Abstract:
There has been increasing interest in using semi-supervised learning to form a classifier. As is well known, the (Fisher) information in an unclassified feature with unknown class label is less (considerably less for weakly separated classes) than that of a classified feature which has known class label. Hence assuming that the labels of the unclassified features are randomly missing or their missing-label mechanism is simply ignored, the expected error rate of a classifier formed from a partially classified sample is greater than that if the sample were completely classified. We propose to treat the labels of the unclassified features as missing data and to introduce a framework for their missingness in situations where these labels are not randomly missing. An examination of several partially classified data sets in the literature suggests that the unclassified features are not occurring at random but rather tend to be concentrated in regions of relatively high entropy in the feature space. Here in the context of two normal classes with a common covariance matrix we consider the situation where the missingness of the labels of the unclassified features can be modelled by a logistic model in which the probability of a missing label for a feature depends on its entropy. Rather paradoxically, we show that the classifier so formed from the partially classified sample may have smaller expected error rate than that if the sample were completely classified.

1 Introduction

We consider the problem of forming a classifier from training data that are not completely classified. That is, the feature vectors \( y_j \) in the training sample have all been observed but their class labels are missing for some of them and so the training data constitute a partially classified sample denoted here by \( x_{PC} \). This problem goes back at least to the mid-seventies (McLachlan, 1975), and it received a boost shortly afterwards with the advent of the EM algorithm (Dempster et al., 1977) which could be applied to carry out maximum likelihood (ML) estimation for a partially classified sample. These days increasing attention is being given to the formation of classifiers on the basis of a partially classified sample (or semi-supervised learning (SSL) as it is referred to in the machine learning literature), particularly in situations where unclassified data are available more freely or more cheaply or both than classified data. Moreover, in some instances in the field of medical diagnosis, a definitive classification can only be made via an invasive procedure that may not be ethical to apply unless there is a high degree of confidence that the patient has the disease for which screening is being performed. There is now a wide literature on SSL techniques (for example, Grandvalet and Bengio (2005) and Berthelot et al. (2019)), which are too numerous to discuss here.

In SSL, it is usually assumed that the labels of the unclassified features are randomly missing or the missing-label mechanism is simply ignored. We propose a joint modelling
framework that introduces a missing-label mechanism for the missing-label indicators which are treated as random variables. Our examination of a number of real datasets shows that the pattern of missing labels is typically related to the difficulty of classification, which can be quantified by the Shannon entropy. This relationship can be captured using a logistic selection model. Full likelihood inference that includes the missing-label mechanism can improve the efficiency of parameter estimation and increase classification accuracy to the extent where it can be greater than if the sample were completely classified.

More specifically, we let $m_j$ be the missing-label indicator being equal to 1 if the $j$th feature vector in the training sample is unclassified; that is, its class label is missing. In the case of a partially classified training sample $x_{PC}$ in the context of the two-class normal discrimination problem, O’Neill (1978) showed that the information about the vector $\beta$ of discriminant function coefficients using the likelihood that ignores the mechanism for the missing labels can be decomposed as

$$I_{PC}^{(ig)}(\beta) = I_{CC}(\beta) - \overline{m}I_{CC}^{(lr)}(\beta),$$

(1)

where $I_{CC}(\beta)$ is the information about $\beta$ in a completely classified sample $x_{CC}$, $I_{CC}^{(chr)}(\beta)$ is the information about $\beta$ under the logistic regression model for the distribution of the class labels given the features in $x_{CC}$, and $\overline{m} = \sum_{j=1}^{n} m_j/n$ is the proportion of unclassified features in the partially classified sample $x_{PC}$. It can be seen from (1) that the loss of information due to the sample being partially classified is equal to $\overline{m}I_{CC}^{(lr)}(\beta)$.

The consequent decrease in the efficiency in estimating the Bayes’ rule can be considerable as illustrated in Table 1 in Section 5.

With our proposed approach, we introduce the random variable $M_j$ corresponding to the realized value $m_j$ for the missing-label indicator for the feature vector $y_j$ and model its distribution to depend on an entropy-based measure. We then consider the estimation of $\beta$ from the partially classified sample $x_{PC}$ on the basis of the so-called full likelihood $L_{PC}^{(full)}(\theta)$ whose logarithm is augmented by the addition of the log likelihood for $\beta$ formed under the proposed logistic model for the missing-label indicator random variable $M_j$. We then show that the information about $\beta$ for the full likelihood formed from the partially classified sample $x_{PC}$ is given by

$$I_{PC}^{(full)}(\beta) = I_{CC}(\beta) - \gamma I_{CC}^{(lr)}(\beta) + I_{PC}^{(miss)}(\beta),$$

(2)

where $I_{CC}^{(chr)}(\beta)$ is the conditional information about $\beta$ under the logistic regression model fitted to the class labels in $x_{CC}$, $I_{PC}^{(miss)}(\beta)$ is the information about $\beta$ in the missing-label indicators $m_j$, and $\gamma$ is the expected proportion of missing class labels in the partially classified sample. It can be seen from (2) that if

$$I_{PC}^{(miss)}(\beta) > \gamma I_{CC}^{(chr)}(\beta),$$

then there is actually an increase in the information about $\beta$ in the partially classified sample over the information $I_{CC}(\beta)$ about $\beta$ in the completely classified sample. Here, the inequality in the above equation is used in the sense that the left-hand side of the equation, minus the right, is positive definite. Following on from Ahfock and McLachlan (2019a), we shall show that under certain conditions on the distribution of the missing labels that the consequent reduction in the asymptotic expected error rate of the Bayes’ rule learnt using the partially classified sample is lower than that of the Bayes’ rule learnt using a completely classified sample. Some Monte Carlo simulations are to be given to support the asymptotic theory.
2 Two-Class Normal Discrimination

In discriminant analysis, the aim is to assign an unclassified entity with \( p \)-dimensional feature vector \( y \) to one of a number of \( g \) classes \( C_1, \ldots, C_g \). It is assumed that the random vector \( Y \) corresponding to \( y \) has density \( f_i(y; \omega_i) \) in \( C_i \), specified up to an unknown vector of parameters \( \omega_i \) \((i = 1, \ldots, g)\). We consider here the case of \( g = 2 \) classes for which \( f_i(y; \omega_i) \) denotes the multivariate normal density with mean \( \mu_i \) and covariance matrix \( \Sigma(i = 1, 2) \). We let \( \theta \) be the vector containing the mixing proportion \( \pi_1 \), the \( 2p \) elements of the means \( \mu_1 \) and \( \mu_2 \), and the \( \frac{1}{2}p(p + 1) \) elements of the common class-covariance matrix \( \Sigma \) known \emph{a priori} to be distinct.

We let \( R(y; \theta) \) denote the Bayes’ (optimal) rule of allocation, where \( R(y; \theta) = h \), that is, \( y \) is allocated to \( C_h \), if

\[
h = \arg \max_i \tau_i(y; \beta),
\]

where

\[
\tau_1(y; \beta) = 1 - \tau_2(y; \beta) = \pi_1 f_1(y; \omega_1) / f_y(y; \theta) = \exp(\beta_0 + \beta_1^T y) / \{1 + \exp(\beta_0 + \beta_1^T y)\}
\]

(3)

is the posterior probability that \( y \) belongs to \( C_1 \) given \( Y = y \); see, for example, \textcite{McLachlan1992}, Chapter 1). Here \( f_y(y; \theta) = \sum_{i=1}^2 \pi_i f_i(y; \omega_i) \) is the marginal (mixture) density of \( Y \) and \( \beta = (\beta_0, \beta_1^T) \) is the vector of discriminant function coefficients, where

\[
\beta_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2), \quad \beta_1 = \Sigma^{-1}(\mu_1 - \mu_2).
\]

It can be seen from (3) that the Bayes’ rule reduces in this case of \( g = 2 \) normal classes with a common covariance matrix to depending only on \( \beta \) with \( R(y; \theta) \) being equal to 1 or 2, according as the discriminant function

\[
d(y; \beta) = \beta_0 + \beta_1^T y
\]

is greater or less than zero.

We henceforth adopt the canonical form

\[
\mu_1 = -\mu_2 = (\frac{1}{2} \Delta, 0, \ldots, 0)^T, \quad \Sigma = I_p,
\]

(4)

where \( \Delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) \) is the Mahalanobis squared distance between the two classes and \( I_p \) is the \( p \times p \) identity matrix.

