Analysis of Estimating the Bayes Rule for Gaussian Mixture Models with a Specified Missing-Data Mechanism

Ziyang Lyu

Abstract Semi-supervised learning (SSL) approaches have been successfully applied in a wide range of engineering and scientific fields. This paper investigates the generative model framework with a missingness mechanism for unclassified observations, as introduced by [1]. We show that in a partially classified sample, a classifier using Bayes’ rule of allocation with a missing-data mechanism can surpass a fully supervised classifier in a two-class normal homoscedastic model, especially with moderate to low overlap and proportion of missing class labels, or with large overlap but few missing labels. It also outperforms a classifier with no missing-data mechanism regardless of the overlap region or the proportion of missing class labels. Our exploration of two- and three-component normal mixture models with unequal covariances through simulations further corroborates our findings. Finally, we illustrate the use of the proposed classifier with a missing-data mechanism on interneuronal and skin lesion datasets.

Keywords Semi-supervised learning · Optimal Bayes’ rule · Missing at random · Entropy · Mixture model

1 Introduction

The field of machine learning has recently witnessed unprecedented success in training extensive, deep neural networks, a feat largely attributable to the availability of voluminous labeled datasets. However, producing these labels is often expensive and time-consuming, requiring specialized expertise. Semi-supervised learning (SSL), as detailed by [5], presents a viable solution to this challenge. A dataset suitable for SSL is composed of a labeled subset $\mathcal{D}^L = \{(y_i, z_i) \mid i = 1, \ldots, n_l\}$ and an unlabeled subset $\mathcal{D}^U = \{y_{n+l} \mid j = 1, \ldots, n_u\}$.

UNSW Data Science Hub, School of Mathematics and Statistics, University of New South Wales, Sydney, Australia
E-mail: ziyang.lyu@unsw.edu.cn
where $y_i$ represents a multi-dimensional observation, and $z_i$ denotes the class membership of $y_i$. Principal approaches encompass generative models $[30,16,10]$, graph-based models $[4,33,38]$, and semi-supervised support vector machines $[35,15,18]$. Normal mixture models have become increasingly common in SSL. [2] provide an insightful literature review on SSL from a statistical perspective. Even though normal mixture models have been the subject of numerous studies, such as those by $[16,30,32,14]$, these works often make the critical assumption that the missing label process can be overlooked for likelihood-based inference $[25,6]$. In contrast, our approach, following $[1]$, treats the labels of unclassified observations as missing data and proposes a framework for their missingness, drawing from Rubin’s seminal work on missingness in incomplete data analysis $[31]$. Here, the probability of label missingness depends on a logistic model with a covariance equivalent to an entropy-based measure. This dependency in the missingness pattern can be harnessed to provide additional information about the optimal classifier as prescribed by Bayes’ rule.

Within this framework, [1] performed an asymptotic analysis showing that, in two-class normal homoscedastic models, partially classified samples can provide more information than fully labeled ones. [21] developed an R package focused on Gaussian mixture models incorporating a mechanism for missing data. However, their research did not clarify how to realize this effect nor did it expand to broader scenarios. Consequently, our research seeks to identify the conditions where a classifier, based on a finite normal mixture model with a missing-data mechanism, surpasses a fully supervised classifier. We build upon [1]’s work to cover two- and three-class normal mixture models, assessed via simulation experiments. Moreover, we corroborate our results by analyzing two real-world datasets.

The structure of this paper is outlined as follows: Section 2 introduces the notation used in this paper. Section 3 details maximum likelihood estimation for completely and partially classified samples, considering both scenarios with and without the missing data mechanism. Section 4 examines the analysis of a two-class normal homoscedastic model. Simulation results for two- and three-class normal mixture models with unequal covariances are presented in Section 5. Section 6 discusses the analysis of two datasets: the interneuronal and skin lesion datasets. Finally, Section 7 provides a summary of our findings and concluding remarks.

2 Notation

To define the notations, we let $y_j$ denote the recorded $p$-dimensional observation, and we let $z_j = (z_{1j}, \ldots, z_{gj})^T$ denote the corresponding $g$-dimensional vector of zero-one indicator variables defining the known class of origin of each, namely, $z_{ij}$ is equal to one if observation $y_j$ belongs to class $i$ and is equal to zero otherwise, $(i = 1, \ldots, g$ and $j = 1, \ldots, n)$. We use $x_{CC} = (x_1^T, \ldots, x_n^T)^T$ with $x_j = (y_j^T, z_j^T)^T$ to denote a set of completely classified samples. For
a partially classified sample, in which some observations have no associated class labels, we let \( m_j \) be the missing-label indicator for the observation \( y_j \), which is equal to 1 if \( z_j \) is missing and to zero if it is available \((j = 1, \ldots, n)\). Accordingly, the partially classified sample \( x_{PC} \) consists of those members \( x_j \) in \( x_{CC} \) with \( m_j = 0 \) and of only the observation \( y_j \) in \( x_{CC} \) with \( m_j = 1 \) (that is, the label \( z_j \) is missing). Henceforth, we will use the subscripts PC and CC to denote partially and completely classified samples, respectively.

A recorded \( p \)-dimensional observation \( y_j \) can arise from one of \( g \) multivariate normal classes,

\[
Y_j \sim N\left(\mu_i, \Sigma_i\right) \quad (i = 1, \ldots, g; j = 1, \ldots, n),
\]

where \( \mu_1, \ldots, \mu_g \) denote the class means, and \( \Sigma_1, \ldots, \Sigma_g \) denote the class covariance matrices. We denote the prior probabilities of the classes \( C_1, \ldots, C_g \) by \( \pi_1, \ldots, \pi_g \), respectively, where \( 1 - \pi_g = \sum_{i=1}^{g-1} \pi_i \). We suppose that the density \( f(y_j) \) of \( Y_j \) can be written as

\[
f(y_j) = \sum_{i=1}^{g} \pi_i f_i(y_j; \omega_i),
\]

where \( \omega_i \) consists of the elements of \( \mu_i \) and the \( \frac{1}{2}p(p + 1) \) distinct elements of \( \Sigma_i \) \((i = 1, \ldots, g)\). The vector \( \theta \) of all unknown parameters is given by \( \theta = (\pi_1, \ldots, \pi_{g-1}, \omega^T)^T \), where \( \omega = (\omega_1^T, \ldots, \omega_g^T)^T \). The classifier/Bayes’ rule of allocation \( R(y_j; \theta) \) assigns an entity with observation \( y_j \) to class \( C_k \) (that is, \( R(y_j; \theta) = k \)), where \( k \) represents the class membership) if

\[
k = \arg \max_i \tau_i(y_j; \theta),
\]

where

\[
\tau_i(y_j; \theta) = \frac{\pi_i f(y_j; \omega_i)}{\sum_{h=1}^{g} \pi_h f(y_j; \omega_h)}
\]

is the posterior probability that \( Y_j \) belongs to \( C_i \) given \( Y_j = y_j \) \((i = 1, \ldots, g \) and \( j = 1, \ldots, n)\).

The optimal (overall) error rate of the Bayes rule for a \( g \)-class normal mixture model is given by

\[
\text{err}(y; \theta) = 1 - \sum_{i=1}^{g} \pi_i \Pr \{ R(y; \theta) = i \mid Z \in C_i \}.
\]

The distribution of \( R(y_j; \theta) \) is very complicated, and manageable analytical expressions were initially obtained only in special cases, such as those addressed by [11][12][24][13]. Here, we write

\[
\Pr \{ R(Y; \theta) \in C_i \mid Z \in C_i \} = \frac{\sum_{j=1}^{n} I_{C_i}(z_j) Q[z_j, R(Y_j; \theta)]}{\sum_{j=1}^{n} I_{C_i}(z_j)},
\]

where \( Q[u, v] = 1 \) if \( u = v \) and \( Q[u, v] = 0 \) otherwise, and \( I_{C_i}(z_j) \) is an indicator function for the \( i \)th class.
3 Methodology

We consider the ML estimation of the vector $\theta$ containing all unknown parameters in the parametric families (2.1). For a partially classified sample $x_{PC}$, we let

$$L_{PC}^{\text{ig}} (\theta; x_{PC}) = \log L_C(\theta) + \log L_{UC}(\theta)$$

(3.1)
denote the log partial likelihood function for $\theta$ formed by ignoring the missing-data mechanism for the missing-label indicators $m_j$, where

$$\log L_C(\theta) = \sum_{j=1}^{n} \sum_{i=1}^{g} z_{ij} \log \{ \pi_i f (y_j; \omega_i) \},$$

$$\text{and} \quad \log L_{UC}(\theta) = \sum_{j=1}^{n} m_j \log \left( \sum_{i=1}^{g} \pi_i f (y_j; \omega_i) \right),$$

denote the log functions for $\theta$ formed on the basis of the classified data and unclassified data, respectively. The log of the likelihood $L_{CC} (\theta; x_{CC})$ for the completely classified sample $x_{CC}$ is given by $\log L_C(\theta)$ with all $m_j = 0$.

$$\log L_{CC} (\theta; x_{CC}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \log \{ \pi_i f (y_j; \omega_i) \}. \quad (3.2)$$

3.1 Missing data mechanism

In this study, we set aside the missing-data mechanism for likelihood inference in cases where missing labels are considered missing completely at random (MCAR), as defined in Rubin (1976). For an in-depth understanding of MCAR and its broader counterpart, missing at random (MAR), see [27]. Under MCAR, the absence of labels is unrelated to both features and labels. In contrast, MAR allows for missingness dependent on features, but not on labels. As noted by [26], in specific MAR instances, such as truncated features, ignoring missing data is justifiable. However, in our MAR context, disregarding missing data would be inappropriate. In fact, incorporating missing data can enhance the effectiveness of the classifier we aim to develop.

