Classification with missing values

Angel Reyero Lobo
Supervisors: Alexis Ayme, Claire Boyer, Aymeric Dieuleveut and Erwan Scornet

Laboratoire de Probabilités, Statistique & Modélisation (LPSM)
Sorbonne Université

July 31, 2023

Abstract

There is a scarcity of methods for predicting outcomes when dealing with missing values. To address this gap, our study focuses on adapting classical classification algorithms to handle missing values. We propose decomposing the Bayes classifier on a pattern-by-pattern basis, thereby defining a specific predictor for each given missing pattern. Firstly, we demonstrate that, under certain assumptions about the distribution of missing patterns, the Linear Discriminant Analysis (LDA) model *preserves* the structure of each individual classifier. Exploiting this property, we establish certain bounds to quantify the information loss compared to the case where complete data is available. Furthermore, we ascertain the convergence rate of diverse data-driven classifiers under various assumptions, ensuring their adaptability to a multitude of real-world scenarios. Secondly, we prove that, even under typical assumptions about the data, the logistic model is not preserved on a pattern-wise basis. The focus then shifts to the study of the perceptron model. In this context, the analysis centers around the linear separability of classes when missing values are present. The objective is to ensure the convergence of the classical perceptron algorithm for a given missing pattern. The consistency of this pattern-by-pattern perceptron, is empirically shown through numerical experiments with datasets that verify the assumptions of the theoretical results.

Keywords: Missing values, linear discriminant analysis(LDA), missclassification error control, missing Completely at random(MCAR), missing not at random(MNAR).

Notations. For $n \in \mathbb{N}$, we denote $[n] = \{1, \ldots, n\}$. We use $\lesssim$ to denote inequality up to a universal constant. For any $x \in \mathbb{R}^d$ and for any set $J \subset [d]$ of indices, we let x_J be the subvector of x composed of the components indexed by J. The abbreviation *p-b-p* refers to *pattern-by-pattern.* The notation $\mathcal{D}_n := \{(X_i, Y_i), i = 1, \ldots, n\}$ represents a sample of size n consisting of pairs of variables X_i along with their respective labels Y_i. $\mathcal{D}_n^\star$ is a sample of size n with missing values only in the input variables X_i. The values $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ respectively designate the greatest and the smallest eigenvalues of any matrix A. We denote $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

Acknowledgements

I would like to express my gratitude to my supervisors Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Their ideas and support have introduced me to this fascinating research field. Additionally, they have equipped me with valuable technical tools and perspectives that I will undoubtedly use in my future work.

I am also immensely thankful to my MSc teachers, especially my supervisors Gilles Blanchard, Christine Keribin, and Marie-Anne Poursat. Not only did they provide me with exceptionally satisfactory teaching, but their dedication also encouraged me to pursue research with enthusiasm.

Lastly, I want to extend my thanks to the *Laboratoire de Probabilités, Statistique et Modélisation* (LPSM) where I felt truly comfortable during my internship.

Contents

1 Preliminaries on supervised statistical learning

The main objective of supervised statistical learning is to predict a target value Y given some new observation X. To accomplish this, we would like to construct a theoretical *predictor* h. To measure the performances of different predictors, various cost functions can be used depending on the nature of the target variable Y. These cost functions c quantify the level of mismatch between the predictions $h(X)$ and the true values Y. As a result, we construct predictors by aiming to minimize $\mathbb{E}\left[c(h(X), Y)\right]$.

Unluckily, to do so, we need to access the unknown conditional distribution of Y given X. To circumvent this difficulty, we build predictors $\widehat{h}$ based on a train set of observations from a sample $\mathcal{D}_n$, which consists of a set of n observations of covariates along with their respective labels, denoted $\mathcal{D}_n := (X_i, Y_i)_{i=1,\dots,n}$. We require the new (or test) and the training observations to be sampled from the same distribution $P_{(X,Y)}$ and to be i.i.d. To construct the predictors $\widehat{h}$, there are primarily two approaches. Firstly, we can attempt to directly minimize an empirical risk that estimates the theoretical risk. Alternatively, we can compute a theoretical predictor that minimizes this risk, and then proceed to estimate this predictor.

1.1 Classification

The goal of classification is to predict the class of a new observation accurately. Therefore, the nature of the label Y is discrete and finite. This problem is present in our daily lives through simple applications such as the spam filter in our mailbox or the filters applied to detect hate speech in social media. It is also applied in other fields such as health sciences to detect diseases or business professions to predict customer interests. In particular, we refer to our predictor h as a *classifier*. For simplicity in calculations, we focus on the scenario of binary classification, where the labels take values in $\{-1, 1\}$.

As shown in Giraud (2021), one way of quantifying the accuracy of a classifier can be given by the probability of misclassification:

$$L_{\text{comp}}(h) := \mathbb{E}\left[\mathbb{1}_{Y \neq h(X)}\right] = \mathbb{P}(Y \neq h(X)).$$

Note that this loss function treats all misclassification errors equally. However, there are situations where misclassifying one class is more critical or dangerous than misclassifying other classes, such as in medical diagnoses. In such cases, asymmetric loss functions can be recommended. Therefore, we would like to find a classifier minimizing this probability of misclassification. As in binary classification, $|Y - h(X)| \in \{0, 2\}$, then,

$$L_{\text{comp}}(h) = \frac{1}{4}\mathbb{E}\left[(Y - h(X))^2\right] = \frac{1}{4}\mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] + \frac{1}{4}\mathbb{E}\left[(\mathbb{E}[Y|X] - h(X))^2\right].$$

This provides the *Bayes classifier*

$$h^{\star}_{\text{comp}}(X) = \operatorname{sign}(\mathbb{E}[Y|X]) \text{ where } \operatorname{sign}(x) = \mathbb{1}_{x>0} - \mathbb{1}_{x\leq 0}. \tag{1}$$

Unfortunately, the distribution $P_{(X,Y)}$ is unknown. As a result, there are multiple approaches to solve this problem. In Section 3, in order to accommodate the missing values to the classification framework, the emphasis is placed on parametric modeling and making assumptions about the distributions of the observations, missing patterns, and labels. However, alternative approaches can be also considered, following suggestions provided in Giraud (2021) for the case of complete data. In that context, assumptions are made on the structure of the classifier, which belongs to a set referred to as a *dictionary*. The classifier is then obtained by minimizing a convexified empirical risk over this set.

1.2 Regression

In contrast to the discrete labels used in the classification context, the regression label Y is continuous. It is also pervasive in various fields, such as electricity consumption forecasting, economic indicators prediction, weather forecasting or the effectiveness of medical methods prediction. In this context, we refer to the predictor as a *regressor*. Many loss functions have been proposed to quantify the level of disagreement between the prediction and the true label. Given a cost function c, we can determine a general mean error introduced by a regressor f through

$$C_{\text{comp}}(f) := \mathbb{E}\left[c(Y, f(X))\right].$$

In the context of regression, the most commonly used cost function is the quadratic loss, i.e., $c(Y, f(X)) = (Y - f(X))^2$. It is worth noting that, similar to the decomposition of the loss function used in the classification framework, the *Bayes regressor* $f^{\star}$ can be obtained decomposing the mean square error giving:

$$f^{\star}(X) = \mathbb{E}[Y|X].$$

In a context of regression, the most simple model is that the underlying function linking Y to X is linear.

Assumption 1 (Linear Model)**.** $Y = \beta_0 + \beta^{\top} X + \epsilon$, with a Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent of X.

The (unknown) model parameters to be estimated are therefore $(\beta_0, \beta) \in \mathbb{R}^{d+1}$. In the missing values framework, Section 2.1 provides a synthetic overview of the work of Ayme et al. (2022) to obtain a minimax optimal estimator under linear assumptions.

2 Prediction with missing inputs

Missing values are a common issue in real-world datasets, arising from various sources. They can result from aggregating values from multiple sources using different metrics, leaving gaps unfilled, insufficient resources to extract expensive information for the entire study, or even sensor failures.

In the context of supervised learning with missing values, we assume that the input observation $X \in \mathbb{R}^d$ is not fully available. Instead, we only have a partial observation masked by a missing pattern M. The missing pattern $M \in \{0,1\}^d$ is a binary vector where each coordinate M_j indicates whether the corresponding entry X_j is observed (0) or missing (1).

Given a specific missing pattern $m \in \{0,1\}^d$, we define $\mathrm{obs}(m)$ as the set of indices where m is 0, representing the observed variables. Conversely, $\mathrm{mis}(m)$ represents the set of indices where m is 1, indicating the missing variables. Using the introduced notation, $X_{\mathrm{obs}(m)}$ refers to the subvector of X consisting of the observed variables corresponding to $\mathrm{obs}(m)$. Similarly, $X_{\mathrm{mis}(m)}$ denotes the subvector of X associated with the missing variables $\mathrm{mis}(m)$. We introduce the following explicit example to our notation.

Example 2.1. Given an observation $X = (6, 3, \mathrm{na}, 3, \mathrm{na})$ with not available values (na) , the associated missing pattern is given by $M = (0, 0, 1, 0, 1)$, the indices of observed covariates are $\mathrm{obs}(M) = (1, 2, 4)$ and the observed covariates are $X_{\mathrm{obs}(M)} = (6, 3, 3)$.

In supervised learning with missing values, instead of predicting the label Y solely from the complete observation X, we aim to predict it from a pair consisting of the masked observation and the missing pattern, denoted as $Z := (X_{\mathrm{obs}(M)}, M)$.

Setting In his leading work, Rubin (1976) established the following three main assumptions made between the distribution of the missing pattern, the observations and the label.

Assumption 2 (Missing Completely At Random(MCAR)). $M \perp\!\!\!\perp X, Y$.

Under MCAR Assumption, we consider the missingness mechanism M to be completely independent of both the explanatory data X and the label Y. This scenario may occur in situations such as forgetting to fill in a form or random failures of sensors.

Assumption 3 (Missing At Random(MAR)). $\forall m \in \mathcal{M}, \mathbb{P}\left(M = m | X, Y\right) = \mathbb{P}\left(M = m | X_{obs(m)}\right)$.

Under the MAR (Missing at Random) assumption, we assume that the missingness mechanism M depends solely on the observed data. This assumption can be applicable in scenarios such as a medical study where, based on the results of an initial analysis, different techniques involving other variables are applied if the initial ones fall outside a confidence interval.

Assumption 4 (Missing Non At Random - MNAR). The missing pattern M depends on the full vector (X, Y).

Under MNAR Assumption, the missing pattern depends on the observed, missing entries and label. This occurs when individuals choose not to provide sensitive information, such as their income, if this information take some specific values (e.g., people with high wages may not want to provide them publicly).

Moreover, other assumptions as the following ones are usually made for the data scenarios (see Ayme et al. (2022)).

Assumption 5 (Independent covariates). The covariates $\{X_j\}_{j\in[d]}$ are mutually independent.

Observe that this assumption does not conflict with the MAR assumption (Assumption 3). Instead, it focuses on the covariates rather than the missing patterns. Consequently, it is feasible to have a MAR assumption regarding the missing pattern while maintaining independent covariates, as shown in the data model of Example 3.2.

Assumption 6 (Gaussian covariates)**.** There exist $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ such that $X \sim \mathcal{N}(\mu, \Sigma)$.

Assumption 7 (Gaussian pattern mixture model-GPMM)**.** For all $m \in \mathcal{M}$ and for all $y \in \{-1, +1\}$, $X_{\text{obs}(m)}|(M = m, Y = y) \sim \mathcal{N}(\mu^{(m,y)}, \Sigma^{(m,y)})$.

Remark that this last model satisfies the MNAR Assumption, since we may observe one different data distribution per missing data pattern.

Problematic/Methodology Classical machine learning algorithms are typically designed to handle complete data and are not capable of accommodating missing values. As a result, numerous studies have focused on imputing these missing cases to address this issue. Once the data is completed, classical algorithms can be applied.

Another alternative approach involves decomposing the Bayes predictor on a pattern-by-pattern basis, training a specific predictor for each missing pattern and leveraging the information provided by them. This approach is employed in Ayme et al. (2022), which is summarized in Section 2.1, and is the chosen approach in this article.

Moreover, there is a paucity of methods for predicting with missing values as most studies are focused on training the algorithms with missing data but applying them to predict the label of a complete data case (see Tony Cai and Zhang (2019)). Indeed, predicting with missing values is a difficult task as the estimated values of the underlying model are not directly applicable to the incomplete data.

2.1 (Regression) Linear prediction with NA

This section presents the results of Ayme et al. (2022).

First, note that using a similar decomposition if the quadratic loss in the complete data case, the Bayes predictor can expressed as $f^\star(Z) = \mathbb{E}[Y|Z]$. Then, decomposing it according to the different missing patterns, we obtain

$$f^\star(Z) = \mathbb{E}[Y|Z] = \mathbb{E}\left[Y|X_{\text{obs}(m)}, M\right] = \sum_{m\in\mathcal{M}} f^\star_m(X_{\text{obs}(m)})\mathbb{1}_{M=m},$$

with $f^\star_m(X_{\text{obs}(m)}) := \mathbb{E}\left[Y|X_{\text{obs}(m)}, M = m\right]$. Then, note that under Assumption 1, $f^\star_m$ can be written as

$$f^\star_m(X_{\text{obs(m)}}) = \beta_0 + \beta^\top_{\text{obs(m)}} X_{\text{obs(m)}} + \beta^\top_{\text{mis}(m)}\mathbb{E}\left[X_{\text{mis(m)}}|X_{\text{obs(m)}}, M = m\right].$$

Thus, $f^\star_m$ remains linear in the observed variables X_{obs}, provided that the function

$$x \mapsto \mathbb{E}\left[X_{\text{mis(m)}}|X_{\text{obs(m)}} = x, M = m\right] \tag{2}$$

is linear. Proposition 2.3 of Ayme et al. (2022) makes a synthetic overview of all the assumptions that allow to obtain a pattern-wise linear Bayes predictor. These assumptions include having independent covariates (Assumption 5) or assuming Gaussianity in the covariates (Assumption 6), along with additional considerations regarding their relationship with the missing pattern (Assumption

2 for complete independency or Assumption 3 for dependence solely on the observed covariates). Furthermore, the links between these assumptions are studied.

Ayme et al. (2022) establish a bound on the excess risk of a truncated version of the least-squares estimator that takes into account only the observations falling into an infinite-norm ball. Unfortunately, this bound exponentially increases with the dimension. Indeed, this result is sharp in the case where all the missing patterns are equiprobable as 2^d regressions are required. Therefore, to improve this bound, they modify the introduced estimator in order to estimate only the predictors corresponding to the most frequent missing patterns. Under assumptions on the distribution of the missing patterns, this pattern frequency thresholded pattern-by-pattern estimator turns out to be minimax optimal.

2.2 (Regression + Classification) The package ToweranNA

It is an R package only used for prediction with missing values. It is based on the tower property or law of total expectation:

$$\mathbb{E}\left[Y|X_{\text{obs}(m)}\right] = \mathbb{E}\left[\mathbb{E}\left[Y|X_{\text{obs}(m)}, X_{\text{mis(m)}}\right]|X_{\text{obs}(m)}\right].$$

They make the following assumption:

Assumption 8 (Toweran-NA)**.** Given a missing pattern m,

$$\mathbb{E}\left[Y|X_{\text{obs}(m)}, M = m\right] = \mathbb{E}\left[Y|X_{\text{obs}(m)}\right].$$

Then, observe that under Assumption 8, the Bayes predictor under the quadratic loss and missing values can be simplified to

$$\begin{aligned} f^\star(Z) &= \mathbb{E}\left[Y|Z\right] \\ &= \mathbb{E}\left[Y|X_{\text{obs}(m)}, M\right] \\ &= \mathbb{E}\left[Y|X_{\text{obs}(m)}\right] && \text{(using Assumption 8)} \\ &= \mathbb{E}\left[\mathbb{E}\left[Y|X_{\text{obs}(m)}, X_{\text{mis(m)}}\right]|X_{\text{obs}(m)}\right] && \text{(using Tower Property)} \\ &= \mathbb{E}\left[f^\star_{\text{comp}}(X)|X_{\text{obs}(m)}\right], \end{aligned}$$

where $f^\star_{\text{comp}}(X)$ is the Bayes predictor with the complete data case. Note that the Bayes predictor under missing values can then be expressed with respect to the Bayes predictor under the complete data case. Based on this expression, Matloff and Mohanty (2023) proposes an algorithm that consists in:

1. train the predictive model over the complete data
2. apply it to the complete data to get $(\hat{Y}_i)_{i \in \text{train} \cap \text{comp}}$
3. for an input X_{new} with missing pattern M_{new}, find the k-nearest neighbours of X_{new} among $(M_{\text{new}} \odot X_i)_{i \in \text{train} \cap \text{comp}}$
4. predict with the average of the corresponding labels.

There are several drawbacks associated with this approach. Firstly, it should be noted that the Bayes predictor is only trained on complete data, resulting in the incomplete training data being wasted. Additionally, in certain cases such as data aggregation from multiple sources, this may be the only available data, making learning through this method impossible. The second drawback stems from the third step, as in high-dimensional statistics the sense of neighborhoods is lost due to the curse of dimensionality.

3 My contributions to classification with NA

This section provides an overview of the contributions made in the research on classification with missing values. Section 3.1 explicitly provides the expression of the pattern-by-pattern Bayes classifier.

3.1 Introduction

In this study, we aim to adapt classical classification algorithms to accommodate missing values, given that they are present in nearly all real-world datasets. We recall that given a missing pattern $m \in \{0,1\}^d$, we denote $\mathrm{obs}(m)$ the set of indices where m equals 0, i.e. the observed values indices and then, $X_{\mathrm{obs}(m)} \in \mathbb{R}^{d-\|m\|_0}$ are the observed values. In the sequel, we denote $Z = (X_{\mathrm{obs}(M)}, M)$. Thus, the objective is to predict a binary label Y from a given Z.

Following the idea of the classical classification, we quantify the accuracy of a classifier using the probability of misclassification given by

$$L(h) := \mathbb{P}(Y \neq h(Z)) \tag{3}$$

Therefore, we would like to find a classifier minimizing this probability of misclassification. As $|Y - h(Z)| \in \{0,2\}$, then,

$$L(h) = \frac{1}{4}\mathbb{E}\left[(Y - h(Z))^2\right] = \frac{1}{4}\mathbb{E}\left[(Y - \mathbb{E}\left[Y|Z\right])^2\right] + \frac{1}{4}\mathbb{E}\left[(\mathbb{E}\left[Y|Z\right] - h(Z))^2\right].$$

Thus, the Bayes predictor is

$$h^\star(Z) := \mathrm{sign}(\mathbb{E}\left[Y|Z\right]) = \mathrm{sign}(\mathbb{E}\left[Y|X_{\mathrm{obs}(M)}, M\right]) \text{ where } \mathrm{sign}(x) = \mathbb{1}_{x\geq 0} - \mathbb{1}_{x<0}.$$

As we have that

$$\mathbb{E}\left[Y|X_{\mathrm{obs}(M)}, M\right] = \sum_{m\in\mathcal{M}} \mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M = m\right] \mathbb{1}_{M=m},$$

then, the Bayes predictor can be written as

$$\begin{aligned} h^\star(Z) &= \mathrm{sign}(\mathbb{E}\left[Y|Z\right]) \\ &= \mathrm{sign}\Big(\sum_{m\in\mathcal{M}} \mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M = m\right] \mathbb{1}_{M=m}\Big) \\ &= \sum_{m\in\mathcal{M}} \mathrm{sign}(\mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M = m\right])\mathbb{1}_{M=m} \\ &= \sum_{m\in\mathcal{M}} h^\star_m(X_{\mathrm{obs}(m)})\mathbb{1}_{M=m} \end{aligned} \tag{4}$$

with

$$h^\star_m(X_{\mathrm{obs}(m)}) := \mathrm{sign}(\mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M = m\right]). \tag{5}$$

Thus, the aim is to compute a function $\widehat{h}_m : \mathbb{R}^{d-\|m\|_0} \to \{-1,1\}$ for each $m \in \{0,1\}^d$ to estimate $h^\star_m$ using an incomplete training sample $\mathcal{D}^\star_n$. Then, in order to account for missing values, we try to adapt classical algorithms to pattern-by-pattern models.

3.2 (Parametric modelling) Linear Discriminant Analysis with missing data

Linear discriminant analysis (LDA) relies on Gaussian assumptions of the distributions of $X|Y = k$ for each class k. This probabilistic model provides an explicit expression for the Bayes predictor $h^\star(X) = \text{sign}(\mathbb{E}[Y|X])$ when working with complete data. The cards are reshuffled when missing data occurs in the input X. We study how to adapt the LDA to missing data in Section 3.2.1, how missing data introduces an error in the risk compared to the complete data case in Section 3.2.2, how to estimate the model in Section 3.2.3, and finally, how to threshold it based on sparsity assumptions in Section 3.2.4.

3.2.1 Setting

First, we start formalizing the assumptions made for the standard LDA.

Assumption 9 (LDA)**.** Let Σ be a positive semi-definite, symmetric matrix of size $d \times d$. Set $\pi_1 = \mathbb{P}(Y = 1)$ and $\pi_{-1} = \mathbb{P}(Y = -1)$ such that $\pi_1, \pi_{-1} > 0$. For each class $k \in \{-1, 1\}$, $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma)$, with $\mu_k \in \mathbb{R}^d$.

Recall that in the complete case, the Bayes predictor reads as

$$h^\star_{\text{comp}}(x) := \text{sign}\left((\mu_1 - \mu_{-1})^\top \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_{-1}}{2} \right) - \log\left(\frac{\pi_{-1}}{\pi_1} \right) \right). \tag{6}$$

For the sake of simplicity, in the following we denote $\Sigma_{\text{obs}(m)} := \Sigma_{\text{obs}(m)\times\text{obs}(m)}$ (and $\Sigma^{-1}_{\text{obs}(m)} = (\Sigma_{\text{obs}(m)})^{-1}$).

When missing data occurs, one can decompose the Bayes predictor with respect to the missing patterns, as in Equation (4). Under MCAR data (Assumption 2), one can characterize $h^\star_m$, defined in (5), corresponding to the Bayes predictor conditional to a missing pattern m.

Proposition 3.1 (MCAR pattern-by-pattern LDA)**.** *Under Assumptions 2 and 9, the pattern-by-pattern Bayes classifier is given by*

$$h^\star_m(x_{\text{obs}(m)}) = \text{sign}\left(\left(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}\right)^\top \Sigma^{-1}_{\text{obs}(m)} \left(x_{\text{obs}(m)} - \frac{\mu_{1,\text{obs}(m)} + \mu_{-1,\text{obs}(m)}}{2} \right) - \log\left(\frac{\pi_{-1}}{\pi_1} \right) \right) \tag{7}$$

with $\pi_k = \mathbb{P}(Y = k)$ *for* $k \in \{-1, 1\}$.

Proof. The proof can be found in Appendix C.1.1 □

The proof of Proposition 3.1 relies mainly on the fact that, using the MCAR assumption, we only need the distribution of $X_{\text{obs}(m)}|Y$, proved to be Gaussian by Lemma B.1, similarly to the complete case. Notice that this does not hold anymore with a MAR missing mechanism, as conditionally to each missing pattern m, the Gaussian distribution of $X_{\text{obs}(m)}|Y, M = m$ may not be preserved. This is illustrated in the following example.

Example 3.2 (LDA+MAR is not pattern-by-pattern LDA). Let $X \in \mathbb{R}^2$ be a random variable satisfying Assumption 9, i.e., such that for each class k, $X|Y = k \sim \mathcal{N}(\mu_k, I_2)$. Let $M = (0, \mathbb{1}_{X_1>0})$ be the MAR missing pattern, where the first coordinate is always observed and the second is only observed if the first coordinate is positive. Note that it satisfies Assumptions 3 and 5. Therefore, $X_{\text{obs}(0,1)}|Y = k, M = (0, 1)$ does not admit a Gaussian distribution (the first coordinate being always positive).

It is now possible to quantify the two types of classification errors of the pattern-by-pattern Bayes classifier.

Proposition 3.3 (P-b-p-missclassification probability of p-b-p LDA)**.** *Under Assumptions 2 and 9, then*

$$\begin{aligned}&\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1\middle|Y=-1\right)\\ =\;& \Phi\left(-\frac{\log\left(\frac{\pi_{-1}}{\pi_1}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})\right\|} - \frac{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right)\end{aligned} \tag{8}$$

and symmetrically,

$$\begin{aligned}&\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1\middle|Y=1\right)\\ =\;& \Phi\left(-\frac{\log\left(\frac{\pi_{1}}{\pi_{-1}}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})\right\|} - \frac{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right)\end{aligned} \tag{9}$$

with Φ the standard Gaussian cumulative function and $\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}} := (\Sigma_{\mathrm{obs}(m)})^{-\frac{1}{2}}$.

