criterion in Section 3.5.
ii. **Exact support recovery:** $\Pr\left(\text{supp}(\beta_k) = S_k^0 \ \forall k = 1,\ldots,K_0\right) \to 1$
iii. **Oracle efficiency:** For each component $k$, $\sqrt{n}\left(\beta_{k,S_k^0} - \beta_{k,S_k^0}^0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{I}_k(\beta_k^0)^{-1}\right)$,

where $\beta_{k,S_k^0}$ denotes the subvector of $\beta_k$ restricted to the true support $S_k^0$, and $\mathbf{I}_k\beta_k^0$ is the

Fisher information matrix for component *k* conditional on correct class membership.

*Proof.* The proof combines results from [9] for penalised mixture models with high-dimensional oracle properties established in [10]. Component separation (D3) ensures that the posterior probabilities $\tau_i$ concentrate on the true component memberships. The restricted eigenvalue condition (D4) guarantees that the penalised estimators within each component satisfy oracle inequalities. Signal strength (D5) prevents true signals from being shrunk to zero. Detailed technical arguments are provided in the supplementary material

### 3.4 Parameter Estimation

#### 3.4.1 Penalised Expectation Maximisation (EM) Algorithm

Estimating the parameters of the MLR-LR model presents two principal challenges: the latent component memberships *Zi* are unobserved, and the LASSO penalty introduces non-differentiability. We address both through a penalised EM algorithm [16], which iteratively constructs and maximises a surrogate complete data penalised log-likelihood. Within each M-step, component-specific coefficients are updated using coordinate descent, exploiting the sparsity induced by the $\ell_1$ penalty.

#### 3.4.2 Complete-Data Penalised Log-Likelihood

Let $Z_{ik} = \mathbb{1}(Z_i = k)$ indicate whether observation *i* belongs to component *k*. If the latent indicators were observed, the complete-data penalised log-likelihood would be:

$$\ell_c(\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left[\log \pi_k + \log f_k(y_i \mid \mathbf{X}_i;\beta_k)\right] - \lambda \sum_{k=1}^{K}\sum_{j=1}^{p} |\beta_{kj}| \quad (3.11)$$

Substituting the component density from (3.2) yields:

$$\ell_c(\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left[\log \pi_k + y_i \log \sigma(\mathbf{X}_i^{\cdot}\beta_k) + (1-y_i)\log\left(1-\sigma(\mathbf{X}_i^{\cdot}\beta_k)\right)\right] - \lambda \sum_{k=1}^{K}\sum_{j=1}^{p} |\beta_{kj}| \quad (3.12).$$

Using the identity
$\log \sigma(\mathbf{X}_i^{\cdot}\beta_k) = \mathbf{X}_i^{\cdot}\beta_k - \log\left(1+\sigma(\mathbf{X}_i^{\cdot}\beta_k)\right)$ and noting that for $y_i \in \{0,1\}$,

$$y_i \log \sigma(\mathbf{X}_i^{\cdot}\beta) + (1-y_i)\log(1-\sigma(\mathbf{X}_i^{\cdot}\beta)) = y_i(\mathbf{X}_i^{\cdot}\beta) - \log(1+\exp(\mathbf{X}_i^{\cdot}\beta)),$$

we obtain the computationally convenient form:

$$\ell_c(\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left[\log \pi_k + y_i(\mathbf{X}_i^{\cdot}\beta_k) - \log\left(1+\exp(\mathbf{X}_i^{\cdot}\beta_k)\right)\right] - \lambda \sum_{k=1}^{K}\sum_{j=1}^{p} |\beta_{kj}| \quad (3.13)$$

#### 3.4.3 Penalised EM Algorithm

The penalised EM algorithm iteratively maximises the expected value of (3.13) conditional on the observed data and current parameter estimates.

**E-step: Expectation**

Given current parameters $\Theta^{(t)}$, compute the posterior probability that observation *i* belongs to component *k*:

$$\tau_{ik}^{(t)} = \mathrm{E}[Z_{ik} \mid y_i, \mathbf{X}_i, \Theta^{(t)}] = \frac{\pi_k^{(t)} f_k(y_i \mid \mathbf{X}_i;\beta_k^{(t)})}{\sum_{\ell=1}^{K} \pi_{\ell}^{(t)} f_\ell(y_i \mid \mathbf{X}_i;\beta_\ell^{(t)})}. \quad (3.14)$$

Expanding the component densities using (3.2):

$$\tau_{ik}^{(t)} = \frac{\pi_k^{(t)} \sigma(\mathbf{X}_i^{\cdot}\beta_k^{(t)})^{y_i}\left(1-\sigma(\mathbf{X}_i^{\cdot}\beta_k^{(t)})\right)^{1-y_i}}{\sum_{\ell=1}^{K} \pi_{\ell}^{(t)} \sigma(\mathbf{X}_i^{\cdot}\beta_\ell^{(t)})^{y_i}\left(1-\sigma(\mathbf{X}_i^{\cdot}\beta_\ell^{(t)})\right)^{1-y_i}}. \quad (3.15)$$

These responsibilities satisfy $\sum_{k=1}^{K} \pi_k = 1$ for each $i$ and represent a soft partitioning of the data across components.

**M-step: Maximisation**

Using the responsibilities $\tau_{ik}^{(t)}$ in place of the unobserved $Z_{ik}$, the expected complete-data penalised log-likelihood is:

$$Q(\Theta|\Theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(t)} \left[\log \pi_k + y_i(\mathbf{X}_i^{\cdot} \beta_k) - \log\left(1+\exp(\mathbf{X}_i^{\cdot} \beta_k)\right)\right] - \lambda \sum_{k=1}^{K}\sum_{j=1}^{p} |\beta_{kj}| \quad (3.16)$$

The M-step maximizes $Q(\Theta|\Theta^{(t)})$ with respect to $\Theta$. Due to the additive structure, the mixing proportions and component coefficients can be updated separately.

**Update for mixing proportions**: For fixed $\{\beta_k\}$, the terms involving $\pi_k$ separate from the rest. Maximizing subject to $\sum_{k=1}^{K} \pi_k = 1$ yields:

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ik}^{(t)} \quad (3.17)$$

**Update for component coefficients:** For each component $k$, the terms involving $\beta_k$ are:

$$Q_k(\beta_k|\Theta^{(t)}) = \sum_{i=1}^{n} \tau_{ik}^{(t)} \left[y_i(\mathbf{X}_i^{\cdot} \beta_k) - \log\left(1+\exp(\mathbf{X}_i^{\cdot} \beta_k)\right)\right] - \lambda \sum_{j=1}^{p} |\beta_{kj}| \quad (3.18)$$

The optimisation in (3.8) has no closed-form solution due to the non-differentiable LASSO penalty. We employ coordinate descent, as described in Section 3.4.2.

The component-specific update is therefore:

$$\beta_k^{(t+1)} = \underset{\beta_k}{\operatorname{argmax}}\, Q_k(\beta_k|\Theta^{(t)}) \quad (3.19)$$

This is a weighted penalised logistic regression problem with weights $\tau_{ik}^{(t)}$. No closed-form solution exists due to the non-differentiable LASSO penalty; we solve it using coordinate descent.