In practice, \( \beta \) has to be estimated from available training data. We let \( x_{CC} = (x_1^T, \ldots, x_n^T)^T \) denote \( n \) independent realizations of \( X = (Y^T, Z)^T \) as the completely classified training data, where \( Z \) denotes the class membership of \( Y \), being equal to 1 if \( Y \) belongs to \( C_1 \), and zero otherwise. We let \( m_j \) be the missing-label indicator being equal to 1 if \( z_j \) is missing and zero if it is available (\( j = 1, \ldots, n \)). Accordingly, the unclassified sample \( x_{PC} \) is given by those members \( x_j \) in \( x_{CC} \) for which \( m_j = 0 \) and only the feature vectors \( y_j \) without their class labels \( z_j \) for those members in \( x_{CC} \) for which \( m_j = 1 \). It should be noted that in our notation to denote the various information matrices \( I(\cdot) \) about a parameter, we only display that parameter in the argument of \( I(\cdot) \), although \( I \)
may depend also on other parameters, including those in the distribution adopted for the missing-label indicators.

With our proposed approach to exploiting the potential information in the missing-label indicators $m_j$, we introduce the random variable $M_j$ corresponding to the realized value $m_j$ for the missing-class label for the feature vector $y_j$ and model its distribution to depend on an entropy-based measure.

3 Mechanism for Missing Class Labels

In many applications, the class labels $z_j$ are often assigned by domain experts, who may not be able to make a confident classification for every feature. As a motivating example for our approach to the formulation of a model for the distribution of the missing-label indicator $M_j$, we present Figure 1, which shows a manually classified flow cytometry dataset from Aghaeepour et al. (2013). Black squares correspond to unclassified features, and the majority of the unclassified features appear to be located near class boundaries. Plots of other such datasets may be found in Ahfock and McLachlan (2019).

A standard approach in semi-supervised learning is to ignore the underlying cause in forming the likelihood from the partially classified dataset. We shall denote this likelihood by $L_{PC}^{(ig)}(\theta)$ with logarithm given by

$$\log L_{PC}^{(ig)}(\theta) = \sum_{j=1}^{n} (1 - m_j) \sum_{i=1}^{2} z_{ij} \log \{ \pi_i f_i(y_j; \omega_i) \} + \sum_{j=1}^{n} m_j \log f_y(y_j; \theta),$$

(5)

where $z_{1j} = 1 - z_{2j} = z_j$ ($j = 1, \ldots, n$).

Note that the log of the likelihood $L_{CC}(\theta)$ for the completely classified sample $x_{CC}$ is given by (5) with all $m_j = 0$.

The missingness of class labels can be ignored in forming the likelihood function for $\theta$ in the case of missing completely at random (MCAR) and for the less restrictive situation
of missing at random (MAR). However, in the latter situation, the (Fisher) information will be affected by ignoring the missingness (McLachlan and Gordon, 1989).

If classification difficulty is a cause of the missing labels, the use of $L_{PC}^{(ig)}(\theta)$ may be suboptimal. In such circumstances, the unlabelled features are likely to lie near class boundaries, and then the pattern of missing labels carries extra information for the estimation of $\theta$ that is not reflected in (5). The missing-data framework pioneered by Rubin (1976) is useful to exploit the potential information in the missing-label pattern in the situation of a partially classified training sample $x_{PC}$. We introduce the missing-label indicator random variable $M_j$ with realized value $m_j (j = 1, \ldots, n)$. An important measure of classification difficulty is the Shannon entropy of the posterior class probabilities. Let $e_j$ denote the entropy for $y_j$,

$$e_j = -\sum_{i=1}^{2} \tau_i(y_j; \beta) \log \tau_i(y_j; \beta).$$

Under our proposed missing-label model, we have that

$$\Pr\{M_j = 1 \mid y_j, z_j\} = \Pr\{M_j = 1 \mid y_j\} = q(y_j; \beta, \xi),$$

where the parameter $\xi$ is distinct from $\beta$.

An obvious choice for the function $q(y_j; \beta, \xi)$ is the logistic model (Molenberghs et al., 2014),

$$q(y_j; \beta, \xi) = \exp \left\{ \xi_0 + \xi_1 e_j \right\} / \left[ 1 + \exp \left\{ \xi_0 + \xi_1 e_j \right\} \right].$$

The expected proportion $\gamma(\Psi)$ of unclassified features in a partially classified sample $x_{PC}$ is given by

$$\gamma(\Psi) = \sum_{j=1}^{n} E(M_j)/n = E[\Pr\{M_j = 1 \mid Y_j\}] = E\{q(Y; \beta, \xi)\},$$

where $\Psi = (\theta^T, \xi^T)^T$.

To simplify the numerical computation in the particular case of only $g = 2$ classes as under consideration, we henceforth replace $e_j$ in (8) by the square of the discriminant function $d(y_j; \beta)$ to give

$$q(y_j; \beta, \xi) = \frac{\exp \{\xi_0 + \xi_1 d(y_j; \beta)^2\}}{1 + \exp \{\xi_0 + \xi_1 d(y_j; \beta)^2\}}.$$
mechanism. Five hundred values of $x$ were simulated from the canonical model with $\pi_1 = \pi_2$ and $\Delta = 2$. The missingness model was then applied to the simulated features with $\xi_0 = 3$ and $\xi_1 = -0.1, -0.5, -1, -2, -5, -10$. Black squares denote unclassified features, red triangles are features in Class $C_1$, and blue circles are features in $C_2$. Moving through the Panels (a) to (f), the unclassified features become more concentrated around the decision boundary as $\xi_1$ decreases. The proportion of unclassified features is different in each panel.

The full likelihood function $L^{\text{full}}_{PC}(\Psi)$ for $\Psi$ that can be formed from the partially classified sample $x_{PC}$ is defined by

$$
\log L^{\text{full}}_{PC}(\Psi) = \log L^{(ig)}_{PC}(\theta) + \log L^{(\text{miss})}_{PC}(\beta, \xi),
$$

where $\log L^{(ig)}_{PC}(\theta)$ is defined by (5) and where

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

is the log likelihood function for $\beta$ and $\xi$ formed on the basis of the missing-label indicators $m_j$ ($j = 1, \ldots, n$).

![Figure 2: Example simulated datasets using the canonical normal discriminant model and the missingness model (10). Here $n = 500, \Delta = 2, \pi_1 = \pi_2, p = 2, \xi_0 = 3$. In panels (a) through (f) $\xi_1 = -0.1, -0.5, 1, -2, -5, -10$, respectively.](image)

We note that there may be an identifiability issue concerning $\beta$ and $\xi$ if $\log L^{(\text{miss})}_{PC}(\beta, \xi)$ given by (10) were to be used on its own for the estimation of $\beta$ and $\xi$. But as it is being combined with $\log L^{(ig)}_{PC}(\theta)$ to form the full log likelihood $\log L^{\text{full}}_{PC}(\Psi)$, $\beta$ and $\xi$ are each identifiable with the use of the latter.
4 Fisher Information

In this section, we derive the Fisher information about $\beta$ in the partially classified sample $x_{PC}$. We reparameterize the two-class normal model by taking

$$\theta = (\theta^T, \beta)^T,$$

where $\theta_1$ contains the elements of $\mu = \pi_1\mu_1 + \pi_2\mu_2$ and the distinct elements of $\Lambda = \Sigma + \pi_1\pi_2(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$. We can now write the vector $\Psi$ of all unknown parameters, including the parameter $\xi$ in the logistic model defined by (10), as

$$\Psi = (\theta^T, \xi^T)^T = (\theta^T_1, \beta^T, \xi^T)^T.$$ 

Theorem 1 (Main Result). The Fisher information about $\beta$ in the partially classified sample $x_{PC}$ via the full likelihood function $L_{PC}^{(full)}(\Psi)$ can be decomposed as

$$I_{PC}^{(full)}(\beta) = I_{CC}(\beta) - \gamma(\Psi)I_{CC}^{(clr)}(\beta) + I_{PC}^{(miss)}(\beta),$$

where $I_{CC}(\beta)$ is the information about $\beta$ in the completely classified sample $x_{CC}$, $I_{CC}^{(clr)}(\beta)$ is the conditional information about $\beta$ under the logistic regression model for the distribution of the class labels given the features in $x_{CC}$, and $I_{PC}^{(miss)}(\beta)$ is the information about $\beta$ in the missing-label indicators under the assumed logistic model for their distribution given their associated features in the partially classified sample $x_{PC}$.