Proof. The proof can be found in Appendix C.1.2 □

Proposition 3.3 establishes the probability of the two types of errors of the Bayes classifier. Several remarks are in order. As in the complete data case, the risk of misclassifying class $k \in \{-1,1\}$ decreases when the probability of class k increases. Indeed, the right-hand side of (8) is decreasing with π_{-1}.

Moreover, in a balanced setting where $\pi_{-1} = \pi_1$, then the probability of misclassification for both classes is the same. In this setting, assuming that the covariance matrix is $\Sigma = \sigma^2 I_d$, the probability of misclassification is

$$\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = k\middle|Y=-k\right) = \Phi\left(-\frac{\left\|\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)}\right\|}{2\sigma}\right). \tag{10}$$

Then, as the distance between the means of the observed covariates increases, the probability decreases, but as σ increases, the probability also increases as the classes are less recognizables. Remark that, if, for two missing patterns m_1 and m_2, $\mathrm{obs}(m_1) \subseteq \mathrm{obs}(m_2)$,

$$\left\|\mu_{1,\mathrm{obs}(m_1)}-\mu_{-1,\mathrm{obs}(m_1)}\right\| \leq \left\|\mu_{1,\mathrm{obs}(m_2)}-\mu_{-1,\mathrm{obs}(m_2)}\right\|, \tag{11}$$

and therefore

$$\mathbb{P}\left(h^\star_{m_1}(X_{\mathrm{obs}(m_1)}) = k\middle|Y=-k\right) \geq \mathbb{P}\left(h^\star_{m_2}(X_{\mathrm{obs}(m_2)}) = k\middle|Y=-k\right).$$

so that the associated risk is larger conditionally to m_1.

On the contrary, assume that $\pi_1 > \pi_{-1}$, and that $\Sigma = \sigma^2 I_d$. Denoting the class 1 as positive samples, and using Equation (11), with $m_2 = \mathbf{0}$, Equation (8) shows that missing data increases the false positive rate. In the sequel, we denote $p_m := \mathbb{P}(M = m)$.

Corollary 3.4 (Bayes Risk of p-b-p LDA)**.** *Under Assumptions 2 and 9, with balanced classes ($\pi_1 = \pi_{-1}$) and a diagonal covariance matrix $\Sigma = \sigma^2 I_d$, the Bayes risk is given by*

$$L(h^\star) = \sum_{m\in\{0,1\}^d} \Phi\left(-\frac{\left\|\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)}\right\|}{2\sigma}\right) p_m. \tag{12}$$

Proof. The proof can be found in Appendix C.1.3. □

Note that when dealing with unbalanced classes and a general covariance structure, one can similarly derive the expression of the Bayes risk; this can be found in Corollary C.1.

3.2.2 Characterizing the error introduced by missing data

In this section, we still focus on the case of balanced classes. We aim at comparing the Bayes risk with missing data (see Equation (12)) to the risk L_{comp} of misclassification with complete data

$$L_{\text{comp}}(h) := \mathbb{P}(h(X) \neq Y). \tag{13}$$

Let $h^\star_{\text{comp}}$ be the Bayes classifier of standard LDA (Assumption 9) with complete inputs, which was recalled in Equation (6). Straightforwardly, the Bayes risk with complete inputs boils down to

$$L_{\text{comp}}(h^\star_{\text{comp}}) = \Phi\left(-\frac{\left\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1})\right\|}{2}\right). \tag{14}$$

Therefore, the goal is to study whether the effect of missing values can be mitigated, at least in high dimension, or equivalently if the Bayes risk with missing values converges to the Bayes risk with complete data as the dimension d increases. To do this, we define the error introduced by missing data, as the difference between the Bayes risk with missing values (3) and the Bayes risk with complete values (13):

$$L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}). \tag{15}$$

Note that, from Corollary C.1 (using that $\pi_1 = \pi_{-1}$) and Equation (14), we have that

$$\begin{aligned} &L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) \\ &= \sum_{m \in \{0,1\}^d} \left(\Phi\left(-\frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|}{2}\right) - \Phi\left(-\frac{\left\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1})\right\|}{2}\right)\right) p_m, \end{aligned} \tag{16}$$

with Φ the c.d.f. of a standard Gaussian variable. Under the MCAR assumption, this quantity is always non-negative, as stated below.

Lemma 3.5. *Under assumptions of balanced classes and covariance matrix $\Sigma = \sigma^2 I_d$,*

$$L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) \geq 0. \tag{17}$$

Proof. Using Expression (16), note that

$$\left\|\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}\right\| \leq \|\mu_1 - \mu_{-1}\|.$$

Conclude using the monotonically increasing nature of Φ. □

To simplify the mathematical analysis, we consider the following set of assumptions.

Assumption 10 (Constant $\mathbb{P}(M_j = 1)$)**.** $\forall j \in \{1, ...d\}, \eta_j := \mathbb{P}(M_j = 1) = \eta < 1$.

Assumption 11 (Constant $(\mu_1 - \mu_{-1})_j$)**.** $\forall j \in \{1, ...d\}, (\mu_1 - \mu_{-1})_j = \pm\mu$, with $\mu > 0$.

Assumption 10 ensures that the missingness probability is the same for each input coordinate. Note that under this assumption and Assumption 2, we are dealing with a missingness mechanism that has been previously considered in multiple research works, such as Loh and Wainwright (2012), which is referred to as missing uniformly and completely at random (MUCAR). Assumption 11 guarantees that the difference between the means of the two classes is the same coordinate by coordinate. This can be achieved w.l.o.g. through a change of coordinates.

In Figure 1, we represent graphically the difference $L(h^\star) - L_{\text{comp}}$ (under Assumptions 10, 11 and $\Sigma = \sigma^2 I_d$) making either the coordinate-wise missingness probability η or the ambient dimension d vary. Note that even as the dimension d increases, η remains constant.

This difference vanishes with increasing dimension d (provided that $\eta \neq 1$). In the following proposition, we give a theoretical upper bound on this quantity. In the sequel, we refer to $\lambda := \mu/\sqrt{\lambda_{\max}(\Sigma)}$ as the signal-to-noise ratio, where $\lambda_{\max}(\Sigma)$ refers to the greatest eigenvalue of the covariance matrix. Indeed, this quantity describes the overlapping of the classes, and therefore the difficulty of the classification task.

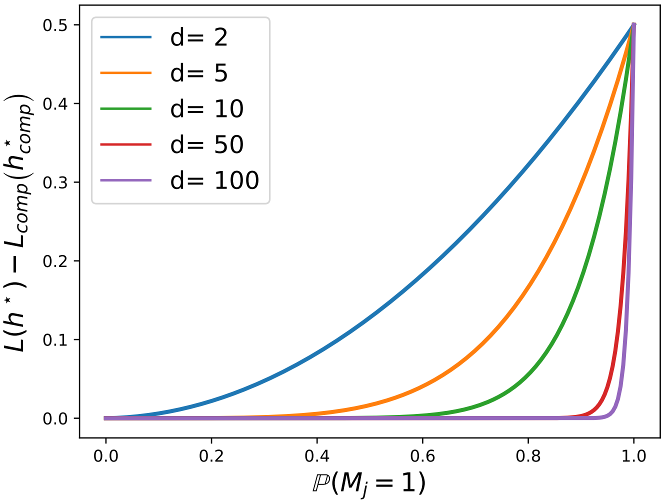


Figure 1: Difference between the Bayes risk for complete data and the Bayes risk for missing data $L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}})$ for different input dimensions with respect to the coordinate-wise missingness probability $\mathbb{P}(M_j = 1)$(with a fixed λ=2). It decreases exponentially with the dimension.

Proposition 3.6 (Bound on $L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}})$). *Under Assumptions 9, 10 and 11, with balanced classes ($\pi_1 = \pi_{-1}$), we have that*

$$\begin{aligned} &L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) \\ &\leq \left(\frac{1}{2} - \Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d + \frac{\mu}{2\sqrt{2\pi}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\left(\left(\eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^d - \eta^d\right)\right. \\ &\qquad \left. - \sqrt{\frac{d}{\lambda_{\max}(\Sigma)}}\left(\eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^{d-1} e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right). \end{aligned}$$

Proof. Appendix C.2.1 □

Proposition 3.6 establishes a bound on the difference between the Bayes risk with missing and complete data that, noticing that $\eta + e^{-\frac{\lambda^2}{8}}(1-\eta) < 1$, it decreases exponentially with dimension (by imposing an assumption of non exponential decrease of the minimum eigenvalue of the covariance matrix with d), as shown in Figure 1. Remark that in order to facilitate the comprehension of this exponential decrease, Corollary 3.8 proposes a simplified version. Note that if there are no missing values, i.e. $\eta = 0$, this upper bound vanishes, and $L_{\text{comp}}(h^\star_{\text{comp}})$ matches $L(h^\star)$.

When the signal-to-noise ratio $\lambda := \mu/\sqrt{\lambda_{\max}(\Sigma)}$ goes to infinity, one should expect the classification rate to be improved. However, it is important to consider the scenario in which all the values are missing, as it imposes a lower bound. Figure 2 shows graphically that (15) tends to this lower bound as the signal-to-noise ratio tends to infinity. In the subsequent corollary, we establish this lower bound, and then the convergence of the upper bound presented in Proposition 3.6 to this lower bound. This demonstrates the tightness of the previous proposition. To prove so, we make the following assumption on the order of the relationship between the eigenvalues as the signal-to-noise ratio tends to the infinity.

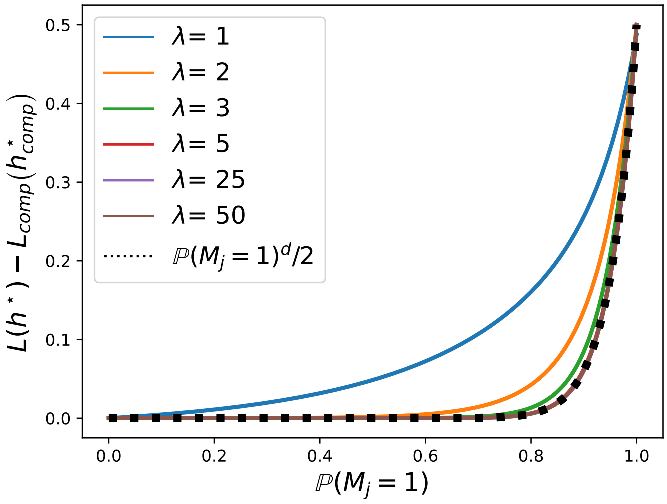


Figure 2: Difference between the Bayes risk for complete data and the Bayes risk for missing data $L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}})$ for different input signals-to-noises ratio with respect to the coordinate-wise missingness probability $\mathbb{P}(M_j = 1)$(with a fixed dimension d=20). It represents Corollary 3.7.

Assumption 12 (Order of $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$)**.** $\sqrt{\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)} = O_\lambda\left(e^{\lambda^2/8}/\lambda\right)$.

It is important to note that this assumption is quite broad and encompasses various scenarios. For instance, it covers cases where $\Sigma = \sigma^2 I_d$ or cases where the covariance matrix remains constant while the signal-to-noise ratio increases due to differences in the means. It also includes cases where the ratio $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ increases, possibly even exponentially.

Corollary 3.7. *Under Assumptions 9, 10, 11 and 12, with balanced classes ($\pi_1 = \pi_{-1}$), we have that*

$$L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) \xrightarrow[\lambda\to\infty]{} \frac{\eta^d}{2},$$

and the bound established in Proposition 3.6 captures this behaviour.

Proof. The proof can be found in Appendix C.2.2. □

To conclude this part, we simplify the upper bound on the Proposition 3.6 to explicitly capture its exponential decay and the dominant term.

Corollary 3.8. *Under Assumptions 9, 10 and 11, with balanced classes ($\pi_1 = \pi_{-1}$), we have that*

$$L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) \lesssim \mu\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\epsilon(\eta,\lambda)^d,$$

with $\epsilon(\eta,\lambda) := \eta + e^{-\frac{\lambda^2}{8}}(1-\eta) < 1$.

Proof. The proof can be found in Appendix C.2.3 □

Nevertheless, it is important to notice that when the dimension increases, the smallest eigenvalue may decrease. Assumptions such as bounding from below and above the eigenvalues of the covariance matrix are commonly extended in high dimensional statistics (see Tony Cai and Zhang (2019) or Cai and Liu (2011)). Here, to ensure that the misclassification error with missing values converges to the misclassification error with complete data, we can make an assumption about the decreasing order of the smallest eigenvalue of Σ. It only has to decrease slower than the exponential of the bound. Thus, convergence is ensured if $\lambda_{\min}(\Sigma)$ is lower bounded, or even if it decreases polynomially. To show the tightness of this upper bound up to a constant, see Figure 3.

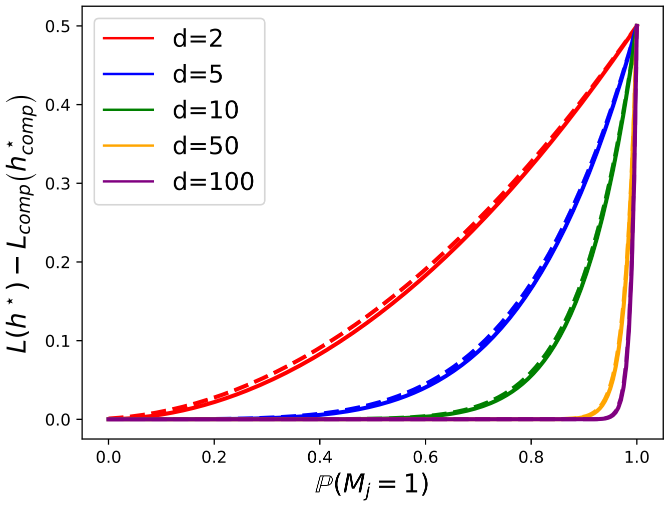


Figure 3: Difference between the Bayes risk for complete data and the Bayes risk for missing data $L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}})$ for different input dimensions with respect to the coordinate-wise missingness probability $\mathbb{P}(M_j = 1)$(with a fixed signal-to-ratio λ=5). Continuous lines represent the true difference and the discontinuous lines represent the bound of Corollary 3.8 divided by 10.

3.2.3 LDA estimation with missing values

In Section 3.2.1, the optimality of the pattern-by-pattern LDA was established. However, in order to construct a data-driven classifier based on the random sample $\mathcal{D}_n$, the parameters of the Gaussian model needs to be estimated. The case of classical LDA with intuitive estimates was addressed in Anderson (2003), where the convergence of the estimates demonstrates the convergence of the data-driven classifier. To treat the complete data case in high dimensional setting, where the dimension d is much larger than the number of samples n, many regularizations have been proposed such as regularized linear discriminant analysis (LDA) of Wu et al. (2009) or the covariance-regularized classification of Witten and Tibshirani (2009). Another approach was proposed by Cai and Liu (2011). The authors made the observation that the Bayes classier depends on the μ_1, μ_{-1} and Σ mainly through the quantity $\beta := \Sigma^{-1}(\mu_1 - \mu_{-1})$, which is called the discriminant direction. Under assumptions of sparsity of β and balanced classes, they proposed a consistent classifier called linear programming discriminant(LPD). Their estimate of the discriminant direction is given by

$$\widehat{\beta}_{\mathrm{LPD}} := \operatorname{argmin}_{\beta} \left\{ \|\beta\|_1 : \text{subject to } \left\| \widehat{\Sigma}\beta - (\widehat{\mu}_1 - \widehat{\mu}_{-1}) \right\|_\infty \leq \lambda_n \right\},$$

where $\widehat{\mu}_1, \widehat{\mu}_{-1}$ and $\widehat{\Sigma}$ are the classical estimates, and λ_n is a tuning parameter. Then, the data-driven classifier is given by

$$\widehat{h}_{\mathrm{LPD}}(X) = \begin{cases} 1 & \text{if} \quad \widehat{\beta}_{\mathrm{LPD}}^\top \left(X - \frac{\widehat{\mu}_1 + \widehat{\mu}_{-1}}{2} \right) > 0 \\ -1 & \text{otherwise.} \end{cases}$$

This is a simple plug-in of the estimates into the theoretical LDA(6). They establish a rate of convergence of $O\left((s\log(d)/n)^{1/2}\right)$ over $L_{\mathrm{comp}}(\widehat{h}_{\mathrm{LPD}}) - L_{\mathrm{comp}}(h^\star_{\mathrm{comp}})$.

On the other hand, we observe that the same constraint of λ_n is imposed for all the coordinates, making an assumption of homocedasticity over the residuals of $\widehat{\Sigma}\beta - (\widehat{\mu}_1 - \widehat{\mu}_{-1})$. Moreover, this parameter λ_n must be tuned empirically through methods such as cross-validation. To resolve these drawbacks, Tony Cai and Zhang (2019) proposed an adaptive algorithm for high dimensional LDA with complete data, called AdaLDA. It is tuning-free and shown to be minimax rate optimal. They construct an element-wise constraint on the $\widehat{\Sigma}\beta - (\widehat{\mu}_1 - \widehat{\mu}_{-1})$ based on a concentration inequality. Furthermore, to accomodate incomplete data, they generalize adaLDA into ADAM under Assumption 2, which is also proven to be minimax rate optimal.

It is important to note that none of these works addresses the problem we want to solve here, which is predicting on new data that may contain missing entries. Indeed, while ADAM deals with missing values to construct the classifier, it predicts values based on complete data. Therefore, the convergence rate bound established for ADAM is over $L_{\mathrm{comp}}(\widehat{h}_{\mathrm{ADAM}}) - L_{\mathrm{comp}}(h^\star_{\mathrm{comp}})$.

Proposition 3.1 provides a pattern-by-pattern Bayes classifier that can be estimated by empirical evaluations of $\pi_k, \mu_{k,\mathrm{obs}(m)}$ and $\Sigma_{\mathrm{obs}(m)}$ for $k \in \{-1, 1\}$ and $m \in \mathcal{M}$. Therefore, given the training set with missing values $(X_i, M_i, Y_i)_{i=1,\ldots n}$, one can naively estimate π_k with the proportion of training samples of class k (the output being always observed). An intuitive way to estimate $\mu_{k,\mathrm{obs}(m)}$ and $\Sigma_{\mathrm{obs}(m)}$ is to compute respectively the empirical mean and covariance over the samples sharing the same missing pattern, i.e.,

$$\widehat{\pi}_{k,m} = \frac{\sum_{i=1}^n \mathbb{1}_{y_i = k} \mathbb{1}_{m_i = m}}{\sum_{i=1}^n \mathbb{1}_{m_i = m}}, \qquad \widehat{\mu}_{k,\mathrm{obs}(m)} = \frac{\sum_{i=1}^n x_i \mathbb{1}_{y_i = k} \mathbb{1}_{m_i = m}}{\sum_{i=1}^n \mathbb{1}_{y_i = k} \mathbb{1}_{m_i = m}},$$

$$\widehat{\Sigma}_{\mathrm{obs}(m)} = \frac{\sum_{i=1}^n (x_i - \widehat{\mu}_{k,\mathrm{obs}(m)})(x_i - \widehat{\mu}_{k,\mathrm{obs}(m)})^\top \mathbb{1}_{y_i = k} \mathbb{1}_{m_i = m}}{\sum_{i=1}^n \mathbb{1}_{y_i = k} \mathbb{1}_{m_i = m}}.$$

Unfortunately, this requires as many estimates as missing patterns, a number that grows exponentially with the dimension. In addition, the number of observations corresponding to a given missing pattern can be low.

In this section, we analyze the convergence rate of the estimated LDA by using a plug-in approach with an intuitive estimate of the mean. We first consider the simplified case of $\Sigma = \sigma^2 I_d$, and then extend it to a general covariance matrix, accounting for biases arising from coordinate asymmetries.

Working with a known covariance matrix $\Sigma = \sigma^2 I_d$. We start by studying the simple case of balanced classes in which the covariance matrix Σ is known and satisfies $\Sigma = \sigma^2 I_d$. In this setting, we propose to use the following mean estimators: for $k \in \{-1, 1\}$ and for $1 \leq j \leq d$,

$$\widehat{\mu}_{k,j} = \frac{\sum_{i=1}^n X_{i,j} \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}} = \frac{\sum_{i=1}^n (X_i \odot (1 - M_i))_j \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}, \tag{18}$$

considering the convention $0/0 = 0$. We observe that under MCAR assumptions, these estimates are built with all the observed inputs, independently of their missing patterns. Firstly, it should be noted that this estimate is asymptotically unbiaised , as demonstrated in the following proposition.

Proposition 3.9. *Let $\mathcal{D}_n$ be a sample satisfying Assumptions 2, 9 and $\eta_j := \mathbb{P}(M_j = 1) < 1$. Let $\widehat{\mu}_{k,j}$ be the estimate defined in 18. Then, we have that*

$$\mathbb{E}\left[\widehat{\mu}_{k,j}\right] \rightarrow \mu_{k,j}.$$

Proof. The proof can be found in Appendix C.3.1 □

With these estimators in mind, we compute the μ-estimated classifier (empirical version of (4)), which results in

$$\widehat{h}(x_{\mathrm{obs}(m)}, m) = \sum_{m' \in \mathcal{M}} \widehat{h}_{m'}(x_{\mathrm{obs}(m)}) \mathbb{1}_{m'=m}, \tag{19}$$

where, by Proposition 3.1, we have

$$\widehat{h}_m(x_{\mathrm{obs}(m)}) = \mathrm{sign}\left(\left(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)} \right)^\top \frac{1}{\sigma^2} \left(x_{\mathrm{obs}(m)} - \frac{\widehat{\mu}_{1,\mathrm{obs}(m)} + \widehat{\mu}_{-1,\mathrm{obs}(m)}}{2} \right) \right). \tag{20}$$

Hereafter, we provide the convergence rate of the generalization error of the corresponding LDA classifier.

Theorem 3.10 (Bound on the μ-estimated p-b-p LDA)**.** *Let $\mathcal{D}_n$ be a sample satisfying Assumptions 2, 9 and 10, with balanced classes and covariance matrix $\Sigma = \sigma^2 I_d$. Then the classifier $\widehat{h}$, defined in (19) with $\widehat{\mu}_1, \widehat{\mu}_{-1}$ estimated as in (18), satisfies the following condition*

$$L(\widehat{h}) - L(h^\star) \leq \frac{2\sqrt{d}}{\sqrt{2\pi}} \left(\frac{\|\mu\|_\infty^2 (1-\eta)}{\sigma^2} \frac{(1+\eta)^n}{2^n} + \frac{4}{n+1} \right)^{\frac{1}{2}}.$$

Then, for n large enough, if $\eta < 1$, we have $L(\widehat{h}) - L(h^\star) \lesssim \sqrt{d/n}$.

Proof. The proof can be found in Appendix C.3.4. □

We observe that Theorem 3.10 provides a convergence rate of the order of $(d/n)^{1/2}$ for the generalization error of the LDA classifier in presence of missing values, when the means of the two classes are not known. Moreover, it is *independent* of the standard deviation σ, as the term depending on it decreases exponentially and is due to the all missing data-case. Finally, we observe that the estimate gets worse as the dimension d increases, but as the number of observations n increases, the bound becomes more accurate.

This upper bound is independent of the missingness probability η. This may be surprising at first sight. To understand this phenomenon, it is important to note that the quantity to be considered is the difference between the misclassification probabilities of the estimated LDA and the incomplete LDA. This incomplete theoretical LDA framework has already been adapted to handle the presence of missing values in the data. Thus, there exists a trade-off between the estimation difficulty and the frequency of estimation. However, it should be noted that this trade-off no longer applies to the difference between misclassification probabilities in the estimation under incomplete data and the complete data case, as demonstrated in the following corollary.

Combining Proposition 3.6 and Theorem 3.10 allows us to establish an upper bound on the error of the LDA classifier with missing values (compared to the Bayes risk for complete data).