### 3.4.4 Coordinate Descent for the Penalised M-step

For each component $k$, we solve (3.19) using cyclic coordinate descent [7]. Let $\tilde{y}_i = y_i$ and $\tilde{w}_i = \tau_{ik}^{(t)}$ denote the working responses and weights. At iteration $s$ of the coordinate descent loop for component $k$, we cycle through each predictor $j$=1,…,$p_j$ =1,…,$p$:

i. Compute the partial residual excluding predictor $j$:

$$r_{ij}^{(s)} = \tilde{y}_i - \sigma\left(\sum_{\ell \neq j} X_{i\ell}\beta_{k\ell}^{(s)} + X_{ij}\beta_{kj}^{(s-1)}\right).$$

ii. Solve the one-dimensional penalised weighted least squares problem:

$$\beta_{kj}^{(s)} = \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \tilde{w}_i \left( r_{ij}^{(s)} \beta X_{ij} - \log(1+e^{\beta X_{ij}}) \right) - \lambda |\beta| \right\}.$$

This step does not admit a closed form; we use a quadratic approximation followed by soft-thresholding, as implemented in standard coordinate descent algorithms for logistic regression (Friedman et al., 2010).

iii. Update $\beta_{kj}^{(s)}$ and proceed to the next coordinate. The coordinate descent loop terminates when the maximum absolute change in coefficients across a complete cycle is less than $\varepsilon_{\text{CD}} = 10^{-4}$.

### 3.4.5 Algorithm Summary and Convergence Control

Algorithm 1: Penalised EM for MLR–LR

**Input**: Data $\{(X_i, Y_i)\}_{i=1}^{n}$, candidate $K$, penalty $\lambda$, tolerance $\varepsilon_{\text{EM}} = 10^{-6}$, max iterations $T_{\max} = 500$, number of random initialisations $M$=10.

**Output:** Parameter estimates $\hat{\Theta}$.

i. Initialisation: For each $M$ initialisation, generate initial component assignments through k-means clustering on $X$ or random partitioning. Set $\pi_k^{(0)}$=1/K, Initialise $\beta_k^{(0)}$ by fitting unpenalised logistic regression on the initial partition.

ii. EM Iterations: For $t$ = 0,1,2,… until convergence:
   a. E-step: Compute $\tau_{ik}^{(t)}$ via (3.15).
   b. M-step:
      i. Update $\pi_k^{(t+1)}$ via (3.17).
      ii. For each $k$, solve (3.19) using coordinate descent (Section 3.4.3) with convergence tolerance $\varepsilon_{\text{CD}} = 10^{-4}$.
   c. Check convergence: if $\frac{|\ell_p(\Theta^{(t+1)}) - \ell_p(\Theta^{(t)})|}{|\ell_p(\Theta^{(t)})| + 0.1} < \varepsilon_{\text{EM}}$, stop.

iii. Selection: Retain the run with the highest penalised log-likelihood across initialisations.

### 3.4.6 Computational Complexity

Each EM iteration requires:

**E-step:** O(Kpn) operations to evaluate the logistic functions and compute responsibilities.

**M-step:** For each component, coordinate descent requires $O(pnT_{CD})$ operations, where $T_{CD}$ is the number of coordinate descent cycles (typically 5-10 for well-conditioned problems). Total M-step cost is $O(KpnT_{CD})$.

With $I_{EM}$ EM iterations, overall complexity is $O(KpnI_{EM}T_{CD})$. In practice $I_{EM}$ is often 20-50 for moderate $K$, making the algorithm feasible for $p$ in the hundreds to low thousands.

### 3.4.7 Multiple Initialisations and Local Maxima

The penalised log-likelihood (3.6) is non-concave, and the EM algorithm may converge to local stationary points. To mitigate this, we employ M=10 random initialisations, combining Clustering-based starts (K-means on $X$ to obtain initial partitions) and Random starts (Random soft assignments drawn from a Dirichlet distribution).

The solution with the highest penalised log-likelihood is retained. This multi-start strategy, while computationally more expensive, substantially increases the probability of locating a near-global optimum.

### 3.4.8 Choice of Mixing Proportions: Constant vs. Covariate-Dependent

In finite mixture models, the mixing proportions $\pi_k$ can be specified either as constants or as functions of covariates. Constant proportions (as in [1, 2]) assume that the probability of belonging to a latent subgroup does not vary with the predictors. Covariate-dependent proportions (as in mixture-of-experts models [24]) model $\pi_k(x)$ via a softmax function of $x$, allowing subgroup membership to be predicted by covariates.

The proposed MLR-LR framework adopts constant mixing proportions for the following reasons:

i. High-dimensional parsimony: Covariate-dependent proportions would require estimating an additional (K-1) × p parameters for the gating network. In high-dimensional settings (p > n), this exacerbates overfitting and complicates sparsity induction, as penalties would need to be applied to both expert and gating parameters.

ii. Theoretical tractability: Constant proportions simplify the identifiability conditions (Proposition 3.1) and the oracle property proofs (Theorem 3.4). Extending these results to covariate-dependent proportions would require additional assumptions on the gating network and the separation of expert and gating parameters.
iii. Focus on subgroup-specific effects: Our primary scientific interest lies in estimating component-specific regression coefficients $\beta_k$, which capture how predictors influence the response differently across latent subgroups. The mixing proportions serve as secondary parameters that weight these components.
iv. Empirical adequacy: In our simulation regimes and the WDBC application, constant proportions produced excellent predictive performance and interpretable component structures, suggesting that covariate-dependent proportions may not be necessary for these settings.

## 3.5 Model Selection

Selecting the number of components $K$ and the penalty parameter λ is fundamental to balancing latent-structure discovery, sparsity, and predictive accuracy. Because $K$ directly controls model dimensionality, selection must jointly account for mixture complexity and regularisation strength. We adopt a BIC-based selection framework, justified by its theoretical consistency for mixture models and penalised likelihoods, and implement it through a two-dimensional grid search. The regularisation parameter λ is selected by minimising the Bayesian Information Criterion (BIC), rather than cross-validation. We adopt BIC for three reasons: (i) theoretical consistency: BIC is known to consistently select the true sparsity pattern and mixture order under the conditions of Theorem 3.4 [23, 20]; (ii) computational efficiency: cross-validation would require refitting the penalised EM algorithm multiple times across λ folds, which is prohibitive given the grid search over K and λ; (iii) BIC's penalty on model complexity naturally balances the LASSO's degrees of freedom against goodness-of-fit without requiring a separate validation set, which is advantageous when sample size is limited.

### 3.5.1 Bayesian Information Criterion for Penalised Mixtures

For a penalised mixture model, the Bayesian Information Criterion (BIC) is defined as

$$\mathrm{BIC}(K,\lambda) = -2\ell_p(\hat{\Theta}_{K,\lambda}) + \mathrm{df}(K,\lambda)\cdot\log n \tag{3.20}$$

where $\hat{\Theta}_{K,\lambda}$ is the penalised MLE for given $(K,\lambda)$ and $\mathrm{df}(K,\lambda)$ represents the degrees of freedom. Following [20], the degrees of freedom for a penalised mixture model count each non-zero parameter

$$\mathrm{df}(K,\lambda) = K + \sum_{k=1}^{K}\left|\{j : \hat{\beta}_{kj} \neq 0\}\right| \tag{3.21}$$

This formulation counts each component's intercept and non-zero slopes as free parameters. Mixing proportions are not penalised and thus do not contribute to the degrees of freedom.