Remark 1. Since $\gamma(\Psi)$ is the probability that $M = 1$, it follows that the second term on the right-hand side of (15), $\gamma(\Psi)I_{CC}^{(clr)}(\beta)$, can be expressed as

$$nE\{-\partial^2 \log f_{Z|Y}(Z | Y; \beta)/\partial \beta \partial \beta^T\} q(Y; \beta, \xi),$$

which is the expected information (under the logistic model) for those class labels $z_j$ in $x_{CC}$ for which their associated features $y_j$ would have missing labels $m_j = 1$ under the assumed model (10) for missingness.

Proof of Theorem 1. From the definition (11) of the full log likelihood function $\log L_{PC}^{(full)}(\Psi)$, we can decompose the information matrix $I_{PC}^{(full)}(\Psi)$ for $\Psi$ as

$$I_{PC}^{(full)}(\Psi) = E\{-\partial^2 \log L_{PC}^{(full)}(\Psi)/\partial \Psi \partial \Psi^T\} = E\{-\partial^2 \log L_{PC}^{(ig)}(\theta)/\partial \Psi \partial \Psi^T\}
+ E\{-\partial^2 \log L_{PC}^{(miss)}(\beta, \xi)/\partial \Psi \partial \Psi^T\} = I_{PC}^{(ig)}(\Psi) + I_{PC}^{(miss)}(\Psi),$$

where

$$I_{PC}^{(ig)}(\Psi) = E\{-\partial^2 \log L_{PC}^{(ig)}(\theta)/\partial \Psi \partial \Psi^T\}$$

and

$$I_{PC}^{(miss)}(\Psi) = E\{-\partial^2 \log L_{PC}^{(miss)}(\beta, \xi)/\partial \Psi \partial \Psi^T\}.\) 

Considering the first term on the right-hand side of (16), we consider its submatrix

$$I_{PC}^{(ig)}(\theta) = E\{-\partial^2 \log L_{PC}^{(ig)}(\theta)/\partial \theta \partial \theta^T\}.$$
It can be expressed as
\[
\begin{align*}
&I_{\text{PC}} = n[1 - \gamma(\Psi)] E \{ -\partial^2 \log f_{yz}(Y; \theta) / \partial \theta \partial \theta^T \mid M = 0 \} \\
&+ n \gamma(\Psi) E \{ -\partial^2 \log f_{y}(Y; \theta) / \partial \theta \partial \theta^T \mid M = 1 \},
\end{align*}
\]
where \( f_{yz}(y, z; \theta) \) denotes the joint density of \( Y \) and \( Z \)
and
\[
 f_y(y; \theta) = \frac{f_{yz}(y, z; \theta)}{f_z(z \mid y; \beta)}
\]
is the marginal density of \( Y \), and where \( f_{z}(z \mid y; \beta) \) is the conditional probability of \( Z \) given \( Y = y \).

On using (21) in (20), we can write \( I_{\text{PC}}(\theta) \) as
\[
I_{\text{PC}} = I_{\text{CC}}(\theta) - \gamma(\Psi) I_{\text{CLR}}(\theta),
\]
which equals
\[
\begin{align*}
&n E \{ -\partial^2 \log f_{yz}(Y, Z; \theta) / \partial \theta \partial \theta^T \mid M = 1 \} \\
&- \gamma(\Psi) n E \{ -\partial^2 \log f_{z}(Z \mid Y; \theta) / \partial \beta \partial \beta^T \mid M = 1 \},
\end{align*}
\]
and so
\[
I_{\text{PC}} = I_{\text{CC}}(\theta) - \gamma(\Psi) I_{\text{CLR}}(\theta),
\]
where
\[
I_{\text{CC}}(\theta) = n E \{ -\partial^2 \log f_{yz}(Y, Z; \theta) / \partial \theta \partial \theta^T \}
= E \{ -\partial^2 \log L_{\text{CC}}(\theta) / \partial \theta \partial \theta^T \}
\]
is the information about \( \theta \) in the completely classified sample and where, corresponding to the partition in (12) of \( \theta \),
\[
I_{\text{CLR}}(\beta) = \left( \begin{array}{cc} O & O \\ O & I_{\text{CC}}(\beta) \end{array} \right),
\]
since the likelihood function for the logistic regression model does not contain \( \theta_1 \). Here
\[
I_{\text{CLR}}(\beta) = n E \{ -\partial^2 \log f_{z}(Z \mid Y; \theta) / \partial \beta \partial \beta^T \mid M = 1 \}
\]
is the expectation conditional on \( M = 1 \) of the negative Hessian of the conditional density of \( Z \) given \( Y \) under the logistic regression model fitted to the completely classified sample.

On considering now the first term on the right-hand side of (22), we have that the information about \( \theta \) in the completely classified sample can be partitioned as
\[
I_{\text{CC}}(\theta) = E \{ -\partial^2 \log L_{\text{CC}}(\theta) / \partial \theta \partial \theta^T \}
= \left( \begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array} \right),
\]
where this partition of $I_{CC}(\theta)$ corresponds to the partition \{12\} of $\theta$. We partition the inverse of $I_{CC}(\theta)$ as

$$I_{CC}^{-1}(\theta) = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

to give the asymptotic covariance matrix of the ML estimator of $\theta$.

It follows that the information matrix for $\beta$ based on the likelihood formed from the completely classified sample is given by the inverse of $A^{22}$,

$$I_{CC}(\beta) = (A^{22})^{-1} = A_{22} - A_{21}A_{11}^{-1}A_{12}. \quad (27)$$

As $L_{PC}^{(ig)}(\theta)$ does not contain $\xi$, it follows from (24) and (27) that the first term on the right-side of (22) for the information matrix $I_{PC}^{(full)}(\Psi)$ can be partitioned corresponding to the partition \{14\} of $\Psi$ as

$$I_{PC}^{(ig)}(\Psi) = \begin{pmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} - \gamma(\Psi)I_{CC}^{(clr)}(\beta) & 0 \\ 0 & 0 & 0 \end{pmatrix}. \quad (28)$$

On considering the other term $I_{PC}^{(miss)}(\Psi)$ on the right-hand side of (16) for the information matrix about $\Psi$ via the full likelihood function $L_{PC}^{(full)}(\Psi)$, it can be partitioned corresponding to the partition \{14\} of $\Psi$ as

$$I_{PC}^{(miss)}(\Psi) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & B_{22} & B_{23} \\ 0 & B_{32} & B_{33} \end{pmatrix}, \quad (29)$$

since $L_{PC}^{(miss)}(\beta, \xi)$ does not contain $\theta_1$.

On using (29) and (28) in (16), we have that the information matrix $I_{PC}^{(full)}(\Psi)$ for $\Psi$ on the basis of the full likelihood $L_{PC}^{(full)}(\Psi)$ fitted to the partially classified sample $x_{PC}$ can be partitioned as

$$I_{PC}^{(full)}(\Psi) = \begin{pmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} - \gamma(\Psi)I_{CC}^{(clr)}(\beta) + B_{22} & B_{23} \\ 0 & B_{32} & B_{33} \end{pmatrix}. \quad (30)$$

Corresponding to this partition of $I_{PC}^{(full)}(\Psi)$, we write it as

$$H = \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & H_{33} \end{pmatrix}, \quad (31)$$

and we let $H^{ij}$ denote the block in $H^{-1}$ corresponding to the block $H_{ij}$ in $H$ ($i, j = 1, 2, 3$).

The inverse of the matrix $H^{22}$ provides the information matrix $I_{PC}^{(full)}(\beta)$ for $\beta$, where $\beta$ is estimated by consideration of the full likelihood function $L_{PC}^{(full)}(\Psi)$. To calculate $H^{22}$, we refine the partition \{31\} of $H$ to

$$H = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \quad (32)$$
where

\[
W_{11} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}, \quad W_{12} = W_{21}^T = \begin{pmatrix} 0 \\ B_{23} \end{pmatrix},
\]

and \( W_{22} = B_{33} \). Using standard results for the inversion of matrices in block form, we have that

\[
\begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} = (W_{11} - W_{12}W_{22}^{-1}W_{21})^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & H_{22} - B_{23}B_{33}^{-1}B_{32} \end{pmatrix}^{-1}.
\]

(33)

Now \( I_{\text{PC}}^{(\text{full})}(\beta) = \{H_{22}\}^{-1} \), which can be calculated from (33) to give

\[
\{H_{22}\}^{-1} = H_{22} - B_{23}B_{33}^{-1}B_{32} = A_{21}A_{11}^{-1}A_{12} = (A_{22} - A_{21}A_{11}^{-1}A_{12}) + \gamma(\Psi)I_{\text{PC}}^{(\text{chr})}(\beta) = I_{\text{CC}}(\beta) - \gamma(\Psi)I_{\text{Psi}}^{(\text{chr})}(\beta) + I_{\text{PC}}^{(\text{miss})}(\beta),
\]

on noting (27) and that

\[
I_{\text{PC}}^{(\text{miss})}(\beta) = B_{22} - B_{23}B_{33}^{-1}B_{32}
\]

is the information about \( \beta \) in the missing-label indicators.