Corollary 3.11. *Let $\mathcal{D}_n$ be a sample satisfying Assumptions 2, 9, 10 and 11, with balanced classes and covariance matrix $\Sigma = \sigma^2 I_d$. Then the classifier $\widehat{h}$, defined in (19) with $\widehat{\mu}_1, \widehat{\mu}_{-1}$ estimated as in (18), satisfies the following condition*

$$\begin{aligned} &L(\widehat{h}) - L_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) \\ &\quad \leq \frac{2\sqrt{d}}{\sqrt{2\pi}} \left(\frac{\|\mu\|_\infty^2 (1-\eta)}{\sigma^2} \frac{(1+\eta)^n}{2^n} + \frac{4}{n+1} \right)^{\frac{1}{2}} \\ &\qquad + \left(\frac{1}{2} - \Phi\left(-\frac{\mu}{2\sigma}\sqrt{d} \right) \right) \eta^d + \frac{\eta\mu\sqrt{d}}{2\sigma\sqrt{2\pi}} \left(\left(\eta + e^{-\frac{\mu^2}{8\sigma^2}} (1-\eta) \right)^{d-1} - \eta^{d-1} \right). \end{aligned}$$

Proof. The proof can be found in Appendix C.3.5. □

First, note that if n tends to infinity, then the generalization error (first term) vanishes, highlighting the consistency of the estimator $\widehat{h}$ ($L(\widehat{h}) \to L(h)$, see Theorem 3.10). When the dimension increases, the rest of the bound vanishes whereas the generalization error suffers from the curse of dimensionality (see Theorem 3.10). Furthermore, note that the learning error $(L(\widehat{h}) - L(h^\star))$ is of the order of $\sqrt{d/n}$ and that the approximation error due to the missing values $(L(h^\star) - L_{\mathrm{comp}}(h^\star_{\mathrm{comp}}))$ is of the order of $\lambda\eta\sqrt{d}\epsilon(\eta, \lambda)^d$, where λ is the signal-to-noise ratio, η the missingness probability, and ϵ is defined in Corollary 3.8. Note that for a sufficiently large d, as the approximation error decreases exponentially, the learning error becomes dominant. Then, with d verifying

$$-\frac{\log(\sqrt{n}\lambda\eta)}{\log(\epsilon(\eta, \lambda))} \lesssim d,$$

we have that the learning error is greater than the approximation error. Then, the missing values are negligible, i.e. the error is mostly made by the approximation and the missing values do not affect in the order of misclassification. Furthermore, we have

$$L(\widehat{h}) - L_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) \lesssim \sqrt{\frac{d}{n}}.$$

Note that it is mostly the quantity $L(h^\star) - L_{\mathrm{comp}}(h^\star_{\mathrm{comp}})$ that suffers from the impact of the missing values, as the dominant term

$$\left(\eta + e^{-\frac{\mu^2}{8\sigma^2}} (1-\eta) \right)^{d-1}$$

increases notably when η increases, i.e., with the presence of missing values. Finally, assuming that $d = o(n)$, the misclassification risk of the estimated LDA with missing values converges to the Bayes risk with complete data.

Working with a known general covariance matrix Σ. In this paragraph we generalize the previous theorem to the case of a general covariance matrix. We define $\rho := \max_{i\in[n]}(\Sigma_{i,i})/\lambda_{\min}(\Sigma)$ the greatest value of the diagonal of the covariance over its smallest eigenvalue, which can be seen as a non-standard condition number of Σ. To cover this case, we perform a similar analysis as before, invoking Lemma C.6 instead of Lemma C.5.

Theorem 3.12 (Bound on the μ-estimated p-b-p LDA with general Σ)**.** *Let $\mathcal{D}_n$ be a sample satisfying Assumptions 2, 9 and 10, with balanced classes. Then the classifier $\widehat{h}$, defined in* (19) *with $\widehat{\mu}_1, \widehat{\mu}_{-1}$ estimated as in* (18)*, satisfies the following condition*

$$L(\widehat{h}) - L(h^\star) \leq \frac{2}{\sqrt{2\pi}} \left(\left(\frac{1+\eta}{2}\right)^n \frac{\|\mu\|_\infty^2\, d(1-\eta)}{\lambda_{\min}(\Sigma)} + \frac{4\rho d}{n} \right)^{\frac{1}{2}},$$

with $\rho := \max_{i\in[n]} \Sigma_{i,i}/\lambda_{\min}(\Sigma)$ the greatest value of the diagonal of the covariance matrix divided by its smallest eigenvalue. Then, for n large enough, we have $L(\widehat{h}) - L(h^\star) \lesssim \sqrt{\rho d/n}$.

Proof. The proof can be found in Appendix C.3.7. □

Note from Theorem 3.10 and Theorem 3.12 that when introducing the general covariance matrix Σ, a factor ρ is introduced to account for the asymmetries of the covariance structure of the inputs. As a result, we can observe varying levels of class distinguishability across different coordinates. When a coordinate with a small variance (consequently, the classes exhibit greater distinctiveness) is missing, then the risk of misclassification increases. Note that the bound from Theorem 3.12 matches the one from Theorem 3.10 when $\Sigma = \sigma^2 I_d$.

3.2.4 LDA estimation under sparsity assumptions

Due to the type of data that has emerged in recent years, the issue of dealing with high dimensionality has become a significant real-world problem. One common assumption in this context is the sparsity assumption, which implies that not all variables are discriminant. In the context of Linear Discriminant Analysis (LDA), one could assume that the difference between the class centers is s-sparse, as formalized in the following assumption.

Assumption 13 (Sparsity)**.** The difference of centers $\mu_1 - \mu_{-1}$ is s-sparse, i.e., the cardinality of the supports $\mathrm{supp}(\mu_1 - \mu_{-1}) := \{j \in [d], \mu_{1,j} - \mu_{-1,j} \neq 0\}$ is at most $s \ll d$.

Under such an assumption, only the directions corresponding to the non-zero components of $\mu_1 - \mu_{-1}$ are relevant for classification purposes. It should be noted that restricted to the remaining components, both classes are centered around 0, and therefore indistinguishable.

One classical approach to handle high dimensions in regression is penalizing the cost function to reflect the sparse prior of the estimated parameters. This is usually done through ℓ_1-norm regularization, solved in practice using soft-thresholding operators.

In the case of LDA classification with equal covariance matrices equal to $\sigma^2 I_d$, we propose a simple and efficient method relying on thresholding the empirical centers, giving thereby more trust to coordinates that have been observed more frequently. More precisely, our estimate is given by

$$\widetilde{\mu}_{k,j} := \widehat{\mu}_{k,j} \mathbb{1}_{\widehat{\mu}_{k,j} > \tau_{k,j}} \tag{21}$$

where

$$\tau_{k,j} := 2\sigma\sqrt{\frac{\log(d)}{N_{k,j}}}. \tag{22}$$

and $\widehat{\mu}_{k,j}$ is defined in (18) and $N_{k,j} := \sum_{i=1}^{n} \mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,j}=0}$ is the number of times that the j-th coordinate is observed for the class k. This estimate can mitigate the curse of dimensionality as shown in the following theorem.

Theorem 3.13 (Bound on the μ-estimated p-b-p LDA under Sparsity Assumptions)**.** *Let $\mathcal{D}_n^\star$ be a sample satisfying Assumptions 2, 9, 10 and 13, with balanced classes and covariance matrix $\Sigma = \sigma^2 I_d$. Then the classifier $\widetilde{h}$, defined in (19) with $\widetilde{\mu}_1, \widetilde{\mu}_{-1}$ estimated as in (21), satisfies the following condition*

$$L(\widetilde{h}) - L(h^\star) \lesssim \left(\frac{s\log(d)}{n} + \frac{\|\mu\|_\infty^2}{\sigma^2}\left(\frac{1+\eta}{2}\right)^n s(1-\eta)\right)^{\frac{1}{2}}.$$

Then, for n large enough, if $\eta < 1$, we have $L(\widetilde{h}) - L(h^\star) \lesssim \sqrt{s\log(d)/n}$.

Proof. The proof can be found in Appendix C.4.2. □

3.2.5 LDA under MNAR assumption (GPMM)

This section focuses on the MNAR assumption (Assumption 4). As discussed in Section 2, under this assumption we assume that we can extract information about the missing values thanks to the missing pattern. This could be the case for missing values due to sensor failure, due to instruments not adapted to extreme values, or due to refusal to answer sensitive questions such as income. Unfortunately, as shown in Example 3.2, the MAR assumption (Assumption 3) is not sufficient to preserve the LDA model restricted to a given missing pattern. To circumvent this issue, we consider a Gaussian Pattern Mixture Model (Assumption 7) coupled with an LDA predictive model, as defined hereafter.

Assumption 14 (GPMM-LDA)**.** For all $m \in \mathcal{M}$ and for all $k \in \{-1, +1\}$, $X_{\mathrm{obs}(m)}|(M = m, Y = k) \sim \mathcal{N}(\mu_{m,k}, \Sigma_m)$.

The Gaussian Pattern Mixture Model first appeared in the context of regression with missing inputs in Le Morvan et al. (2020b,a); Ayme et al. (2022) .

Setting In the following, we study a p-b-p LDA with GPMM inputs (see Proposition 3.1 for MCAR inputs). Then, we quantify the classification error of the p-b-p Bayes classifier (see Proposition 3.3 for MCAR inputs). Note that the proofs of this section, provided in Appendix C.5, are very similar to the ones studied in the MCAR case. In the sequel, we denote $\pi_{m,k} := \mathbb{P}(Y = k, M = m)$.

Proposition 3.14 (Bayes classifier for GPMM-LDA)**.** *Under Assumption 14, the pattern-by-pattern Bayes classifier is given by*

$$h_m^\star(x_{\mathrm{obs}(m)}) = \\ \operatorname{sign}\left((\mu_{m,1} - \mu_{m,-1})^\top \Sigma_m^{-1}\left(x_{\mathrm{obs}(m)} - \frac{\mu_{m,1} + \mu_{m,-1}}{2}\right) - \log\left(\frac{\pi_{m,-1}}{\pi_{m,1}}\right)\right)$$

with $\pi_{m,k} := \mathbb{P}(Y = k, M = m)$ for $k \in \{-1, 1\}$ and $m \in \{0, 1\}^d$.

Proof. The proof can be found in Appendix C.5.1 □

Note that this proof primarily relies on the Gaussianity assumption of $X_{\mathrm{obs}(m)}|Y, M = m$, instead of that of $X_{\mathrm{obs}(m)}|Y$ in the MCAR case.

Furthermore, it is important to note that under this modeling, the proportion of classes $\pi_{m,-1}/\pi_{m,1}$ used in the p-b-p classifier is specific to class m and not global for all the missing patterns, as in the MCAR setting.

To control the generalization error $L(h^\star)$ of $h^\star$, we can proceed similarly as in Corollary C.1, i.e.,

$$\begin{aligned}
L(h^\star) &= \mathbb{P}\left(h^\star(X_{\mathrm{obs}(M)}, M) \neq Y\right) \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star(X_{\mathrm{obs}(m)}, M) \neq Y \middle| M = m\right) p_m \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m\right) \mathbb{P}(Y = 1|M = m) p_m \\
&\quad + \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m\right) \mathbb{P}(Y = -1|M = m) p_m \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m\right) \pi_{m,1} \\
&\quad + \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m\right) \pi_{m,-1}.
\end{aligned}$$

Contrary to the MCAR case, the conditioning to the missing pattern cannot be dropped anymore. It is now possible to quantify the two types of classification errors of the pattern-by-pattern Bayes classifier in the GPMM-LDA setting.

Proposition 3.15 (P-b-p-missclassification probability of p-b-p GPMM-LDA)**.** *Under Assumptions 2 and 9, then*

$$\begin{aligned}
&\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m\right) \\
&= \Phi\left(-\frac{\log\left(\frac{\pi_{m,-1}}{\pi_{m,1}}\right)}{\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \mu_{m,-1})\right\|} - \frac{\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \mu_{m,-1})\right\|}{2}\right)
\end{aligned} \tag{23}$$

and symmetrically,

$$\begin{aligned}
&\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m\right) \\
&= \Phi\left(-\frac{\log\left(\frac{\pi_{m,1}}{\pi_{m,-1}}\right)}{\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \mu_{m,-1})\right\|} - \frac{\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \mu_{m,-1})\right\|}{2}\right)
\end{aligned} \tag{24}$$

with Φ the standard Gaussian cumulative function.

Proof. The proof can be found in Appendix C.5.2 □

Next, following the approach taken with the MCAR assumption, we assume that all classes are uniformly distributed given a specific missing pattern, i.e. $\mathbb{P}(Y = k|M = m) = 1/2$.

Estimation Using a training set $\mathcal{D}_n^\star$ (with missing data), the objective is to obtain a plug-in estimator of the Bayes classifier for the GPMM-LDA setting. To obtain theoretical guarantees, we consider an overly simplified setting in which the covariance matrix for each missing pattern is assumed to be known: only the means need to be estimated. Under Assumption 14, we observe that for a given missing pattern m, one has to estimate 2 mean parameters $\mu_{m,1}$ and $\mu_{m,-1}$. This can be done simply by computing empirical means over the observations that share the same missing pattern and class:

$$\widehat{\mu}_{m,k} := \frac{\sum_{i=1}^n X_i \mathbb{1}_{Y_i=k} \mathbb{1}_{M_i=m}}{\mathbb{1}_{Y_i=k} \mathbb{1}_{M_i=m}}. \tag{25}$$

Nevertheless, there are several drawbacks to this approach. It suffers from a high computational complexity: one needs to estimate as many mean parameters as twice the number of missing patterns, which can scale as 2^d. This may be accompanied by an overfitting issue as there may not be enough samples specifically to each missing pattern to ensure a good estimation. To address this problem, inspired by the thresholded p-b-p linear regression proposed by Ayme et al. (2022), we propose to estimate only the missing patterns that are more likely to occur. Therefore, the final estimate is given by:

$$\widetilde{\mu}_{m,k} := \widehat{\mu}_{m,k} \mathbb{1}_{\frac{N_{m,k}}{n} > \tau} , \tag{26}$$

with $\tau := \sqrt{d/n}$ and $N_{m,k} := \sum_{i=1}^n \mathbb{1}_{M_i=m} \mathbb{1}_{Y_i=k}$ the number of observations of the class k with m as missing pattern. Note that this estimate is only useful when $d < n$.

Theorem 3.16 (MNAR p-b-p LDA estimation)**.** *Under Assumption 14, the plug-in classifier based on* (26) *satisfies*

$$L(\widetilde{h}) - L(h^\star) \leq \left(\frac{4}{\sqrt{2\pi}} + \frac{8}{\sqrt{\pi}} \frac{d \left\lVert \mu \right\rVert_\infty}{\sqrt{\lambda_{\min}(\Sigma)}} \right) \sum_{m \in \{0,1\}^d} \tau \wedge p_m$$

Proof. The proof can be found in Appendix C.5.5. □

Denoting

$$\mathfrak{C}_p(\tau) := \sum_{m \in \{0,1\}^d} \tau \wedge p_m,$$

we recover the missing pattern distribution complexity used in Ayme et al. (2022). This bound remains small in scenarios where the set of frequent missing patterns has a small cardinality (refer to Lemma 4.2 of Ayme et al. (2022)). For more comprehensive details on this complexity, we suggest referring to Section 4.2 of Ayme et al. (2022).

3.2.6 LDA under the MNAR (self-masking assumption)

In this section, we address another missing not at random mechanism. Based on this assumption, we can provide a practical method that circumvents the exponential number of estimates typically introduced by MNAR mechanisms. To accomplish this, we begin by presenting a general decomposition of the pattern-by-pattern Bayes classifier, while only assuming the Gaussianity of the underlying model for both classes (Assumption 9).

Proposition 3.17 (General p-b-p bayes classifier)**.** *Under LDA assumption (Assumption 9), the pattern-by-pattern Bayes classifier can be decomposed as*

$$h^\star_m(X_{\mathrm{obs}(m)}) = \mathrm{sign}\left(\left(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right)^\top \Sigma^{-1}_{\mathrm{obs}(m)} \left(x_{\mathrm{obs}(m)} - \frac{\mu_{1,\mathrm{obs}(m)} + \mu_{-1,\mathrm{obs}(m)}}{2}\right)\right.$$
$$\left. - \log\left(\frac{\mathbb{P}\left(M = m|Y = -1, X_{\mathrm{obs}(m)}\right)\pi_{-1}}{\mathbb{P}\left(M = m|Y = 1, X_{\mathrm{obs}(m)}\right)\pi_1}\right)\right)$$

Proof. The proof can be found in Appendix C.6.1. □

First, notice that if we add the assumption of independence between the missing pattern and the covariates and label (Assumption 2), then this pattern-by-pattern Bayes classifier is consequently the one given in Proposition 3.1. Then, notice that contrary to Section 3.2.5, $\mu_{k,\mathrm{obs}(m)}$ and $\Sigma_{\mathrm{obs}(m)}$ represent the projected parameters of the underlying mean and covariance of the class, rather than being specific to each missing pattern. Finally, notice that in order to derive a more explicit expression for the pattern-by-pattern Bayes classifier, we need to introduce specific assumptions concerning the distribution of $M = m|Y = k, X$. A MNAR setting already considered in the literature(see Le Morvan et al. (2020b); Sportisse (2021)) is when the missingness of a value depends solely on the value itself and the class (hence the name of self-masked missing values).

Assumption 15 (Self-masking)**.** Given an $m \in \mathcal{M}$, then

$$\mathbb{P}(M = m|X, Y = k) = \prod_{j=1}^{d} \mathbb{P}\left(M_j = m_j|X_j, Y = k\right).$$

Consequently, based on this assumption, the missingness of a coordinate depends only on the value of that coordinate and its corresponding class. This is an adaptation of the missing scenario considered by Le Morvan et al. (2020b) and Ayme et al. (2022) in the context of regression.

Another prevalent method used to handle missing not at random mechanisms involves considering monotone mechanisms. To accommodate this assumption, various strategies have been employed, such as logistic missing random. Additionally, Miao et al. (2015) provides necessary and sufficient conditions for model identifiability under the logistic missing mechanism. We say that a model is identifiable if the distribution of the underlying model can be uniquely determined by the observed distribution. Nevertheless, this research only applies on missingness related to the label variable Y and completeness of the covariates X, which are the complementary assumptions of the ones followed in this article.

Furthermore, our goal is to extend this monotone missingness mechanism to handle missing values whenever they fall either within a relatively small range or within a relatively large range. This situation arises, for instance, with sensors that only function within specific intervals, such as a speed sensor or a thermometer.

Assumption 16 (Extreme missingness)**.** For every $j \in [d]$ and $k \in \{-1, 1\}$, there exists two cut-offs $\beta_{k,j,1}, \beta_{k,j,2} \in \mathbb{R}$ such that $\mathbb{P}\left(M_j = 0|X_j, Y = k\right) = \mathbb{1}_{\beta_{k,j,1} \le X_j \le \beta_{k,j,2}}$.

As a result, any values outside this permissible interval are always missing and the ones inside are always observed.

Note that under Gaussian assumptions by class, self-masking and extreme missingness (resp. Assumptions 9, 15 and 16), we have that the coordinate-by-coordinate distribution of the observed

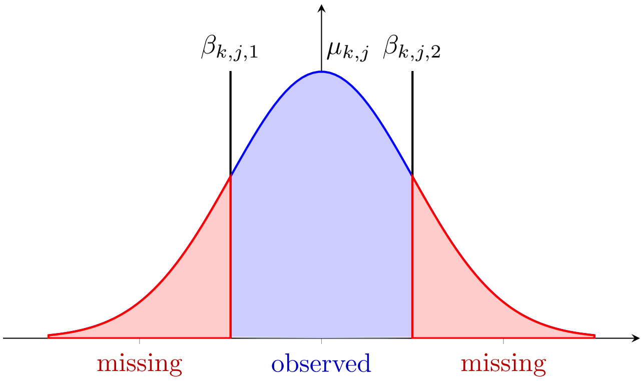


Figure 4: Underlying j-th coordinate distribution of $X|Y = k$. Note that we only have access to the observations between $\beta_{k,j,1}$ and $\beta_{k,j,2}$.

covariates are distributed as a truncated Gaussian, as shown in Figure 4. To see this, given $j \in [d]$ and $k \in \{-1, 1\}$, we have that

$$\mathbb{P}(X_j|Y = k, M_j = 0) = \frac{\mathbb{P}(M_j = 0|X_j, Y = k)}{\mathbb{P}(M_j = 0)}\mathbb{P}(X_j|Y = k) = \frac{\mathbb{1}_{\beta_{k,j,1} \leq X_j \leq \beta_{k,j,2}}}{\mathbb{P}(M_j = 0)}\mathbb{P}(X_j|Y = k),$$

with $X_j|Y = k \sim \mathcal{N}(\mu_{k,j}, \Sigma_j)$ by Lemma B.1.

The corollary presented below simplifies the p-b-p Bayes classifier described in Proposition 3.17 by applying the *extreme self-masking* assumptions on the missingness distribution.

Corollary 3.18 (Extreme self-masking p-b-p Bayes classifier)**.** *Under Assumption 9 on the complete data distribution and Assumptions 15 and 16 on missing pattern, the pattern-by-pattern Bayes classifier can be expressed as*

$$h^\star_m(X_{\mathrm{obs}(m)}) = \mathrm{sign}\Bigg(\left(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right)^\top \Sigma^{-1}_{\mathrm{obs}(m)}\left(x_{\mathrm{obs}(m)} - \frac{\mu_{1,\mathrm{obs}(m)} + \mu_{-1,\mathrm{obs}(m)}}{2}\right) - \log\left(\frac{\pi_{-1}}{\pi_1}\prod_{j \notin \mathrm{obs}(m)} \frac{\mathbb{P}\left(M_j = 1|Y = -1\right)}{\mathbb{P}\left(M_j = 1|Y = 1\right)}\right)\Bigg)$$

Proof. The proof can be found in Appendix C.6.2. □

Estimation It should be noted that under this missingness mechanism, the parameters of the underlying model cannot be estimated solely using the observed variables. For instance, due to the refusal-to-answer behavior of rich individuals in surveys, the average wage appear lower than it actually is. As a consequence, calculating the mean as done in previous sections by the empirical mean over the observed covariates is no longer feasible due to the introduction of bias. It is worth noting that for a Gaussian distribution, the mean is equivalent to the *median*. However, estimating the mean through the median would also introduce bias under these assumptions. Additionally, the *mode* serves as an estimator for the mean in the case of a Gaussian distribution. In contrast to previous estimates, under some assumptions for the permissible interval, the mode is preserved for the observed distribution, enabling us to estimate it and consequently estimate the mean of the underlying distribution. An illustrative example is presented in Figure 5.

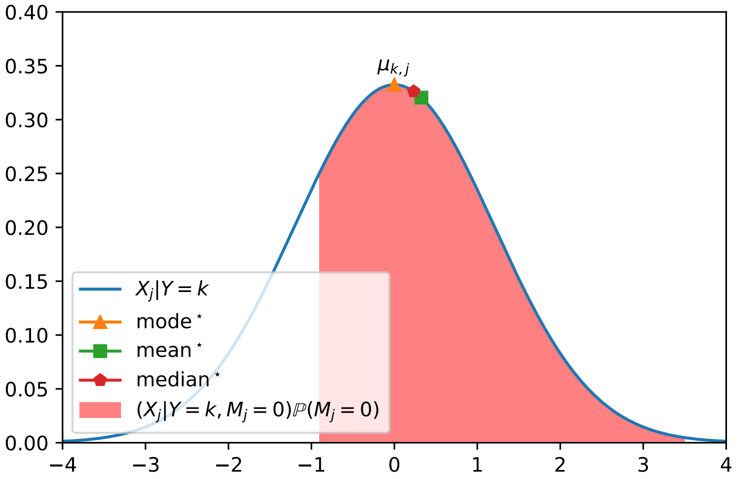


Figure 5: Example of data modeled under Assumptions 9, 15 and 16. The permissible interval is $[-0.9, 3.5]$, and the observed distribution, scaled by the missingness probability, is highlighted in the red area. The superscript $\star$ denotes that the variable is computed for the observed distribution.

As demonstrated in Corollary 3.18, the p-b-p Bayes classifier can be represented using the parameters of the complete model. Thus, our goal is to estimate these underlying parameters. To achieve this, we begin with the mean, which will be estimated coordinate-by-coordinate. Building on the previously discussed intuition, we utilize the estimated mode of the observed distribution, which aligns with the mode of the complete distribution, i.e., the mean. In order to accomplish this, we take advantage of the symmetry of the Gaussian distribution and the fact that the modal region, which corresponds to the region of maximum density, is centered around the mean. As a result, we construct moving intervals, count the number of observations within each interval, and select the middle-point of the interval that maximizes this count.