### 3.5.2 Grid Search Implementation

We perform a two-dimensional grid search over the number of components $K$ and the LASSO penalty parameter λ.

**a) Mixture order grid**

Let $k=\{1,2,\ldots,K_{\max}\}$ where

$$K_{\max} = \min\left(8,\sqrt{n}\right) \tag{3.22}$$

This heuristic balances several considerations which includes:

**Sufficient observations per component:** The bound $K_{\max} = \min\left(8, \sqrt{n}\right)$ ensures that the average number of observations per component is at least $\sqrt{n}$ . While not a theoretical guarantee, this heuristic (standard in applied mixture modelling) provides a practical safeguard against components estimated from too few observations, which can lead to unstable parameter estimates and spurious cluster detection.

**Computational tractability:** An absolute upper bound of 8 prevents excessive grid search evaluations and overfitting in moderate sample sizes.

**Empirical practice:** Applied mixture modelling typically restricts *K* to a small, plausible set [21, 1].

**b) Penalty parameter grid**

For each candidate *K*, we consider a logarithmically spaced grid of $\lambda$ values

$$\mathrm{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}, \quad \lambda_m = \lambda_{\min}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{(m-1)/(M-1)}, \quad m = 1, \ldots, M \tag{3.23}$$

with *M*=50 grid points. The logarithmic spacing ensures finer resolution at smaller $\lambda$ values, where sparsity patterns change most rapidly.

The maximum penalty $\lambda_{\max}$ is the smallest value for which all coefficients are shrunk to zero for a given component. Following standard practice in penalised regression [22], we estimate

$$\lambda_{\max} = \max_j \left| \frac{1}{n} \sum_{i=1}^{n} \tau_{ik} x_{ij} (y_i - \bar{y}) \right| \tag{3.24}$$

where $\tau_{ik}$ are the posterior responsibilities from a preliminary EM run (or uniform weights in the homogeneous case). This quantity approximates the threshold at which the penalty dominates the score function, forcing all coefficients to zero. We the set

$$\lambda_{\min} = 10^{-3} \cdot \lambda_{\max} \tag{3.25}$$

allowing near-unpenalised estimation when the data support complex component structures.

### 3.5.3 Model Selection Procedure

For each pair $(K, \lambda) \in k \times \mathrm{L}$ ,

i. **Run Algorithm 1** (penalised EM) with multiple random initialisations (Section 3.4.6).
ii. **Retain the solution** $\hat{\Theta}_{K,\lambda}$ achieving the highest penalised log-likelihood across initialisations.
iii. **Compute** BIC(K,λ) using (3.20) and (3.21).

The optimal model is selected as

$$(\hat{K}, \hat{\lambda}) = \underset{K \in \mathrm{K}, \lambda \in \mathrm{L}}{\operatorname{argmin}} \mathrm{BIC}(K, \lambda) \tag{3.26}$$

This two-dimensional minimisation jointly determines both the number of latent components and the regularisation strength, directly accounting for their interplay.

### 3.5.4 Theoretical Justification for the $\lambda$ Grid

**Proposition 3.3 (Upper Bound for $\lambda$)** Let $\hat{\beta}_k(\lambda)$ denote the penalised MLE for component *k*. If $\lambda \geq \lambda_{\max}$ as defined in (3.24), then $\hat{\beta}_k(\lambda) = 0$ for all k.

*Proof sketch.* The Karush-Kuhn-Tucker conditions for the penalised logistic regression problem require that for each j, the subgradient equation includes the term λ·sign $(\beta_{kj})$. When λ exceeds the maximum absolute value of the score function at $\beta_{kj}=0$, the only feasible solution is $\beta_{kj}=0$. The expression in (3.24) directly computes this maximum score. The

logarithmic grid in (3.23) ensures that we explore the entire regularisation path from complete sparsity ($\lambda_{\max}$) to near-unpenalised estimation ($\lambda_{\min}$), with sufficient resolution in regions where sparsity patterns change.

### 3.5.5 Practical Considerations

The penalised log-likelihood is non-concave, and different initialisations may converge to distinct local optima. For each grid point, we employ M=10 random starts (Section 3.4.6) and retain the best solution. This substantially increases the probability of locating a near-global optimum. The total number of model fits is |K|×|L|×M. With |K|≤8, |L|=50, and M=10, this yields up to 4,000 EM runs. While computationally intensive, the procedure is parallel and feasible on modern computing clusters. Under the regularity conditions of Theorem 3.4, BIC is consistent for selecting both the mixture order and the sparsity pattern [23, 20]. That is, as n→∞, the probability that BIC selects the true ($K_0$, $\lambda_0$) approaches one.

## 3.6 Simulation Studies

### 3.6.1 Objectives

The simulation study assesses the finite-sample performance of the proposed MLR–LR model under controlled data-generating mechanisms that reflect varying degrees of latent heterogeneity, sparsity, and predictor dependence. The primary objectives are:

i. To evaluate whether explicitly modelling unobserved population structure yields improved predictive and inferential performance relative to standard LASSO-regularised logistic regression.
ii. To assess the method's robustness under model misspecification, including scenarios where latent structure is absent or weak.
iii. To verify that the theoretical properties established in Section 3.3 manifest in finite samples.

The simulation design mirrors the theoretical assumptions developed earlier, ensuring that empirical results complement and validate the theoretical framework.

### 3.6.2 Data-Generating Mechanisms

Synthetic datasets are generated from finite mixtures of logistic regression models according to the following procedure:

i. Latent class assignment: For each observation $i=1,\ldots,n$, draw a latent class indicator $Z_i \in \{1,\ldots,K_0\}$ from a multinomial distribution with mixing proportions $\pi_1^0,\ldots,\pi_{K_0}^0$.
ii. Covariate generation: Depending on the simulation regime, predictor vectors $X_i \in \mathbb{R}^p$ are generated either from a single multivariate Gaussian distribution: $X_i \sim N(0,\Sigma)$ or a mixture of Gaussian distributions: $X_i \sim N(\mu Z_i, \Sigma Z_i)$ to induce component-specific covariate distributions.
iii. Response generation: Conditional on $Z_i = k$, generate the binary response

$$Y_i \sim \text{Bernoulli}\left(\sigma(\mathbf{X}_i^{\cdot} \beta_k^0)\right),$$

where $\beta_k^0$ is the true coefficient vector for component $k$.

iv. Sparsity structure: Each $\beta_k^0$ has exactly $s_0$=25 non-zero coefficients, with indices randomly selected. Non-zero coefficients are drawn uniformly from $\{\pm\beta_{\min}, \pm\beta_{\max}\}$ to control signal strength.
v. Dimensionality: We set sample size n=200 and predictor dimension p=250,

yielding a high-dimensional regime with $p>n$.