**Remark 2.** Note that the contribution \( I_{\text{PC}}^{(\text{miss})}(\beta) \) to the full information matrix would be equal to \( B_{22} \) if \( \xi \) were known, so the term

\[
B_{23}B_{33}^{-1}B_{32}
\]

can be viewed as the loss of information about \( \beta \) by virtue of \( \xi \) not being known and having to be estimated as well as \( \beta \).

5 **Asymptotic Relative Efficiencies**

We let (i) \( \hat{\beta}_{\text{CC}} \) denote the maximum likelihood (ML) estimate of \( \beta \) by consideration of the likelihood function \( L_{\text{CC}}(\theta) \) that can be formed from the completely classified sample \( x_{\text{CC}} \); (ii) \( \hat{\beta}_{\text{PC}}^{(\text{ig})} \) denote the ML estimate of \( \beta \) on the basis of the likelihood function \( L_{\text{PC}}^{(\text{ig})}(\theta) \) formed from the partially classified sample \( x_{\text{PC}} \) by ignoring the missingness in the labels of the unclassified features; (iii) \( \hat{\beta}_{\text{PC}}^{(\text{full})} \) denote the ML estimate of \( \beta \) by consideration of the full likelihood function \( L_{\text{PC}}^{(\text{full})}(\Psi) \).

We let \( \hat{R}_{\text{CC}}, \hat{R}_{\text{PC}}^{(\text{ig})}, \) and \( \hat{R}_{\text{PC}}^{(\text{full})} \) denote the estimated Bayes’ rule obtained by plugging in the estimates \( \hat{\beta}_{\text{CC}}, \beta_{\text{PC}}^{(\text{ig})}, \) and \( \hat{\beta}_{\text{PC}}^{(\text{full})} \), respectively, for \( \beta \) in the Bayes’ rule \( R(y; \beta) \).

The overall error rate of the Bayes’ rule \( R(y; \beta) \) is denoted by \( \text{err}(\beta) \) (the optimal error rate). The conditional error rates of the estimated Bayes’ rules \( R_{\text{CC}}, R_{\text{PC}}^{(\text{ig})}, \) and \( R_{\text{PC}}^{(\text{full})} \) are denoted by \( \text{err}(\hat{\beta}_{\text{CC}}), \text{err}(\hat{\beta}_{\text{PC}}^{(\text{ig})}), \) and \( \text{err}(\hat{\beta}_{\text{PC}}^{(\text{full})}) \), respectively. The asymptotic relative
Table 1: Asymptotic relative efficiency of $\hat{R}_{\text{PC}}^{(\text{ig})}$ compared to $\hat{R}_{\text{CC}}$

| $\pi_1$ | $\Delta = 1$ | $\Delta = 2$ | $\Delta = 3$ | $\Delta = 4$ |
|--------|--------------|--------------|--------------|--------------|
| 0.1    | 0.0036       | 0.0591       | 0.2540       | 0.5585       |
| 0.2    | 0.0025       | 0.0668       | 0.2972       | 0.6068       |
| 0.3    | 0.0027       | 0.0800       | 0.3289       | 0.6352       |
| 0.4    | 0.0038       | 0.0941       | 0.3509       | 0.6522       |
| 0.5    | 0.0051       | 0.1008       | 0.3592       | 0.6580       |

efficiency (ARE) of the rule $\hat{R}_{\text{PC}}^{(\text{full})}$ compared to the rule $\hat{R}_{\text{CC}}$ based on the completely classified sample is defined as

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

where the expectation in the numerator and denominator of (34) is taken over the distribution of the estimators of $\beta$ and is expanded up to terms of the first order.

Under the assumption that the class labels are missing completely at random, [Ganesalingam and McLachlan (1978)] derived the ARE of $\hat{R}_{\text{PC}}^{(\text{ig})}$ compared to $\hat{R}_{\text{CC}}$,

$$\text{ARE}(\hat{R}_{\text{PC}}^{(\text{ig})}) = \frac{E\{\text{err}(\hat{\beta}_{\text{CC}})\} - \text{err}(\beta)}{E\{\text{err}(\hat{\beta}_{\text{PC}}^{(\text{ig})})\} - \text{err}(\beta)},$$

in the case of a completely unclassified sample ($\gamma = 1$) for univariate features ($p = 1$). Their results are listed in Table 1 for $\Delta = 1$, 2, and 3. [O’Neill (1978)] extended their result to multivariate features and for arbitrary $\gamma$. His results showed that this ARE was not sensitive to the values of $p$ and does not vary with $p$ for equal class prior probabilities. Not surprisingly, it can be seen from Table 1 that the ARE of $\hat{R}_{\text{PC}}^{(\text{ig})}$ for a totally unclassified sample is low, particularly for classes weakly separated as represented by $\Delta = 1$ in Table 1.

In other work on the ARE of $\hat{R}_{\text{PC}}^{(\text{ig})}$ compared to $\hat{R}_{\text{CC}}$, [McLachlan and Scot (1995)] evaluated it where the unclassified univariate features had labels missing at random (MAR) due to truncation of the features.

Here the focus is on the ARE of $\hat{R}_{\text{PC}}^{(\text{full})}$ where additional information on $\beta$ from the missing-data mechanism is incorporated into the full likelihood function to yield the full ML estimator $\hat{\beta}_{\text{PC}}^{(\text{full})}$ on the basis of the partially classified sample $x_{\text{PC}}$.

We now sketch the derivation of the ARE of $\hat{R}_{\text{PC}}^{(\text{full})}$. We let $\hat{\beta}$ denote a generic estimator of $\beta$ that satisfies

$$\sqrt{n} \left( \frac{\hat{\beta}_0 - \beta_0}{\beta_1 - \beta_1} \right) \xrightarrow{D} N(0, V),$$

as $n \to \infty$, and that the first and second moments also converge. Then the first order expansion of the so-called excess error rate, that is, the expected error rate $\text{err}(\hat{\beta})$ over the optimal rate $\text{err}(\beta)$ for the estimated Bayes’ rule $R(y; \hat{\beta})$, can be expanded as

$$E\{\text{err}(\hat{\beta})\} - \text{err}(\beta) = n^{-1} \text{tr}(JV) + o(1/n),$$

(36)
where

\[ J = \tfrac{1}{2}[\nabla \nabla^T \text{err}(\hat{\beta})]_{\hat{\beta} = \beta} \]

and \( \nabla = (\partial/\partial \hat{\beta}_1, \ldots, \partial/\partial \hat{\beta}_p)^T \).

In deriving the ARE of logistic regression, Efron (1975) showed under the canonical for (4) adopted here for the two-class normal discrimination model that the expansion (36) reduces to

\[ E\{\text{err}(\hat{\beta})\} - \text{err}(\beta) = \frac{\pi_1 \phi(\Delta^*; 0, 1)}{2\Delta n} [v_{00} - \frac{\lambda}{\Delta^2} v_{01} + v_{22} + \ldots + v_{pp}] + o(1/n), \]  

(37)

where \( \lambda = \log(\pi_1/\pi_2) \), \( \Delta^* = \frac{1}{2} \Delta - \lambda/\Delta \), and \( \phi(y; \mu, \sigma^2) \) denotes the normal density with mean \( \mu \) and variance \( \sigma^2 \). Here \( v_{jk} = (V)_{jk} \), where the columns and rows in \( V \) are indexed from zero to \( p \).

The following theorem gives the ARE of \( \hat{R}_{PC}^{(\text{full})} \) compared to \( \hat{R}_{CC} \) in the case of equal prior probabilities \( \pi_1 = \pi_2 \).