To be more precise, let us consider an interval middle-length denoted by $\tau > 0$. We define

$$G_{\tau,k,j}(x) := \mathbb{P}(x - \tau \leq X_j \leq x + \tau), \tag{27}$$

where $X_j \sim \mathcal{N}(\mu_{k,j}, \Sigma_{k,j})$. It is important to note that in the complete data case, $\mathrm{argmax}_{x\in\mathbb{R}} G_{\tau,k,j}(x) = \mu_{k,j}$ for any value of τ. This property remains valid for the observed distribution as long as the τ-interval of the mean does not intersect the permissible boundaries. To ensure this property, we introduce the following assumption:

Assumption 17 (τ-permissible boundaries)**.** For all $k \in \{-1, 1\}, j \in [d]$, then $\min(\mu_{k,j} - \beta_{k,j,1}, \beta_{k,j,2} - \mu_{k,j}) > \tau$.

In practice, as demonstrated later, we will select a decreasing τ that approaches 0 as the number of samples increases. Consequently, this assumption is not overly restrictive, as long as the probability of missing data is smaller than 0.5. For a sufficiently large value of n, the assumption is guaranteed to hold.

Our strategy involves estimating the function $G_{\tau,k,j}$ and selecting the maximizer of this function as the estimate of the mean. Therefore, we introduce the following function estimate:

$$\widehat{G}_{\tau,k,j}(x) := \frac{1}{\sum_{i=1} \mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,j}=0}} \sum_{i=1}^{n} \mathbb{1}_{x-\tau\leq X_{i,j}\leq x+\tau}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,j}=0}.$$

For the sake of simplicity, in the sequel we will separate both classes, having $n_{j,k}$ *observed* coordinates of each class. Then, the estimate can be defined as

$$\widehat{G}_{\tau,k,j}(x) := \frac{1}{n_{j,k}} \sum_{i=1}^{n_{j,k}} \mathbb{1}_{x-\tau\leq X_{i,j}\leq x+\tau}. \tag{28}$$

Our estimate will be

$$\widehat{\mu}_{k,j} := \mathrm{argmax}_x \widehat{G}_{\tau,k,j}(x). \tag{29}$$

Related works(Non-parametric density estimation). As τ decreases, $\widehat{G}_{\tau,k,j}$ can be interpreted as a non-parametric density estimation with a rectangular kernel. Numerous studies have explored the establishment of an optimal theoretical bandwidth (for example, see Li and He (2018)). However, in our case, the bandwidth selection is aimed at ensuring that the maximizer of our density estimate converges to the mean, as shown in the following theorem.

As usual, our focus lies in establishing convergence rates for our estimates. This is the objective of the following proposition, which concerns our mean estimate under the *extreme self-masking* setting:

Theorem 3.19 (Consistency of the *extreme self-masking* mean estimate)**.** *Given* $\tau_{n_{k,j}} := \sqrt[4]{\log(n_{k,j})/n_{k,j}}$ *and given a dataset with missing values* $\mathcal{D}_n^\star$ *where the underlying model follows Assumption 9 and the missing pattern follows Assumptions 15, 16 and 17, then*

$$\mathbb{E}\left[|\mu_{k,j} - \widehat{\mu}_{k,j}|\right] \lesssim \sqrt[4]{\frac{\log(n_{k,j})}{n_{k,j}}} \frac{\Sigma_j^3}{A_{k,j}} \exp\left(\frac{9A_{k,j}^2}{8\Sigma_j^2}\right).$$

with $\widehat{\mu}_{k,j}$ *defined in* (29) *and* $A_{k,j} := \max\left(\mu_{k,j} - \beta_{k,j,1}, \beta_{k,j,2} - \mu_{k,j}\right)$.

Proof. The proof can be found in Appendix C.6.3. □

Theorem 3.19 ensures the consistency of our estimate. Note that when there are missing values only on one side, the quantity $A_{k,j}$ is not defined. Hence, to accommodate this theorem, we can utilize Lemma A.1 and ensure that it is bounded with high probability.

In the sequel we propose an algorithm without theoretical background.

Notice that by applying the same concept as explained for the mean, it becomes unfeasible to estimate the covariance matrix using the empirical covariance without introducing a bias. Hence, we propose an alternative estimation method of the covariance matrix that takes advantage of the fact that the mean of a Gaussian distribution is also its median. By sorting the sample data, our estimate effectively divides the complete sample, i.e. the observed and the missing values, in half on both sides . Therefore, we suggest a covariance estimation approach based on quantiles, sound when the covariance matrix is diagonal.

Let's begin by analyzing the theoretical quantiles for a random variable X, which follows a distribution $\mathcal{F}$. We denote $q_\alpha^{\mathcal{F}}$ the quantile of order α if it satisfies $\mathbb{P}(X \leq q_\alpha^{\mathcal{F}}) \leq \alpha$ and $\mathbb{P}(X \geq q_\alpha^{\mathcal{F}}) \geq 1 - \alpha$.

However, since $q_\alpha^{\mathcal{F}}$ is unknown, one can use empirical quantiles instead. To do so, we start by sorting the sample. Let's consider a specific coordinate j and a class k, with $n_{k,j}$ observations. We rearrange the observations as $(X_{k,j,(1)}, ..., X_{k,j,(n_{k,j})})$ in ascending order, such that $X_{k,j,(l)} < X_{k,j,(l+1)}$ for all $l \in [n_{k,j} - 1]$. It's important to recognize that this given order is not the true order due to the missing values. In order to obtain the *real* order, we need to incorporate the missing values that are smaller than $\beta_{k,j,1}$. Let's denote by $s_{k,j}$ the number of missing values of the class k that are on the left-hand side of the distribution:

$$s_{k,j} := \sum_{i=1}^{n_k} \mathbb{1}_{X_{i,j} < \beta_{k,j,1}},$$

with n_k the number of instances of the class k (missing or observed). Then, sorting the j-th coordinate of the complete data of the class k, we have

$$(\underbrace{X^\star_{k,j,(1)}, \dots, X^\star_{k,j,(s_{k,j})}}_{missing}, \underbrace{X^\star_{k,j,(s_{k,j}+1)}, \dots, X^\star_{k,j,(s_{k,j}+n_{k,j})}}_{observed}, \underbrace{X^\star_{k,j,(s_{k,j}+n_{k,j}+1)}, \dots, X^\star_{k,j,(n_k)}}_{missing}),$$

where $X^\star_{k,j,(l)} \leq X^\star_{k,(l+1)}$ for $l \in [n_k - 1]$, or equivalently,

$$(\underbrace{X^\star_{k,j,(1)}, \dots, X^\star_{k,j,(s_{k,j})}}_{missing}, \underbrace{X_{k,j,(1)}, \dots, X_{k,j,(n_{k,j})}}_{observed}, \underbrace{X^\star_{k,j,(s_{k,j}+n_{k,j}+1)}, \dots, X^\star_{k,j,(n_k)}}_{missing}).$$

Unfortunately, since the value is missing, we cannot determine the precise shift. We lack information about whether the value is missing because it is smaller than $\beta_{k,j,1}$ or greater than $\beta_{k,j,2}$. To compensate for this unknown shift, we locate the index l_0 such that $X_{k,j,(l_0)} \leq \widehat{\mu}_{k,j} < X_{k,j,(l_0+1)}$. Note that this index always exists due to the way we have defined the estimate, ensuring that it lies between $X_{k,j,(1)}$ and $X_{k,j,(n_{k,j})}$.

To compute this quantity, we can observe that the sum of l_0 and $s_{k,j}$, divided by the total number of observations $n_{k,j}$,should equalize 0.5, indicating that half of the samples fall on each side of $\widehat{\mu}_{k,j}$. Therefore, we estimate it as $\widehat{s}_{k,j} = 0.5 n_{k,j} - l_0$.

We define our $\alpha-$quantile estimate for $\alpha \in [0.5 - \lfloor l_0/n_k \rfloor, 0.5 + \lfloor (n_{k,j} - l_0)/n_k \rfloor]$ as

$$\widehat{q}_\alpha^{\mathcal{N}(\mu_{k,j},\Sigma_j)} := X_{k,j,(\left\lfloor n_{k,j}\left(\alpha - \frac{\widehat{s}_{k,j}}{n_k}\right)\right\rfloor)}. \tag{30}$$

Note that

$$\Phi\left(\frac{q_\alpha^{\mathcal{N}(\mu_{k,j},\Sigma_j)} - \mu_{k,j}}{\Sigma_j}\right) = \alpha.$$

Using this property, we can calculate Σ_j by substituting the previously discussed estimates. For a given $\alpha \in [0.5 - \lfloor l_0/n_k \rfloor, 0.5 + \lfloor (n_{k,j} - l_0)/n_k \rfloor]$, we denote $\widehat{\Sigma}_j^\alpha$ the estimate through $\widehat{q}_\alpha^{\mathcal{N}(\mu_{k,j},\Sigma_j)}$, i.e. $\widehat{\Sigma}_j^\alpha := \left(\widehat{q}_\alpha^{\mathcal{N}(\mu_{k,j},\Sigma_j)} - \widehat{\mu}_{k,j}\right)/\Phi^{-1}(\alpha)$. Observe that we can only estimate this for the *available* quantiles, which are the ones observed. We have then for $\alpha \in \mathcal{H}_k := \{\widehat{s}_{k,j} + i(n_{k,j}/n_k)\}_{i\in[n]}$. In order to regularize our estimate, we can take the mean over the set of observations from both classes:

$$\widehat{\Sigma}_j = \frac{\sum_{\alpha_{-1}\in\mathcal{H}_{-1}} \widehat{\Sigma}_j^{\alpha_{-1}} + \sum_{\alpha_1\in\mathcal{H}_1} \widehat{\Sigma}_j^{\alpha_1}}{n}. \tag{31}$$

Finally, we can see from Corollary 3.18 that the only remaining quantities to be estimated are π_k and $\mathbb{P}(M_j = 1|Y = k)$ for $k \in -1, 1$. We propose estimating them straightforwardly by computing their respective proportions:

$$\widehat{\pi}_k := \frac{\sum_{i=1}^n \mathbb{1}_{Y_i=k}}{n} \qquad \widehat{\mathbb{P}}(M_j = 1|Y = k) := \frac{\sum_{i=1}^n \mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,j}=1}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k}}. \tag{32}$$

In summary, we present a practical LDA approach that accommodates MNAR assumptions while avoiding the typical exponential number of parameters associated with MNAR caused by the number of missing patterns, as demonstrated in Section 3.2.5, through the adoption of a self-masking assumption. This LDA method capitalizes on the fact that the available data follows a truncated Gaussian distribution, enabling robust parameter estimation for the pattern-by-pattern Bayes classifier.

Overall pattern-by-pattern LDA with extreme self-masked missing values reads as the pattern-by-pattern Bayes classifier given in Corollary 3.18 plugged-in by the following estimates: $\widehat{\mu}_{k,j}$ given at (29) with $\tau_{n_{k,j}} = \sqrt[4]{\log(n_{k,j})/n_{k,j}}$; $\widehat{\Sigma}_j$ given at (31); $\widehat{\pi}_k$ and $\widehat{\mathbb{P}}(M_j = 1|Y = k)$ both given at (32).

3.3 (Semi-parametric modelling) Logistic Regression

First, recall the Bayes classifier for the complete data case given in Equation (1). Then, note that we have that

$$\mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X) - \mathbb{P}(Y = -1|X) = 2\mathbb{P}(Y = 1|X) - 1.$$

Thus, the Bayes classifier can be expressed as a difference of indicator functions that threshold the probability of belonging to class 1 given the observation x:

$$h^\star(x) = \text{sign}(\mathbb{P}(Y = 1|X = x) - \frac{1}{2}) = \mathbb{1}_{\mathbb{P}(Y=1|X=x)\geq 0.5} - \mathbb{1}_{\mathbb{P}(Y=1|X=x)<0.5}.$$

Therefore, as it only depends on $Y|X$, a common approach is to assume a parametric model for the distribution of $Y|X$. An extended model to study is the logistic regression, based on the following assumption:

Assumption 18 (Logistic model)**.**

$$\begin{aligned} &Y_i|X_i = x_i \sim \mathcal{B}(\eta(x_i)) \\ &\text{where } \eta(x_i) := \mathbb{P}\left(Y_i = 1|X_i = x_i\right) = \frac{1}{1 + \exp(-x_i^\top \beta)}. \end{aligned}$$

Hence, similarly to the study of Ayme et al. (2022) commented in Section 2.1, in the following, we study if under Assumption 18 on the complete underlying model and other type of assumptions presented in Section 2, the pattern-by-pattern Bayes predictor conserves this logistic structure. To study this, we define the pattern-by-pattern probability of belonging to the class 1, knowing the missing pattern and the observed covariates, analogous to the complete case., i.e.

$$\eta_m(x) := \mathbb{P}(Y = 1|X_{\text{obs}(M)} = x, M = m). \tag{33}$$

The logistic model is valid pattern by pattern, if, for each $m \in \{0, 1\}^d$, there exists $\beta^\star_m$ such that η_m follows the structure given in Assumption 18, that is η_m follows a logistic model. Unfortunately, it is not true in general, as we see in the following proposition.

Proposition 3.20 (P-b-p predictor not logistic even for MCAR model)**.** *Under MCAR assumption(Assumption 2) and logistic model assumption on the underlying model of $Y|X$ (Assumption 18), we do not necessarily preserve the logistic model restricted to missing patterns working with observed covariates.*

Proof. The proof can be found in Appendix D.1 □

In Ayme et al. (2022), as discussed in Section 2.1, the authors have made some general assumptions on the joint distribution (X, M) that ensure the preservation of the linear assumption made for the linear model on the pattern-by-pattern model. We may be tempted to think that this is also the case for the logistic model. They used assumptions such as MCAR Gaussian model or independent covariates. Unfortunately, as shown in the following case, even under both conditions, this result is not preserved for the logistic model.

Proposition 3.21 (P-b-p predictor not logistic even for MCAR Gaussian model with Independent Covariates)**.** *Under Assumptions MCAR(Assumption 2), logistic model on the underlying model of $Y|X$ (Assumption 18), Gaussian covariates (Assumption 6) and independent covariates(Assumption 5), we do not necessarily preserve the logistic model restricted to missing patterns working with observed covariates.*

Proof. The proof can be found in Appendix D.2. □

Thus, unlike LDA, this semi-parametric approach is too restrictive on the complete underlying model and it is not possible for the pattern-by-pattern classifiers to *inherit* this structure. Then, in the following section, a more general approach is chosen: the *perceptron* algorithm.

3.4 (Non parametric modelling) Perceptron

Another common approach to solve the classification problem is the *perceptron* algorithm. The goal is to find an hyperplane which separates our labeled data. The classification is performed according to the signed Euclidean distance of the input point to the perceptron hyperplane. To find such a separating hyperplane, we iteratively update this hyperplane depending on the misclassification errors. The convergence of the method is ensured under the separability of the observations. Therefore, if we aim at adapting the perceptron to the case of missing values, it is natural to investigate whether this data separability is preserved.

The objective of the study conducted by Bandeira et al. (2014) was to investigate the preservation of linear separability between two convex sets under random projections. They sought to identify the smallest dimensional subspace where this property remained valid with a high probability. This particular problem is referred to as the *rare eclipse problem.* It is worth noting that missing values can be viewed as the result of projecting complete data onto a random subspace of incomplete data. Specifically, if we denote the missing pattern as $M \in \{0,1\}^d$ and the observed data as $X \in \mathbb{R}^d$, then $X_{\text{obs}(M)} \in \mathbb{R}^{d-\|M\|_0}$ represents the randomly projected subspace with a size equal to the number of observed values. Unlike the Gaussian projection covered in Bandeira et al. (2014), the case of missing values involves a binomial projection. Therefore, the main focus of this section is to establish a high probability of maintaining linear separability in the presence of missing values to ensure the applicability of a *pattern-by-pattern perceptron.*

Linear Separability in the complete data case We say that the points $(X_i, Y_i)_{i=1,\dots,n} \in \left(\mathbb{R}^d, \{-1,+1\}\right)$ are linearly separable if there exists an hyperplane parameterized by $w^\star$ and the intercept $b^\star$, such that

$$\forall i \in \{1, ..., n\}, Y_i \left(X_i^\top w^\star + b^\star\right) > 0.$$

Let $\mathcal{W} = \{(w^\star, b^\star) \in \mathbb{R}^d \times \mathbb{R}, \forall i \in \{1, ..., n\}, Y_i \left((w^\star)^\top X_i + b^\star\right) > 0\}$ be the set of separating hyperplanes for the underlying complete data.

Linear Separability in the missing data case We want to see if the linear separability condition can be preserved when missing variables are imputed by 0. The available data is $(X_i \odot (1 - M_i), M_i, Y_i)_{i=1,\dots,n} \in \left(\mathbb{R}^d \times \{0,1\}^d \times \{-1,+1\}\right)^n$ where if the j-th component of M_i is 1 $M_{ij} = 1$, then the j-th component of X_i is missing. We say that the linear separability in the imputed-by-0 data is preserved if $\forall m \in \{0,1\}^d$, $\exists w_{(m)} \in \mathbb{R}^d, \exists b_{(m)} \in \mathbb{R}$ such that

$$\forall i \text{ s.t. } M_i = m, \quad Y_i \left(w_{(m)}^\top (1 - M_i) \odot X_i + b_{(m)}\right) > 0.$$

Define

$$\mathcal{W}_{\text{mis},(m)} = \{(w,b) \in \mathbb{R}^d \times \mathbb{R}, \forall i \quad \text{s.t.} \quad M_i = m, \quad Y_i \left(w^\top (1 - M_i) \odot X_i + b\right) > 0\}$$

the set of separating hyperplanes when missing values occur.

When analyzing the preservation of the linear separability in the incomplete and imputed-by-0 datasets, we actually ask

$$\mathcal{W} \neq \emptyset \quad \overset{?}{\Longrightarrow} \quad \forall m \in \{0,1\}^d, \quad \mathcal{W}_{\text{mis},(m)} \neq \emptyset.$$

Unfortunately, this is not the case as shown in the following example.

Example 3.22 (Not Linear Separability of incomplete data). Suppose that we only have two points $X_1, X_2 \in \mathbb{R}^d$ where $x_2 = (x_{1,1}, \dots, x_{1,(k-1)}, x_{2,k}, x_{1,(k+1)}, \dots, x_{1,d})$ with $x_{1,k} \neq x_{2,k}$. We have $y_2 = -y_1$. We also suppose that $m_{1,k} = m_{2,k} = 1$ and $m_1 = m_2$. Then, $\mathcal{W}$ is not empty, but $\mathcal{W}_{mis}$ is empty as $(1 - m_1) \odot x_1 = (1 - m_2) \odot x_2$, thus for any $w \in \mathbb{R}^d$ if $y_1 w^\top (1 - m_1) \odot x_1 > 0$ then $y_2 w^\top (1 - m_2) \odot x_2 = -y_1 w^\top (1 - m_1) \odot x_1 < 0$, or the symmetric case.

This negative results shows that we cannot expect the separability to directly extend to the missing data case. In the following, we investigate particular models on the data generation that would enable to obtain positive results.

Linear separability in the missing data case of two separable balls Suppose the normed vector space $(\mathbb{R}^d, \|\cdot\|_p)$ with $p > 0$. In this section, we define a model to randomly draw separable data in dimension d. To do so, consider the balls B_1 and B_2 resp. centered at C_1, C_2 and of respective radius R_1, R_2.

Assumption 19. Coordinates of $C_1 - C_2$ are i.i.d..

The assumption is verified if, for example, the coordinates of C_1 and C_2 are i.i.d. and independent. A simple case is when the two centers are drawn from isotropic independent Gaussian distribution $C_1 \sim \mathcal{N}(\mu_1, \lambda_1 I_d), C_2 \sim \mathcal{N}(\mu_2, \lambda_2 I_d)$. Another simple example is when both centers are drawn independently and uniformly as $C_1 \sim \mathcal{U}(a_1, b_1)^{\otimes d}, C_2 \sim \mathcal{U}(a_2, b_2)^{\otimes d}$.

Assumption 20. For all $j \in \{1, \dots, d\}$, $\mathbb{E}\left[(C_1 - C_2)_j^p\right] < \infty$.

This assumption is achieved for all p such that $0 < p < \infty$ by several distributions such as sub-Gaussian or subexponential. It is also obviously satisfied by bounded distributions, such as the uniform distribution. Thus, there are several distributions that satisfy both Assumption 19 and Assumption 20. Once the centers C_1 and C_2 are given, one should draw R_1 and R_2 to ensure the separability of the data. This is the purpose of the following assumption.

Assumption 21 (Uniform Radius)**.** Conditional to the centers C_1 and C_2, the radii R_1 and R_2 are uniformly distributed as

$$(R_1 | (C_1, C_2), R_2 | (C_1, C_2)) \sim \mathcal{U}(0, \frac{1}{2} \|C_1 - C_2\|_p)^{\otimes 2}. \tag{34}$$

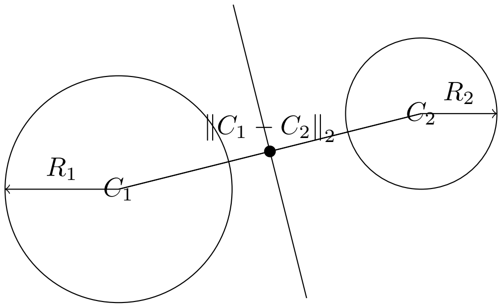


Figure 6: Example of data modelled satisfying Assumptions 19, 20 and 21.

For example, with $p = 2$ and $d = 2$ we graphically represent in Figure 6 the balls centered at C_1, C_2 and with a respective radius R_1, R_2 small enough so that each ball does not touch the separating hyperplane defined in the middle of the two centers.

For the sequel, we consider random missing patterns, corresponding to a fixed number s of missing variables.

Assumption 22 (Uniform s-missing patterns)**.** $M \sim \mathcal{U}(\{m \in \{0,1\}^d, \|m\|_0 = s\})$.

Therefore, the missing pattern M is sampled uniformly among missing patterns admitting s missing values in total. We notice that for any ball $B \in \mathbb{R}^d$ with center C, and any missing pattern M with $s := \|M\|_0$ missing values, the projected set $\Pi_M(B) = \operatorname{diag}(1 - M)B \subset \mathbb{R}^d$, obtained by 0-imputation, can be identified to a ball $B_{\mathrm{obs}(M)}$ of $\mathbb{R}^{d-s}$. Subsequently, we propose a lemma that gives us a new characterization of linear separability in the case of convex compact sets.

Lemma 3.23 (Linear separability for convex compact sets)**.** *Consider two convex disjoint compact sets C_1, C_2. Then, they are linearly separable.*

Proof. The proof can be found in Appendix E.1. □

Then, using this lemma, to verify the linear separability between two balls it is possible to verify that two balls with the same radius are disjoint.

Proposition 3.24 (Asymptotic separability of two balls with the same radius)**.** *Consider the data model provided by Assumptions 19, 20, 21 and $R_2 = R_1$. Under Assumption 22, call $\gamma \geq 0$ such that $\gamma := \lim_{d\to\infty} \frac{s}{d}$. Then,*

$$\lim_{d\to+\infty} \mathbb{P}\left(B_{1,\mathrm{obs}(M)} \cap B_{2,\mathrm{obs}(M)} = \emptyset\right) = \sqrt[p]{1-\gamma}. \tag{35}$$

Therefore, this proposition guarantees that the probability that the projected balls does not intersect converges to $\sqrt[p]{1-\gamma}$, where γ is the asymptotic ratio of missing values. Note that when $s = o_{d\to\infty}(d)$, i.e., s grows with d but not too fast, the separability of the balls is ensured with probability 1.

Another important observation is that this asymptotic separability probability improves when p increases. This may be seen as a trade off between the probability bound and the radius possible amplitude. Effectively, note that for any vector $x \in \mathbb{R}^d$, the function $\|x\|_. : p \mapsto \|x\|_p$ is non-increasing. Then, as the radius is distributed as $R_1|(C_1, C_2) \sim \mathcal{U}(0, \frac{1}{2}\|C_1 - C_2\|_p)$, the maximum radius decreases with the p and the balls are more separated.

To proof this proposition, we start by proving a lemma to characterizes the separability of two balls.