We introduce a random variable $M_j$ corresponding to the realized value $m_j$ of the missing-label indicator for the observation $y_j$. The missing-data mechanism of [31] is specified in the present context as

$$\Pr \{ M_j = 1 \mid y_j, z_j \} = \Pr \{ M_j = 1 \mid y_j \} = q (y_j; \theta, \xi) \quad (j = 1, \ldots, n), \quad (3.3)$$

where $\xi = (\xi_0, \xi_1)^T$ is distinct from $\theta$. The conditional probability $q (y_j; \theta, \xi)$ is taken to be a logistic function of the Shannon entropy $en_j (y_j; \theta)$, yielding

$$q (y_j; \theta, \xi) = \frac{\exp \{ \xi_0 + \xi_1 \log en_j (y_j; \theta) \}}{1 + \exp \{ \xi_0 + \xi_1 \log en_j (y_j; \theta) \}}, \quad (3.4)$$
where \( cn_j(y_j; \theta) = - \sum_{i=1}^{g} \tau_i(y_j; \theta) \log \tau_i(y_j; \theta) \). We use \( \gamma \) to denote the expected proportion of missing class labels in the partially classified sample given by

\[
\gamma = \frac{n}{\sum_{j=1}^{n} E(M_j)/n} = E[Pr\{ M_j = 1 | y_j \}] = E[q(Y; \theta, \xi)]. \tag{3.5}
\]

We let \( \Psi = (\theta^T, \xi^T)^T \) be the vector of all unknown parameters. To construct the full likelihood function \( f^{(\text{full})}(\Psi) \) from the partially classified sample \( x_{\text{PC}} \), we need expressions for the classified and unclassified observation \( y_j \), respectively. Now, for a classified observation \( y_j \), it follows that

\[
f(y_j, z_j, m_j = 0) = f(z_j) f(y_j | z_j) Pr\{ M_j = 0 | y_j, z_j \}
= \prod_{i=1}^{g} \{ \pi_i f_i(y_j; \omega_i) \}^{z_{ij}} \{ 1 - q(y_j; \theta, \xi) \}, \tag{3.6}
\]

while for an unclassified observation \( y_j \), we have that

\[
f(y_j, m_j = 1) = f(y_j) Pr\{ M_j = 1 | y_j \} = \{ \sum_{i=1}^{g} \pi_i f_i(y_j; \omega_i) \} q(y_j; \theta, \xi). \tag{3.7}
\]

Then, the full likelihood function can be expressed as

\[
f^{(\text{full})}(\Psi) = L_{\text{PC}}(\theta, \xi; x_{\text{PC}}) = \prod_{j=1}^{n} f^{(1-m_j)}(y_j, z_j, m_j = 0) f^{m_j}(y_j, m_j = 1). \tag{3.8}
\]

Thus, we can write the full log-likelihood function log \( f^{(\text{full})}(\Psi) \) as

\[
\log f^{(\text{full})}(\Psi; x_{\text{PC}}) = \log L^{(\text{ig})}_{\text{PC}}(\theta; x_{\text{PC}}) + \log L^{(\text{miss})}_{\text{PC}}(\theta, \xi; x_{\text{PC}}), \tag{3.9}
\]

where

\[
\log L^{(\text{miss})}_{\text{PC}}(\theta, \xi; x_{\text{PC}}) = \sum_{j=1}^{n} \{ (1 - m_j) \log \{ 1 - q(y_j; \theta, \xi) \} + m_j \log q(y_j; \theta, \xi) \} \tag{3.10}
\]

is the log-likelihood function formed on the basis of the missing-label indicator \( m_j \).

We let \( \theta_{\text{CC}}, \theta_{\text{PC}}^{(\text{ig})} \) and \( \theta_{\text{PC}}^{(\text{full})} \) be the estimates of \( \theta \) formed by consideration of the likelihood \( \log L^{(\text{ig})}_{\text{PC}}(\theta; x_{\text{PC}}) \) and \( \log L^{(\text{full})}_{\text{PC}}(\theta, \xi; x_{\text{PC}}) \), respectively. Similarly, we let \( R(\hat{\theta}_{\text{CC}}), R(\hat{\theta}_{\text{PC}}^{(\text{ig})}) \) and \( R(\hat{\theta}_{\text{PC}}^{(\text{full})}) \) denote the estimated Bayes’ rule obtained by plugging in the estimates \( \hat{\theta}_{\text{CC}}, \hat{\theta}_{\text{PC}}^{(\text{ig})} \) and \( \hat{\theta}_{\text{PC}}^{(\text{full})} \), respectively. Henceforth, we abbreviate the above notation as \( R \) with superscripts (ig) and (full) and subscripts CC and PC where appropriate.
4 Two-class normal homoscedastic model

Given the complexity of the optimal error rate expression, providing a theoretical analysis of likelihood inference with a missing-data mechanism is challenging in a general case. Thus, we study a particular case of $g=2$ classes under the assumption of equal-covariance ($\Sigma_1 = \Sigma_2 = \Sigma$). To simplify the numerical computations, [1] gave a Taylor series approximation of the log entropy $\log \{ \text{en}_j (y_j; \theta) \}$, shown as

$$
\log \{ \text{en}_j (y_j; \theta) \} = -\frac{1}{2} \log(2) + \frac{d^2 (y_j; \beta)}{8 \log(2)} + O(d^4 (y_j; \beta)),
$$

where $d(y_j; \beta) = \beta_0 + (\mu_1 - \mu_2)^T y_j \cdot (\mu_1 - \mu_2)$.

(4.1)

This expression provides a linear relationship between the negative log entropy $\log \{ \text{en}_j (y_j; \theta) \}$ and the square of the discriminant function $d^2 (y_j; \beta)$, neither of which is large. Therefore, the negative log entropy in the conditional probability (3.4), can be replaced by the square of the discriminant function

$$
q (y_j; \beta, \xi) = \exp \left\{ \xi_0 + \xi_1 d^2 (y_j; \beta) \right\},
$$

(4.3)

In convenient canonical form, let $\mu_1 = -\mu_2 = (\Delta/2, 0, \ldots, 0)^T$ and $\Sigma = I_p$, where $\Delta = [((\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2))^{1/2}$ is the Mahalanobis distance between the two classes. Therefore, $\beta_0 = \log (\pi_1/\pi_2)$ and $\beta_1 = (\Delta, 0, \ldots, 0)^T$ in the discriminant function (4.1).

Lemma 1 Given a two-class normal homoscedastic model in the canonical form in the same case of equal prior probabilities $\pi_1 = \pi_2$, the entropy of observation $y$ increases as the squared Mahalanobis distance decreases (equivalently the overlap region becomes smaller).

Proof In the canonical form, the discriminant function can be simplified as $d(y, \Delta) = \Delta y_1$, where $y_1$ is the first element of $y$. Under the two-class normal homoscedastic model, we can write the posterior probabilities as $\tau_1 (y; \Delta) = \exp (\Delta y_1) / \{ 1 + \exp (\Delta y_1) \}$. Then, we obtain the entropy by

$$
en(y; \Delta) = \log \{ 1 + \exp (\Delta y_1) \} - \Delta y_1 \exp (\Delta y_1) / (1 + \exp (\Delta y_1)).
$$

Since the first derivative of $en(y; \Delta)$ is given by

$$
en' (y; \Delta) = -\frac{\Delta y_1^2}{\{ 1 + \exp (|\Delta y_1|) \}^2} < 0,
$$

$en(y; \Delta)$ is a monotonically decreasing function of $\Delta$. 