### 3.6.3 Simulation Regime

We consider six simulation regimes designed to test the method under diverse conditions:

| Regime | $K_0$ | $\pi_k^0$ | $X$ distribution | Signal strength |
|---|---|---|---|---|
| R1: Homogeneous | 1 | $\pi_1$=1 | $N(0,I)$ | Strong |
| R2: Heterogeneous effects | 2 | (0.5, 0.5) | $N(0,I)$ | Strong |
| R3: Joint heterogeneity | 2 | (0.5, 0.5) | $N(\mu_k,I)$ with $\mu_1=(1,1,0,\ldots)$, $\mu_2=(-1,-1,0,\ldots)$ | Strong |
| R4: Overlapping components | 2 | (0.5, 0.5) | $N(0,I)$ | Moderate |
| R5: Correlated predictors | 2 | (0.5, 0.5) | AR(1) correlation, $\rho=0.5$ | Strong |
| R6: Weak signal | 2 | (0.5, 0.5) | N(0,I)N(0,I) | Weak |

These regimes span favourable conditions (R2, R3), challenging scenarios (R4, R5), and misspecification (R1, R6).

### 3.6.4 Replication and Evaluation Protocol

Each regime is independently replicated R = 200 times. Within each replication:

i**. Data splitting:** Data are randomly partitioned into training (70%) and test (30%) sets, stratified by the outcome variable to preserve class proportions. For simulation studies, training (n = 200), validation (n = 100), and test (n = 100) sets are generated from the same data-generating mechanism.

ii. **Standardisation:** Predictor means and standard deviations are computed using the training data only. These parameters are then applied to standardise training, validation, and test sets to prevent information leakage.

iii. **Feature reduction (high-dimensional settings):** When the number of predictors exceeds 100, LASSO pre-filtering is applied on the training data to select the top 50 most informative features. This step reduces dimensionality while preserving predictive signals.

iv. **Model fitting:** MLR-LR is fitted over the grid K×L defined in Section 3.5.2, using Algorithm 1 with 10 random initialisations per grid point.

v. **Model selection:** The optimal (K, λ) is selected via BIC (Section 3.5.1) evaluated on training data.

vi. **Benchmark fitting:** Competing methods (Section 3.6.7) are fitted with hyperparameters tuned via 5-fold cross-validation on training data.

vii. **Evaluation:** All performance metrics (Section 3.6.6) are computed on the held-out test set.

viii. **Seed control:** Random seeds are fixed per replication (2024 + r for replication r) to ensure reproducibility.

Results are averaged across replications, with Monte Carlo standard errors reported.

### 3.6.6 Algorithmic Details and Convergence Control

To ensure numerical stability and comparability across regimes, the algorithm terminates when $\frac{|\ell_p(\Theta^{(t+1)})-\ell_p(\Theta^{(t)})|}{|\ell_p(\Theta^{(t)})|+0.1}<10^{-6}$, or after $T_{max}$ =500 iterations. Within each M-step, coordinate descent stops when the maximum absolute change in coefficients across a complete cycle is less than $10^{-4}$, and for each grid point, we use 10 random starts.

These settings are held fixed across all simulation regimes and competing methods to ensure fair comparison.

### 3.6.6 Performance Metrics

Performance is evaluated using the metrics defined in Section 3.7.2:

Area Under the ROC Curve (AUC) (threshold-free measure of discriminative ability),

Misclassification Rate at 0.5 threshold, Recall, Precision, F1 Score and Log-loss.

### 3.6.7 Competing Methods

We benchmark MLR-LR against five competing methods (detailed in Section 3.7.1):

| Method | Description | Tuning |
|---|---|---|
| LASSO-LR | Homogeneous logistic LASSO | $\lambda$ via 5-fold CV |
| Decision Tree | CART | cp via 5-fold CV |
| Random Forest | 500 trees | mtry via 5-fold CV |
| Gradient Boosting | XGBoost | learning rate, max depth via 5-fold CV |

All methods receive the same standardised training data and are evaluated on identical test sets.

### 3.6.8 Reproducibility

To ensure full reproducibility, we fixed random seed (set.seed(2024 + r)) per replication , full simulation code will be provided as supplementary material, all preprocessing, tuning, and model fitting confined to training data; test data used once and Simulations run on a high-performance cluster (specifications in Appendix)

This protocol strictly prevents information leakage and ensures that reported results reflect genuine generalisation performance.

## 4. RESULTS AND DISCUSSIONS

### 4.1 Introduction

This chapter presents a comprehensive empirical evaluation of the proposed Mixture of Logistic Regression with LASSO Regularisation (MLR-LR) model. The evaluation is designed to rigorously validate each theoretical claim established in Section 3.3:

Predictive performance: Does modelling latent heterogeneity improve classification accuracy?

Mixture order consistency (Theorem 3.4(i)): Does BIC correctly identify the true number of components?

Sparsity recovery (Theorem 3.3): Does the LASSO penalty successfully distinguish relevant from irrelevant predictors?

Oracle efficiency (Theorem 3.4(iii)): Do the non-zero coefficient estimators achieve asymptotic normality?

We begin with controlled simulation studies across the six regimes defined in Section 3.6.3, followed by applications to real-world datasets. Performance is assessed using the full suite of metrics specified in Section 3.7.2: AUC, misclassification rate, recall, precision, F1 score, and log-loss. The proposed MLR-LR model is compared against all five benchmark methods listed in Section 3.7.1: homogeneous LASSO-LR, unpenalised mixture logistic regression (MLR), decision trees (CART), random forests (RF), and gradient boosting machines (GBM).

All results are aggregated over R = 200 independent replications, with Monte Carlo standard deviations reported in parentheses. The evaluation protocol strictly follows the reproducibility guidelines of Section 3.6.8: all preprocessing (standardisation) is performed on training data only, and all metrics are computed on held-out test sets. Random seeds are fixed per replication to ensure exact reproducibility.

### 4.2 Simulation Results

#### 4.2.1 Predictive Performance under Joint Heterogeneity (R3)

We first examine the joint heterogeneity regime (R3), where both predictor distributions and regression coefficients vary across two latent components. This scenario most closely mirrors the modelling assumptions underlying MLR-LR and provides a baseline for assessing its advantages relative to all benchmark methods. Table 4.1 presents the complete results for this regime.

Table 4.1: Predictive Performance in the Joint Heterogeneity Regime (R3): All Benchmark Methods.

| Model | AUC | Misclassification | Recall | Precision | F1 | Log-Loss |
|---|---|---|---|---|---|---|
| LASSO-LR | 0.589 (0.094) | 0.428 (0.076) | 0.581 (0.112) | 0.564 (0.098) | 0.572 (0.104) | 0.679 (0.051) |
| MLR | 0.612 (0.088) | 0.411 (0.082) | 0.603 (0.108) | 0.589 (0.094) | 0.596 (0.099) | 0.668 (0.058) |
| Decision Tree | 0.574 (0.102) | 0.445 (0.091) | 0.562 (0.121) | 0.541 (0.112) | 0.551 (0.115) | 0.698 (0.072) |
| Random Forest | 0.628 (0.081) | 0.398 (0.079) | 0.621 (0.098) | 0.608 (0.088) | 0.614 (0.092) | 0.665 (0.055) |
| GBM | 0.631 (0.079) | 0.395 (0.077) | 0.624 (0.095) | 0.611 (0.086) | 0.617 (0.089) | 0.663 (0.053) |
| **MLR–LR** | **0.635 (0.077)** | **0.393 (0.084)** | **0.628 (0.093)** | **0.615 (0.085)** | **0.621 (0.088)** | 0.671 (0.060) |

Note: Values show mean (standard deviation) across 200 replications. Bold indicates best performance in each column. All models tuned via 5-fold CV (benchmarks) or BIC (MLR-LR, MLR).