Lemma 3.25 (Separability characterization)**.** *Consider the balls B_1 and B_2 resp. centered at C_1, C_2 and of respective radius R_1, R_2. They are disjoint for the p-norm if and only if $R_1 + R_2 < \|(C_1 - C_2)\|_p$.*

Proof of Lemma 3.25. The proof can be found in Appendix E.2. □

By utilizing this characterization, note that we can redefine the linear separability of two balls as the condition where the distance between their centers is greater than the sum of their individual radii. In the context of our projected balls, we observe that

$$\begin{aligned}\mathbb{P}\left(B_{1,obs(M)} \cap B_{2,obs(M)} = \emptyset\right) &= \mathbb{P}\left(R_1 + R_2 < \left\|C_{1,obs(M)} - C_{2,obs(M)}\right\|_p\right) \\ &= \mathbb{P}\left(R_1 + R_2 < \|\Pi_M(C_1) - \Pi_M(C_2)\|_p\right) \\ &= \mathbb{P}\left(R_1 + R_2 < \|(1-M) \odot (C_1 - C_2)\|_p\right) \end{aligned} \tag{36}$$

Proof of Proposition 3.24. This proof can be found in Appendix E.3 □

In the remainder, we fix $p = 2$ (the Euclidean norm). In addition, the centers C_1 and C_2 are no longer considered random variables, but are given fixed values c_1 and c_2. Therefore, Assumption 21 is adapted to

Assumption 23 (Uniform Radius)**.** Radii R_1 and R_2 are uniformly distributed as $R_1, R_2 \sim \mathcal{U}(0, \frac{1}{2}\|c_1 - c_2\|_p)^{\otimes 2}$.

Assumption 24 (M independent of R_1, R_2)**.** $M \perp\!\!\!\perp R_1, R_2$

This assumption could be seen as an Assumption 2, because the missing pattern does not depend on the radii, nor does it depend on the class. Then, the missing pattern is completely independent of the data. Recall that we denote $\eta_j := \mathbb{P}(M_j = 1)$ the probability of missingness of the coordinate j. Re-using Assumption 10, it guarantees that the probability of being missing is the same for each coordinate, and set to η. We note that the distribution from the Assumption 22 satisfies the two previous assumptions, then the Assumptions 10 and 24 are more general. Another example of distribution for M satisfying the assumptions is $M \sim \mathcal{B}(\epsilon) \otimes \cdots \otimes \mathcal{B}(\epsilon)$ for a given $\epsilon \in [0, 1]$.

Proposition 3.26 (Separability of two balls with different radius)**.** *Given two fixed centers c_1 and c_2, consider the data model provided by Assumptions 23 with R_1 independent of R_2. Under Assumptions 10 and 24 for the missing pattern, then,*

$$\mathbb{P}\left(B_{1,obs(M)} \cap B_{2,obs(M)} = \emptyset\right) \geq 1 - \eta.$$

Proof. This proof can be found in Appendix E.4. □

Then, if the probability of missing values on each coordinate remains low, then there is a high probability of maintaining separability.

Therefore, Proposition 3.26 proposes a finite distance bound in contrast with the dimension asymptotic bound proposed by the Proposition 3.24. In addition, the assumptions are more general as the radii are not the same and the distribution of the missing pattern is even more general. Furthermore, there is no assumption about the centers c_1 and c_2, they are just two given points.

We note that there are some cases where this boundary is accurate. This is the extreme case where the given centers differ from only one coordinate j_0. In this case, the balls $B_{1,\mathrm{obs}(m)}, B_{2,\mathrm{obs}(m)}$ do not collapse if and only if $j_0 \in \mathrm{obs}(m)$, i.e. $m_{j_0} = 0$. Using Assumption 10, we have that

$$\mathbb{P}\left(B_{1,obs(M)} \cap B_{2,obs(M)} = \emptyset\right) = \mathbb{P}(M_{j_0} = 0) = 1 - \mathbb{P}(M_{j_0} = 1) = 1 - \eta$$

However, the bound is not exact as shown in 3.24 because it is asymptotically lower since there is no square root.

Numerical experiments To generate a dataset of observations that satisfy the assumptions outlined in Proposition 3.26, we follow the following procedure:

- We begin by fixing the centers c_1 and c_2. These centers are arbitrarily chosen from a Gaussian distribution.
- Next, we determine the radii of the balls based on the conditions specified in Assumption 23.
- To obtain an observation from class k, then located on ball C_k, we start by centering it to his respective class, c_k.
- We then select the direction L of the observation relative to c_k and normalise it.
- Finally, we uniformly choose the distance of the observation from c_k, ensuring that it falls within the constraints of the ball, i.e., it is smaller than the radius R_k of ball C_k.

An example of a dataset generated according to this distribution can be found in Figure 7. A more concise representation of the dataset is as follows: MCAR Balls: $M \sim \mathcal{B}(p)^{\otimes d}$, $(R_1, R_2) \sim \mathcal{U}(0, \frac{1}{2}\|c_1 - c_2\|)^{\otimes 2}$, $L \sim \mathcal{N}(0, I)$, $r'_k \sim \mathcal{U}(0, R_k)$ and $X|(Y = k) = c_k + r'_k \frac{L}{\|L\|}$.

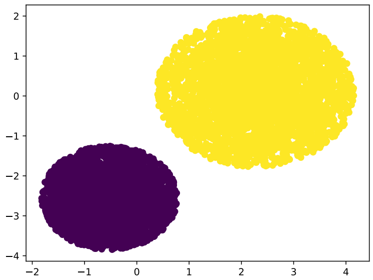
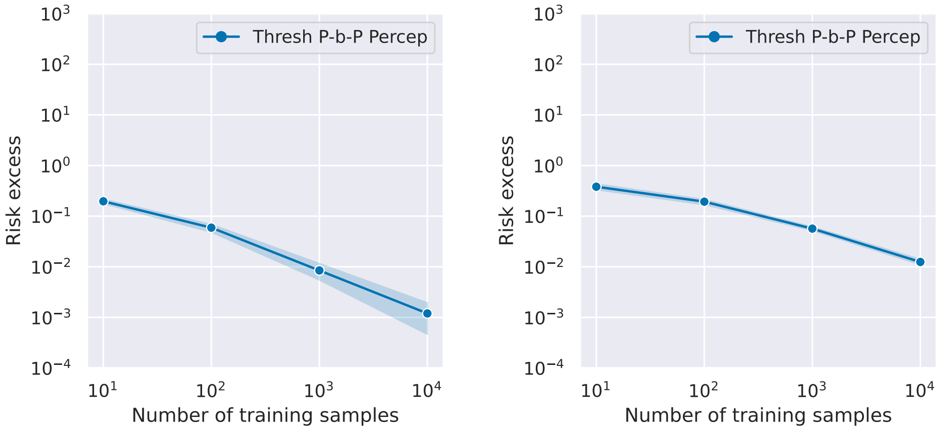


Figure 7: Example of data set modelling for the numerical experiments.

The empirical consistency of this method is demonstrated in Figure 8. In addition, we can observe from Proposition 3.24 that it is desirable for the dimension to increase in order to ensure the high probability separability of the balls (for the case where of missingness probability increases negligibly with the dimension). However, a notable drawback of this approach, as is the case with pattern-by-pattern methods in general, is the curse of dimensionality. Specifically, it is important to note that the number of possible missing patterns is 2^d. Therefore, as the dimensionality becomes larger, there are numerous missing patterns for which there are insufficient samples to accurately estimate a specific predictor. Hence, following a similar approach to Ayme et al. (2022), we adopt a strategy where the predictor is estimated only for the most recent missing patterns. This can be seen as a type of regularization in order to avoid a model overfitting. Consequently, in order to achieve the same level of accuracy, a substantial increase in the number of observations is required. As illustrated in the figure, this results in a slower convergence rate as the dimensionality increases.



(a) MCAR $d = 5$

(b) MCAR $d = 10$

Figure 8: Excess risk w.r.t. the number of training samples. The curve represents the averaged excess risk over 100 repetitions within a 95% confidence interval.

4 Conclusion

To address the paucity of methods for prediction with missing values in the context of classification, we propose a pattern-by-pattern approach. This approach involves employing a specific predictor for each missing pattern to avoid the bias usually induced by imputation methods commonly used in such scenarios.

The primary contribution of this paper lies in the application of linear discriminant analysis (LDA). Initially, we address the missing completely at random (MCAR) assumption. We observe that the pattern-by-pattern Bayes classifier utilizes only the projection of the underlying parameters, enabling us to avoid the exponential increase in parameters typically introduced by the pattern-by-pattern approach. By studying the exponential decay with the dimension of the error introduced by the missing values, we concentrate on estimating the parameters of the underlying model. We control the estimation error in the classical context and then adapt it through simple thresholding to the high-dimensional context.

Next, we consider the missing not at random (MNAR) assumption. The first result is general, with the bound depending on the complexity of the missing pattern distribution. By applying a threshold to limit the number of missing patterns to be estimated, it becomes feasible to utilize this method in cases where the number of likely missing patterns is small. Subsequently, we make an assumption on the missing pattern based on a real-world observation: usually, the missing values are extreme values. Under this assumption, we demonstrate that the number of parameters does not exponentially increase as we primarily need to estimate the underlying model. Then, we propose an estimation of the parameters based on a rectangular kernel estimation of the density. Future work should focus on extending this result with more accurate kernels and establishing bounds on the covariance estimate.

References

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, 3rd edition, July 2003. ISBN 978-0-471-36091-9.

Sylvain Arlot. Fondamentaux de l'apprentissage statistique. In Myriam Maumy-Bertrand, Gilbert Saporta, and Christine Thomas-Agnan, editors, *Apprentissage statistique et données massives.* Editions Technip, May 2018. URL https://hal.science/hal-01485506.

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Near-optimal rate of consistency for linear models with missing values. In *International Conference on Machine Learning*, pages 1211–1243. PMLR, 2022.

Afonso S. Bandeira, Dustin G. Mixon, and Benjamin Recht. Compressive classification and the rare eclipse problem, 2014.

Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis, 2011.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Christophe Giraud. *Introduction to high-dimensional statistics.* CRC Press, 2021.

Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gaël Varoquaux. NeuMiss networks: differentiable programming for supervised learning with missing values. In *NeurIPS 2020 - 34th Conference on Neural Information Processing Systems*, Vancouver / Virtual, Canada, December 2020a. URL https://hal.archives-ouvertes.fr/hal-02888867.

Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gaël Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR, 2020b.

Zhen-Wei Li and Ping He. Data-based optimal bandwidth for kernel density estimation of statistical samples. *Communications in Theoretical Physics*, 70(6):728, dec 2018. doi: 10.1088/0253-6102/70/6/728. URL https://doi.org/10.1088%2F0253-6102%2F70%2F6%2F728.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3), jun 2012. doi: 10.1214/12-aos1018. URL https://doi.org/10.1214%2F12-aos1018.

Norm Matloff and Pete Mohanty. *toweranNA: A Method for Handling Missing Values in Prediction Applications*, 2023. URL https://github.com/matloff/toweranNA. R package version 0.1.0.

Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with nonignorable missing data, 2015.

DONALD B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Aude Sportisse. *Handling heterogeneous and MNAR missing data in statistical learning frameworks : imputation based on low-rank models, online linear regression with SGD, and model-based clustering.* Theses, Sorbonne Université, June 2021. URL https://theses.hal.science/tel-03722429.

T. Tony Cai and Linjun Zhang. High Dimensional Linear Discriminant Analysis: Optimality, Adaptive Algorithm and Missing Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):675–705, 06 2019. ISSN 1369-7412. doi: 10.1111/rssb.12326. URL https://doi.org/10.1111/rssb.12326.

Daniela M. Witten and Robert Tibshirani. Covariance-regularized regression and classification for high-dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009. doi: 10.1111/j.1467-9868.2009.00699.x.

Michael C. Wu, Jing Liu, Yufeng Chen, Jun Wang, and Hongyu Zhao. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009. doi: 10.1093/bioinformatics/btp019.

A Inequalities

Lemma A.1 (Gaussian concentration inequality)**.** *Given $Z \sim \mathcal{N}(\mu, \sigma^2)$, then*

$$\mathbb{P}\left(|Z-\mu| \geq x\right) \leq 2\text{exp}\left(\frac{-x^2}{2\sigma^2}\right).$$

Lemma A.2 (Hoeffding's inequality)**.** *Consider a sequence $(X_k)_{1\leq k\leq n}$ of independent real-valued random variables satisfying, for two sequences $(a_k)_{1\leq k\leq n}$, $(b_k)_{1\leq k\leq n}$ of real numbers such that $a_k < b_k$ for all k,*

$$\forall k, \qquad \mathbb{P}(a_k \leq X_k \leq b_k) = 1.$$

Let

$$S_n = \sum_{i=1}^{n} X_i - \mathbb{E}\left[X_i\right].$$

Then, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(\lambda S_n)\right] \leq \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2\right).$$

Let's begin by a useful lemma on binomial law, which can be found in (Devroye et al., 2013, Lemma A2 p 587).

Lemma A.3. *Let $B \sim \mathcal{B}(p,n)$, we have*

$$\frac{1}{1+np} \leq \mathbb{E}\left[\frac{1}{1+B}\right] \leq \frac{1}{p(n+1)} \tag{37}$$

and

$$\mathbb{E}\left[\frac{\mathbb{1}\{B>0\}}{B}\right] \leq \frac{2}{p(n+1)}. \tag{38}$$

Proof.

- Lower bound of (37): we use Jensen inequality

$$\frac{1}{1+np} = \frac{1}{1+\mathbb{E}B} \leq \mathbb{E}\left[\frac{1}{1+B}\right].$$

- Upper bound of (37),

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1+B}\right] &= \sum_{i=0}^{n}\binom{n}{i}\frac{1}{1+i}p^i(1-p)^{n-i} \\
&= \sum_{i=0}^{n}\frac{n!}{i!(n-i)!(1+i)}p^i(1-p)^{n-i} \\
&= \frac{1}{(n+1)\,p}\sum_{i=0}^{n}\frac{(n+1)!}{(i+1)!(n+1-i-1)!}p^{i+1}(1-p)^{n-i} \\
&= \frac{1}{(n+1)\,p}\sum_{i=0}^{n}\binom{n+1}{i+1}p^{i+1}(1-p)^{n+1-i-1} \\
&\leq \frac{1}{(n+1)\,p},
\end{aligned}$$

using binomial formula.

- For (38), we use that $1/x \leq 2/(x+1)$ on $x \geq 1$ and previous result.

□

Following the same idea, we can establish an upper bound on the square in the following lemma.

Lemma A.4. *Given an $B \sim \mathcal{B}(n,p)$, we have that*

$$\mathbb{E}\left[\frac{1}{(1+B)^2}\right] \leq \frac{2}{(n+1)(n+2)p^2} \tag{39}$$

and

$$\mathbb{E}\left[\frac{\mathbb{1}_{B>0}}{B^2}\right] \leq \frac{8}{(n+1)(n+2)p^2} \tag{40}$$

Proof.

- For (39) do the following development:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{(1+B)^2}\right] &= \sum_{i=0}^{n} \binom{n}{i} \frac{1}{(1+i)^2} p^i (1-p)^{n-i} \\
&= \frac{1}{p(n+1)} \sum_{i=0}^{n} \frac{(n+1)!}{i!(n-i)!} \frac{1}{(1+i)^2} p^{i+1}(1-p)^{n-i} \\
&= \frac{1}{p(n+1)} \sum_{i=0}^{n} \frac{(n+1)!}{(i+1)!(n+1-(i+1))!} \frac{1}{(1+i)} p^{i+1}(1-p)^{n+1-(i+1)} \\
&= \frac{1}{p(n+1)} \sum_{j=1}^{n+1} \frac{(n+1)!}{j!(n+1-j)!} \frac{1}{j} p^j (1-p)^{n+1-j} \\
&= \frac{1}{p(n+1)} \sum_{j=1}^{n+1} \frac{(n+1)!}{j!(n+1-j)!} \frac{1}{j+1} \frac{j+1}{j} p^j (1-p)^{n+1-j} \\
&\leq \frac{2}{p(n+1)} \sum_{j=1}^{n+1} \frac{(n+1)!}{j!(n+1-j)!} \frac{1}{j+1} p^j (1-p)^{n+1-j} \\
&= \frac{2}{p^2(n+1)(n+2)} \sum_{j=1}^{n+1} \frac{(n+2)!}{(j+1)!(n+2-(j+1))!} p^{j+1}(1-p)^{n+2-(j+1)} \\
&= \frac{2}{p^2(n+1)(n+2)} \sum_{k=2}^{n+2} \frac{(n+2)!}{k!(n+2-k)!} p^k (1-p)^{n+2-k} \\
&\leq \frac{2}{p^2(n+1)(n+2)}.
\end{aligned}$$

- For (40), we use that $1/x \leq 2/(x+1)$ on $x \geq 1$ and previous result.

□

Lemma A.5 (Diagonal trace inequality)**.** *Given a symmetric matrix $A \in \mathcal{M}_{n,n}(\mathbb{R})$ and a diagonal matrix $B = (b_i)_{i,i} \in \mathcal{M}_{n,n}(\mathbb{R})$ where all the terms are bounded by a constant $C \in \mathbb{R}$, we have that*

$$\operatorname{tr}(ABA) \leq C\operatorname{tr}(A^2).$$

Proof. Lets rewrite the product of the matrices block-by-block, where $A_i \in \mathcal{M}_{n,1}(\mathbb{R})$ are the columns of A:

$$
\begin{aligned}
&\operatorname{tr}\left(\begin{bmatrix} A_1 & A_2 & A_3 & \cdots & A_n \end{bmatrix} \begin{bmatrix} b_1 & 0 & 0 & \cdots & 0 \\ 0 & b_2 & 0 & \cdots & 0 \\ 0 & 0 & b_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & b_n \end{bmatrix} \begin{bmatrix} A_1^\top \\ A_2^\top \\ A_3^\top \\ \vdots \\ A_n^\top \end{bmatrix}\right) \\
&= \operatorname{tr}\left(\begin{bmatrix} b_1 A_1 & b_2 A_2 & b_3 A_3 & \cdots & b_n A_n \end{bmatrix} \begin{bmatrix} A_1^\top \\ A_2^\top \\ A_3^\top \\ \vdots \\ A_n^\top \end{bmatrix}\right) \\
&= \operatorname{tr}\left(\sum_{i=1}^n b_i A_i A_i^\top\right) \\
&= \sum_{i=1}^n b_i \operatorname{tr}\left(A_i A_i^\top\right) \\
&\leq C \sum_{i=1}^n \operatorname{tr}\left(A_i A_i^\top\right) \\
&= C\operatorname{tr}(A^2)
\end{aligned}
$$

□

The subsequent lemma, which provides a bound on the maximum of sub-Gaussian random variables, has been derived from Section 8.2 of Arlot (2018).

Lemma A.6 (Maximum of sub-Gaussian variables)**.** *Given $Z_1, ..., Z_k$ sub-Gaussian random variables with variance factor v, i.e.*

$$\forall k \in [K], \qquad \mathbb{E}\left[Z_k\right] = 0 \qquad \textit{and} \qquad \forall \lambda \in \mathbb{R}, \qquad \log\left(\mathbb{E}\left[\exp \lambda Z_k\right]\right) \leq \frac{v\lambda^2}{2},$$

then

$$\mathbb{E}\left[\max_{i \in [K]} Z_k\right] \leq \sqrt{2v\log(K)}.$$

B Gaussian Vectors

Lemma B.1. *Given a missing pattern $m \in \{0,1\}^d$ and a Gaussian vector $X \sim \mathcal{N}(\mu, \Sigma)$, then the vector with missing values $X_{\mathrm{obs}(m)}$ is still a Gaussian vector and $X_{\mathrm{obs}(m)} \sim \mathcal{N}(\mu_{\mathrm{obs}(m)}, \Sigma_{\mathrm{obs}(m)\times\mathrm{obs}(m)})$.*

Proof. Since X is a Gaussian vector, every linear combination of its coordinates is a Gaussian variable. In particular, every linear combination of the subset $\mathrm{obs}(m)$ of coordinates is a Gaussian variable, then $X_{\mathrm{obs}(m)}$ is a Gaussian vector.

To prove the second statement, for a given $u \in \mathbb{R}^{d-\|m\|_0}$, we will denote $u' \in \mathbb{R}^d$ the imputed-by-0 vector, i.e. $u'_j = 0$ if $m_j = 1$ and $u'_j = u_i$ with $i = j - \sum_{k=1}^{j} m_k$ otherwise. Then,

$$\begin{aligned}
\forall u \in \mathbb{R}^{d-\|m\|_0}, \qquad \Psi_{X_{\mathrm{obs}(m)}}(u) &= \mathbb{E}\left[\exp(iu^\top X_{\mathrm{obs}(m)})\right] \\
&= \mathbb{E}\left[\exp(i(u')^\top X)\right] \\
&= \exp(i(u')^\top \mu - \frac{1}{2}(u')^\top \Sigma(u')) && (X \sim \mathcal{N}(\mu, \Sigma)) \\
&= \exp(iu^\top \mu_{\mathrm{obs}(m)} - \frac{1}{2}u^\top \Sigma_{\mathrm{obs}(m)\times\mathrm{obs}(m)}u)
\end{aligned}$$

□

C Proofs of Section 3.2 (LDA)

C.1 Proofs of Section 3.2.1

C.1.1 Proof of Proposition 3.1

Proof. Expanding (5),

$$\begin{aligned}
h^\star_m(X_{\mathrm{obs}(m)}) &= \mathrm{sign}(\mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M=m\right]) \\
&= \mathrm{sign}\left(\mathbb{P}\left(Y=1\middle|X_{\mathrm{obs}(m)}, M=m\right) - \mathbb{P}\left(Y=-1\middle|X_{\mathrm{obs}(m)}, M=m\right)\right). \qquad (41)
\end{aligned}$$

Remark that

$$\begin{aligned}
\mathbb{P}\left(Y=k\middle|X_{\mathrm{obs}(m)} \in dx, M=m\right) &= \frac{\mathbb{P}\left(Y=k, X_{\mathrm{obs}(m)} \in dx\middle|M=m\right)}{\mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\middle|M=m\right)} \\
&= \frac{\mathbb{P}\left(Y=k, X_{\mathrm{obs}(m)} \in dx\right)}{\mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\right)} && \text{(using Assumption 2)} \\
&= \frac{\mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\middle|Y=k\right)\pi_k}{\mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\right)},
\end{aligned}$$

so that

$$\begin{aligned}
h^\star_m(X_{\mathrm{obs}(m)} \in dx) &= \mathrm{sign}\left(\mathbb{P}\left(Y=1\middle|X_{\mathrm{obs}(m)} \in dx, M=m\right) - \mathbb{P}\left(Y=-1\middle|X_{\mathrm{obs}(m)} \in dx, M=m\right)\right) \\
&= \mathrm{sign}\left(\frac{\mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\middle|Y=1\right)\pi_1}{\mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\right)} - \frac{\mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\middle|Y=-1\right)\pi_{-1}}{\mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\right)}\right) \\
&= \mathrm{sign}\left(\mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\middle|Y=1\right)\pi_1 - \mathbb{P}\left(X_{\mathrm{obs}(m)} \in dx\middle|Y=-1\right)\pi_{-1}\right).
\end{aligned}$$

Therefore, the objective is to study whether

$$\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=1\right)\pi_1 > \mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=-1\right)\pi_{-1}$$

or equivalently,

$$\frac{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=1\right)}{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=-1\right)} > \frac{\pi_{-1}}{\pi_1}$$

or again

$$\log\left(\frac{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=1\right)}{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=-1\right)}\right) > \log\left(\frac{\pi_{-1}}{\pi_1}\right).$$

Under LDA model (Assumption 9), the objective is to determine the distribution of $X_{\mathrm{obs}(m)}|Y=k$ for each $m \in \{0,1\}^d$. To this end, Lemma B.1 proves that the projection of a Gaussian vector onto a subset of coordinates preserves the Gaussianity with projected parameters. Hence, $X_{\mathrm{obs}(m)}|Y=k \sim \mathcal{N}(\mu_{k,\mathrm{obs}(m)}, \Sigma_{\mathrm{obs}(m)})$ and therefore,

$$\begin{aligned}
&\log\left(\frac{\mathbb{P}\left(X_{\mathrm{obs}(m)}=x\middle|Y=1\right)}{\mathbb{P}\left(X_{\mathrm{obs}(m)}=x\middle|Y=-1\right)}\right)\\
&\quad= \log\left(\frac{(\sqrt{2\pi})^{-(d-\|m\|_0)}\sqrt{\det(\Sigma^{-1}_{\mathrm{obs}(m)})}\exp\left(-\frac{1}{2}(x-\mu_{1,\mathrm{obs}(m)})^\top\Sigma^{-1}_{\mathrm{obs}(m)}(x-\mu_{1,\mathrm{obs}(m)})\right)}{(\sqrt{2\pi})^{-(d-\|m\|_0)}\sqrt{\det(\Sigma^{-1}_{\mathrm{obs}(m)})}\exp\left(-\frac{1}{2}(x-\mu_{-1,\mathrm{obs}(m)})^\top\Sigma^{-1}_{\mathrm{obs}(m)}(x-\mu_{-1,\mathrm{obs}(m)})\right)}\right)\\
&\quad= -\frac{1}{2}(x-\mu_{1,\mathrm{obs}(m)})^\top\Sigma^{-1}_{\mathrm{obs}(m)}(x-\mu_{1,\mathrm{obs}(m)}) + \frac{1}{2}(x-\mu_{-1,\mathrm{obs}(m)})^\top\Sigma^{-1}_{\mathrm{obs}(m)}(x-\mu_{-1,\mathrm{obs}(m)})\\
&\quad= \mu^\top_{1,\mathrm{obs}(m)}\Sigma^{-1}_{\mathrm{obs}(m)}x\\
&\qquad -\frac{1}{2}\mu^\top_{1,\mathrm{obs}(m)}\Sigma^{-1}_{\mathrm{obs}(m)}\mu_{1,\mathrm{obs}(m)} - \mu^\top_{-1,\mathrm{obs}(m)}\Sigma^{-1}_{\mathrm{obs}(m)}x + \frac{1}{2}\mu^\top_{-1,\mathrm{obs}(m)}\Sigma^{-1}_{\mathrm{obs}(m)}\mu_{-1,\mathrm{obs}(m)}\\
&\quad= (\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})^\top\Sigma^{-1}_{\mathrm{obs}(m)}x\\
&\qquad -\frac{1}{2}\left((\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})^\top\Sigma^{-1}_{\mathrm{obs}(m)}\mu_{1,\mathrm{obs}(m)} + (\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})^\top\Sigma^{-1}_{\mathrm{obs}(m)}\mu_{-1,\mathrm{obs}(m)}\right)\\
&\quad= (\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})^\top\Sigma^{-1}_{\mathrm{obs}(m)}\left(x-\frac{\mu_{1,\mathrm{obs}(m)}+\mu_{-1,\mathrm{obs}(m)}}{2}\right).
\end{aligned}$$

The result follows by applying the sign function to the last expression.