Since $\theta$ corresponds to $\beta$ as defined in (4.2), Bayes’ rule $R(\hat{\theta})$ is redefined as $R(\hat{\beta})$ in terms of the estimates $\hat{\beta}$. Likewise, the true error rate is also redefined as $\text{err}(\beta)$. The asymptotic relative efficiencies (AREs) comparing the rule $R(\hat{\beta}_{PC}^{(\text{full})})$ with $R(\hat{\beta}_{CC})$ and $R(\hat{\theta}_{PC}^{(\text{ig})})$ are defined as

$$\text{ARE}_{\beta_{PC}^{(\text{full})}, \beta_{CC}} = \frac{AE\{\text{err}(\hat{\beta}_{CC})\} - \text{err}(\beta)}{AE\{\text{err}(\hat{\beta}_{PC}^{(\text{full})})\} - \text{err}(\beta)},$$

and

$$\text{ARE}_{\beta_{PC}^{(\text{full})}, \beta_{PC}^{(\text{ig})}} = \frac{AE\{\text{err}(\hat{\beta}_{PC}^{(\text{ig})})\} - \text{err}(\beta)}{AE\{\text{err}(\hat{\beta}_{PC}^{(\text{full})})\} - \text{err}(\beta)},$$

respectively, where $AE\{\text{err}(\hat{\beta}_{CC})\}$, $AE\{\text{err}(\hat{\beta}_{PC}^{(\text{ig})})\}$ and $AE\{\text{err}(\hat{\beta}_{PC}^{(\text{full})})\}$ denote the expansions of the expected error rates of $E\{\text{err}(\hat{\beta}_{CC})\}$, $E\{\text{err}(\hat{\beta}_{PC}^{(\text{ig})})\}$ and $E\{\text{err}(\hat{\beta}_{PC}^{(\text{full})})\}$, respectively, up to and including terms of the first order in $1/n$.

Comparing the performance of Bayes’ rule estimates shows the superiority of those from $\hat{\beta}_{PC}^{(\text{full})}$ over $\hat{\beta}_{CC}$, especially when its ARE is 1 or greater. Analogously, an enhanced performance can be observed in Bayes’ rule estimations achieved using $\hat{\beta}_{PC}^{(\text{full})}$ when compared with those using $\hat{\beta}_{PC}^{(\text{ig})}$, under the condition that the ARE between $\beta_{PC}^{(\text{full})}$ and $\beta_{PC}^{(\text{ig})}$ is at least 1. As demonstrated by [1] in their second theorem, the ARE of $\beta_{PC}^{(\text{full})}$ and $\beta_{CC}$ was computed under the circumstance of equivalent prior probabilities $\pi_1 = \pi_2$. This calculation depends on the variables $\Delta$ and $\xi$. Simulations further showed cases where the ARE of $\beta_{PC}^{(\text{full})}$ to $\beta_{CC}$ was 1 or higher. This implies an improved performance of Bayes’ rule estimations achieved via $\hat{\beta}_{PC}^{(\text{full})}$ in comparison to those obtained through $\hat{\beta}_{CC}$. In the following, we elucidate the expression for the ARE between $\beta_{PC}^{(\text{full})}$ and $\beta_{PC}^{(\text{ig})}$ under the equivalent scenario of equal prior probabilities, namely $\pi_1 = \pi_2$.

**Theorem 1** Under the missing-label model defined in (4.3), the ARE of $R(\hat{\beta}_{PC}^{(\text{full})})$ compared to $R(\hat{\beta}_{PC}^{(\text{ig})})$ in the case of $\pi_1 = \pi_2$ is given by

$$\text{ARE}_{\beta_{PC}^{(\text{full})}, \beta_{PC}^{(\text{ig})}} = \frac{(4 + \Delta^2)^{-1} - c_0}{(4 + \Delta^2)^{-1} - c_0},$$

for all $p$, where

$$\frac{(4 + \Delta^2)^{-1} - c_0}{(4 + \Delta^2)^{-1} - c_0},$$
\[ b_0 = \int_{-\infty}^{\infty} 4\xi^2_{i} \Delta^2 y_i^2 q_1 (y_1) (1 - q (y_1)) f_{g_i} (y_1) \, dy_1, \]
\[ c_0 = \int_{-\infty}^{\infty} \phi (y_1) \exp \left\{ -\Delta^2 / 8 \right\} \frac{1}{2} \exp \left\{ \Delta y_i / 2 \right\} + \frac{1}{2} \exp \left\{ -\Delta y_i / 2 \right\} \, dy_1, \]
\[ d_0 = \int_{-\infty}^{\infty} \tau_1 (y_1) \tau_2 (y_1) q_1 (y_1) \gamma^{-1} f_{g_i} (y_1) \, dy_1, \]

\[ \text{with} \]
\[ \tau_i (y_1) = \Pr (Z_i = 1 \mid (Y)_1 = y_1) \quad (i = 1, 2), \]
\[ q_1 (y_1; \Delta, \xi) = \Pr \{ M = 1 \mid (Y)_1 = y_1 \}, \]
\[ f_{g_i} (y_1; \Delta, \pi_1) = \pi_1 \phi \left( y_1; \Delta / 2, 1 \right) + \left( 1 - \pi_1 \right) \phi \left( y_1; -\Delta / 2, 1 \right). \]

**Proof** [9] gave the first order expansion of the expected excess error rate of the plug-in form of Bayes’ rule using the estimator \( \hat{\beta} \) of \( \beta \), where \( \sqrt{n} (\hat{\beta} - \beta) \) converges in distribution to \( N(0, \mathbf{V}) \), as \( n \to \infty \), and where \( \mathbf{V} \) is a \((p + 1) \times (p + 1)\) variance matrix of \( \sqrt{n} (\hat{\beta} - \beta) \), the first and second order moments also converge. The expectation of the so-called excess error rate can be expanded as follows:

\[ \mathbb{E} \{ \text{err} (\hat{\beta}) \} - \text{err} (\beta) = \frac{\pi_1 \phi \left( \Delta^*; 0, 1 \right)}{2 \Delta n} w + o(1/n), \]

where \( \Delta^* = \frac{1}{2} \Delta - \lambda / \Delta \) with \( \lambda = \log (\pi_1 / \pi_2) \), \( \phi (y; \mu, \sigma^2) \), denotes the normal density with mean \( \mu \) and variance \( \sigma^2 \), and

\[ w = v_{00} - \frac{2 \lambda}{\Delta} v_{01} + \frac{\lambda^2}{\Delta^2} v_{11} + \sum_{i=2}^{p} v_{ii}, \]

with \( v_{ij} \) representing the element in the \( i \) th row and \( j \) th column of \( \mathbf{V} \).

It has been shown [1] that the information matrix \( \mathbf{I}_{\text{PC}}^{(\text{full})} (\beta) \) can be decomposed as follow

\[ \mathbf{I}_{\text{PC}}^{(\text{full})} (\beta) = \mathbf{I}_{\text{CC}} (\beta) - \gamma \mathbf{I}_{\text{CC}}^{(\text{cl})} (\beta) + \mathbf{I}_{\text{PC}}^{(\text{miss})} (\beta), \]

where \( \gamma \) is the expected proportion of missing class labels in the partially classified sample, \( \mathbf{I}_{\text{CC}} (\beta) \) is the information about \( \beta \) contained in \( \mathbf{x}_{\text{CC}} \), \( \mathbf{I}_{\text{CC}}^{(\text{cl})} (\beta) \) is the conditional information about \( \beta \) under the logistic regression model for the distribution of the class labels given \( \mathbf{x}_{\text{CC}} \), and \( \mathbf{I}_{\text{PC}}^{(\text{miss})} (\beta) \) is the information about \( \beta \) contained in the missing-label indicators under the assumed logistic model for their distribution given \( \mathbf{x}_{\text{PC}} \). In the case of \( \pi_1 = \pi_2 \), the information matrix is a diagonal matrix, \( \mathbf{I}_{\text{PC}}^{(\text{full})} (\beta) = \text{diag} \{ u_0, u_1, u_0, \ldots, u_0 \} \); we have \( \lambda = 0 \) and

\[ w = v_{00} + \sum_{i=2}^{p} v_{ii}. \]