Table 4.1 reveals several important patterns. First, MLR-LR achieves the highest discriminative ability, with AUC = 0.635, outperforming the next best method (GBM, 0.631) by a small but consistent margin.

Second, ensemble methods (RF, GBM) are strong competitors, reflecting their flexibility in capturing complex patterns. However, they do not provide interpretable component structures or variable selection.

Third, unpenalised MLR underperforms relative to MLR-LR (AUC 0.612 vs. 0.635), confirming the value of LASSO regularisation in high-dimensional settings (p = 250 > n = 200).

Fourth, decision trees perform worst, as expected given the high-dimensional linear structure of the data-generating mechanism.

Fifth, log-loss remains competitive across all methods, with MLR–LR, RF, and GBM all within 0.01 of each other.

These results provide strong evidence that explicitly modelling latent heterogeneity, combined with sparsity-inducing regularisation, yields predictive advantages even relative to powerful black-box ensemble methods.

#### 4.2.2 Performance Across All Simulation Regimes

Table 4.2 presents comprehensive results across all six simulation regimes for all six performance metrics. Due to space constraints, we show AUC, F1 score, and log-loss as representative

**Table 4.2: Predictive Performance Across All Simulation Regimes: Selected Metrics**

| Regime | Model | AUC | F1 | Log-Loss |
|---|---|---|---|---|
| **R1: Homogeneous sparsity** | LASSO–LR | 0.823 (0.042) | 0.782 (0.051) | 0.525 (0.038) |
| | MLR-LR | 0.814 (0.051) | 0.771 (0.059) | 0.524 (0.042) |
| **R2: Coefficient heterogeneity** | LASSO-LR | 0.596 (0.088) | 0.548 (0.096) | 0.686 (0.055) |
| | **MLR-LR** | **0.624 (0.081)** | **0.579 (0.088)** | 0.687 (0.058) |
| **R3: Joint heterogeneity** | LASSO–LR | 0.589 (0.094) | 0.572 (0.104) | 0.672 (0.051) |
| | **MLR–LR** | **0.635 (0.077)** | **0.621 (0.088)** | 0.671 (0.060) |
| **R4: Overlapping mixtures** | LASSO–LR | 0.815 (0.038) | 0.774 (0.048) | 0.522 (0.035) |
| | **MLR–LR** | 0.809 (0.044) | 0.766 (0.054) | 0.535 (0.041) |
| **R5: Correlated predictors** | LASSO-LR | 0.620 (0.082) | 0.578 (0.089) | 0.656 (0.059) |
| | **MLR-LR** | **0.651 (0.076)** | **0.611 (0.083)** | **0.648 (0.062)** |
| **R6: Weak signals** | LASSO-LR | 0.523 (0.112) | 0.491 (0.124) | 0.708 (0.067) |
| | **MLR-LR** | **0.577 (0.098)** | **0.548 (0.109)** | 0.702 (0.071) |

*Note: Bold indicates best performance in each regime.*

In regimes with true heterogeneity (R2, R3, R5, R6), MLR-LR consistently achieves the highest or tied-highest AUC and F1 scores. The improvements over LASSO–LR range from 0.028 (R2) to 0.054 (R6), confirming the value of modelling latent structure.

Regimes without heterogeneity (R1, R4): MLR-LR performs comparably to LASSO-LR and ensemble methods, demonstrating that it does not overfit when no mixture structure exists. In R1, MLR-LR selects K = 1 in 94% of replications (see Table 4.3), effectively collapsing to the homogeneous model.

Comparison with ensembles: Random Forest and GBM are strong performers across all regimes, often matching or approaching MLR–LR's predictive accuracy. However, they do not provide interpretable component-specific coefficients, explicit variable selection, statistical inference (standard errors, confidence intervals) and theoretical guarantees (identifiability, consistency, oracle properties).

MLR-LR thus offers a unique combination of predictive performance and statistical interpretability.

Unpenalised MLR consistently underperforms MLR-LR in high-dimensional regimes (R2-R6), highlighting the critical role of LASSO regularisation for stability and variable selection. The slight underperformance of MLR-LR in R1 (homogeneous sparsity) is expected and desirable: when no latent heterogeneity exists, the mixture model incurs a small penalty for estimating extra parameters. BIC correctly selects K=1 in 94% of replications (Table 4.3), demonstrating that MLR-LR adapts gracefully to the absence of heterogeneity without overfitting.

### 4.2.3 Recovery of Latent Structure and Sparsity

Table 4.3 reports the model's ability to recover both the true number of components and the

correct sparsity pattern, directly validating the theoretical claims of Section 3.3.

**Table 4.3: Recovery of Latent Structure and Sparsity Patterns**

| Regime | $P(\hat{I} = K_0)$ | TNR (%) | FPR (%) | CI Coverage (%) |
|---|---|---|---|---|
| R1: Homogeneous sparsity | 0.94 | 91.2 (3.8) | 8.8 | — |
| R2: Coefficient heterogeneity | 0.88 | 87.6 (4.2) | 5.2 | 93.4 |
| R3: Joint heterogeneity | 0.92 | 89.3 (3.9) | 4.7 | 94.1 |
| R4: Overlapping mixtures | 0.56 | 78.4 (6.1) | 8.1 | 88.2 |
| R5: Correlated predictors | 0.84 | 85.9 (4.8) | 6.3 | 91.8 |
| R6: Weak signals | 0.79 | 82.1 (5.4) | 7.5 | 89.5 |

*Note: TNR = true negative rate (% of truly zero coefficients correctly estimated as zero). FPR = false positive rate (% of truly non-zero coefficients incorrectly shrunk to zero). CI coverage = coverage probability of 95% confidence intervals for non-zero coefficients. Standard deviations for TNR in parentheses.*

**Mixture Order Consistency (Theorem 3.4(i))**

In regimes with well-separated components (R1-R3, R5), the probability of correctly selecting $K_0$ exceeds 0.84, providing strong empirical support for Theorem 3.4(i). Performance degrades predictably in R4, where overlapping components make identification intrinsically difficult, yet even here, BIC selects $K_0$ more often than not (0.56). In R6 (weak signals), the rate remains respectable at 0.79, demonstrating robustness even when signal strength is low.

These results align with the theoretical prediction that BIC consistently estimates the true mixture order under the conditions of Theorem 3.2.

**Sparsity Recovery (Theorem 3.3)**

Across all regimes, MLR-LR correctly identifies 78-91% of truly zero coefficients (TNR), with false positive rates below 9%. These figures directly validate Theorem 3.3, confirming that the LASSO penalty successfully induces sparsity without excessive shrinkage of true signals.

The trade-off between TNR and FPR is most favourable in well-separated regimes (R1-R3) and degrades slightly under weak separation or weak signals. This pattern is expected: when components overlap or signals are faint, distinguishing truly zero from near-zero coefficients becomes more challenging.

Oracle Efficiency (Theorem 3.4(iii))

For regimes where the true non-zero coefficients are known (R2-R6), we constructed 95% confidence intervals using the estimated asymptotic variance from the Fisher information matrix. Coverage probabilities range from 88.2% to 94.1%, approaching the nominal 95% level in well-behaved regimes (R2, R3, R5). This provides empirical support for the oracle efficiency claim of Theorem 3.4(iii): conditioned on correct component membership and support recovery, the estimators for non-zero coefficients achieve asymptotic normality with near-nominal coverage.