□

C.1.2 Proof of Proposition 3.3

Proof. W.l.o.g., we focus only on $\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)})=1\middle|Y=-1\right)$, and cover the other case by symmetry. Using Proposition 3.1, we have that $\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)})=1\middle|Y=-1\right)$ is equal to

$$\mathbb{P}\left(\left(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)}\right)^\top\Sigma^{-1}_{\mathrm{obs}(m)}\left(X_{\mathrm{obs}(m)}-\frac{\mu_{1,\mathrm{obs}(m)}+\mu_{-1,\mathrm{obs}(m)}}{2}\right)-\log\left(\frac{\pi_{-1}}{\pi_1}\right)>0\middle|Y=-1\right)$$

Call $N := \Sigma^{-\frac{1}{2}}_{\mathrm{obs}(m)}(X_{\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})$ and remark that $N|Y=-1 \sim \mathcal{N}(0, Id_{d-\|m\|_0})$ as shown in Lemma B.1. Fix $\gamma := \Sigma^{-\frac{1}{2}}_{\mathrm{obs}(m)}(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})$, so that

$$\begin{aligned}
\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)})=1\middle|Y=-1\right) &= \mathbb{P}\left(\gamma^\top N-\frac{1}{2}\|\gamma\|^2>\log\left(\frac{\pi_{-1}}{\pi_1}\right)\middle|Y=-1\right)\\
&= \mathbb{P}\left(\frac{\gamma^\top N}{\|\gamma\|}>\frac{1}{2}\|\gamma\|+\frac{1}{\|\gamma\|}\log\left(\frac{\pi_{-1}}{\pi_1}\right)\middle|Y=-1\right)\\
&= \Phi\left(-\frac{1}{2}\|\gamma\|-\frac{1}{\|\gamma\|}\log\left(\frac{\pi_{-1}}{\pi_1}\right)\right).
\end{aligned}$$

□

C.1.3 Proof of Corollary 3.4

Proof.

$$\begin{aligned}
L(h^\star) &= \mathbb{P}\left(h^\star(X_{\mathrm{obs}(M)}, M) \neq Y\right) \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star(X_{\mathrm{obs}(m)}, M) \neq Y \middle| M = m\right) p_m \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) \neq Y\right) p_m && \text{(using Assumption 2)} \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1\right) \mathbb{P}(Y=1) p_m \\
&\quad + \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1\right) \mathbb{P}(Y=-1) p_m \\
&= \sum_{m\in\{0,1\}^d} \Phi\left(-\frac{\left\|\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right\|}{2\sigma}\right) \mathbb{P}(Y=1) p_m \\
&\quad + \Phi\left(-\frac{\left\|\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right\|}{2\sigma}\right) \mathbb{P}(Y=-1) p_m && \text{(using Expression (10))} \\
&= \sum_{m\in\{0,1\}^d} \Phi\left(-\frac{\left\|\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right\|}{2\sigma}\right) p_m
\end{aligned}$$

□

C.1.4 Generalized Corollary 3.4

Corollary C.1 (Bayes Risk of p-b-p LDA)**.** *Under Assumptions 2 and 9, the Bayes risk is given by*

$$\begin{aligned}
L(h^\star) = \sum_{m\in\{0,1\}^d} \Phi &\left(-\frac{\log\left(\frac{\pi_{-1}}{\pi_1}\right)}{\left\|\Sigma^{-\frac{1}{2}}_{\mathrm{obs}(m)}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|} - \frac{\left\|\Sigma^{-\frac{1}{2}}_{\mathrm{obs}(m)}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right) \pi_{-1} p_m \\
+ \Phi &\left(-\frac{\log\left(\frac{\pi_1}{\pi_{-1}}\right)}{\left\|\Sigma^{-\frac{1}{2}}_{\mathrm{obs}(m)}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|} - \frac{\left\|\Sigma^{-\frac{1}{2}}_{\mathrm{obs}(m)}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right) \pi_1 p_m
\end{aligned}$$

Proof.

$$\begin{aligned}
L(h^\star) &= \mathbb{P}\left(h^\star(X_{\mathrm{obs}(M)}, M) \neq Y\right) \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star(X_{\mathrm{obs}(m)}, M) \neq Y \middle| M=m\right) p_m \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) \neq Y\right) p_m && \text{(using Assumption 2)} \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y=1\right) \mathbb{P}(Y=1) p_m \\
&\quad + \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y=-1\right) \mathbb{P}(Y=-1) p_m \\
&= \sum_{m\in\{0,1\}^d} \Phi\left(-\frac{\log\left(\frac{\pi_1}{\pi_{-1}}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|} - \frac{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right)\pi_1 p_m \\
&\quad + \Phi\left(-\frac{\log\left(\frac{\pi_{-1}}{\pi_1}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|} - \frac{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right)\pi_{-1} p_m \\
& && \text{(using Expressions (8) and (9))}
\end{aligned}$$

□

C.2 Proofs of Section 3.2.2

C.2.1 Proof of Proposition 3.6

Proof. Using Assumption 11, we have that

$$\left\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1})\right\| \leq \frac{\|\mu_1 - \mu_{-1}\|}{\sqrt{\lambda_{\min}(\Sigma)}} = \mu\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}$$

$$\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\| \geq \frac{\left\|\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right\|}{\sqrt{\lambda_{\max}(\Sigma)}} = \mu\sqrt{\frac{d - \|m\|_0}{\lambda_{\max}(\Sigma)}}$$

Retaking Equation (16), we rewrite the difference as

$$\begin{aligned}
L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) &= \sum_{m\in\{0,1\}^d} \left(\Phi\left(- \frac{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\|}{2} \right) - \Phi\left(- \frac{\left\| \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_{-1}) \right\|}{2} \right) \right) p_m \\
&\leq \sum_{m\in\{0,1\}^d} \left(\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - \|m\|_0}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) p_m \\
&= \sum_{i=0}^{d} \sum_{\{m\in\{0,1\}^d \mid \|m\|_0 = i\}} \left(\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - i}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) p_m \\
&= \sum_{i=0}^{d} \left(\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - i}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) \binom{d}{i} \eta^i (1-\eta)^{d-i} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(using Assumption 10)} \\
&= \mathbb{E}\left[\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - B}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right] \qquad (42)
\end{aligned}$$

where $B \sim \mathcal{B}(d, \eta)$. The decomposition of this last expression gives us

$$\begin{aligned}
&L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) \\
&\leq \mathbb{E}\left[\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - B}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \middle| B = d \right] \mathbb{P}(B = d) \\
&\quad + \mathbb{E}\left[\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - B}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \middle| B \neq d \right] \mathbb{P}(B \neq d) \\
&= \left(\frac{1}{2} - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) \eta^d + \mathbb{E}\left[\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - B}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \middle| B \neq d \right] (1 - \eta^d) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (43)
\end{aligned}$$

Therefore, our first step is to bound the difference $\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right) - \Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)$.

We denote $Q(x) := \int_x^\infty e^{-\frac{t^2}{2}} dt$. Using the symmetry of the standard Gaussian distribution, we notice that

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \int_{-x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} Q(-x) \qquad (44)$$

To simplify the expressions, we denote $t_B := \frac{\mu}{2}\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}$ and $t := \frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}$. Obviously, $0 < t_B \leq t$, Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\middle|B\neq d\right]\\
&=\mathbb{E}\left[\frac{1}{\sqrt{2\pi}}\left(Q\left(\frac{\mu}{2}\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)-Q\left(\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\middle|B\neq d\right]\\
&=\frac{1}{\sqrt{2\pi}}\mathbb{E}\left[Q(t_B)-Q(t)|B\neq d\right]. \qquad (45)
\end{aligned}$$

Using the theorem of intermediate value, we have that there exists a $c\in(t_B,t)$ such that

$$Q(t_B)-Q(t)=-Q'(c)(t-t_B)=e^{-\frac{c^2}{2}}(t-t_B)\leq e^{-\frac{t_B^2}{2}}(t-t_B), \qquad (46)$$

where we have used that the function $f(x)=e^{-\frac{x^2}{2}}$ is decreasing for $x\geq 0$.

Hence, using Equations (45) and (46), we have that

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left(-\frac{\mu\sqrt{d-B}}{2\sqrt{\lambda_{\max}(\Sigma)}}\right)-\Phi\left(-\frac{\mu\sqrt{d}}{2\sqrt{\lambda_{\min}(\Sigma)}}\right)\middle|B\neq d\right]\\
&=\frac{1}{\sqrt{2\pi}}\mathbb{E}\left[Q(t_B)-Q(t)|B\neq d\right]\\
&\leq\frac{1}{\sqrt{2\pi}}\mathbb{E}\left[e^{-\frac{t_B^2}{2}}(t-t_B)\middle|B\neq d\right]\\
&=\frac{\mu}{2\sqrt{2\pi}}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\middle|B\neq d\right]. \qquad (47)
\end{aligned}$$

Observe that

$$\begin{aligned}
&\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right]\\
&=\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\middle|B\neq d\right]\mathbb{P}(B\neq d)+\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\mathbb{P}(B=d),
\end{aligned}$$

so that

$$\begin{aligned}
&\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\middle|B\neq d\right]\\
&=\frac{1}{\mathbb{P}(B\neq d)}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right]-\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\frac{\mathbb{P}(B=d)}{\mathbb{P}(B\neq d)}. \qquad (48)
\end{aligned}$$

Hence, start by developing the expectation

$$
\begin{aligned}
&\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right] \\
&=\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d}{\lambda_{\max}(\Sigma)}}+\sqrt{\frac{d}{\lambda_{\max}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right] \\
&=\sqrt{d}\left(\frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}-\frac{1}{\sqrt{\lambda_{\max}(\Sigma)}}\right)\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\right]+\frac{1}{\sqrt{\lambda_{\max}(\Sigma)}}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{d}-\sqrt{d-B}\right)\right] \\
&=\sqrt{d}\left(\frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}-\frac{1}{\sqrt{\lambda_{\max}(\Sigma)}}\right)\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\right]+\frac{1}{\sqrt{\lambda_{\max}(\Sigma)}}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\frac{d-d+B}{\sqrt{d}+\sqrt{d-B}}\right)\right] \\
&\leq\sqrt{d}\left(\frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}-\frac{1}{\sqrt{\lambda_{\max}(\Sigma)}}\right)\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\right]+\frac{1}{\sqrt{\lambda_{\max}(\Sigma)d}}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}B\right].
\end{aligned}
$$

Observe that, on the one hand,

$$
\begin{aligned}
\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\right] &= \sum_{i=0}^{d}\binom{d}{i}e^{-\frac{\mu^2(d-i)}{8\lambda_{\max}(\Sigma)}}\eta^i\left(1-\eta\right)^{d-i} \\
&= \sum_{i=0}^{d}\binom{d}{i}\eta^i\left(e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}\left(1-\eta\right)\right)^{d-i} \\
&= \left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}\left(1-\eta\right)\right)^{d}. \qquad (49)
\end{aligned}
$$

Observe that, on the other hand,

$$
\begin{aligned}
\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}B\right] &= \sum_{i=0}^{d}\binom{d}{i}e^{-\frac{\mu^2(d-i)}{8\lambda_{\max}(\Sigma)}}i\eta^i(1-\eta)^{d-i} \\
&= \sum_{i=1}^{d}\binom{d}{i}i\eta^i\left(e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}\left(1-\eta\right)\right)^{d-i} \\
&= \eta d\sum_{i=1}^{d}\frac{(d-1)!}{(i-1)!(d-1-(i-1))!}\eta^{i-1}\left(e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}\left(1-\eta\right)\right)^{d-1-(i-1)} \\
&= \eta d\left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}\left(1-\eta\right)\right)^{d-1}. \qquad (50)
\end{aligned}
$$

Therefore, we have that

$$
\begin{aligned}
&\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right]\\
&\le \sqrt{d}\left(\frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}-\frac{1}{\sqrt{\lambda_{\max}(\Sigma)}}\right)\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\right]+\frac{1}{\sqrt{\lambda_{\max}(\Sigma)d}}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}B\right]\\
&= \sqrt{d}\left(\frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}-\frac{1}{\sqrt{\lambda_{\max}(\Sigma)}}\right)\left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^d \qquad \text{(using Equation (49))}\\
&\qquad +\frac{1}{\sqrt{\lambda_{\max}(\Sigma)d}}\eta d\left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^{d-1} \qquad \text{(using Equation (50))}\\
&= \frac{\sqrt{d}}{\sqrt{\lambda_{\min}(\Sigma)}}\left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^d\\
&\qquad -\sqrt{\frac{d}{\lambda_{\max}(\Sigma)}}\left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^{d-1}e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta).
\end{aligned}
\tag{51}
$$

To conclude, we combine all the previous steps as follows:

$$
\begin{aligned}
&L(h^\star)-L_{\mathrm{comp}}(h^\star_{\mathrm{comp}})\\
&=\left(\frac{1}{2}-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d+\mathbb{E}\left[\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)-\Phi\left(\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\middle| B\neq d\right](1-\eta^d)\\
&\qquad\qquad \text{(using Equation (43))}\\
&\le\left(\frac{1}{2}-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d+\frac{\mu}{2\sqrt{2\pi}}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\middle| B\neq d\right](1-\eta^d)\\
&\qquad\qquad \text{(using Inequality (47))}\\
&=\left(\frac{1}{2}-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d\\
&\quad+\frac{\mu}{2\sqrt{2\pi}}\left(\frac{1}{\mathbb{P}(B\neq d)}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right]-\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\frac{\mathbb{P}(B=d)}{\mathbb{P}(B\neq d)}\right)(1-\eta^d)\\
&\qquad\qquad \text{(using Equation (48))}\\
&=\left(\frac{1}{2}-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d\\
&\quad+\frac{\mu}{2\sqrt{2\pi}}\left(\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right]-\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\eta^d\right)\\
&\le\left(\frac{1}{2}-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d+\frac{\mu}{2\sqrt{2\pi}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^d\right.\\
&\quad\left.-\sqrt{\frac{d}{\lambda_{\max}(\Sigma)}}\left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^{d-1}e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)-\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\eta^d\right)\\
&\qquad\qquad \text{(using Inequality (51))}
\end{aligned}
$$

□

C.2.2 Proof of Corollary 3.7

Proof. Begin by establishing a lower bound on the difference that converges to $\eta^d/2$: To do so, start by using the decomposition showed in Equation (16), and that all the terms are positive:

$$
\begin{aligned}
&L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) \\
&= \sum_{m\in\{0,1\}^d} \left(\Phi\left(-\frac{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\|}{2} \right) - \Phi\left(-\frac{\left\| \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_{-1}) \right\|}{2} \right) \right) p_m \\
&\hfill \text{(using Equality (16))} \\
&\geq \left(\Phi\left(0\right) - \Phi\left(-\frac{\left\| \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_{-1}) \right\|}{2} \right) \right) \eta^d. \qquad \text{(using only } m = \mathbf{1}\text{)}
\end{aligned}
$$

Now, using Assumption 11, we have that $\left\| \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_{-1}) \right\| \geq \frac{d\mu}{\sqrt{\lambda_{\max}(\Sigma)}}$, then

$$
\begin{aligned}
\left(\Phi\left(0\right) - \Phi\left(-\frac{\left\| \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_{-1}) \right\|}{2} \right) \right) \eta^d &\geq \left(\Phi\left(0\right) - \Phi\left(-\frac{d\mu}{2\sqrt{\lambda_{\max}(\Sigma)}} \right) \right) \eta^d \\
&= \left(\frac{1}{2} - \Phi\left(-\frac{d\lambda}{2} \right) \right) \eta^d \xrightarrow[\lambda\to\infty]{} \frac{\eta^d}{2}.
\end{aligned}
$$

For the second part of the proof, first notice that when $\lambda := \mu/\sqrt{\lambda_{\max}(\Sigma)}$ goes to the infinity, $\mu/\sqrt{\lambda_{\min}(\Sigma)}$ does the same. Then, note that we have

$$
\left| \Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right| \leq \Phi\left(-\frac{\lambda}{2}\sqrt{d} \right) \xrightarrow[\lambda\to\infty]{} 0,
$$

we also have

$$
\begin{aligned}
\left| \mu\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left(\left(\eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}} (1-\eta) \right)^d - \eta^d \right) \right| &= \mu\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left(\sum_{i=0}^{d} \binom{d}{i} \eta^{d-i} e^{-\frac{i\lambda^2}{8}} (1-\eta)^i - \eta^d \right) \\
&= \frac{\mu}{\sqrt{\lambda_{\max}(\Sigma)}} \sqrt{\frac{d\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}} \left(\sum_{i=1}^{d} \binom{d}{i} \eta^{d-i} e^{-\frac{i\lambda^2}{8}} (1-\eta)^i \right) \\
&= \lambda\sqrt{\frac{d\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}} \left(\sum_{i=1}^{d} \binom{d}{i} \eta^{d-i} e^{-\frac{i\lambda^2}{8}} (1-\eta)^i \right) \xrightarrow[\lambda\to\infty]{} 0, \\
&\qquad \text{(using Assumption 12)}
\end{aligned}
$$

and finally we have

$$
\lambda\sqrt{d} \left(\eta + e^{-\frac{\lambda^2}{8}} (1-\eta) \right)^{d-1} e^{-\frac{\lambda^2}{8}} (1-\eta) \xrightarrow[\lambda\to\infty]{} 0.
$$

Then, combining these three limits we have the convergence. □

C.2.3 Proof of Corollary 3.8

Proof. Begin by taking the upper bound established in Proposition 3.6 and removing the negative terms:

$$
\begin{aligned}
&L(h^{\star}) - L_{\text{comp}}(h^{\star}_{\text{comp}}) \\
&\leq \left(\frac{1}{2} - \Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d + \frac{\mu}{2\sqrt{2\pi}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\left(\left(\eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^d - \eta^d\right)\right. \\
&\qquad \left. -\sqrt{\frac{d}{\lambda_{\max}(\Sigma)}}\left(\eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^{d-1} e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right) \\
&\leq \frac{1}{2}\eta^d + \frac{\mu}{2\sqrt{2\pi}}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\left(\eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^d .
\end{aligned}
$$

Finally, conclude using that $\eta < \epsilon(\eta, \lambda)$ and noticing that as the dimension increases, the quantity $d/\lambda_{\min}(\Sigma)$ diverges, so the first term is negligible compared with the second one. □

C.3 Proofs of Section 3.2.3

C.3.1 Proof of Proposition 3.9

Proof. Let $\mathcal{A}_{k,j} := \{\forall i \in \{1, ..., n\}, (Y_i = -k \vee M_{i,j} = 1) = 1\}$ denote the event that no samples of class k are observed at the j-th coordinate. Let $\mathcal{A}^c_{k,j}$ be its complementary. Note that

$$
\mathbb{P}(\mathcal{A}_{k,j}) = \Pi_{i=1}^n \mathbb{P}(Y_i = -k \vee M_{i,j} = 1) = \left(\frac{\eta+1}{2}\right)^n .
$$

Then,

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\mu}_{k,j}\right] &= \mathbb{E}\left[\widehat{\mu}_{k,j} \middle| \mathcal{A}_{k,j}\right]\mathbb{P}\left(\mathcal{A}_{k,j}\right) + \mathbb{E}\left[\widehat{\mu}_{k,j} \middle| \mathcal{A}^c_{k,j}\right]\mathbb{P}\left(\mathcal{A}^c_{k,j}\right) \\
&= \mathbb{E}\left[\widehat{\mu}_{k,j} \middle| \mathcal{A}^c_{k,j}\right]\mathbb{P}\left(\mathcal{A}^c_{k,j}\right) && \text{(using that } \widehat{\mu}_{k,j} := 0 \text{ if } \mathcal{A}_{k,j}) \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n X_{i,j}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,j}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,j}=0}} \middle| \mathcal{A}^c_{k,j}\right]\mathbb{P}\left(\mathcal{A}^c_{k,j}\right) \\
&= \sum_{i=1}^n \mathbb{E}\left[\frac{X_{i,j}}{1 + \sum_{l\neq i}^n \mathbb{1}_{Y_l=k}\mathbb{1}_{M_{l,j}=0}} \middle| Y_i = k, M_{i,j} = 0\right]\mathbb{P}\left(Y_i = k, M_{i,j} = 0 \middle| \mathcal{A}^c_{k,j}\right)\mathbb{P}\left(\mathcal{A}^c_{k,j}\right) \\
&= \sum_{i=1}^n \mathbb{E}\left[\frac{X_{i,j}}{1 + \sum_{l\neq i}^n \mathbb{1}_{Y_l=k}\mathbb{1}_{M_{l,j}=0}} \middle| Y_i = k\right]\mathbb{P}\left(Y_i = k\right)\mathbb{P}\left(M_{i,j} = 0\right) && \text{(using Assumption 2)} \\
&= n\mathbb{E}\left[\frac{X_{1,j}}{1 + \sum_{l=2}^n \mathbb{1}_{Y_l=k}\mathbb{1}_{M_{l,j}=0}} \middle| Y_1 = k\right]\mathbb{P}\left(Y_1 = k\right)\mathbb{P}\left(M_{1,j} = 0\right) && \text{(using the exchangeability)} \\
&= n\mathbb{E}\left[X_{1,j} | Y_1 = k\right]\mathbb{E}\left[\frac{1}{1 + \sum_{l=2}^n \mathbb{1}_{Y_l=k}\mathbb{1}_{M_{l,j}=0}}\right]\mathbb{P}\left(Y_1 = k\right)\mathbb{P}\left(M_{1,j} = 0\right) \\
& && \text{(using the independence)} \\
&= n\mu_{k,j}\mathbb{E}\left[\frac{1}{1+B}\right]\frac{1-\eta}{2} && \left(\text{with } B \sim \mathcal{B}\left(n-1, \tfrac{1-\eta}{2}\right)\right)
\end{aligned}
$$

Now, using the left hand-side of inequality (37), we have that

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mu}_{k,j}\right] &= n\mu_{k,j}\mathbb{E}\left[\frac{1}{1+B}\right]\frac{1-\eta}{2} \\
&\geq \mu_{k,j}\frac{n}{1+(n-1)\frac{1-\eta}{2}}\frac{1-\eta}{2} \\
&= \mu_{k,j}\left(1+\frac{-\frac{1+\eta}{2}}{1+(n-1)\frac{1-\eta}{2}}\right) \xrightarrow[n\to\infty]{} \mu_{k,j}.
\end{aligned}$$

Using the other inequality (37), we have

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mu}_{k,j}\right] &= n\mu_{k,j}\mathbb{E}\left[\frac{1}{1+B}\right]\frac{1-\eta}{2} \\
&\leq n\mu_{k,j}\frac{1}{n\frac{1-\eta}{2}}\frac{1-\eta}{2} = \mu_{k,j}
\end{aligned}$$

□

C.3.2 General lemmas for LDA misclassification control.