Since the asymptotic covariance matrix \( \mathbf{V} \) is given by \( n \times \{ \mathbf{I} (\beta) \}^{-1} \), we obtain
where

\[ u_0 = (4 + \Delta^2)^{-1} - \gamma d_0 + b_0, \]
\[ b_0 = \int_{-\infty}^{\infty} 4 \xi^2 (\Delta^2 y_1^2 + 2\lambda \Delta + \lambda^2) q_1(y_1)(1 - q(y_1)) f_{y_1}(y_1) \, dy_1, \]
\[ d_0 = \int_{-\infty}^{\infty} \tau_1(y_1) \tau_2(y_1) q_1(y_1) \gamma^{-1} f_{y_1}(y_1) \, dy_1. \]

\[ \text{[20]} \]

assumed that the missingness of the labels did not depend on the data and showed that the information matrix \( I_{PC}^{(ig)}(\beta) \) can be decomposed as follows

\[ I_{PC}^{(ig)}(\beta) = I_{CC}(\beta) - \gamma I_{CC}^{(tr)}(\beta), \]

where \( I_{CC}^{(tr)}(\beta) \) is the information about \( \beta \) contained in the inverse of the asymptotic covariance matrix of the \( \beta \) obtained from the marginal likelihood. For \( \pi_1 = \pi_2, I_{PC}^{(ig)}(\beta) = \text{diag}\{a_0, a_1, a_0, \ldots, a_0\} \). Similarly, we obtain

\[ E \left\{ \text{err} \left( \hat{\beta}_{PC}^{(ig)} \right) \right\} - \text{err} (\beta) = \frac{\pi_1 p \phi(\Delta/2; 0, 1)}{2a_0} + o(1/n), \]

where

\[ a_0 = (4 + \Delta^2)^{-1} - c_0, \quad \text{and} \quad c_0 = \int_{-\infty}^{\infty} \frac{\phi(y_1) \exp \{-\Delta^2/8\}}{1/2 \exp \{-\Delta y_1/2\} + 1/2 \exp \{-\Delta y_1/2\}} \, dy_1. \]

Our examination of the two-class normal homoscedastic model, even with equal prior probabilities (\( \pi_1 = \pi_2 \)), unveils a substantial complexity in the expressions for \( \text{ARE}(\hat{\beta}_{PC}^{(full)}, \beta_{CC}) \) and \( \text{ARE}(\hat{\beta}_{PC}^{(full)}, \beta_{PC}^{(ig)}) \). Consequently, elucidating the relationship among \( \text{ARE}(R(\hat{\theta}_{PC}^{(full)}), \Delta) \), and \( \xi \) remains an intricate task. In an attempt to demystify this complexity, we take the approach of fixing \( \xi_0 \) and \( \xi_1 \), then systematically exploring the range of \( \Delta \) values that yield \( \text{ARE}(\hat{\beta}_{PC}^{(full)}, \beta_{CC}) \geq 1 \) and \( \text{ARE}(\hat{\beta}_{PC}^{(full)}, \beta_{PC}^{(ig)}) \geq 1 \). Figures 1 and 2 present the interrelation between \( \text{ARE} \) and \( \Delta \) under assorted combinations of \( \xi_0 \in \{0.1, 0.5, 1\} \) and \( \xi_1 \in \{-0.1, -0.5, -1, -3, -5\} \), given the prior probability equalities, \( \pi_1 = \pi_2 \). Even for the cases with unequal prior probabilities, specifically \( \pi_1 \in \{0.1, 0.2, 0.3, 0.4\} \), we observed trends that are analogous to those displayed in Figures 1 and 2. The findings from this investigation are elaborated in the supplementary.

As seen in Figure 1, as \( \Delta \) increases under different combinations of \( \xi_0 \) and \( \xi_1 \), the value of \( \text{ARE}(\hat{\beta}_{PC}^{(full)}, \beta_{CC}) \) initially increases from 0 to a maximum and then decreases and converges to 1. This behavior can be interpreted as follows.
Fig. 1 Plot of the asymptotic relative efficiency $\text{ARE}(\beta_{\text{PC}}^{(\text{full})}, \beta_{\text{CC}})$ versus the squared root of the squared Mahalanobis distance between the two classes, $\Delta$, for $\pi_1 = \pi_2$.

– When the squared Mahalanobis distance between the two classes is too large (a small overlap region, e.g., Figure 3c), the value of $\text{ARE}(\beta_{\text{PC}}^{(\text{full})}, \beta_{\text{CC}})$ converges to 1, so there is no difference between the estimated Bayes’ rules $R(\hat{\beta}_{\text{PC}}^{(\text{full})})$ and $R(\hat{\beta}_{\text{CC}})$.

– When the squared Mahalanobis distance between the two classes is too small (a large overlap region, e.g., Figure 3a), the value of $\text{ARE}(\beta_{\text{PC}}^{(\text{full})}, \beta_{\text{CC}})$ is smaller than 1, indicating that $R(\hat{\beta}_{\text{PC}}^{(\text{full})})$ performs worse than $R(\hat{\beta}_{\text{CC}})$.

– When the squared Mahalanobis distance between the two classes is moderate (a moderate overlap region, e.g., Figure 3b), the value of $\text{ARE}(\beta_{\text{PC}}^{(\text{full})}, \beta_{\text{CC}})$ is greater than 1, which means that $R(\hat{\beta}_{\text{PC}}^{(\text{full})})$ performs better than $R(\hat{\beta}_{\text{CC}})$.

Moreover, when we focus on a partially classified sample with a large overlap region (such as $\Delta = 0.25$), we find that when $\xi_0$ is fixed, the smaller the $\xi_1$ value is, the larger the ARE value will be, implying that when $\Delta$ is fixed, a reduction in the proportion of missing class labels will cause the error rate of $R(\hat{\beta}_{\text{PC}}^{(\text{full})})$ to be close to or even better than that of $R(\hat{\beta}_{\text{CC}})$.

Figure 2 shows that as $\Delta$ increases under different combinations of $\xi_0$ and $\xi_1$, the value of $\text{ARE}(\beta_{\text{PC}}^{(\text{full})}, \beta_{\text{PC}}^{(\text{ig})})$ initially increases from 1 to a maximum and then decreases and converges to 1, indicating that the performance of $R(\hat{\beta}_{\text{PC}}^{(\text{full})})$ is no worse than that of $R(\hat{\beta}_{\text{PC}}^{(\text{ig})})$.

– When the squared Mahalanobis distance between the two classes is too large or too small (small or large overlap region, e.g., Figures 3a and 3c), the value of $\text{ARE}(\beta_{\text{PC}}^{(\text{full})}, \beta_{\text{PC}}^{(\text{ig})})$ converges to 1, so there is no difference between the estimated Bayes’ rules $R(\hat{\beta}_{\text{PC}}^{(\text{full})})$ and $R(\hat{\beta}_{\text{PC}}^{(\text{ig})})$.

– When the squared Mahalanobis distance between the two classes is moderate, (a moderate overlap region, e.g., Figure 3b), the value of $\text{ARE}(\beta_{\text{PC}}^{(\text{full})}, \beta_{\text{PC}}^{(\text{ig})})$ is greater than 1, which means that $R(\hat{\beta}_{\text{PC}}^{(\text{full})})$ performs better than $R(\hat{\beta}_{\text{PC}}^{(\text{ig})})$.

Moreover, when we focus on a partially classified sample with a large overlap region (such as $\Delta = 0.25$), we find that when $\xi_0$ is fixed, the smaller the
Estimating the Bayes Rule under a Specified Missing-Data Mechanism

The plot shows the asymptotic relative efficiency (ARE) of estimating the Bayes rule under a specified missing-data mechanism. The ARE is plotted against the squared root of the squared Mahalanobis distance between the two classes, $\Delta$, for $\pi_1 = \pi_2$.

A lower $\xi_1$ value is, the larger the ARE value will be, implying that when $\Delta$ is fixed, a reduction in the proportion of missing class label will cause the error rate of $R(\hat{\beta}_{PC}^{(full)})$ to be much smaller than that of $R(\hat{\beta}_{PC}^{(ig)})$.

### Figure 2
Plot of the asymptotic relative efficiency $\text{ARE}(\beta_{PC}^{(full)}, \beta_{PC}^{(ig)})$ versus the squared root of the squared Mahalanobis distance between the two classes, $\Delta$, for $\pi_1 = \pi_2$.

### Figure 3
Simulated scalar observations in the case of $n = 5000$, $\Delta \in \{0.25, 2.5, 10\}$, and $\pi_1 = \pi_2$ (black represents class 1; red represents class 2; blue represents unclassified observations with $\xi = (1, 0.5)^T$). (a) Large overlap region, (b) Moderate overlap region, and (c) Small overlap region represent the simulated scalar observations when $\pi_1 = \pi_2$. (d) to (f) show the same overlap regions as (a) to (c) but include unclassified observations (blue) represented by $\xi = (1, 0.5)^T$. 

---

0
1000 2000 3000 4000 5000
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
$\Delta = 0.25$

0
1000 2000 3000 4000 5000
-6 -4 -2 0 2 4 6
$\Delta = 2.5$

0 1000 2000 3000 4000 5000
-20 -15 -10 -5 0 5 10 15
$\Delta = 10$
When $\xi$ is fixed, equation (4.3) is a monotonically decreasing function of $\Delta$, meaning that the proportion of missing class labels in the partially classified sample decreases as $\Delta$ increases with fixed $\xi$. In other words, when $\xi$ is fixed, large/moderate/small overlap leads to a relatively high/median/low proportion of missing class labels under the missing-label mechanism (e.g., Figures 3d, 3e and 3f). This model is adopted with the understanding that in many datasets, the unclassified observations tend to fall in regions of overlap between the classes in the feature space.