The slight under-coverage in R4 and R6 reflects the difficulty of these regimes and suggests that larger sample sizes may be needed for the asymptotic approximation to take full effect.

### 4.2.4 Computational Considerations

The penalised EM algorithm (Algorithm 1) converged in all 1,200 simulation runs (6 regimes × 200 replications). Convergence criteria were set to $\varepsilon\_EM = 10^{-6}$ for the relative change in penalised log-likelihood and $\varepsilon\_CD = 10^{-4}$ for coordinate descent, as specified in Section 3.4.4.

**Table 4.4: Computational Performance**

| Regime | EM Iterations | Time per Replication (s) | Convergence Rate (%) |
|---|---|---|---|
| R1 | 24.3 (5.2) | 12.8 (2.1) | 100 |
| R2 | 38.7 (8.4) | 19.4 (3.8) | 100 |
| R3 | 41.2 (9.1) | 21.6 (4.2) | 100 |
| R4 | 29.5 (6.8) | 16.2 (3.1) | 100 |
| R5 | 36.4 (7.9) | 18.5 (3.6) | 100 |
| R6 | 44.8 10.2) | 23.1 (4.8) | 100 |
| **Average** | **35.8** | **18.6** | **100** |

*Note: Mean (standard deviation) across 200 replications. Times measured on 2.8 GHz Intel Xeon processor.*

Average computation time per replication was 18.6 seconds, with more challenging regimes (R3, R6) requiring slightly longer due to slower EM convergence. The grid search over ($K$, $\lambda$) with 20 random initialisations accounted for approximately 75% of runtime.

Multiple initialisations proved essential: the best initialisation yielded penalised log-likelihood values on average 5.8% higher than the median run, and in 12% of cases, the default initialisation would have converged to a suboptimal local maximum. This confirms the importance of the multi-start strategy described in Section 3.4.6.

### 4.2.5 Summary of Simulation Findings

The simulation studies provide comprehensive empirical validation of the MLR–LR framework across all theoretical claims:

1. Predictive performance**:** MLR–LR consistently matches or exceeds all five benchmark methods when latent heterogeneity exists, while avoiding overfitting in homogeneous regimes.
2. Mixture order consistency (Theorem 3.4(i))**:** BIC correctly identifies the true number of components with high probability (0.79-0.94) in well-behaved regimes.
3. Sparsity recovery (Theorem 3.3): The LASSO penalty achieves true negative rates of 78-91% with false positive rates below 9%, confirming consistent variable selection.
4. Oracle efficiency (Theorem 3.4(iii)): Confidence intervals for non-zero coefficients achieve near-nominal coverage (88–94%) when conditioned on correct recovery.
5. Computational reliability**:** The penalised EM algorithm with multi-start initialisation converges reliably across all regimes, with average runtime of 18.6 seconds per replication.

These findings, obtained under the exact protocols specified in the methodology, position MLR-LR as a rigorously validated tool for high-dimensional heterogeneous binary classification.

## 4.3 Real Data Application

To complement the controlled simulation studies, the proposed Mixture of Logistic Regression with LASSO Regularisation (MLR-LR) model is evaluated on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [26],

a benchmark dataset widely used for evaluating binary classification methods. This analysis demonstrates how the proposed methodology performs under realistic data complexity, including correlated predictors, natural noise, and potential latent subpopulations.

### 4.3.1 Dataset Description

The WDBC dataset comprises features computed from digitised images of Fine Needle Aspirate (FNA) biopsies of breast masses. Each record corresponds to a patient, with the outcome indicating whether the tumour is malignant (1) or benign (0). The dataset contains 569 observations with 30 continuous predictors that describe various morphological properties of cell nuclei, including radius, texture, perimeter, smoothness, compactness, concavity, and related measures. These predictors are known to be highly correlated and exhibit heterogeneity across tumour types, making the dataset an excellent test bed for mixture-based models with sparsity. The data are sourced from the UCI Machine Learning Repository. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?resource=download.

The response distribution shows moderate class imbalance: 357 benign (62.7%) and 212 malignant (37.3%) cases, with a predictor-to-sample ratio of approximately 0.05:1, placing the dataset in a low-dimensional regime where the primary challenge is multicollinearity rather than dimensionality.

### 4.3.2 Clustering Tendency

Before applying mixture models, clustering tendency was assessed using the Hopkins statistic computed on the training data after PCA-based dimension reduction. Table 4.5 presents the clustering diagnostics.

**Table 4.5:** Clustering Diagnostics for the Wisconsin Breast Cancer Dataset

| Measure | Result |
|---|---|
| Hopkins Statistic | 0.726 |
| Optimal Clusters (Gap) | 2 |
| Silhouette Analysis | Best at $K$=2 |

From Table 4.5, the Hopkins statistic was computed to evaluate whether the data exhibit an inherent clustering tendency. A value of 0.726 was obtained, which exceeds the conventional threshold of 0.7, indicating meaningful clustering structure. Clustering quality was further assessed using silhouette analysis and the gap statistic. Both methods supported the existence of latent grouping. The silhouette index showed that clustering quality improves up to $K$=2, beyond which there is little gain. Similarly, the gap statistic also identified two clusters as optimal. These findings confirm the presence of heterogeneity in the dataset, thereby supporting the adoption of a mixture modelling framework.

### 4.3.3 Multicollinearity and Dimensionality Diagnostics

Variance Inflation Factor (VIF) analysis was carried out to assess the severity of multicollinearity among the predictors.

**Table 4.6:** Variance Inflation Factor Results

| **Predictor** | **VIF** |
|---|---|
| radius mean | 3806 |
| area mean | 348 |
| concavity mean | 70.8 |
| fractal dimension mean | 15.8 |
| perimeter se | 70.4 |
| compactness se | 15.4 |
| symmetry se | 5.18 |
| texture worst | 18.6 |
| smoothness worst | 10.9 |

| | |
|---|---|
| concave points worst | 36.8 |
| texture mean | 11.88 |
| smoothness mean | 8.19 |
| concave points mean | 60.04 |
| radius se | 75.46 |
| area se | 41.16 |
| concavity se | 15.7 |
| fractal dimension se | 9.72 |
| perimeter worst | 405 |
| compactness worst | 36.98 |
| symmetry worst | 9.52 |
| perimeter mean | 3786 |
| compactness mean | 50.51 |
| symmetry mean | 4.22 |
| texture se | 4.21 |
| smoothness se | 4.03 |
| concave points se | 11.52 |
| radius worst | 799.1 |
| area worst | 337.2 |
| concavity worst | 31.97 |
| fractal dimension worst | 18.86 |

Table 4.6 presents the VIF results. Out of the 30 predictors, 6 exhibited extreme VIF (VIF > 100), 17 exhibited severe values (VIF ≥ 10), 4 exhibited high VIF (5 ≤ VIF < 10), and 3 variables with VIF < 5. These findings indicate severe multicollinearity, thereby justifying the use of the LASSO regularisation technique, which can simultaneously perform estimation and variable selection.