**Theorem 2.** Under the missing-label model defined by (10), the ARE of \( \hat{R}_{PC}^{(\text{full})} \) compared to \( \hat{R}_{CC} \) is given in the case of \( \pi_1 = \pi_2 \) by

\[ \text{ARE}(\hat{R}_{PC}^{(\text{full})}) = 4(1 + \Delta^2/4)u_0 \]

(38)

for all \( p \), where

\[ u_0 = 1/\{4(1 + \Delta^2/4)\} - \gamma d_0 + b_0, \]

(39)

\[ b_0 = \int_{-\infty}^{\infty} 4\xi^2 \Delta^2 q_1(y_1)(1 - q(y_1)) f_{y_1}(y_1) dy_1, \]

\[ d_0 = \int_{-\infty}^{\infty} \tau_1(y_1) \tau_2(y_1) q_1(y_1) \gamma^{-1} f_{y_1}(y_1) dy_1, \]

and where

\[ \tau_1(y_1) = \text{pr}\{Z = 1 \mid (Y)_1 = y_1\} \quad (i = 1, 2), \]

\[ q_1(y_1; \Delta, \xi) = \text{pr}\{M = 1 \mid (Y)_1 = y_1\}, \]

\[ f_{y_1}(y_1; \Delta, \pi_1) = \pi_1 \phi(y_1; \Delta/2, 1) + (1 - \pi_1) \phi(y_1; -\Delta/2, 1). \]

In the above definitions of \( b_0 \) and \( d_0 \), we have suppressed the dependence of \( \tau_1(y_1), q_1(y_1), \) and \( f_{y_1}(y_1) \) on \( \Delta, \pi_1, \) and \( \xi \).

**Proof of Theorem 2.** To derive the ARE of \( \hat{R}_{PC}^{(\text{full})} \), we have to calculate the first order expansions of the numerator and denominator of the right-hand side of (34). Now the first order expansion of the numerator of (34) has been given by Efron (1975) for arbitrary values of \( \pi_1, \Delta, \) and \( p \) under the adopted canonical form. It is given for \( \pi_1 = \pi_2 \) by

\[ E\{\text{err}(\hat{\beta}_{\text{CC}})\} - \text{err}(\beta) = \frac{p \phi(\Delta/2; 0, 1)(1 + \Delta^2/4)}{\Delta n} + o(1/n). \]

(40)

To obtain the denominator of (34) under the adopted canonical form, we apply the following result of Efron (1975, Theorem 1), who developed it in the course of deriving
the ARE of logistic regression under the canonical form (4) adopted here for the two-class normal discrimination model.

Let \( \hat{\beta} \) be an estimator of \( \beta \) for which \( \sqrt{n}(\hat{\beta} - \beta) \) converges in distribution to the \( N(0, V) \) distribution, as \( n \to \infty \), and that the first and second order moments also converge. Then the expectation of the so-called excess error rate can be expanded as

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

where

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

and where \( \lambda = \log(\pi_1/\pi_2) \), \( \Delta^* = \frac{1}{2}\Delta - \lambda/\Delta \), and \( \phi(y; \mu, \sigma^2) \) denotes the normal density with mean \( \mu \) and variance \( \sigma^2 \). Here \( v_{jk} = (V)_{jk} \), where the columns and rows in \( V \) are indexed from zero to \( p \).

In order to apply the result (41) for \( \hat{\beta} \) equal to the full ML estimator \( \hat{\beta}_{PC}^{(\text{full})} \) of \( \beta \), we need to invert \( (1/n) \) times the information matrix \( I_{PC}^{(\text{full})}(\beta) \) for \( \beta \) given by (15) in Theorem 1. This evaluation is simplified in the case of \( \pi_1 = \pi_2 \) on noting several of the submatrices of the matrices in (15) become diagonal. On inverting \( I_{PC}^{(\text{full})}(\beta) \), we find that when \( \pi_1 = \pi_2 \),

\[
v_{jj} = 1/u_0, \quad j \neq 1,
\]

where \( u_0 \) is defined by (39). Substituting into (42), it follows that for \( \pi_1 = \pi_2 \),

\[
E\{\text{err}(\hat{\beta}_{PC}^{(\text{full})})\} - \text{err}(\beta) = \frac{p\phi(\Delta/2; 0, 1)}{4n\Delta u_0} + o(1/n),
\]

where we have used the fact that \( \lambda = 0 \) when \( \pi_1 = \pi_2 \). The ratio of the right-hand side of (40) to that of (43) gives the ARE. This completes the proof of Theorem 2. The extension of this theorem to the case of unequal prior probabilities is given in the Appendix.

In the case of \( \pi_1 = \pi_2 \), Table 2 gives the ARE of \( \hat{R}_{PC}^{(\text{full})} \) compared to \( R_{CC} \) for various combinations of the parameters \( \Delta, \xi_0 \), and \( \xi_1 \), the results applying for all values of \( p \). It can be seen for most of the combinations in Table 2 that the ARE of \( \hat{R}_{PC}^{(\text{full})} \) is greater than one, being appreciably greater than one for some combinations of the parameters. For example, for \( \Delta = 1 \) (representing classes close together) or \( \Delta = 2 \) (classes moderately separated), the ARE is not less than 15.48 for any combination with \( \xi_0 = 2 \) or 3 and \( \xi_1 = -5 \) or -10, being as high as 40.4 for \( \Delta = 1 \), \( \xi_0 = 5 \), \( \xi_1 = -10 \). This shows that the asymptotic expected excess error rate using the partially classified sample \( x_{PC} \) can be much lower than the corresponding excess rate using the completely classified sample \( x_{CC} \). The contribution to the Fisher information from the missingness mechanism can be relatively very high if \( |\xi_1| \) is large, as the location of the unclassified features in the feature space provides information about regions of high uncertainty, and hence where the absolute value of the discriminant function \( |d(y_j; \hat{\beta})| \) should be small. Consistent with this, it can be seen in Table 1 that as \( \xi_1 \) decreases, the ARE of \( \hat{R}_{PC}^{(\text{full})} \) increases for fixed \( \xi_0 \) and \( \Delta \).

In the Appendix, we give the general expression for the ARE of \( \hat{R}_{PC}^{(\text{full})} \) for \( \pi_1 \neq \pi_2 \). We find that this ARE is not sensitive to the value \( \pi_1 \) in the range \((0.2, 0.8)\), so that Theorem 2 can provide useful guidelines for arbitrary prior probabilities.
ξ_0 = 1.5
ξ_0 = 3
ξ_0 = 5

Table 2: Asymptotic relative efficiency of \( \hat{R}_{PC}^{(full)} \) for \( \pi_1 = \pi_2 \) (applicable for all \( p \))

| \( \Delta \) | \( \xi_0 = 1.5 \) | \( \xi_0 = 3 \) | \( \xi_0 = 5 \) |
|------------|-----------|-----------|-----------|
| 1          | 0.2       | 0.1       | 0.01      |
| 1.5        | 0.4       | 0.2       | 0.1       |
| 2          | 0.8       | 0.5       | 0.3       |
| 2.5        | 1.3       | 1.2       | 0.9       |
| 3          | 1.6       | 1.9       | 1.9       |

In Table 3, we have listed the probability of a missing label for each combination of the parameters in Table 2. If a feature \( y_j \) is on the decision boundary, then \( d(y_j; \beta) = 0 \) and the conditional probability of a missing label is equal to

\[
\Pr\{M_j = 1 \mid y_j\} = 1 / \{1 + \exp(-\xi_0)\}
\]

This probability is equal to 0.82, 0.95, and 0.99 for \( \xi_0 = 1.5, 3, \) and 5, respectively, which are the values of \( \xi_0 \) used in Table 2.

6 Simulations

We conducted a simulation to assess to what extent the asymptotic results of the previous section apply in practice. For each of the combinations of the parameters in Table 1, we generated \( B = 1000 \) samples of \( X = (Y^T, Z)^T \) to form the completely classified sample \( x_{CC} \) and the partially classified sample \( x_{PC} \). On each replication, the estimates \( \hat{\beta}_{CC} \) and \( \hat{\beta}_{PC}^{(full)} \) were computed using a quasi-Newton algorithm, along with the conditional error rates, \( \text{err}(\hat{\beta}_{CC}) \) and \( \text{err}(\hat{\beta}_{PC}^{(full)}) \). We let \( \text{err}(\hat{\beta}_{CC}^{(b)}) \) and \( \text{err}(\hat{\beta}_{PC}^{(full,b)}) \) denote the conditional error rate of \( \hat{R}_{CC} \) and of \( \hat{R}_{PC}^{(full)} \), respectively, on the \( b \)th replication. The relative efficiency (RE) of \( \hat{R}_{PC}^{(full)} \) compared to \( \hat{R}_{CC} \) was estimated by

\[
\text{RE}(\hat{R}_{PC}^{(full)}) = \frac{B^{-1} \sum_{b=1}^{B} \{\text{err}(\hat{\beta}_{CC}^{(b)}) - \text{err}(\hat{\beta})\}}{B^{-1} \sum_{b=1}^{B} \{\text{err}(\hat{\beta}_{PC}^{(full,b)}) - \text{err}(\hat{\beta})\}}.
\]

The nonparametric bootstrap with 1000 resamples was used to assess the variability of the estimates (Efron and Tibshirani 1986).