Remark 1 Within the framework of the missing-data mechanism, we prefer to apply the estimated Bayes’ rule obtained using the ML estimator $\hat{\beta}^{(\text{full})}_{PC}$ instead of $\hat{\beta}^{(\text{ig})}_{PC}$ for a partially classified sample, regardless of the extent of overlap or the proportion of missing class labels. For a completely classified sample, the estimated Bayes’ rule obtained using the ML estimator $\hat{\beta}^{(\text{full})}_{PC}$ is preferred over that obtained using $\hat{\beta}^{(\text{ig})}_{C}$, either when both the overlap region and the proportion of missing class labels are moderate or small, or when the overlap region is large but the proportion of missing class labels is relatively small.

5 Simulation study

To our knowledge, the ARE analysis is limited to the specific case of a two-class normal homoscedastic model because of the calculation’s complexity. Exploring more general situations is challenging as it requires theoretical discernment of the conditions where Bayes’ rule, estimated using a full likelihood ML estimator, is valid. Therefore, we broaden our investigation through simulation studies to evaluate whether our preliminary conclusions hold in more varied contexts.

5.1 Two-class normal model with unequal covariance matrices

We continue to study the two-class normal model but extend it to unequal covariance matrices. We consider proportional covariance matrices; i.e., we assume $\Sigma_1 = \Sigma$ and $\Sigma_2 = \lambda \Sigma$. This model has been studied by [8,11,13,22,23]. The error rate of the Bayes’ rule is evaluated for various combinations of the parameters in canonical form, where $\mu_1 = 0$, $\mu_2 = (\Delta, 0, \ldots, 0)^T$, and $\Sigma = I$. [22,23] introduced a parameter $D = \Delta/(1 + \sqrt{\lambda})$ as a measure of the degree of separation of two populations (when $\lambda = 1$, there is a common covariance, $D = \Delta/2$), which is defined as the Euclidean distance between $\mu_1$ and the best linear discriminant hyperplane that yields equal error rates for $\pi_1$ and $\pi_2$ (after they have been transformed into canonical form).

Although our model’s extension to unequal covariance matrices is modest, deriving the expression for Asymptotic Relative Efficiency (ARE) as conducted in Theorem 1 is too complex. Consequently, we use simulation calculations to
investigate the relationship between simulated relative efficiency, the degree of population separation, and the proportion of missing class labels.

The first step is to examine the correlation between population separation and missing class labels. For each simulation combination, we generated $B = 1000$ samples (replications) with a sample size of $n = 200$ and an equal probability of choosing each population, i.e., $\pi_1 = \pi_2$. For each combination of $p \in \{2, 3, 4\}$, $\lambda \in \{2, 3, 4\}$, $\Delta \in \{0.5, 1, 1.5, 2, \ldots, 12.5\}$, and $\xi \in \{(3, 1)^T, (3, 4)^T\}$, the observations and missing label indicators in each sample were generated following the equations (2.1) and (3.4), respectively. Based on this simulation setup, we observe that the degree of separation, $D = \Delta/(1 + \sqrt{\lambda})$, varies roughly from 0.2 to 4. Figure 4 indicates that for all combinations, the average proportion of missing class labels across 1000 samples diminishes as the degree of separation between the two populations augments from high overlap to negligible overlap.

The second step is to examine how relative efficiency correlates with the proportion of missing class labels, we employed the same simulation setup, but only selected $\Delta \in \{1, 2, 3\}$. In each replication, the estimates $\hat{\theta}_{CC}^g$, $\hat{\theta}_{PC}^{(g)}$, and $\hat{\theta}_{PC}^{(full)}$ were computed using the general log-likelihood function, (3.2), (3.1), and (3.9), respectively. Since we cannot analytically calculate the conditional error rate (2.4), we utilized a holdout sample of size $0.2 \times n$ in each simulation trial to assess it.

![Fig. 4 Plots of the degree of separation of two populations versus the average proportion of missing class labels over 1000 samples with sample size $n = 200$ and $\pi_1 = \pi_2$.](image)

Then, the relative efficiency (RE) of $R(\hat{\theta}_{PC}^{(full)})$ compared to $R(\hat{\theta}_{CC})$ was estimated as

$$
\text{RE} \left\{ R(\hat{\theta}_{PC}^{(full)}), \hat{\theta}_{CC} \right\} = \frac{B^{-1} \sum_{b=1}^{B} \left\{ \text{err}(\hat{\theta}_{CC}^{(b)}) - \text{err}(\theta) \right\}}{B^{-1} \sum_{b=1}^{B} \left\{ \text{err}(\hat{\theta}_{PC}^{(full,b)}) - \text{err}(\theta) \right\}}, \tag{5.1}
$$
and the RE of $R(\hat{\theta}_{PC}^{(full)})$ compared to $R(\hat{\theta}_{IG}^{(ig)})$ was estimated as

$$
\text{RE} \left\{ R(\hat{\theta}_{PC}^{(full)}, \hat{\theta}_{PC}^{(ig)}) \right\} = \frac{B^{-1} \sum_{b=1}^{B} \left\{ \text{err}(\hat{\theta}_{PC}^{(ig,b)}) - \text{err}(\theta) \right\}}{B^{-1} \sum_{b=1}^{B} \left\{ \text{err}(\hat{\theta}_{PC}^{(full,b)}) - \text{err}(\theta) \right\}}.
$$

(5.2)

Table 1 Simulated relative efficiency $\text{RE} \left\{ R(\hat{\theta}_{PC}^{(full)}, \hat{\theta}_{PC}^{(ig)}) \right\}$ with $g = 2, \pi_1 = \pi_2, \mu_1 = 0, \mu_2 = (\Delta, 0, \ldots, 0)^T, \Sigma_1 = I$ and $\Sigma_2 = \lambda I$ for $n = 200$ and the missing labels $M_j$ generated with a missing label probability of $q(y_j^{(b)}, \theta, \xi)$ in the 6th trial (the proportion of missing class labels in parentheses)

| $\lambda$ | $\Delta$ | $\xi = (3, 1)^T$ | $\xi = (3, 4)^T$ |
|----------|----------|----------------|----------------|
| $p = 2$  | $p = 3$  | $p = 4$  | $p = 2$  | $p = 3$  | $p = 4$  |
| 1        | 0.21(0.90) | 0.27(0.89) | 0.28(0.89) | 1.20(0.66) | 1.71(0.60) | 1.71(0.60) |
| 2        | 1.07(0.65) | 1.09(0.62) | 1.09(0.62) | 1.71(0.20) | 1.58(0.18) | 1.58(0.18) |
| 3        | 0.28(0.87) | 0.53(0.84) | 0.57(0.82) | 1.52(0.61) | 1.54(0.46) | 1.54(0.46) |
| 4        | 0.35(0.67) | 1.34(0.62) | 1.34(0.62) | 3.10(0.24) | 1.39(0.19) | 1.39(0.19) |

The nonparametric bootstrap with 1000 re-samples was used to assess the variability of the estimates. All bootstrap standard errors are mostly small relative to their estimand.

Table 2 Simulated relative efficiency $\text{RE} \left\{ R(\hat{\theta}_{PC}^{(full)}, \hat{\theta}_{PC}^{(ig)}) \right\}$ with $g = 2, \pi_1 = \pi_2, \mu_1 = 0, \mu_2 = (\Delta, 0, \ldots, 0)^T, \Sigma_1 = I$ and $\Sigma_2 = \lambda I$ for $n = 200$ and the missing labels $M_j$ generated with a missing label probability of $q(y_j^{(b)}, \theta, \xi)$ in the 6th trial (the proportion of missing class labels in parentheses)

| $\lambda$ | $\Delta$ | $\xi = (3, 1)^T$ | $\xi = (3, 4)^T$ |
|----------|----------|----------------|----------------|
| $p = 2$  | $p = 3$  | $p = 4$  | $p = 2$  | $p = 3$  | $p = 4$  |
| 1        | 0.37(0.84) | 0.61(0.79) | 0.75(0.75) | 1.25(0.55) | 1.82(0.35) | 1.79(0.36) |
| 2        | 0.50(0.77) | 0.94(0.69) | 1.08(0.69) | 1.50(0.42) | 2.46(0.27) | 1.59(0.27) |
| 3        | 0.46(0.67) | 1.09(0.60) | 1.09(0.60) | 3.04(0.26) | 1.58(0.17) | 2.03(0.17) |
In an analysis employing a multi-way ANOVA on Tables 1 and 2, we assessed the impact of variables $\lambda$, $\Delta$, $p$, and $\xi$ on the proportion of missing class labels. The results indicate that only the factor $\Delta$ exerted a statistically significant influence on the outcome variable ($p$-value < 0.001). The variable $\xi$ approached statistical significance with a $p$-value of 0.073, corroborating the findings presented in Figure 4.