Lemma C.2 ($\widehat{\mu}$ misclassification probability)**.** *Given a sample satisfying Assumptions 2 and 9, with balanced classes, then*

$$\begin{aligned}
&\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)})=1\middle|Y=-1,\mathcal{D}_n\right) \\
&=\Phi\left(\frac{\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\text{obs}(m)}-\widehat{\mu}_{-1,\text{obs}(m)})\right)^\top\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\left(\mu_{-1,\text{obs}(m)}-\frac{\widehat{\mu}_{1,\text{obs}(m)}+\widehat{\mu}_{-1,\text{obs}(m)}}{2}\right)}{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\text{obs}(m)}-\widehat{\mu}_{-1,\text{obs}(m)})\right\|}\right) \qquad (52)
\end{aligned}$$

and symmetrically,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)})=-1\middle|Y=1,\mathcal{D}_n\right) \\
&=\Phi\left(-\frac{\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\text{obs}(m)}-\widehat{\mu}_{-1,\text{obs}(m)})\right)^\top\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\left(\mu_{1,\text{obs}(m)}-\frac{\widehat{\mu}_{1,\text{obs}(m)}+\widehat{\mu}_{-1,\text{obs}(m)}}{2}\right)}{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\text{obs}(m)}-\widehat{\mu}_{-1,\text{obs}(m)})\right\|}\right) \qquad (53)
\end{aligned}$$

with Φ the standard Gaussian cumulative function.

Proof. We start with $\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)})=1\middle|Y=-1,\mathcal{D}_n\right)$. Using the Equation (20), we have that $\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)})=1\middle|Y=-1,\mathcal{D}_n\right)$ is equal to

$$\mathbb{P}\left(\left(\widehat{\mu}_{1,\text{obs}(m)}-\widehat{\mu}_{-1,\text{obs}(m)}\right)^\top\Sigma_{\text{obs}(m)}^{-1}\left(X_{\text{obs}(m)}-\frac{\widehat{\mu}_{1,\text{obs}(m)}+\widehat{\mu}_{-1,\text{obs}(m)}}{2}\right)>0\middle|Y=-1,\mathcal{D}_n\right)$$

Call $N:=\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(X_{\text{obs}(m)}-\mu_{-1,\text{obs}(m)})$ and remark that $N|Y=-1\sim\mathcal{N}(0,Id_{d-\|m\|_0})$ as shown in Lemma B.1. Note that $X_{\text{obs}(m)},Y$ are independent from $\mathcal{D}_n$, then $N|Y=-1\sim N|Y=-1,\mathcal{D}_n$. We denote $\widehat{\gamma} := \Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})$. Thus,

$$\begin{aligned}\mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = 1 \middle| Y = -1, \mathcal{D}_n\right) &= \mathbb{P}\left(\widehat{\gamma}^\top N + \widehat{\gamma}^\top \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left(\mu_{-1,\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2}\right) > 0 \middle| Y = -1, \mathcal{D}_n\right)\\ &= \mathbb{P}\left(\frac{\widehat{\gamma}^\top N}{\|\widehat{\gamma}\|} > -\frac{\widehat{\gamma}^\top}{\|\widehat{\gamma}\|} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left(\mu_{-1,\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2}\right) \middle| Y = -1, \mathcal{D}_n\right)\\ &= \Phi\left(\frac{\widehat{\gamma}^\top}{\|\widehat{\gamma}\|} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left(\mu_{-1,\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2}\right)\right)\end{aligned}$$

Now we prove the second part of the lemma. Using the Equation (20), we have that $\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) = -1 \middle| Y = 1, \mathcal{D}_n\right)$ is equal to

$$\mathbb{P}\left(\left(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\right)^\top \Sigma_{\text{obs}(m)}^{-1} \left(X_{\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2}\right) < 0 \middle| Y = 1\right)$$

Rename $N := \Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(X_{\text{obs}(m)} - \mu_{1,\text{obs}(m)})$ and remark that $N|Y = 1 \sim \mathcal{N}(0, Id_{d-\|m\|_0})$ as shown in Lemma B.1. We redenote $\widehat{\gamma} := \Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})$. Thus,

$$\begin{aligned}\mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = 1 \middle| Y = -1, \mathcal{D}_n\right) &= \mathbb{P}\left(\widehat{\gamma}^\top N + \widehat{\gamma}^\top \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left(\mu_{1,\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2}\right) < 0 \middle| Y = 1, \mathcal{D}_n\right)\\ &= \mathbb{P}\left(\frac{\widehat{\gamma}^\top N}{\|\widehat{\gamma}\|} < -\frac{\widehat{\gamma}^\top}{\|\widehat{\gamma}\|} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left(\mu_{1,\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2}\right) \middle| Y = 1, \mathcal{D}_n\right)\\ &= \Phi\left(-\frac{\widehat{\gamma}^\top}{\|\widehat{\gamma}\|} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left(\mu_{1,\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2}\right)\right)\end{aligned}$$

□

Lemma C.3. *Given a dataset $\mathcal{D}_n$ satisfying Assumptions 2 and 9, with which we can build estimates of the mean for each class, denoted as $\widehat{\mu}_1$ and $\widehat{\mu}_{-1}$, we can state that for the classifier $\widehat{h}_m$ defined in Equation (20),*

$$\begin{aligned}&\left|\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) = 1 \middle| Y = -1, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = 1 \middle| Y = -1\right)\right|\\ &\quad\leq \frac{3}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\| + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)})\right\| \qquad (54)\end{aligned}$$

and symmetrically,

$$\begin{aligned}&\left|\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) = -1 \middle| Y = 1, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = -1 \middle| Y = 1\right)\right|\\ &\quad\leq \frac{3}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(-\widehat{\mu}_{1,\text{obs}(m)} + \mu_{1,\text{obs}(m)})\right\| + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\| \qquad (55)\end{aligned}$$

Proof. We begin with Inequality (54). We have that

$$
\begin{aligned}
&\left|\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)})=1\middle|Y=-1,\mathcal{D}_n\right)-\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)})=1\middle|Y=-1\right)\right|\\
&\quad=\left|\Phi\left(\frac{\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)}-\widehat{\mu}_{-1,\mathrm{obs}(m)})\right)^\top\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mu_{-1,\mathrm{obs}(m)}-\frac{\widehat{\mu}_{1,\mathrm{obs}(m)}+\widehat{\mu}_{-1,\mathrm{obs}(m)}}{2}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)}-\widehat{\mu}_{-1,\mathrm{obs}(m)})\right\|}\right)\right.\\
&\qquad\qquad\left.-\Phi\left(-\frac{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right)\right|,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(using Proposition 3.3 and Lemma C.2)}
\end{aligned}
$$

using that Φ is Lipschitz,

$$
\begin{aligned}
&\leq\frac{1}{\sqrt{2\pi}}\left|\frac{\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)}-\widehat{\mu}_{-1,\mathrm{obs}(m)})\right)^\top\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mu_{-1,\mathrm{obs}(m)}-\frac{\widehat{\mu}_{1,\mathrm{obs}(m)}+\widehat{\mu}_{-1,\mathrm{obs}(m)}}{2}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)}-\widehat{\mu}_{-1,\mathrm{obs}(m)})\right\|}\right.\\
&\qquad\left.+\frac{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right|
\end{aligned}
$$

or equivalently,

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \left| \frac{\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right)^{\top} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2}\right)}{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|} \right. \\
&\qquad \left. + \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}\big)\right\|}{2} \right| \\
&= \frac{1}{\sqrt{2\pi}} \left| \frac{\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right)^{\top} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \big(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)}{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|} \right. \\
&\qquad \left. + \frac{\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right)^{\top} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \big(\widehat{\mu}_{-1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)}\big)}{2\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|} + \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}\big)\right\|}{2} \right| \\
&= \frac{1}{\sqrt{2\pi}} \left| \frac{\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right)^{\top} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \big(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)}{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|} \right. \\
&\qquad \left. - \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|}{2} + \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m))}\right\|}{2} \right| \\
&\leq \frac{1}{\sqrt{2\pi}} \left| \frac{\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right)^{\top} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \big(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)}{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|} \right| \\
&\qquad + \frac{1}{\sqrt{2\pi}} \left| - \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|}{2} + \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m))}\right\|}{2} \right| \\
&\leq \frac{1}{\sqrt{2\pi}} \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\| \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|}{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|} \\
&\hspace{20em} \text{(Using Cauchy-Schwarz Inequality)} \\
&\qquad + \frac{1}{\sqrt{2\pi}} \left| - \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|}{2} + \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}\big)\right\|}{2} \right| \\
&\leq \frac{1}{\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\| + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(-\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)} + \mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}\big)\right\| \\
&\hspace{20em} \text{(Using the Triangle Inequality)} \\
&\leq \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\big)\right\|}{\sqrt{2\pi}} + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)}\big)\right\| \\
&\qquad + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\big(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}\big)\right\| \qquad \text{(Using the Triangle Inequality)}
\end{aligned}
$$

and finally we regroup the terms to get the result:

$$= \frac{3\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{-1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right\|}{2\sqrt{2\pi}} + \frac{1}{2\sqrt{2\pi}}\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \widehat{\mu}_{1,\mathrm{obs}(m)})\right\|.$$

Similarly, for to prove the Inequality (55), we do the following steps:

$$\begin{aligned}
&\left|\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = -1\middle|Y = 1, \mathcal{D}_n\right) - \mathbb{P}\left(h_m^\star(X_{\mathrm{obs}(m)}) = -1\middle|Y = 1\right)\right| \\
&= \left|\Phi\left(-\frac{\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right)^\top \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mu_{1,\mathrm{obs}(m)} - \frac{\widehat{\mu}_{1,\mathrm{obs}(m)}+\widehat{\mu}_{-1,\mathrm{obs}(m)}}{2}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right\|}\right)\right. \\
&\qquad \left. -\Phi\left(-\frac{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right)\right| \qquad \text{(Using Equalities (10) and (53))} \\
&\leq \frac{1}{\sqrt{2\pi}}\left|-\frac{\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right)^\top \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mu_{1,\mathrm{obs}(m)} - \frac{\widehat{\mu}_{1,\mathrm{obs}(m)}+\widehat{\mu}_{-1,\mathrm{obs}(m)}}{2}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right\|}\right. \\
&\qquad \left. +\frac{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|}{2}\right| \qquad \text{(Using that } \Phi \text{ is Lipschitz)}
\end{aligned}$$

or equivalently,

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \left| -\frac{\left(\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right)^{\top} \Sigma^{-\frac{1}{2}}_{\text{obs}(m)} \left(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{1,\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)}+\widehat{\mu}_{-1,\text{obs}(m)}}{2}\right)}{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\|} \right. \\
&\qquad \left. + \frac{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|}{2} \right| \\
&\leq \frac{1}{\sqrt{2\pi}} \left| -\frac{\left(\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right)^{\top} \Sigma^{-\frac{1}{2}}_{\text{obs}(m)} \left(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)}\right)}{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\|} \right| \\
&\qquad + \frac{1}{\sqrt{2\pi}} \left| -\frac{\left(\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right)^{\top} \Sigma^{-\frac{1}{2}}_{\text{obs}(m)} \left(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}\right)}{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\|} \right. \\
&\qquad \left. + \frac{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|}{2} \right| \qquad \text{(Using the Triangle Inequality)} \\
&\leq \frac{1}{\sqrt{2\pi}} \frac{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\| \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)})\right\|}{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\|} \\
&\qquad \text{(Using Cauchy-Schwarz Inequality)} \\
&\qquad + \frac{1}{\sqrt{2\pi}} \left| -\frac{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\|}{2} + \frac{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|}{2} \right| \\
&\leq \frac{1}{\sqrt{2\pi}} \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)})\right\| + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(-\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)} + \mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\| \\
&\qquad \text{(Using the Triangle Inequality)} \\
&\leq \frac{1}{\sqrt{2\pi}} \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)})\right\| + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(-\widehat{\mu}_{1,\text{obs}(m)} + \mu_{1,\text{obs}(m)})\right\| \\
&\qquad + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\| \qquad \text{(Using the Triangle Inequality)} \\
&= \frac{3}{2\sqrt{2\pi}} \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(-\widehat{\mu}_{1,\text{obs}(m)} + \mu_{1,\text{obs}(m)})\right\| + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|.
\end{aligned}
$$

□

Lemma C.4. *Given a dataset $\mathcal{D}_n$ satisfying Assumptions 2 and 9,and a estimate of the mean for each class, then the classifier $\widehat{h}_m$ defined in Equation (19) verifies that*

$$
\begin{aligned}
&L(\widehat{h}) - L(h^{\star}) \\
&\qquad \leq \sum_{m \in \mathcal{M}} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(-\widehat{\mu}_{1,\text{obs}(m)} + \mu_{1,\text{obs}(m)})\right\| + \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\| \right] \right) p_m.
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&L(\widehat{h}) - L(h^\star)\\
&\quad = \mathbb{P}\left(\widehat{h}(X_{\text{obs}(M)}, M) \neq Y\right) - \mathbb{P}\left(h^\star(X_{\text{obs}(M)}, M) \neq Y\right)\\
&\quad = \sum_{m\in\mathcal{M}} \left(\mathbb{P}\left(\widehat{h}(X_{\text{obs}(M)}, M) \neq Y\middle|M=m\right) - \mathbb{P}\left(h^\star(X_{\text{obs}(M)}, M) \neq Y\middle|M=m\right)\right)p_m\\
&\quad = \sum_{m\in\mathcal{M}} \left(\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) \neq Y\middle|M=m\right) - \mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) \neq Y\middle|M=m\right)\right)p_m \qquad \text{(using (19))}\\
&\quad = \sum_{m\in\mathcal{M}} \left(\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) \neq Y\right) - \mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) \neq Y\right)\right)p_m \qquad \text{(using Assumption 2)}\\
&\quad = \sum_{m\in\mathcal{M}} \pi_{-1}\left(\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) = 1\middle|Y=-1\right) - \mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = 1\middle|Y=-1\right)\right)p_m\\
&\qquad\quad + \sum_{m\in\mathcal{M}} \pi_{1}\left(\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) = -1\middle|Y=1\right) - \mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = -1\middle|Y=1\right)\right)p_m.\\
&\quad = \sum_{m\in\mathcal{M}} \frac{1}{2}\left(\mathbb{E}\left[\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) = 1\middle|Y=-1,\mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = 1\middle|Y=-1\right)\right]\right)p_m\\
&\qquad\quad + \sum_{m\in\mathcal{M}} \frac{1}{2}\left(\mathbb{E}\left[\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) = -1\middle|Y=1,\mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = -1\middle|Y=1\right)\right]\right)p_m\\
&\quad \leq \sum_{m\in\mathcal{M}} \frac{1}{4\sqrt{2\pi}}\left(\mathbb{E}\left[3\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\| + \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)})\right\|\right]\right)p_m\\
&\qquad\quad + \sum_{m\in\mathcal{M}} \frac{1}{4\sqrt{2\pi}}\left(\mathbb{E}\left[3\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(-\widehat{\mu}_{1,\text{obs}(m)} + \mu_{1,\text{obs}(m)})\right\| + \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|\right]\right)p_m\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(using Lemma C.3)}\\
&\quad = \sum_{m\in\mathcal{M}} \frac{1}{\sqrt{2\pi}}\left(\mathbb{E}\left[\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(-\widehat{\mu}_{1,\text{obs}(m)} + \mu_{1,\text{obs}(m)})\right\| + \left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|\right]\right)p_m.
\end{aligned}
$$

□

It is worth noting that, at this juncture, neither the structure of the estimate nor the structure of the covariance matrix have been incorporated. We demonstrate the application of this inequality in the case of a covariance matrix $\Sigma = \sigma^2 I_d$, in the case of a general covariance matrix and its utility for other types of estimates.

C.3.3 Lemma for Theorem 3.10

Lemma C.5. *Given an $m \in \mathcal{M}$ and a $k \in \{-1, 1\}$, we have that*

$$\frac{1}{\sigma}\mathbb{E}\left[\left\|\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}\right\|\right] \leq \left(\frac{\left\|\mu_{k,\text{obs}(m)}\right\|^2}{\sigma^2}\left(\frac{1+\eta}{2}\right)^n + \frac{4(d - \|m\|_0)}{(1-\eta)(n+1)}\right)^{\frac{1}{2}}$$

with $\widehat{\mu}_{k,\text{obs}(m)}$ defined in (18).

Proof. Let $\mathcal{A}_{k,j} := \{\forall i \in \{1,...,n\}, (Y_i = -k \vee M_{i,j} = 1) = 1\}$ denote the event that no samples of class k are observed at the j-th coordinate. Let $\mathcal{A}_{k,j}^c$ be its complementary. Note that

$$\mathbb{P}(\mathcal{A}_{k,j}) = \Pi_{i=1}^n \mathbb{P}(Y_i = -k \vee M_{i,j} = 1) = \left(\frac{\eta+1}{2}\right)^n .$$

Then, observe that

$$\begin{aligned}
&\frac{1}{\sigma}\mathbb{E}\left[\left\|\widehat{\mu}_{k,\mathrm{obs}(m)} - \mu_{k,\mathrm{obs}(m)}\right\|\right] \\
&\leq \mathbb{E}\left[\frac{\left\|\widehat{\mu}_{k,\mathrm{obs}(m)} - \mu_{k,\mathrm{obs}(m)}\right\|^2}{\sigma^2}\right]^{\frac{1}{2}} && \text{(Using Jensen inequality)} \\
&= \mathbb{E}\left[\sum_{j\in\mathrm{obs}(m)} \frac{\left(\widehat{\mu}_{k,j} - \mu_{k,j}\right)^2}{\sigma^2}\right]^{\frac{1}{2}} \\
&= \left(\sum_{j\in\mathrm{obs}(m)} \mathbb{E}\left[\frac{\left(\widehat{\mu}_{k,j} - \mu_{k,j}\right)^2}{\sigma^2}\middle|\mathcal{A}_{k,j}\right]\mathbb{P}(\mathcal{A}_{k,j}) + \sum_{j\in\mathrm{obs}(m)} \mathbb{E}\left[\frac{\left(\widehat{\mu}_{k,j} - \mu_{k,j}\right)^2}{\sigma^2}\middle|\mathcal{A}_{k,j}^c\right]\mathbb{P}(\mathcal{A}_{k,j}^c)\right)^{\frac{1}{2}} \\
&= \left(\sum_{j\in\mathrm{obs}(m)} \frac{\left(0 - \mu_{k,j}\right)^2}{\sigma^2}\left(\frac{\eta+1}{2}\right)^n + \sum_{j\in\mathrm{obs}(m)} \mathbb{E}\left[\frac{\left(\widehat{\mu}_{k,j} - \mu_{k,j}\right)^2}{\sigma^2}\middle|\mathcal{A}_{k,j}^c\right]\left(1 - \left(\frac{\eta+1}{2}\right)^n\right)\right)^{\frac{1}{2}} \\
&= \left(\frac{\left\|\mu_{k,\mathrm{obs}(m)}\right\|^2}{\sigma^2}\left(\frac{\eta+1}{2}\right)^n + \sum_{j\in\mathrm{obs}(m)} \mathbb{E}\left[\frac{\left(\frac{\sum_{i=1}^n X_{i,j}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}{\sum_{i=1}^n \mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}} - \mu_{k,j}\right)^2}{\sigma^2}\middle|\mathcal{A}_{k,j}^c\right]\left(1 - \left(\frac{\eta+1}{2}\right)^n\right)\right)^{\frac{1}{2}} \\
& && \text{(Using Estimate (18))} \\
&= \left(\frac{\left\|\mu_{k,\mathrm{obs}(m)}\right\|^2}{\sigma^2}\left(\frac{\eta+1}{2}\right)^n + \sum_{j\in\mathrm{obs}(m)} \mathbb{E}\left[\left(\frac{\sum_{i=1}^n \frac{X_{i,j}-\mu_{k,j}}{\sigma}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}{\sum_{i=1}^n \mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}\right)^2\middle|\mathcal{A}_{k,j}^c\right]\left(1 - \left(\frac{\eta+1}{2}\right)^n\right)\right)^{\frac{1}{2}}
\end{aligned} \tag{56}$$

Note that for each $i_1, i_2 \in \{1, ...n\}$ with $i_1 \neq i_2$,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\frac{X_{i_1,j}-\mu_{k,j}}{\sigma}\mathbb{1}_{M_{i_1,j}=0}\mathbb{1}_{Y_{i_1}=k}}{\sum_{i=1}^{n}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}\frac{\frac{X_{i_2,j}-\mu_{k,j}}{\sigma}\mathbb{1}_{M_{i_2,j}=0}\mathbb{1}_{Y_{i_2}=k}}{\sum_{i=1}^{n}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}\middle|\mathcal{A}_{k,j}^c\right] \\
&\quad= \mathbb{E}\left[\frac{\frac{X_{i_1,j}-\mu_{k,j}}{\sigma}}{1+\sum_{i\neq i_1}^{n}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}\frac{\frac{X_{i_2,j}-\mu_{k,j}}{\sigma}\mathbb{1}_{Y_{i_2}=k}\mathbb{1}_{M_{i_2,j}=0}}{1+\sum_{i\neq i_1}^{n}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}\middle|Y_{i_1}=k, M_{i_1,j}=0, \mathcal{A}_{k,j}^c\right]\mathbb{P}(Y_{i_1}=k, M_{i_1,j}=0|\mathcal{A}_{k,j}^c) \\
&\quad= \mathbb{E}\left[\frac{X_{i_1,j}-\mu_{k,j}}{\sigma}\middle|Y_{i_1}=k, M_{i_1,j}=0\right] \\
&\qquad\qquad \mathbb{E}\left[\frac{\frac{X_{i_2,j}-\mu_{k,j}}{\sigma}\mathbb{1}_{Y_{i_2}=k}\mathbb{1}_{M_{i_2,j}=0}}{\left(1+\sum_{i\neq i_1}^{n}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}\right)^2}\middle|Y_{i_1}=k, M_{i_1,j}=0\right]\mathbb{P}(Y_{i_1}=k, M_{i_1,j}=0|\mathcal{A}_{k,j}^c) \\
&\hspace{30em}\text{(using the independence)} \\
&\quad= \mathbb{E}\left[\frac{X_{i_1,j}-\mu_{k,j}}{\sigma}\middle|Y_{i_1}=k\right]\mathbb{E}\left[\frac{\frac{X_{i_2,j}-\mu_{k,j}}{\sigma}\mathbb{1}_{Y_{i_2}=k}\mathbb{1}_{M_{i_2,j}=0}}{\left(1+\sum_{i\neq i_1}^{n}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}\right)^2}\right]\mathbb{P}(Y_{i_1}=k, M_{i_1,j}=0|\mathcal{A}_{k,j}^c) \\
&\hspace{22em}\text{(using Assumption 2 and independence)} \\
&\quad= 0. \hspace{18em}\text{(Using that } \tfrac{X_{i_1,j}-\mu_{k,j}}{\sigma}|Y_{i_1}=k \sim \mathcal{N}(0,1)\text{)}
\end{aligned}
$$

Then, we have that the products of the mixed terms in the sum have expectation zero. All that is left is the expectation of the squares of the terms. Thus, we have that

$$
\mathbb{E}\left[\left(\frac{\sum_{i=1}^{n}\frac{X_{i,j}-\mu_{k,j}}{\sigma}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}{\sum_{i=1}^{n}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}\right)^2\middle|\mathcal{A}_{k,j}^c\right] = \sum_{i=1}^{n}\mathbb{E}\left[\frac{\left(\frac{X_{i,j}-\mu_{k,j}}{\sigma}\right)^2\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}{\left(\sum_{i=1}^{n}\mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}\right)^2}\middle|\mathcal{A}_{k,j}^c\right].
$$