### 4.3.4 Model Selection and Estimation

Following the model selection protocol established in Section 3.5, a two-dimensional grid search was conducted over the number of components $K \in \{2, 3, 4\}$ and penalty parameter $\lambda$ on a logarithmic grid of 25 values. For each $(K, \lambda)$ pair, the penalised EM algorithm (Algorithm 1) was run with 10 random initialisations to mitigate local maxima, and the solution achieving the highest penalised log-likelihood was retained. Model selection was performed using the BIC criterion defined in Equation (3.20). Table 4.5 presents the BIC values for selected $(K,\lambda)$ combinations, with the optimal model highlighted.

**Table 4.7:** Model Selection Results for the Breast Cancer Dataset

| K | Λ | BIC | K | Λ | BIC |
|---|---|---|---|---|---|
| 2 | 0.000788 | 169.12 | 3 | 0.000788 | 246.88 |
| 2 | 0.00105 | 171.67 | 3 | 0.00105 | 249.4 |
| 2 | 0.0014 | 174.6 | 3 | 0.0014 | 246.3 |
| 2 | 0.001867 | 171.95 | 3 | 0.001867 | 237.6 |
| 2 | 0.00249 | 181.82 | 3 | 0.00249 | 247.44 |
| 2 | 0.003321 | 169.66 | 3 | 0.003321 | 234.24 |
| 2 | 0.004429 | 168.35 | 3 | 0.004429 | 233.7 |
| **2** | **0.005906** | **167.44** | 3 | 0.005906 | 221.23 |
| 2 | 0.007875 | 167.47 | 3 | 0.007875 | 215.3 |
| 2 | 0.010502 | 168.88 | 3 | 0.010502 | 216.69 |
| 2 | 0.014004 | 172.02 | 3 | 0.014004 | 213.83 |
| 2 | 0.018675 | 183.14 | 3 | 0.018675 | 218.92 |
| 2 | 0.024903 | 184.21 | 3 | 0.024903 | 225.9 |
| 2 | 0.033209 | 200.52 | 4 | 0.000788 | 318.63 |
| 2 | 0.044285 | 203.2 | 4 | 0.00105 | 327.1 |
| 2 | 0.059055 | 211.44 | 4 | 0.0014 | 311.98 |
| 2 | 0.078752 | 237.02 | 4 | 0.001867 | 321.23 |
| 2 | 0.105017 | 254.58 | 4 | 0.00249 | 312.98 |

*Note: Only a subset of results is shown for brevity. The full grid comprised 75 combinations (3 K × 25 λ).*

The optimal model is identified as a two-component mixture ($K=2$) with penalty parameter $\lambda$=**0.00591**, achieving the minimum BIC of 167.44. This selection aligns with the clustering diagnostics, which suggested two latent subgroups. The BIC curves reveal that models with $K$=3 or $K$=4 yield substantially higher BIC values, confirming that a two-component solution provides the optimal balance between goodness-of-fit and model complexity. These results provide empirical support for Theorem 3.4(i), demonstrating that BIC consistently selects the true number of components when latent structure exists.

### 4.3.5 Parameter Estimates and Interpretation

Table 4.6 presents the coefficient estimates from the optimal MLR-LR model ($K=2$, $\lambda$=0.005906). To enhance interpretability, only predictors with non-zero coefficients in at least one component are shown.

**Table 4.8:** Parameter Estimates from Optimal MLR-LR Model ($K = 2$, $\lambda = 0.00591$)

| Feature | Cluster 1 ($\pi$ = 0.467) | Cluster 2 ($\pi$ = 0.533) |
|---|---|---|
| Intercept | -0.0256 | -0.5892 |
| radius se | 2.668 ★★ | - |
| concavity mean | 1.627 ★★ | 1.096 ★★ |
| concave points worst | 1.609 ★★ | 0.763 ★ |
| radius worst | 1.593 ★★ | 4.456 ★★★ |
| texture worst | 0.649 ★ | 0.913 ★ |
| fractal dimension se | -0.614 ★ | - |
| smoothness worst | 0.319 | 0.951 ★ |
| concavity worst | 0.186 | 0.422 |
| symmetry worst | 0.137 | 0.457 |
| texture mean | 0.123 | 0.028 |

(Only non-zero predictors shown; zeros omitted for clarity).

Note: Stars indicate relative magnitude: ★ $|\beta| < 1$, ★★ $1 \leq |\beta| < 3$, ★★★ $|\beta| \geq 3$. Only non-zero predictors shown; zeros omitted for clarity.

From Table 4.8, the fitted MLR-LR model reveals two latent clusters with distinct predictor profiles, providing clinically interpretable insights into breast cancer heterogeneity:

Cluster 1 (46.7% of observations): This subgroup is characterised by strong structural predictors indicative of aggressive malignant features. The largest coefficients include radius se ($\beta$ = 2.668), concavity mean ($\beta$ = 1.627), concave points worst ($\beta$ = 1.609), and radius worst ($\beta$ = 1.593). The negative coefficient for fractal dimension se ($\beta$ = –0.614) suggests a protective effect in this subgroup. The concentration of strong coefficients in this cluster suggests it captures tumours with pronounced malignant characteristics.

Cluster 2 (53.3% of observations): This larger subgroup exhibits a different pattern of risk factors. The most notable feature is radius worst with a very large coefficient ($\beta$ = 4.456), indicating this feature is a dominant risk factor in this subgroup. Other important predictors include Xconcavity mean ($\beta$ = 1.096), texture worst ($\beta$ = 0.913), and smoothness worst ($\beta$ = 0.951). The negative intercept (–0.5892) suggests a baseline probability of malignancy below 50% when all predictors are at their means, aligning with the dataset's benign majority.

The coefficient magnitudes illustrate the LASSO shrinkage effect: within each cluster, coefficients for less important predictors are shrunk exactly to zero, yielding sparse and interpretable component-specific models. This directly validates Theorem 3.3, demonstrating consistent variable selection in a real-data context.

### 4.3.6 Comparative Performance against Alternative Models

Following the evaluation protocol established in Section 3.7, the proposed MLR-LR model is compared against benchmark methods: LASSO logistic regression, decision trees, random forests (RF), and gradient boosting machines (XGBoost). Table 4.9 presents the complete set of performance metrics computed on the held-out test set.

**Table 4.9: Cross-Tabulation of Cluster Assignments vs True Class Labels**

| True Class | Cluster 1 | Cluster 2 | Total |
|---|---|---|---|
| Benign | 42 (11.8%) | 315 (88.2%) | 357 |

| True Class | Cluster 1 | Cluster 2 | Total |
|---|---|---|---|
| Malignant | 170 (80.2%) | 42 (19.8%) | 212 |
| Total | 212 (37.3%) | 357 (62.7%) | 569 |

Interpretation: Cluster 1 contains 80.2% of malignant cases but also 11.8% of benign cases. Cluster 2 contains 88.2% of benign cases but also 19.8% of malignant cases. This demonstrates that the clusters are not merely recapitulating the malignant/benign distinction. The 42 malignant cases in Cluster 2 represent a clinically important subgroup-malignant tumors with feature profiles resembling benign tissue.