Table 1 reports the results for the RE of $R(\hat{\theta}_{PC}^{(\text{full})})$ compared to $R(\hat{\theta}_{CC})$. For $\xi = (3, 1)^T$, we can see that when $\lambda$ is fixed and $\Delta$ increases (or the degree of separation of the two populations increases), the proportion of missing class labels drops (from approximately 0.9 to approximately 0.6), consistent with Figure 4 and all simulated RE values increase accordingly, except the value for $p = 2$, $\lambda = 4$ and $\Delta = 3$. In other words, when excessive overlap decreases to moderate overlap with a relatively high proportion of missing class labels, the error rate of $R(\hat{\theta}_{PC}^{(\text{full})})$ decreases to close to that of $R(\hat{\theta}_{CC})$, and in some cases, the error rate of $R(\hat{\theta}_{PC}^{(\text{full})})$ is even smaller than that of $R(\hat{\theta}_{CC})$. A possible explanation for why the value for $p = 2$, $\lambda = 4$ and $\Delta = 3$ does not grow is that there is still excessive overlap ($D = 1$) with a relatively high proportion of missing class labels and a dimension equal 2. For $\xi = (3, 4)^T$, the proportion of missing class labels decreases (from approximately 0.66 to approximately 0.17) when $\lambda$ is fixed and $\Delta$ increases (or the degree of separation of the two populations increases), which is similarly consistent with Figure 4. All simulated RE values are greater than 1, meaning that all classifiers $R(\hat{\theta}_{PC}^{(\text{full})})$ perform better than $R(\hat{\theta}_{CC})$ when the size of the overlap region changes from large to moderate with a relatively low proportion of missing class labels. Another finding is that when $\lambda$ is fixed and $\Delta$ increases, some of simulated RE values initially increase and then decrease, which may be interpreted as cases in which, with a relatively small proportion of missing labels, the overlap first reaches the most appropriate overlap and then further decreases. This phenomenon is consistent with the results in Figure 2. Therefore, within the framework of the missing-data mechanism, for a two-class normal mixture model with unequal covariance matrices, the classifier $R(\hat{\theta}_{PC}^{(\text{full})})$ performs better than $R(\hat{\theta}_{CC})$ when the overlap region ranges from moderate to small with the proportion of missing class labels from relatively high to low, or when the overlap region is large but the proportion of missing class labels is relatively low.

Table 2 reports the results for the RE of $R(\hat{\theta}_{PC}^{(\text{full})})$ compared to $R(\hat{\theta}_{PC}^{(ig)})$. We find that all RE values are greater than 1, indicating that the classifier $R(\hat{\theta}_{PC}^{(\text{full})})$ performs better than $R(\hat{\theta}_{PC}^{(ig)})$. For $\xi = (3, 1)^T$ (a relatively high missing label proportion between 0.6 and 0.9), when $\lambda$ is fixed, the higher the $D$ value is, the higher the simulated RE value, implying that the classifier $R(\hat{\theta}_{PC}^{(\text{full})})$ performs much better than $R(\hat{\theta}_{PC}^{(ig)})$ as the overlap region changes from large to moderate. For $\xi = (3, 4)^T$ (a relatively low missing label proportion between 0.17 and 0.66), when $\lambda$ is fixed and $\Delta$ increases, some simulated RE values increase then decrease while others only decrease, which may be
interpreted to indicate that under a relatively low proportion of missing class labels, the overlap must first reach the most appropriate value in some cases before decreasing further, while in other cases, the overlap starts at the most appropriate value and then decreases. This phenomenon is consistent with the results in Figure 2. Therefore, for a two-class normal mixture model with unequal covariance matrices, the classifier \( R(\hat{\theta}_{PC}^{(\text{full})}) \) performs better than \( R(\hat{\theta}_{PC}^{(ig)}) \) regardless of the extent of overlap, and the performance gap between \( R(\hat{\theta}_{PC}^{(full)}) \) and \( R(\hat{\theta}_{PC}^{(ig)}) \) reaches its maximum when the partially classified sample has a moderate overlap region.

5.2 Three-class normal model with unequal covariance matrices

In this simulation, we consider a three-class normal model with unequal covariance matrices, for which is not straightforward to obtain the Euclidean distances between the means or the best linear discriminant hyperplanes (the effort to do so is outside the scope of this study). Thus, we simply explore the relationship between the simulated efficiency value and the proportion of missing class labels. We constructed \( \pi = (\pi_1, \pi_2, \pi_3) \), \( \mu_1 = (-1, 0, \ldots, 0)^T \), \( \mu_2 = \mu_3 = (1, 0, \ldots, 0)^T \), and \( \Sigma_i = \lambda_i I(i = 1, 2, 3) \). For each combination of \( \pi \) and \( \lambda \) (\( \pi \in \{(0.5, 0.35, 0.15)^T, (1/3, 1/3, 1/3)^T\}; \lambda_1 \in \{0.5, 1.5, 2.5\}; \lambda_2 \in \{1, 2, 3\}; \lambda_3 \in \{2, 4, 6\} \)), we generated \( B = 1000 \) samples with sample size \( n = 200 \) and \( p = 2 \). Note that if \( \lambda_1 = \lambda_2 = 2 \), observations would be actually generated from a two-component normal model. Therefore, we do not consider this simulation setting. For the partially classified sample \( x_{PC} \), the missing label indicators were generated using the missing label probability in 3.4, where we set \( \xi = (3, 7)^T \). In each replication, the estimates \( \hat{\theta}_{CC}, \hat{\theta}_{PC}^{(ig)} \), and \( \hat{\theta}_{PC}^{(full)} \) were computed using the general log-likelihood function, (3.2), (3.1) and (3.3), respectively, and we generated a holdout sample of size \( 0.2 \times n \), to assess the error rate via (2.4).

The results for the RE of \( R(\hat{\theta}_{PC}^{(\text{full})}) \) compared to \( R(\hat{\theta}_{CC}) \) and \( R(\hat{\theta}_{PC}^{(ig)}) \) are summarized in Table 3 and Table 4, respectively.

Table 3 reports that when the proportion of missing class labels is relatively high (over 0.65 in the case of \( \pi = (0.5, 0.35, 0.15)^T \) and over 0.74 in the case of \( \pi_1 = \pi_2 = \pi_3 \)), the values for the RE of \( R(\hat{\theta}_{PC}^{(\text{full})}) \) compared to \( R(\hat{\theta}_{CC}) \) are less than 1, namely, the classifier \( R(\hat{\theta}_{CC}) \) performs better than \( R(\hat{\theta}_{PC}^{(\text{full})}) \); when the proportion of missing class labels is relatively low (below 0.6 in the cases of both \( \pi = (0.5, 0.35, 0.15)^T \) and \( \pi_1 = \pi_2 = \pi_3 \)), most values for the RE of \( R(\hat{\theta}_{PC}^{(\text{full})}) \) compared to \( R(\hat{\theta}_{CC}) \) are greater than 1, indicating that the classifier \( R(\hat{\theta}_{PC}^{(\text{full})}) \) performs better than \( R(\hat{\theta}_{CC}) \).

Table 4 reports the results for the RE of \( R(\hat{\theta}_{PC}^{(\text{full})}) \) compared to \( R(\hat{\theta}_{PC}^{(ig)}) \). We find that the proportions of the missing label are between 0.25 and 0.9,
and all values of the RE of $R(\tilde{\theta}_{PC}^{(\text{full})})$ compared to $R(\tilde{\theta}_{PC}^{(\text{ig})})$ are smaller than 1, which implies that the classifier $R(\tilde{\theta}_{PC}^{(\text{full})})$ outperforms $R(\tilde{\theta}_{PC}^{(\text{ig})})$ on a partially classified sample.

Although we have simply explored the relationship between the simulated efficiency value and the proportion of missing class labels in a three-class normal model with unequal covariance matrices, the results are consistent with the conclusions from Figure 1 and 2 and Section 5.1.
6 Application

A multitude of practical applications, ranging from medical diagnostics to neuroscience investigations, primarily rely on human expertise for assigning class labels. This methodology of manual dataset annotation gives rise to a systematic mechanism for missingness. Intriguingly, when such missingness is accurately deciphered and leveraged, it can yield significant insights. Our research provides compelling evidence based on two authentic datasets: specifically, the Interneuron dataset and the Skin Lesion dataset. Recent literature, including works by [17,36,20,37], highlights the growing interest in these datasets due to their potential to advance understanding in neuroscience and dermatology, respectively. Our investigations reveal a significant correlation between the mechanism for missingness, responsible for the absence of some labels, and the concept of Shannon entropy. This novel finding has multidisciplinary implications across numerous fields, including but not limited to statistics, computer science, and healthcare. Furthermore, we demonstrated that a classifier accounting for this missingness mechanism exhibits superior performance compared to a classifier that disregards it. Intriguingly, it even surpasses the performance of a classifier built on a completely classified sample. These findings have considerable potential to benefit researchers and practitioners across diverse fields by improving classification accuracy and enhancing data utilization. By effectively addressing the issue of missingness in datasets, we catalyze the exploration of new horizons for data-driven insights and research potential. An example with simulated data, intended to demonstrate the application of our methodology, is available in the supplementary material.