Tables 4 and 5 report the results with the bootstrap standard errors in parentheses. It can be seen in the case of \( n = 500 \) that there is very close agreement between the ARE
of $\hat{R}_{PC}^{(\text{full})}$ and its simulated values for the various combinations of $\Delta$, $\xi_0$, and $\xi_1$ in Table 5. As one would expect, the agreement is not as close for the smaller sample size $n = 100$, but there is still good agreement for most of the combinations of the parameters in Table 4. The simulated value of the ARE of $\hat{R}_{PC}^{(\text{full})}$ for $n = 100$ is less than its actual value for nearly all of the combinations in Table 4 with $\xi_1 \leq -0.5$, indicating that the gain in efficiency for finite samples is not as high as given asymptotically for these combinations.

One of them for which the agreement between the ARE of $\hat{R}_{PC}^{(\text{full})}$ and its simulated value is not close is $\Delta = 3$ with $\xi_0 = 3$, $\xi_1 = -10$, where the ARE is 12.8 but its simulated value is 4.4. A possible explanation for this is that for this combination of the parameters the probability $\gamma$ that a feature vector will have a missing label is very low at 0.06, so in a sample of size $n = 100$ the estimation of $\xi_1$ has to be based on a sample with few values of the missing-label indicator variable equal to 1.

7 Discussion

The analysis of partially classified data often involves additional considerations relative to completely classified data; see, for example, Chapelle et al. (2010). Partially classified data can arise in situations where classifications are made by subjective judgement, and there is uncertainty on the best assignment for a number of instances in the training set. From a statistical point of view, the propensity of high entropy features to remain unclassified represents an extra source of information for learning a classification rule. More formally, the Fisher information in a partially classified sample will include a contribution from the missing data mechanism under mild assumptions (Rubin, 1976). We have shown that in the case of two-class normal discriminant analysis, the Fisher information about the vector of discriminant function coefficients in the partially classified dataset can be much greater than in a completely classified dataset where the relationship between classification difficulty and the probability of a missing label is strong. As a consequence, the asymptotic expected error rate of the classifier trained using $x_{PC}$ can be smaller than the expected error rate of the classifier trained using $x_{CC}$. We observed this theoretical superefficiency in our Monte Carlo simulations. We have focused on a simple logistic selection model to give mathematical insight into this phenomenon. Generic model checking and diagnostic tools can be used to assess the goodness of fit of a proposed missingness model. Further work will involve the mathematical and empirical study of more complex models. The distance of unlabelled observations from the separating hyperplane has also been identified as an important quantity for semi-supervised learning with support vector machines (Vapnik, 1998), and this is a possible direction to follow to extend the proposed methodology to nonlinear models. The likelihood contribution of the missingness model can also be viewed as a regularisation term that includes the unlabelled observations, placing it within a general paradigm in semi-supervised learning (Berthelot et al., 2019). This perspective may also help to understand the behaviour of the full likelihood if the missingness mechanism is misspecified.
\[ \xi = \begin{pmatrix} 0.2 & 0.0 (0.1) & 0.3 (0.02) & 0.7 (0.04) & 1.0 (0.06) & 1.6 (0.09) \\ 0.1 (0.01) & 0.1 (0.005) & 0.3 (0.04) & 1.2 (0.07) & 1.5 (0.1) \\ 0.05 (0.01) & 0.1 (0.01) & 0.06 (0.01) & 0.8 (0.05) & 2.0 (0.1) \\ 0.01 & 0.1 & 0.3 & 0.9 & 1.9 \\ 0.1 (0.01) & 0.9 (0.2) & 3.8 (0.2) & 4.9 (0.3) & 4.8 (0.3) \\ 0.4 & 2.2 & 4.4 & 5.5 & 5.5 \end{pmatrix} \]

Table 4: Simulated relative efficiency of \( \hat{R}^{(\text{full})}_{\pi_1\pi_2} \) with \( \pi_1 = \pi_2 \) for \( n = 100, p = 1 \) (Bootstrap standard errors are in parentheses). The grey shaded rows give for ease of comparison the values of the ARE from Table 2

| \( \Delta \) | \( \xi_0 = 1.5 \) | \( \xi_0 = 3 \) | \( \xi_0 = 5 \) |
|---|---|---|---|
| 1 | 1.5 | 2 | 2.5 | 3 | 1 | 1.5 | 2 | 2.5 | 3 | 1 | 1.5 | 2 | 2.5 | 3 |
| \( \xi = -0.1 \) | 0.2 | 0.0 (0.1) | 0.3 (0.02) | 0.7 (0.04) | 1.0 (0.06) | 1.6 (0.09) | 0.1 (0.01) | 0.1 (0.005) | 0.3 (0.04) | 1.2 (0.07) | 1.5 (0.1) | 0.05 (0.01) | 0.1 (0.01) | 0.06 (0.01) | 0.8 (0.05) | 2.0 (0.1) |
| \( \xi = -0.1 \) | 0.2 & 0.1 & 0.3 & 0.7 & 1.0 & 1.6 & 0.1 & 0.2 & 0.5 & 1.2 & 1.9 & 0.01 & 0.1 & 0.3 & 0.9 & 1.9 |
| \( \xi = -0.5 \) | 0.9 (0.1) & 2.1 (0.1) & 2.9 (0.2) & 2.9 (0.2) & 2.6 (0.2) | 0.2 (0.02) & 2.5 (0.2) & 3.7 (0.2) & 4.3 (0.3) & 3.1 (0.2) | 0.1 (0.01) & 0.9 (0.2) & 3.8 (0.2) & 4.9 (0.3) & 4.8 (0.3) |
| \( \xi = -1 \) | 1.5 (0.2) & 2.6 (0.3) & 3.1 (0.3) & 3.2 (0.2) & 2.9 (0.2) | 1.0 & 2.7 & 4.0 & 4.3 & 4.1 | 0.4 & 2.2 & 4.4 & 5.5 & 5.5 |
| \( \xi = -1 \) | 3.5 (0.2) & 4.6 (0.3) & 4.4 (0.3) & 3.7 (0.2) & 3.0 (0.2) | 1.8 (0.3) & 5.5 (0.4) & 6.5 (0.4) & 5.8 (0.4) & 4.0 (0.3) | 0.2 (0.02) & 5.3 (0.3) & 6.8 (0.4) & 6.6 (0.5) & 5.4 (0.4) |
| \( \xi = -5 \) | 14.8 (1) & 10.7 (0.6) & 8.1 (0.5) & 5.9 (0.4) & 3.5 (0.2) | 19.3 (1) & 16.8 (1) & 10.6 (0.7) & 7.4 (0.5) & 4.5 (0.3) | 23.6 (2) & 17 (1) & 13.2 (0.9) & 8.4 (0.6) & 4.6 (0.3) |
| \( \xi = -10 \) | 22.1 (2) & 16.5 (1) & 11.5 (0.7) & 6.7 (0.5) & 3.2 (0.2) | 33.3 (2) & 19.9 (1) & 15.2 (1) & 8.8 (0.6) & 4.2 (0.3) | 39.6 (3) & 25.4 (2) & 16.5 (1) & 8.1 (0.5) & 4.2 (0.3) |

Table 5: Simulated relative efficiency of \( \hat{R}^{(\text{full})}_{\pi_1\pi_2} \) with \( \pi_1 = \pi_2 \) for \( n = 500, p = 1 \) (Bootstrap standard errors are in parentheses). The grey shaded rows give for ease of comparison the values of the ARE from Table 2
References

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T., Brinkman, R., Gottardo, R., Scheuermann, R.H., FlowCAP Consortium and Dream Consortium (2013). Critical assessment of automated flow cytometry data analysis techniques. Nature Methods 10, 228.

Ahfock, D. and McLachlan, G.J. (2019a). Missing label patterns in semi-supervised learning. Abstract of Seminar Presented on 29th August, 2019, at the 16th Conference of the International Federation of Classification Societies, Thessaloniki. In Book of Abstracts of IFCS 2019. URL: https://ifcs.gr/wp-content/uploads/2019/08/IFCS2019_ConferenceBook-Preview-2.pdf.