**Table 4.10:** Comparative Performance on WDBC Dataset.

| Model | AUC | F1 | Recall | Precision | M.rate | Log-Loss |
|---|---|---|---|---|---|---|
| LASSO-LR | 0.9769 | 0.9273 | 0.895 | 0.9623 | 0.0468 | 0.177 |
| Decision Tree | 0.884 | 0.8403 | 0.877 | 0.8065 | 0.1111 | 0.799 |
| Random Forest | 0.9777 | 0.8772 | 0.877 | 0.8772 | 0.0819 | 0.175 |
| XGBoost | 0.9808 | 0.8889 | 0.912 | 0.8667 | 0.076 | 0.177 |
| **MLR-LR** | **0.99** | **0.955** | **0.93** | **0.982** | **0.0292** | **0.12** |

From Table 4.10, the proposed MLR-LR model achieved the best overall performance across all metrics. M.rate stands for Misclassification rate. Note: Random Forest's lower F1 score (0.8772) reflects its higher misclassification rate (0.0819) compared to LASSO (0.0468) and MLR-LR (0.0292), likely due to overfitting in the presence of multicollinearity.

### 4.3.7 Summary of Real Data Findings

The real data analysis provides strong empirical validation of the MLR-LR framework:

i. Heterogeneity confirmed: The Hopkins statistic (0.726) and clustering diagnostics confirm the presence of latent subgroups, validating the mixture modelling approach. Both silhouette and gap statistics identified two clusters as optimal.
ii. Regularisation justified: Severe multicollinearity (VIF values exceeding 3800) demonstrates the necessity of LASSO regularisation for stable estimation and variable selection, resulting in sparse, interpretable component-specific models.
iii. Optimal model selection: BIC correctly selected a two-component model ($K = 2$) with $\lambda = 0.005906$, achieving the minimum BIC value of 167.44 across 75 grid combinations.
iv. Interpretable structure recovered: The two-component solution reveals clinically meaningful subgroups with distinct risk factor profiles, enhancing model interpretability.
v. Superior predictive performance: MLR-LR achieves the highest AUC (0.9903), F1 (0.9550), recall (0.9298), precision (0.9815), and lowest misclassification rate (0.0292) and log-loss (0.1213), outperforming all benchmark methods including LASSO, Random Forest, and XGBoost across all six metrics.
vi. Theoretical validation: The results empirically support:

**Proposition 3.1:** Distinct components are identifiable and clinically meaningful

**Theorem 3.3:** LASSO successfully induces sparsity

**Theorem 3.4(i):** BIC correctly selects the true number of components ($K = 2$)

**Theorem 3.4(iii):** The estimator achieves near-oracle efficiency (high precision and recall)

These findings complement the simulation studies and position MLR-LR as a rigorously validated tool for high-dimensional heterogeneous binary classification in real-world applications.

### 4.3.8 Limitations

First, while the dataset exhibits strong clustering tendency (Hopkins = 0.726), the sample size (n = 569) is modest; larger datasets would provide stronger generalisability. Second, LASSO pre-filtering was applied to manage multicollinearity; while this aligns with the methodological framework, alternative dimension reduction strategies could be explored. Third, the interpretability of the two clusters aligns with known breast cancer biology, but external validation with clinical outcomes would strengthen the findings. Fourth, while MLR-LR outperforms all benchmarks, the primary advantage lies in interpretability and theoretical guarantees rather than dramatic predictive gains. Fifth, while the WDBC dataset demonstrates the method's effectiveness in low-dimensional settings with severe multicollinearity, it does not represent the ultra-high-dimensional regime ($p >> n$) simulated in Section 4.2. Future work should apply MLR-LR to genuinely high-dimensional genomic data where p exceeds n by orders of magnitude.

## 5. Conclusions

This study addressed a central limitation in binary classification: the inability of standard models to simultaneously capture latent heterogeneity and perform variable selection in high-dimensional settings. The Introduction established three expectations: (i) that modelling latent subpopulations would improve predictive performance, (ii) that $\ell_1$-regularisation would enable effective sparsity recovery, and (iii) that the proposed estimator would exhibit desirable asymptotic properties under high-dimensional regimes. The results confirm these expectations. Across heterogeneous regimes, the proposed Mixture Logistic Regression with LASSO Regularisation (MLR-LR) consistently outperformed homogeneous LASSO logistic regression, indicating that explicitly modelling latent heterogeneity yields measurable gains in predictive accuracy. In homogeneous settings, the method effectively reduced to a single component, demonstrating robustness to model misspecification. The LASSO penalty achieved reliable variable selection, with high true negative rates and controlled false discoveries across all regimes, in line with the theoretical sparsity recovery results. In addition, the empirical behaviour of the estimator aligned with theoretical guarantees on consistency and stability, supporting the validity of the proposed

framework in high-dimensional settings. Although the real dataset is low-dimensional, the simulation results confirm the method's effectiveness when the number of predictors exceeds the sample size. The real data application further demonstrates that the model identifies latent subgroups with distinct coefficient structures, capturing meaningful heterogeneity beyond simple class separation while maintaining strong predictive performance.

Overall, the findings establish clear compatibility between the initial research objectives and the empirical results, confirming that the proposed framework successfully addresses the identified methodological gap.

Future work may extend the framework to multi-class outcomes, incorporate non-linear effects, and develop more scalable optimisation procedures for large-scale data. Applications in domains such as medical diagnosis, defence analytics, and high-dimensional biological studies provide promising directions for further development.

## References

[1] McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley.

[2] Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. John Wiley & Sons.

[3] Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Springer.

[4] Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science*. Cambridge University Press.

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

[6] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58(1), 267-288.

[7] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1-22.

[8] Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70(1), 53-71.

[9] Khalili, A., & Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479), 1025-1038.

[10] Städler, N., Bühlmann, P., & van de Geer, S. (2010). $\ell_1$-penalization for mixture regression models. *TEST*, 19(2), 209-256.

[11] Städler, N., Bühlmann, P., & van de Geer, S. (2012). $\ell_1$-penalization for mixture regression models. *Journal of Statistical Planning and Inference, 142*(12), 3192-3205.

[12] Khalili, A. (2011). An overview of the new feature selection methods in finite mixture of regression models. *Journal of the Iranian Statistical Society*, 10(2), 201-235.

[13] Morvan, M., Bertrand, A., Doutrelon, P., & Bruyère, F. (2021). Penalized mixture of logistic regressions for the prediction of nonalcoholic steatohepatitis. *Annals of Applied Statistics*, 15(2), 1043-1061.

[14] Zhu, J., & Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3), 427-443.

[15] Grün, B., & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, 51(11), 5247-5252.

[16] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1-22.

[17] Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4), 1265-1269.

[18] Allman, E., Matias, C., & Rhodes, J. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A), 3099-3132.

[19] van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

[20] Pan, W., & Shen, X. (2007). Penalized model-based clustering with applications to variable selection. *Journal of Machine Learning Research*, 8, 1145-1164.

[21] Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611-631.

[22] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The Lasso and generalizations*. Chapman and Hall/CRC.

[23] Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics*, 62(1), 49-66.

[24] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79-87.

[25] Hopkins, B., & Skellam, J. G. (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(2), 213-227.

[26] Wolberg, W. H. (1995). *Breast Cancer Wisconsin (Diagnostic) Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[27] Huynh, B. T. (2020). *Estimation and feature selection in high-dimensional mixtures-of-experts models* [Doctoral dissertation].

[28] Javanmard, A., Shao, S., & Bien, J. (2025). Prediction sets for high-dimensional mixture of experts models. *Journal of the Royal Statistical Society: Series B*, 87(3), 850-871.