6.1 Interneuron dataset

There is currently no unique catalog of cortical GABAergic interneuron types. [28] asked 48 leading neuroscientists to classify 320 interneurons by inspecting images of their morphology. Each neuroscientist used a web application to classify the interneurons, and each image was resized to $32 \times 32$ pixels. A total of 48 neuroscientists participated in the study. Here, we consider only the 42 neuroscientists who fully classified all 320 neurons. We consider 2 classes in this dataset: transcolumnar and intracolumnar. We discard those observations for which more than or equal to half of the 42 neuroscientists viewed an entry as not applicable to a given interneuron. Thus, we consider a total of $n = 304$ observations with a dimension of 1024. The ground truth for each image was generated by agreement among at least half of the 42 neuroscientists.

We select five neuroscientists who reviewed the images and 3D visualizations to determine whether an interneuron was transcolumnar or intracolumnar. We form a consensus labeling using the individual neuroscientists’ labels. Observations for which all five neuroscientists agreed are treated as labeled data, corresponding to $n_l = 173$. Observations for which fewer than five neuroscientists agreed are treated as unlabeled data, corresponding to $n_u = 131$. 
The dimensionality of the dataset is reduced to 3 features via principal component analysis (PCA) prior to classification. The decision to select $p = 3$ principal components was guided by the analysis of a scree plot, a graphical representation of the eigenvalues derived from the correlation or covariance matrix. The 'elbow' in the plot, where the subsequent eigenvalues diminish and become approximately equal, was identified at $p = 3$, hence substantiating our choice.

We fit a skew t-mixture model [19] to estimate the classification entropy of each observation. Figure 5 (a) compares the empirical cumulative distribution functions of the estimated entropy distributions in the labeled and unlabeled groups. Figure 5 (b) presents a Nadaraya–Watson kernel estimate of the probability of missing class labels. From Figure 5 (a), we find that the unlabeled observations typically have higher entropy than the labeled observations. Figure 5 (b) shows that the estimated probability of missing class labels decreases as the negative log entropy increases. This is in accordance with the relation of equation (3.4). The higher the entropy is, the higher the probability of missing class labels. Therefore, we fit a two-class normal model with unequal covariance matrices to compare the three classifiers $R(\hat{\theta}_{CC})$, $R(\hat{\theta}_{PC}^{(k)})$ and $R(\hat{\theta}_{PC}^{(full)})$.

We apply fivefold cross-validation (CV) to test the error rates of the three classifiers, splitting the interneuron dataset into a test set and a training set.

![Fig. 5](image)

**Fig. 5** Analysis of the interneuron dataset with regard to the relationship between the entropy and the labeled and unlabeled observations.

The dataset is randomly split into five subsets of equal sizes. Of the five subsets, a single subset is retained as the validation data for testing the model, and the remaining four subsets are used as the training data. This CV process...
Table 5 Summary statistics for the interneuron dataset. There are $n = 304$ observations with dimensions of $p = 3$. There are $g = 2$ classes (transcolumnar and intracolumnar).

| Classifier | $n$ | $p$ | $g$ | $n_c$ (classified) | $n_u$ (unclassified) | 5-fold CV error rate |
|------------|-----|-----|-----|------------------|---------------------|------------------|
| $R(\hat{\theta}_{CC})$ | 304 | 0   | 0   | 0                | 0                   | 0.2466           |
| $R(\hat{\theta}_{PC}^{(ig)})$ | 304 | 3   | 2   | 173              | 131                 | 0.3406           |
| $R(\hat{\theta}_{PC}^{(full)})$ | 173 | 131 |      | 0.2629           |

is repeated five times, with each of the five subsets used exactly once as the validation data. The five results are then averaged to produce a single estimate. The error rate summary is given in Table 5 and we find that the fivefold CV error rate from the partially classified sample based on the missingness mechanism is smaller than that obtained when ignoring the missingness mechanism and close to the fivefold CV error rate from the completely classified sample.

6.2 Skin lesion dataset

Dermatoscopic images can be accessed and downloaded from Harvard Dataverse provided by [34]. We select 3 classes e.g., benign keratosis (BKL), dermatofibroma (DF), and vascular skin lesions (VASC). The dataset includes a set of ground–truth labels of two types. Some labels have been confirmed by histopathology (performed by specialized dermatopathologists), while the rest have been confirmed either by follow-up, expert consensus, or in-vivo confocal microscopy. We treat the observations diagnosed by dermatopathologists, as labeled data and the rest of the cases as data with missing labels. Thus, the subset we consider contains $n = 1356$ observations, with $n_l = 888$ labeled observations and $n_u = 468$ unlabeled observations. To reduce the dimensionality of the dataset, we employed sparse linear discriminant analysis (LDA) [7] to select a subset of four features for model-based clustering. Employing LDA for dimensionality reduction in a skin lesion dataset was motivated by the availability of ground-truth labels. These labels served as response variables in LDA, facilitating the selection of four discriminative features, thereby reducing feature redundancy and enhancing subset quality.

Similarly, we fit a t-mixture model [19] to estimate the classification entropy of each observation. Figure 6(a) compares box-plots of the estimated entropies in the labeled and unlabeled groups. Figure 6(b) presents a Nadaraya–Watson kernel estimate of the probability of missing class labels. From Figure 6(a), we find that the unlabeled observations typically have higher entropy than the labeled observations. Figure 6(b) shows the estimated probability of missing class labels decreases as the negative log entropy increases. This is in accordance with the relation of equation (3.4). Therefore, we fit a three-class normal model with unequal covariance matrices to compare the three classifiers $R(\hat{\theta}_{CC}), R(\hat{\theta}_{PC}^{(ig)})$ and $R(\hat{\theta}_{PC}^{(full)})$. 
Estimating the Bayes Rule under a Specified Missing-Data Mechanism

Table 6 Summary statistics for the skin lesion dataset. There are $n = 1356$ observations with $p = 4$ dimensions. There are $g = 3$ classes (BKL, DF, and VASC).

| Classifier | $n$ | $p$ | $g$ | $n_c$ (classified) | $n_u$ (unclassified) | 5-fold CV error rate |
|------------|-----|-----|-----|-------------------|----------------------|---------------------|
| $R(\hat{\theta}_{CC})$ | 1356 | 0   | 0   | 0                 | 0                    | 0.1998              |
| $R(\hat{\theta}_{PC}^{(ig)})$ | 1356 | 4   | 3   | 888               | 468                  | 0.2005              |
| $R(\hat{\theta}_{PC}^{(full)})$ | 888  | 468 | 0.1969 |

We apply fivefold CV to test the error rates of three classifiers, splitting the skin lesion dataset into a test set and a training set. The error rate summary is given in Table 6. We find that the error rate from the partially classified sample based on the missingness mechanism is smaller than that obtained when the missingness mechanism is ignored and even better than that from the completely classified sample. But the difference between the error rates are very small, suggesting that the unlabeled observations must be easy to classify in the iterative fitting of the EM algorithm and entropies in Figure 6(a) are relatively low which can explain that $R(\hat{\theta}_{PC}^{(full)})$ does not do much better than $R(\hat{\theta}_{PC}^{(ig)})$ and is close to $R(\hat{\theta}_{CC})$.

Note that in using cross-validation to estimate the error rate of the various rules, the rule $R(\hat{\theta}_{CC})$ has an advantage over $R(\hat{\theta}_{PC})$ in their application to the unclassified feature vectors since $R(\hat{\theta}_{CC})$ is formed using the true values of the labels for the unclassified features. Bearing this in mind, the full rule based on the partially classified sample does well to be fairly close to that of the rule based on the completely classified sample in the former Interneuron dataset and even better in this Skin Lesion dataset.