Ahfock, D. and McLachlan, G.J. (2019b). On missing data patterns in semi-supervised learning. arXiv ePrint arXiv:1904.02883.

Berthelot, D, Carlini, N, Goodfellow, I, Papernot, N, Oliver, A and Raffel, C (2019). Mixmatch: A holistic approach to semi-supervised learning. in ‘Advances in Neural Information Processing Systems’.

Chapelle, O, Schlköpf, B and Zien, A (2010). Semi-Supervised Learning. Cambridge, MA: The MIT Press.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society: Series B (Methodological) 39, 1–22.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association 70, 892–898.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1, 54–75.

Ganesalingam, S. and McLachlan, G.J. (1978). The efficiency of a linear discriminant function based on unclassified initial samples. Biometrika 65, 658–665.

Grandvalet, Y. and Bengio, Y. (2005). Semi-supervised learning by entropy minimization. in ‘Advances in Neural Information Processing Systems’. pp. 529–536.

McLachlan, G.J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. Journal of the American Statistical Association 70, 365–369.

McLachlan, G.J. (1992). Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley.

McLachlan, G.J. and Gordon, R.D. (1989). Mixture models for partially unclassified data: a case study of renal venous renin in hypertension. Statistics in Medicine 8, 1291–1300.

McLachlan, G.J. and Scot, D. (1995). Asymptotic relative efficiency of the linear discriminant function under partial nonrandom classification of the training data. Journal of Statistical Computation and Simulation 52, 415–426.
Appendix

We consider here under the canonical form (4) of the model the evaluation of the information matrices in the expression for the information matrix $I^{\text{full}}_{PC}(\beta)$, which has to be carried out to obtain the ARE of the rule $\hat{R}^{(\text{full})}_{PC}$ based on the full ML estimator of $\beta$ formed from the partially classified sample $x_{PC}$. We also provide more details on the proof of Theorem 2, in particular, the extension of Theorem 2 to the case of unequal prior probabilities.

The information matrix $I_{CC}(\beta)$

It is shown in Efron [1975] that the matrix $I_{CC}(\beta)$ has the following structure

$$I_{CC}(\beta) = \begin{pmatrix}
a_0 & a_1 & 0 & 0 & \cdots & 0 \\
a_1 & a_2 & 0 & 0 & \cdots & 0 \\
0 & 0 & a_3 & 0 & \cdots & 0 \\
\vdots & \vdots & 0 & \ddots & \cdots & 0 \\
0 & 0 & 0 & 0 & \cdots & a_3 \\
0 & 0 & 0 & 0 & 0 & a_3
\end{pmatrix}, \quad (45)$$

where

$$(a_0, a_1, a_2)^{-1} = \frac{1}{\pi_1 \pi_2} \begin{pmatrix}
1 + \Delta^2/4 & -(\pi_2 - \pi_1)\Delta/2 \\
-(\pi_2 - \pi_1)\Delta/2 & 1 + 2\pi_1 \pi_2 \Delta^2
\end{pmatrix}, \quad (46)$$

and $a_3 = \pi_1 \pi_2 (1 + \Delta^2 \pi_1 \pi_2)^{-1}$. If $\pi_1 = \pi_2$, $a_1$ is zero and so the matrix $I_{CC}(\beta)$ is diagonal, and in addition $a_0 = a_3 = \{4(1 + \Delta^2/4)\}^{-1}$.

The information matrix $I^{\text{clr}}_{CC}(\beta)$

The conditional distribution of $Y$ given that $M = 1$ can be expressed as

$$f_{\text{miss}}(y \mid M = 1) = \frac{q_1(y_1)f_{\text{un}}(y_1)}{\gamma} \prod_{i=2}^{p} \phi(y_i; 0, 1). \quad (47)$$

The matrix $I^{\text{clr}}_{CC}(\beta)$ is given by the integral

$$I^{\text{clr}}_{CC}(\beta) = \int_{R^p} \begin{pmatrix} y \end{pmatrix} (1 - y y^T) \tau_1(y_1)\tau_2(y_1)f_{\text{miss}}(y \mid M = 1) \, dy.$$
Using the independence of the variables in \( Y \) in the conditional distribution (47), the matrix has the structure

\[
I_{CC}^{(\text{clr})}(\beta) = \begin{pmatrix}
  d_0 & d_1 & 0 & 0 & \cdots & 0 \\
  d_1 & d_2 & 0 & 0 & \cdots & 0 \\
  0 & 0 & d_0 & 0 & \cdots & 0 \\
  \vdots & \vdots & 0 & \ddots & \ddots & \cdots & 0 \\
  0 & 0 & 0 & 0 & d_0 & 0 \\
  0 & 0 & 0 & 0 & 0 & d_0 \\
\end{pmatrix},
\]

(48)

where

\[
d_k = \int_{-\infty}^{\infty} y_1^k \tau_1(y_1)\tau_2(y_1)q_1(y_1)\gamma^{-1}f_{y_1}(y_1) \, dy_1, \quad (k = 0, 1, 2),
\]

(49)

and the functions \( \tau_1, \tau_2, q_1 \), and \( f_{y_1} \) are as given in Theorem 1. For \( \pi_1 = \pi_2 \), \( d_1 \) is zero and so the matrix \( IC_{CC}^{(\text{clr})}(\beta) \) is diagonal.

The information matrix \( I_{PC}^{(\text{miss})}(\beta) \)

Using the independence of the variables in \( Y \), the matrix \( B_{22} \) has the following structure

\[
B_{22} = \begin{pmatrix}
  b_0 & b_1 & 0 & 0 & \cdots & 0 \\
  b_1 & b_2 & 0 & 0 & \cdots & 0 \\
  0 & 0 & b_0 & 0 & \cdots & 0 \\
  \vdots & \vdots & 0 & \ddots & \ddots & \cdots & 0 \\
  0 & 0 & 0 & 0 & b_0 & 0 \\
  0 & 0 & 0 & 0 & 0 & b_0 \\
\end{pmatrix}.
\]

(50)

The elements of \( B_{22} \) are given by

\[
b_0 = \int_{-\infty}^{\infty} 4\xi_1^2(\Delta^2y_1^2 + 2\lambda\Delta + \lambda^2)q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1,
\]

\[
b_1 = \int_{-\infty}^{\infty} \xi_1^2(2\lambda + 2\Delta y_1)q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1,
\]

\[
b_2 = \int_{-\infty}^{\infty} \xi_1^2(2\lambda + 2\Delta y_1)q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1.
\]

The matrix \( B_{23} \) has the following structure

\[
B_{23} = \begin{pmatrix}
  r_0 & r_1 \\
  r_2 & r_3 \\
  0 & 0 \\
  \vdots & \vdots \\
  0 & 0 \\
\end{pmatrix}.
\]

(51)
The nonzero elements of $B_{33}$ are given by

$$r_0 = \int_{-\infty}^{\infty} \xi_1(2\lambda + 2\Delta y_1)q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1,$$

$$r_1 = \int_{-\infty}^{\infty} \xi_1(\lambda + \Delta y_1)^2(2\lambda + 2\Delta y_1)q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1,$$

$$r_2 = \int_{-\infty}^{\infty} \xi_1(2\lambda y_1 + 2\Delta y_1^2)q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1,$$

$$r_3 = \int_{-\infty}^{\infty} \xi_1(\lambda + \Delta y_1)^2(2\lambda y_1 + 2\Delta y_1^2)q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1.$$

For $\pi_1 = \pi_2$, $r_0$ and $r_1$ are both equal to zero as they are equal to the integral of an odd function over the real line. The information matrix for the estimation of $\xi$, $B_{33}$, can be written as

$$B_{33} = \begin{pmatrix} s_0 & s_1 \\ s_1 & s_2 \end{pmatrix},$$

where

$$s_0 = \int_{-\infty}^{\infty} q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1,$$

$$s_1 = \int_{-\infty}^{\infty} (\lambda + \Delta y_1)^2q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1,$$

$$s_2 = \int_{-\infty}^{\infty} (\lambda + \Delta y_1)^4q_1(y_1)(1 - q_1(y_1))f_{y_1}(y_1) \, dy_1.$$

As the lower block of $B_{33}$ given by (51) is the zero matrix for all $\pi_1$, only the top left two-by-two block of $B_{33}B_{33}^{-1}B_{33}$ will be nonzero. Let

$$\begin{pmatrix} w_0 & w_1 \\ w_1 & w_2 \end{pmatrix} = \begin{pmatrix} r_0 & r_1 \\ r_2 & r_3 \end{pmatrix} \begin{pmatrix} s_0 & s_1 \\ s_1 & s_2 \end{pmatrix}^{-1} \begin{pmatrix} r_0 & r_1 \\ r_2 & r_3 \end{pmatrix}^T. \quad (53)$$

In general,

$$I_{PC}^{(\text{miss})}(\beta) = \begin{pmatrix} b_0 - w_0 & b_1 - w_1 & 0 & 0 & \cdots & 0 \\ b_1 - w_1 & b_2 - w_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & b_0 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \cdots & 0 \\ 0 & 0 & 0 & 0 & b_0 & 0 \\ 0 & 0 & 0 & 0 & 0 & b_0 \end{pmatrix}. \quad (54)$$

For $\pi_1 = \pi_2$, $h_0$ and $h_1$ are both zero, and so then $w_0$ and $w_1$ are also both zero, leading then to the matrix $I_{PC}^{(\text{miss})}(\beta)$ being diagonal.

**Asymptotic covariance matrix of $\hat{\beta}_{PC}^{(\text{full})}$**

Let

$$\begin{pmatrix} h_0 & h_1 \\ h_1 & h_2 \end{pmatrix} = \begin{pmatrix} a_0 & a_1 \\ a_1 & a_2 \end{pmatrix} - \gamma(\Psi) \begin{pmatrix} d_0 & d_1 \\ d_1 & d_2 \end{pmatrix} + \begin{pmatrix} b_0 & b_1 \\ b_1 & b_2 \end{pmatrix} - \begin{pmatrix} w_0 & w_1 \\ w_1 & w_2 \end{pmatrix}, \quad (55)$$

$$u_0 = a_3 - \gamma(\Psi)d_0 + b_0, \quad (56)$$

20
where the constants $a_0, a_1, a_2, \text{ and } a_3$ are given in (46), the constants $d_0, d_1$ and $d_2$ are given in (49), $b_0, b_1, b_2$ are given in (50) and $w_0, w_1, w_2$ are given in (53). The general form of the information matrix is

$$I_{\text{PC}}^{(\text{full})}(\beta) = \begin{pmatrix} h_0 & h_1 & 0 & 0 & \cdots & 0 \\ h_1 & h_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & u_0 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & u_0 \\ 0 & 0 & 0 & 0 & 0 & u_0 \end{pmatrix}.$$  \hspace{1cm} (57)

As mentioned earlier, for $\pi_1 = \pi_2$ a number of useful simplifications can be made. The matrices $I_{\text{CC}}(\beta), I_{\text{CC}}^{(\text{clr})}(\beta)$, and $I_{\text{PC}}^{(\text{miss})}(\beta)$ are diagonal, and $r_0 = u_0$ and $r_1 = 0$. For $\pi_1 = \pi_2$ the information matrix reduces to

$$I_{\text{PC}}^{(\text{full})}(\beta) = \begin{pmatrix} u_0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & a_2 - \gamma(\Psi) d_2 + b_2 - w_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & u_0 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & u_0 \\ 0 & 0 & 0 & 0 & 0 & u_0 \end{pmatrix}. \hspace{1cm} (58)$$

The asymptotic covariance matrix is given by $V$ is given by $n\{I_{\text{PC}}^{(\text{full})}(\beta)\}^{-1}$. For $\pi_1 = \pi_2$, $v_{jj} = 1/u_0$ for $j = 0, 2, 3, \ldots, p$.

**Extension of Theorem 2 to Arbitrary Prior Probabilities**

We refer to the result (41) given by Efron (1975) for the first order expansion of the expected excess error rate of the plug-in form of the Bayes’ rule using the estimator $\hat{\beta}$ of $\beta$ where $\sqrt{n}(\hat{\beta} - \beta)$ converges in distribution to the $N(0, V)$ distribution, as $n \to \infty$ and where the first and second order moments also converge.

The expectation of the so-called excess error rate can be expanded as

$$E\{\text{err}(\hat{\beta})\} - \text{err}(\beta) = \frac{\pi_1 \phi(\Delta^*; 0, 1)}{2\Delta n} w + o(1/n), \hspace{1cm} (59)$$

where

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

and where $\lambda = \log(\pi_1/\pi_2), \Delta^* = \frac{1}{2} \Delta - \lambda/\Delta$, and $\phi(y; \mu, \sigma^2)$ denotes the normal density with mean $\mu$ and variance $\sigma^2$. Here $v_{jk} = (V)_{jk}$, where the columns and rows in $V$ are indexed from zero to $p$.

Let

$$Q_1 = \frac{1}{\pi_1 \pi_2} \left( 1 - \frac{\lambda}{\Delta} \right) \begin{pmatrix} 1 + \Delta^2/4 & -(\pi_2 - \pi_1) \Delta/2 \\ -(\pi_2 - \pi_1) \Delta/2 & 1 + 2\pi_1 \pi_2 \Delta^2 \end{pmatrix} \begin{pmatrix} 1 - \lambda/\Delta \end{pmatrix}^T,$$

$$Q_2 = \frac{1}{\pi_1 \pi_2} (1 + \pi_1 \pi_2 \Delta^2).$$
Using the expansion of the error rate in (41), the first order approximation to the expected error rate of \( \hat{R}_{CC} \) is

\[
E\{\text{err}(\hat{\beta}_{CC})\} - \text{err}(\beta) = \frac{\pi_1 \phi(\Delta^*; 0, 1)}{2\Delta n} \{Q_1 + (p - 1)Q_2\} + o(1/n). \tag{60}
\]

Let

\[
\begin{align*}
H &= \begin{pmatrix} a_0 & a_1 \\ a_1 & a_2 \end{pmatrix} - \gamma(\Psi) \begin{pmatrix} d_0 & d_1 \\ d_1 & d_2 \end{pmatrix} + \begin{pmatrix} b_0 & b_1 \\ b_1 & b_2 \end{pmatrix} - \begin{pmatrix} w_0 & w_1 \\ w_1 & w_2 \end{pmatrix}, \\
u_0 &= a_3 - \gamma(\Psi)d_0 + b_0,
\end{align*}
\tag{61}
\]

where the constants \( a_0, a_1, a_2, \) and \( a_3 \) are given in (46), the constants \( d_0, d_1, \) and \( d_2 \) are given in (49), \( b_0, b_1, b_2 \) are given in (50) and \( w_0, w_1, w_2 \) are given in (53). Define

\[
Q_3 = \left(1 - \lambda/\Delta \right) H^{-1} \left(1 - \lambda/\Delta \right)^T, \\
Q_4 = 1/u_0.
\]

Using the expansion (41), the first order approximation to the expected error rate of \( \hat{R}_{PC}^{(\text{full})} \) is

\[
E\{\text{err}(\hat{\beta}_{PC}^{(\text{full})})\} - \text{err}(\beta) = \frac{\pi_1 \phi(\Delta^*; 0, 1)}{2\Delta n} \{Q_3 + (p - 1)Q_4\} + o(1/n), \tag{63}
\]

which gives the denominator for arbitrary \( \pi_1 \) in the formula (34) for the ARE. Evaluation of \( Q_3 \) involves some effort, as we need to determine each of the constants appearing in the matrix \( H \) in (61).

Taking the ratio of (60) to (63) and ignoring terms of \( o(1/n) \) gives the asymptotic relative efficiency of \( \hat{R}_{PC}^{(\text{full})} \) to \( \hat{R}_{CC} \),

\[
\text{ARE}\left(\hat{R}_{PC}^{(\text{full})}\right) = \frac{Q_1 + (p - 1)Q_2}{Q_3 + (p - 1)Q_4}, \tag{64}
\]

Evaluation of (64) involves many calculations due to the number of terms in \( Q_3 \). The general form (64) simplifies if \( \pi_1 = \pi_2 \), as then \( Q_1 = Q_2 = 4(1 + \Delta^2/4) \) and \( Q_3 = Q_4 = 1/u_0 \). The asymptotic relative efficiency when \( \pi_1 = \pi_2 \) then collapses to the more interpretable form,

\[
\text{ARE}\left(\hat{R}_{PC}^{(\text{full})}\right) = \frac{pQ_2}{pQ_4} = 4(1 + \Delta^2/4)u_0, \tag{65}
\]

which holds for all \( p \).