7 Discussion

In the present study, we have delved into a statistical semi-supervised learning (SSL) methodology predicated on a finite normal mixture model, taking into account the MAR concept of missing data. The missingness mechanism is articulated as a logistic function to model the conditional probability of a label being missing, contingent on the log entropy of the observation. We have scrutinized the conditions under which the classifier $R_{PC}^{(full)}$ is more advantageous to use than $R_{CC}$ for a two-class normal homoscedastic model. Our findings suggest that the classifier $R_{PC}^{(full)}$ on a partially classified sample can surpass the completely supervised learning classifier $R_{CC}$ when the size of the overlap region ranges from moderate to small and the proportion of missing class labels ranges from relatively high to low. Alternatively, this can also occur when the overlap region is large but the proportion of missing labels is relatively low. Furthermore, the classifier $R_{PC}^{(full)}$ consistently outperforms $R_{PC}^{(ig)}$ irrespective of the extent of overlap or the proportion of missing class labels. We have conducted simulation experiments to extend the analysis to
two-class and three-class normal mixture models with unequal covariance matrices, and the results further corroborate the above analysis. We have also compared the three classifiers on two real datasets, and the results align with our expectations and the theoretical results.

While our study provides preliminary evidence in support of the efficacy of the proposed semi-supervised learning approach, it is important to acknowledge the limitation of the size of the real-world datasets utilized. The datasets were manually labeled by domain experts, a process that is both labor-intensive and costly, thus limiting the scale of data available for our study. Moreover, we designed our analyses to focus on a particular type of missingness mechanism. To mitigate this limitation, we included a simulated example based on synthetic datasets generated to mimic the specific missingness mechanism in the supplementary. The results from these additional experiments are consistent with those derived from the real-world datasets, thereby enhancing the robustness and credibility of our findings. Future research would benefit from the inclusion of larger and more diverse datasets to validate the generalizability and robustness of the proposed semi-supervised learning method.

In light of the current research landscape, there has been a recent resurgence of interest in SSL in the machine learning community due to significant empirical progress on benchmark image and text classification datasets. However, from a statistical perspective, the theoretical understanding of these modern SSL algorithms is still limited [2]. One of the most intuitive approaches to SSL is self-training, an iterative method for learning with alternating steps between generating pseudo-labels for the unlabeled observations and then
training a classifier using both the labeled and pseudo-labeled data. Theoretical analysis of SSL has been limited, but recent work by [1] provided an asymptotic basis on how to increase the accuracy of the commonly used linear discriminant function formed from a partially classified sample in SSL. This increase in accuracy can be of sufficient magnitude for this SSL-based classifier to have a smaller error rate than that if it were formed from a completely classified sample. Given the scarcity of theoretical analysis in this area, future work will encompass mathematical and empirical studies of more complex models. This will include exploring the potential of the missingness mechanism to provide additional information and improve classification accuracy, as well as investigating the conditions under which these gains can be achieved.

References
1. Ahfock, D., McLachlan, G.J.: An apparent paradox: a classifier based on a partially classified sample may have smaller expected error rate than that if the sample were completely classified. Statistics and Computing pp. 1–12 (2020)
2. Ahfock, D., McLachlan, G.J.; Semi-supervised learning of classifiers from a statistical perspective: A brief review. Econometrics and Statistics 26, 124–138 (2023)
3. Bartlett, M., Please, N.; Discrimination in the case of zero mean differences. Biometrika 50(1/2), 17–21 (1963)
4. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory, pp. 92–100 (1998)
5. Chapelle, O., Schölkopf, B., Zien, A.: Semi-supervised learning, adaptive computation and machine learning. MIT Press, Cambridge, MA, USA. Cited in page(s) 21(1), 2 (2010)
6. Chawla, N.V., Karakoulas, G.: Learning from labeled and unlabeled data: An empirical study across techniques and domains. Journal of Artificial Intelligence Research 23, 331–366 (2005)
7. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics 53(4), 406–413 (2011)
8. Côme, E., Oukhellou, L., Denoeux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern recognition 42(3), 334–348 (2009)
9. Efron, B.: The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association 70(352), 892–898 (1975)
10. Fujino, A., Ueda, N., Saito, K.: Semisupervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(3), 424–437 (2008)
11. Gilbert, E.S.: The effect of unequal variance-covariance matrices on fisher’s linear discriminant function. Biometrics pp. 505–515 (1969)
12. Han, C.P.: Distribution of discriminant function when covariance matrices are proportional. The Annals of Mathematical Statistics 40(3), 979–985 (1969)
13. Hawkins, D.M., Raath, E.L.: An extension of geisser’s discrimination model to proportional covariance matrices. The Canadian Journal of Statistics/La Revue Canadienne de Statistique pp. 261–270 (1982)
14. Huang, J.T., Hasegawa-Johnson, M.: Semi-supervised training of gaussian mixture models by conditional entropy minimization. In: Eleventh Annual Conference of the International Speech Communication Association (2010)
15. Joachims, T., et al.: Transductive inference for text classification using support vector machines. In: Icml, vol. 99, pp. 200–209 (1999)
16. Kim, S.C., Kang, T.J.: Texture classification and segmentation using wavelet packet frame and gaussian mixture model. Pattern recognition 40(4), 1207–1221 (2007)
17. Kołosowska, K., Schratt, G., Winterer, J.: microRNA-dependent regulation of gene expression in gabaergic interneurons. Frontiers in Cellular Neuroscience 17 (2023)
18. Lachnit, G.R., Cristianini, N., Bartlett, P., Ghahoui, L.E., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. Journal of Machine learning research 5(Jan), 27–72 (2004)
19. Lee, S., McLachlan, G.J.: Finite mixtures of multivariate skew t-distributions: some recent and new results. Statistics and Computing 24(2), 181–202 (2014)
20. Lucius, M., All, J.D., All, J.A.D., Belvisi, M., Radizza, L., Lanfranconi, M., Lorenzetti, V., Galmarini, C.M.: Deep neural frameworks improve the accuracy of general practitioners in the classification of pigmented skin lesions. Diagnostics 10 (2020)
21. Liu, Z., Ahfock, D., Thompson, R., McLachlan, G.J.: gmmsslm: Semi-supervised gaussian mixture modeling with a missing data mechanism in r. arXiv preprint arXiv:2302.12306 (2023)
22. Marks, S., Dunn, O.J.: Discriminant functions when covariance matrices are unequal. Journal of the American Statistical Association 69(346), 555–559 (1974)
23. McLachlan, G.J.: Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. Journal of the American Statistical Association 70(350), 365–369 (1975)
24. McLachlan, G.J.: Some expected values for the error rates of the sample quadratic discriminant functions. Australian Journal of Statistics 17(3), 161–165 (1975)
25. McLachlan, G.J.: Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. Journal of the American statistical association 72(358), 403–406 (1977). DOI 10.1080/01621459.1977.10481009
26. McLachlan, G.J., Gordon, R.: Mixture models for partially unclassified data: a case study of renal venous renin in hypertension. Statistics in Medicine 8(10), 1291–1300 (1989). DOI 10.1002/sim.4780081012
27. Mealli, F., Rubin, D.B.: Clarifying missing at random and related definitions, and implications when coupled with exchangeability. Biometrika 102(4), 995–1000 (2015)
28. Mihaljević, B., Benavides-Piccione, R., Bielza, C., Larrañaga, P., DeFelipe, J.: Classification of gabaergic interneurons by leading neuroscientists. Scientific data 6(1), 1–6 (2019)
29. O’Neill, T.J.: Normal discrimination with unclassified observations. Journal of the American Statistical Association 73(364), 821–826 (1978)
30. Pan, W., Shen, X., Jiang, A., Hebbel, R.P.: Semi-supervised learning via penalized mixture model with application to microarray sample classification. Bioinformatics 22(19), 2388–2395 (2006)
31. Rubin, D.B.: Inference and missing data. Biometrika 63(3), 581–592 (1976)
32. Szczurek, E., Biecek, P., Tiuryn, J., Vingron, M.: Introducing knowledge into differential expression analysis. Journal of Computational Biology 17(8), 953–967 (2010)
33. Szummer, M., Jaakkola, T.: Partially-labeled classification with markov random walks. Advances in neural information processing systems 14 (2001)
34. Tschandl, P.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions (2018)
35. Vapnik, V.: The support vector method of function estimation pp. 55–85 (1998)
36. Volo, M.D., Destexhe, A.: Optimal responsiveness and information flow in networks of heterogeneous neurons. Scientific Reports 11, 17611 (2021). DOI 10.1038/s41598-021-96745-2
37. Wang, Y., Wang, T.t., Montero-Pedrazuela, A., Guadaño-Ferraz, A., Rausell, E.: Thyroid Hormone Transporters MCT8 and OATP1C1 Are Expressed in Pyramidal Neurons and Interneurons in the Adult Motor Cortex of Human and Macaque Brain. International Journal of Molecular Sciences 24 (2023)
38. Zhou, D., Bousquet, O., Lal, T., Weston, J., Schölkopf, B.: Learning with local and global consistency. Advances in neural information processing systems 16 (2003)