For a given $i \in \{1, ..., n\}$, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\left.\frac{\left(\frac{X_{i,j}-\mu_{1,j}}{\sigma}\right)^2 \mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}{\left(\sum_{l=1}^{n} \mathbb{1}_{M_{l,j}=0}\mathbb{1}_{Y_l=k}\right)^2}\right| \mathcal{A}_{k,j}^c\right] \\
&= \mathbb{E}\left[\left.\frac{\left(\frac{X_{i,j}-\mu_{k,j}}{\sigma}\right)^2}{\left(1+\sum_{l\neq i} \mathbb{1}_{M_{l,j}=0}\mathbb{1}_{Y_l=k}\right)^2}\right| Y_i = k, M_{i,j} = 0, \mathcal{A}_{k,j}^c\right] \mathbb{P}\left(M_{i,j} = 0, Y_i = k \middle| \mathcal{A}_{k,j}^c\right) \\
&= \mathbb{E}\left[\left.\frac{\left(\frac{X_{i,j}-\mu_{k,j}}{\sigma}\right)^2}{\left(1+\sum_{l\neq i} \mathbb{1}_{M_{l,j}=0}\mathbb{1}_{Y_l=k}\right)^2}\right| Y_i = k, M_{i,j} = 0\right] \frac{\mathbb{P}(M_{i,j} = 0, Y_i = k, \mathcal{A}_{k,j}^c)}{\mathbb{P}(\mathcal{A}_{k,j}^c)} \\
&= \mathbb{E}\left[\left.\frac{\left(\frac{X_{i,j}-\mu_{k,j}}{\sigma}\right)^2}{\left(1+\sum_{l\neq i} \mathbb{1}_{M_{l,j}=0}\mathbb{1}_{Y_l=k}\right)^2}\right| Y_i = k\right] \frac{\mathbb{P}(M_{i,j} = 0)\mathbb{P}(Y_i = k)}{\left(1-\left(\frac{1+\eta}{2}\right)^n\right)} \\
&\hspace{20em} \text{(Using Assumption 2 \& } (X_{i,\mathrm{obs}(M_i)}, M_i, Y_i) \text{ i.i.d)} \\
&= \frac{1-\eta}{2\left(1-\left(\frac{1+\eta}{2}\right)^n\right)} \mathbb{E}\left[\left.\frac{\left(\frac{X_{i,j}-\mu_{k,j}}{\sigma}\right)^2}{\left(1+\sum_{l\neq i} \mathbb{1}_{M_{l,j}=0}\mathbb{1}_{Y_l=k}\right)^2}\right| Y_i = k\right] \qquad \text{(Using Assumption 10 \& } \pi_k = \tfrac{1}{2}) \\
&= \frac{1-\eta}{2\left(1-\left(\frac{1+\eta}{2}\right)^n\right)} \mathbb{E}\left[\left.\left(\frac{X_{i,j}-\mu_{k,j}}{\sigma}\right)^2\right| Y_i = k\right] \mathbb{E}\left[\frac{1}{\left(1+\sum_{l\neq i} \mathbb{1}_{M_{l,j}=0}\mathbb{1}_{Y_l=k}\right)^2}\right] \\
&\hspace{24em} \text{(Using } (X_{i,\mathrm{obs}(M_i)}, M_i, Y_i) \text{ i.i.d)} \\
&= \frac{1-\eta}{2\left(1-\left(\frac{1+\eta}{2}\right)^n\right)} \mathbb{E}\left[\frac{1}{\left(1+\sum_{l\neq i} \mathbb{1}_{M_{l,j}=0}\mathbb{1}_{Y_l=k}\right)^2}\right] \qquad \text{(Using that } \left.\tfrac{X_{i,j}-\mu_{k,j}}{\sigma}\right| Y_i = k \sim \mathcal{N}(0,1))
\end{aligned}
$$

Note that $\sum_{l\neq i} \mathbb{1}_{M_{l,j}=0}\mathbb{1}_{Y_l=k} \sim \mathcal{B}(n-1, \frac{1-\eta}{2})$, then using the Lemma A.4, we have that

$$
\begin{aligned}
\mathbb{E}\left[\left.\frac{\left(\frac{X_{i,j}-\mu_{k,j}}{\sigma}\right)^2 \mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}}{\left(\sum_{i=1}^{n} \mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}\right)^2}\right| \mathcal{A}_{k,j}^c\right] &\leq \frac{1-\eta}{2\left(1-\left(\frac{1+\eta}{2}\right)^n\right)} \frac{2}{\left(\frac{1-\eta}{2}\right)^2 n(n+1)} \\
&= \frac{4}{\left(1-\left(\frac{1+\eta}{2}\right)^n\right)(1-\eta)n(n+1)}.
\end{aligned}
$$

Finally, with the combination of the previous in (56), we have,

$$
\begin{aligned}
&\frac{1}{\sigma}\mathbb{E}\left[\left\|\widehat{\mu}_{k,\mathrm{obs}(m)} - \mu_{k,\mathrm{obs}(m)}\right\|\right] \\
&\quad\leq \left(\frac{\left\|\mu_{k,\mathrm{obs}(m)}\right\|^2}{\sigma^2} \frac{(\eta+1)^n}{2^n} + \left(1 - \frac{(\eta+1)^n}{2^n}\right) \sum_{j\in\mathrm{obs}(m)} \sum_{i=1}^{n} \frac{4}{\left(1-\left(\frac{1+\eta}{2}\right)^n\right)(1-\eta)n(n+1)} \right)^{\frac{1}{2}} \\
&\quad= \left(\frac{\left\|\mu_{k,\mathrm{obs}(m)}\right\|^2}{\sigma^2} \frac{(1+\eta)^n}{2^n} + \frac{4(d-\|m\|_0)}{(1-\eta)(n+1)} \right)^{\frac{1}{2}}
\end{aligned}
$$

□

C.3.4 Proof of Theorem 3.10

Proof. All the lemmas used in this proof (and their corresponding proofs) can be found in Appendix C.3.2 and Appendix C.3.3.

We start by decomposing this difference:

$$
\begin{aligned}
&L(\widehat{h}) - L(h^\star) \\
&\quad= \sum_{m\in\mathcal{M}} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\frac{1}{\sigma}\left\|-\widehat{\mu}_{1,\mathrm{obs}(m)} + \mu_{1,\mathrm{obs}(m)}\right\| + \frac{1}{\sigma}\left\|\widehat{\mu}_{-1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right\| \right] \right) p_m .
\end{aligned}
$$

(using Lemma C.4)

Therefore, we have that

$$
\begin{aligned}
&\leq \sum_{m\in\mathcal{M}} \frac{1}{\sqrt{2\pi}} \left(\frac{\left\|\mu_{1,\mathrm{obs}(m)}\right\|^2}{\sigma^2} \frac{(1+\eta)^n}{2^n} + \frac{4(d-\|m\|_0)}{(1-\eta)(n+1)} \right)^{\frac{1}{2}} p_m \\
&\quad + \sum_{m\in\mathcal{M}} \frac{1}{\sqrt{2\pi}} \left(\frac{\left\|\mu_{-1,\mathrm{obs}(m)}\right\|^2}{\sigma^2} \frac{(1+\eta)^n}{2^n} + \frac{4(d-\|m\|_0)}{(1-\eta)(n+1)} \right)^{\frac{1}{2}} p_m \qquad \text{(using Lemma C.5)} \\
&\leq \sum_{m\in\mathcal{M}} \frac{2}{\sqrt{2\pi}} \left(\frac{\|\mu\|_\infty^2 (d-\|m\|_0)}{\sigma^2} \frac{(1+\eta)^n}{2^n} + \frac{4(d-\|m\|_0)}{(1-\eta)(n+1)} \right)^{\frac{1}{2}} p_m \\
&= \frac{2}{\sqrt{2\pi}} \mathbb{E}\left[\left(\frac{\|\mu\|_\infty^2 (d-B)}{\sigma^2} \frac{(1+\eta)^n}{2^n} + \frac{4(d-B)}{(1-\eta)(n+1)} \right)^{\frac{1}{2}} \right] \qquad \text{(where } B \sim \mathcal{B}(d,\eta)\text{)} \\
&\leq \frac{2}{\sqrt{2\pi}} \mathbb{E}\left[\left(\frac{\|\mu\|_\infty^2 (d-B)}{\sigma^2} \frac{(1+\eta)^n}{2^n} + \frac{4(d-B)}{(1-\eta)(n+1)} \right) \right]^{\frac{1}{2}} \qquad \text{(using Jensen Inequality)} \\
&= \frac{2}{\sqrt{2\pi}} \left(\left(\frac{\|\mu\|_\infty^2 d(1-\eta)}{\sigma^2} \frac{(1+\eta)^n}{2^n} + \frac{4d}{(n+1)} \right) \right)^{\frac{1}{2}},
\end{aligned}
$$

which is lower than the bound given in the theorem. Finally, we observe that if $\eta < 1$, i.e. not all the values are missing, as $(1+\eta)/2 < 1$, then there is an n_0 such that

$$\forall n \geq n_0, \qquad \frac{2\sqrt{d}}{\sqrt{2\pi}} \left(\left(\frac{\|\mu\|_\infty^2 \left(1-\eta\right)}{\sigma^2} \frac{\left(1+\eta\right)^n}{2^n} + \frac{4}{n+1} \right) \right)^{\frac{1}{2}} \lesssim \sqrt{\frac{d}{n}}.$$

□

C.3.5 Proof of Corollary 3.11

Proof of Corollary 3.11. Besides, from Proposition 3.6 and Theorem 3.10 we have that

$$\begin{aligned}
&L(\widehat{h}) - L_{\text{comp}}(h^\star_{\text{comp}}) = L(\widehat{h}) - L(h^\star) + L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) \\
&\leq \frac{2\sqrt{d}}{\sqrt{2\pi}} \left(\frac{\|\mu\|_\infty^2 \left(1-\eta\right)}{\sigma^2} \frac{\left(1+\eta\right)^n}{2^n} + \frac{4}{n+1} \right)^{\frac{1}{2}} + \left(\frac{1}{2} - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) \eta^d \\
&\quad + \frac{\mu}{2\sqrt{2\pi}} \left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left(\left(\eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}} \left(1-\eta\right) \right)^d - \eta^d \right) \right. \\
&\quad \left. - \sqrt{\frac{d}{\lambda_{\max}(\Sigma)}} \left(\eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}} \left(1-\eta\right) \right)^{d-1} e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}} \left(1-\eta\right) \right) \\
&= \frac{2\sqrt{d}}{\sqrt{2\pi}} \left(\frac{\|\mu\|_\infty^2 \left(1-\eta\right)}{\sigma^2} \frac{\left(1+\eta\right)^n}{2^n} + \frac{4}{n+1} \right)^{\frac{1}{2}} + \left(\frac{1}{2} - \Phi\left(-\frac{\mu}{2\sigma} \sqrt{d} \right) \right) \eta^d \\
&\quad + \frac{\mu\sqrt{d}}{2\sigma\sqrt{2\pi}} \left(\left(\eta + e^{-\frac{\mu^2}{8\sigma^2}} \left(1-\eta\right) \right)^d - \eta^d - \left(\eta + e^{-\frac{\mu^2}{8\sigma^2}} \left(1-\eta\right) \right)^{d-1} e^{-\frac{\mu^2}{8\sigma^2}} \left(1-\eta\right) \right) \\
&= \frac{2\sqrt{d}}{\sqrt{2\pi}} \left(\frac{\|\mu\|_\infty^2 \left(1-\eta\right)}{\sigma^2} \frac{\left(1+\eta\right)^n}{2^n} + \frac{4}{n+1} \right)^{\frac{1}{2}} + \left(\frac{1}{2} - \Phi\left(-\frac{\mu}{2\sigma} \sqrt{d} \right) \right) \eta^d \\
&\quad + \frac{\eta\mu\sqrt{d}}{2\sigma\sqrt{2\pi}} \left(\left(\eta + e^{-\frac{\mu^2}{8\sigma^2}} \left(1-\eta\right) \right)^{d-1} - \eta^{d-1} \right).
\end{aligned}$$

□

C.3.6 Lemma for Theorem 3.12

Lemma C.6. *Given an $m \in \mathcal{M}$ and a $k \in \{-1, 1\}$, we have that*

$$\mathbb{E}\left[\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)} \right) \right\| \right] \leq \left(\left(\frac{1+\eta}{2} \right)^n \frac{\|\mu\|_\infty^2 \left(d - \|m\|_0\right)}{\lambda_{\min}\left(\Sigma\right)} + \frac{4\rho(d - \|m\|_0)}{\left(n+1\right)\left(1-\eta\right)} \right)^{\frac{1}{2}}$$

with $\widehat{\mu}_{k,\text{obs}(m)}$ defined in (18) and $\rho := \max_{i \in [n]} \Sigma_{i,i}/\lambda_{\min}(\Sigma)$ the greatest value of the diagonal of the covariance matrix divided by its smallest eigenvalue.

Proof. Firstly, applying Jensen Inequality, we have that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right\|\right] \leq \mathbb{E}\left[\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right\|^2\right]^{\frac{1}{2}} \\
&= \mathbb{E}\left[\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right)^\top \left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right)\right]^{\frac{1}{2}} \\
&= \mathbb{E}\left[\mathrm{tr}\left(\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right)^\top \left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right)\right)\right]^{\frac{1}{2}} \\
&= \mathbb{E}\left[\mathrm{tr}\left(\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right) \left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right)^\top\right)\right]^{\frac{1}{2}} \\
&= \mathrm{tr}\left(\mathbb{E}\left[\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right) \left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right)^\top\right]\right)^{\frac{1}{2}} \\
&= \mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mathbb{E}\left[\left(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)}\right)\left(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)}\right)^\top\right]\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)^{\frac{1}{2}}.
\end{aligned}$$

Hence, begin by examining the matrix provided by

$$\mathcal{C}(k,m) := \mathbb{E}\left[\left(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)}\right)\left(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)}\right)^\top\right].$$

By adopting this approach, we can solve the following problem

$$\mathbb{E}\left[\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right\|\right] \leq \mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mathcal{C}(k,m)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)^{\frac{1}{2}}. \tag{57}$$

In order to accomplish this, we calculate the coordinate $\mathcal{C}(k,m)_{r,l}$, where $r,l \in \mathrm{obs}(m)$. This can be done by decomposing it into two cases: when $r = l$ and when $r \neq l$.

- $\underline{r = l}$ The objective is
$$\mathbb{E}\left[(\widehat{\mu}_{k,l}-\mu_{k,l})^2\right].$$
To compute this quantity, we split up in to the case $\mathcal{A}_{k,l}$ where $\mu_{k,l}$ cannot be estimated, i.e. there is no sample of the class k where the l-th coordinate is observed, and the complementary case $\mathcal{A}_{k,l}^c$. Formally,
$$\mathcal{A}_{k,l} := \{\forall i \in \{1, ...n\}, \quad (Y_i = -k \vee M_{i,l} = 1) = 1\}. \tag{58}$$
Note that
$$\mathbb{P}(\mathcal{A}_{k,l}) = \Pi_{i=1}^n P(Y_i = -k \text{ or } M_{i,l} = 1) = \left(\frac{1+\eta}{2}\right)^n.$$
Then, we have that
$$\begin{aligned}
&\mathbb{E}\left[(\widehat{\mu}_{k,l}-\mu_{k,l})^2\right] = \mathbb{E}\left[(\widehat{\mu}_{k,l}-\mu_{k,l})^2\middle|\mathcal{A}_{k,l}\right]\mathbb{P}\left(\mathcal{A}_{k,l}\right) + \mathbb{E}\left[(\widehat{\mu}_{k,l}-\mu_{k,l})^2\middle|\mathcal{A}_{k,l}^c\right]\mathbb{P}\left(\mathcal{A}_{k,l}^c\right) \\
&= (0-\mu_{k,l})^2\left(\frac{1+\eta}{2}\right)^n + \mathbb{E}\left[(\widehat{\mu}_{k,l}-\mu_{k,l})^2\middle|\mathcal{A}_{k,l}^c\right]\left(1-\left(\frac{1+\eta}{2}\right)^n\right) \\
&= \mu_{k,l}^2\left(\frac{1+\eta}{2}\right)^n + \mathbb{E}\left[\left(\frac{\sum_{i=1}^n (X_{i,l}-\mu_{k,l})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}\right)^2\middle|\mathcal{A}_{k,l}^c\right]\left(1-\left(\frac{1+\eta}{2}\right)^n\right)
\end{aligned}$$

Now, we can employ the same approach utilized in the proof of Lemma C.5.
Then, this is equivalent to

$$
\begin{aligned}
&= \mu_{k,l}^2 \left(\frac{1+\eta}{2}\right)^n + \sum_{i=1}^n \mathbb{E}\left[\frac{(X_{i,l}-\mu_{k,l})^2 \mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}{\left(\sum_{i=1}^n \mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}\right)^2}\middle|\mathcal{A}_{k,l}^c\right]\left(1-\left(\frac{1+\eta}{2}\right)^n\right) \\
&= \mu_{k,l}^2 \left(\frac{1+\eta}{2}\right)^n \\
&\quad + \sum_{i=1}^n \mathbb{E}\left[\frac{(X_{i,l}-\mu_{k,l})^2}{\left(1+\sum_{j\neq i}^n \mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}\right)^2}\middle|\mathcal{A}_{k,l}^c, Y_i=k, M_{i,l}=0\right]\mathbb{P}\left(Y_i=k, M_{i,l}=0\middle|\mathcal{A}_{k,l}^c\right)\left(1-\left(\frac{1+\eta}{2}\right)^n\right) \\
&= \mu_{k,l}^2 \left(\frac{1+\eta}{2}\right)^n + \sum_{i=1}^n \mathbb{E}\left[\frac{(X_{i,l}-\mu_{k,l})^2}{\left(1+\sum_{j\neq i}^n \mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}\right)^2}\middle|Y_i=k, M_{i,l}=0\right]\frac{1-\eta}{2} \\
&= \mu_{k,l}^2 \left(\frac{1+\eta}{2}\right)^n + \sum_{i=1}^n \mathbb{E}\left[\frac{(X_{i,l}-\mu_{k,l})^2}{\left(1+\sum_{j\neq i}^n \mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}\right)^2}\middle|Y_i=k\right]\frac{1-\eta}{2} \\
&\hspace{20em}\text{(using Assumption 2)} \\
&= \mu_{k,l}^2 \left(\frac{1+\eta}{2}\right)^n + n\mathbb{E}\left[\frac{(X_{1,l}-\mu_{k,l})^2}{\left(1+\sum_{j\neq 1}^n \mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}\right)^2}\middle|Y_1=k\right]\frac{1-\eta}{2} \\
&\hspace{20em}\text{(using the exchangeability)} \\
&= \mu_{k,l}^2 \left(\frac{1+\eta}{2}\right)^n + n\mathbb{E}\left[(X_{1,l}-\mu_{k,l})^2\middle|Y_1=k\right]\mathbb{E}\left[\frac{1}{\left(1+\sum_{j\neq 1}^n \mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}\right)^2}\right]\frac{1-\eta}{2} \\
&\hspace{20em}\text{(using the independence)} \\
&= \mu_{k,l}^2 \left(\frac{1+\eta}{2}\right)^n + n\Sigma_{l,l}\mathbb{E}\left[\frac{1}{\left(1+\sum_{j\neq 1}^n \mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}\right)^2}\right]\frac{1-\eta}{2}.
\end{aligned}
$$

In the sequel, we denote

$$A(n,\eta) := \mathbb{E}\left[\frac{1}{(1+B)^2}\right],$$

where $B \sim \mathcal{B}(n-1,(1-\eta)/2)$. Then, we have that

$$\mathcal{C}(k,m)_{l,l} = \mu_{k,l}^2 \left(\frac{1+\eta}{2}\right)^n + n\Sigma_{l,l}A(n,\eta)\frac{1-\eta}{2}. \tag{59}$$

- $\underline{r \neq l}$ The objective is

$$\mathcal{C}(k,m)_{r,l} = \mathbb{E}\left[(\widehat{\mu}_{k,r}-\mu_{k,r})(\widehat{\mu}_{k,l}-\mu_{k,l})\right].$$

In order to estimate each component r and l of the mean, note that three events are possible:

- There are no samples available to estimate any of the means. For each sample, it either belongs to the other class or is missing at both coordinates. Formally, we denote this event as follows:

  $$\mathcal{A}_{k,l,r} := \{\forall i \in \{1, ...n\}, \qquad (Y_i = -k \vee (M_{i,r} = 1 \wedge M_{i,l} = 1)) = 1\}.$$

  Observe that the probability of this event is

  $$\begin{aligned}\mathbb{P}(\mathcal{A}_{k,l,r}) &= \mathbb{P}\left(\{\forall i \in \{1, ...n\}, \qquad (Y_i = -k \vee (M_{i,r} = 1 \wedge M_{i,l} = 1)) = 1\}\right)\\ &= \Pi_{i=1}^n \mathbb{P}\left(\{(Y_i = -k \vee (M_{i,r} = 1 \wedge M_{i,l} = 1)) = 1\}\right)\\ &= (\mathbb{P}(Y_i = -k) + \mathbb{P}(M_{i,r} = 1 \cap M_{i,l} = 1) - \mathbb{P}(Y_i = -k \cap M_{i,r} = 1 \cap M_{i,l} = 1))^n\\ &= \left(\frac{\eta^2+1}{2}\right)^n. \end{aligned} \tag{60}$$

  Note that if $\eta < 1$, then this probability vanishes exponentially. Furthermore, in this case we have that

  $$\mathbb{E}\left[(\widehat{\mu}_{k,r} - \mu_{k,r})\,(\widehat{\mu}_{k,l} - \mu_{k,l})|\mathcal{A}_{k,l,r}\right] = \mu_{k,r}\mu_{k,l}. \tag{61}$$

- Estimates for both means can be obtained since there are available samples of class k at coordinates l and r. This can be formalized as

  $$\mathcal{B}_{k,l,r} := \{\exists i \in \{1, ..., n\}, \quad (Y_i = k \wedge M_{i,l} = 0) = 1\} \cap \{\exists i \in \{1, ...n\}, \quad (Y_i = k \wedge M_{i,r} = 0) = 1\}.$$

  To compute this probability, observe that

  $$\begin{aligned}\mathbb{P}\left(\mathcal{B}_{k,l,r}\right) = 1 - \mathbb{P}&\left(\{\forall i \in \{1, ..., n\}, \quad (Y_i = -k \vee M_{i,l} = 1) = 1\}\right.\\ &\left.\cup \{\forall i \in \{1, ...n\}, \quad (Y_i = -k \vee M_{i,r} = 1) = 1\}\right)\\ = 1 &- \mathbb{P}\left(\{\forall i \in \{1, ..., n\}, \quad (Y_i = -k \vee M_{i,l} = 1) = 1\}\right)\\ &- \mathbb{P}\left(\{\forall i \in \{1, ...n\}, \quad (Y_i = -k \vee M_{i,r} = 1) = 1\}\right)\\ &+ \mathbb{P}\left(\{\forall i \in \{1, ..., n\}, \quad (Y_i = -k \vee (M_{i,l} = 1 \wedge M_{i,r} = 1)) = 1\}\right),\end{aligned}$$

  where the last probability was already computed for $\mathcal{A}_{k,l,r}$. On the other hand, remark that

  $$\mathbb{P}\left(\{\forall i \in \{1, ...n\}, \quad (Y_i = -k \vee M_{i,r} = 1) = 1\}\right) = \left(\frac{1+\eta}{2}\right)^n.$$

  Then, we have that

  $$\mathbb{P}\left(\mathcal{B}_{k,l,r}\right) = 1 - 2\left(\frac{1+\eta}{2}\right)^n + \left(\frac{\eta^2+1}{2}\right)^n. \tag{62}$$

  Note that if $\eta < 1$, this probability tends to 1 as n increases.

Moreover, we have that

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{\mu}_{k,r}-\mu_{k,r})\left(\widehat{\mu}_{k,l}-\mu_{k,l}\right)|\mathcal{B}_{k,l,r}\right]\\
&\quad=\mathbb{E}\left[\left(\frac{\sum_{i=1}^{n}(X_{i,r}-\mu_{k,r})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}\right)\left(\frac{\sum_{i=1}^{n}(X_{i,l}-\mu_{k,l})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}\right)\middle|\mathcal{B}_{k,l,r}\right]\\
&\quad=\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\left(\frac{(X_{i,r}-\mu_{k,r})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}\right)\left(\frac{(X_{j,l}-\mu_{k,l})\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}\right)\middle|\mathcal{B}_{k,l,r}\right]\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{(X_{i,r}-\mu_{k,r})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}\right)\left(\frac{(X_{i,l}-\mu_{k,l})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}\right)\middle|\mathcal{B}_{k,l,r}\right]\\
&\qquad+\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathbb{E}\left[\left(\frac{(X_{i,r}-\mu_{k,r})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}\right)\left(\frac{(X_{j,l}-\mu_{k,l})\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}\right)\middle|\mathcal{B}_{k,l,r}\right].
\end{aligned}$$

Observe that this second term is sum of null terms. Effectively, we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{(X_{i,r}-\mu_{k,r})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}\right)\left(\frac{(X_{j,l}-\mu_{k,l})\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}\right)\middle|\mathcal{B}_{k,l,r}\right]\\
&\quad=\mathbb{E}\left[\left(\frac{(X_{i,r}-\mu_{k,r})}{1+\mathbb{1}_{M_{j,r}=0}+\sum_{s\neq i,j}^{n}\mathbb{1}_{Y_s=k}\mathbb{1}_{M_{s,r}=0}}\right)\left(\frac{(X_{j,l}-\mu_{k,l})}{1+\mathbb{1}_{M_{i,l}=0}+\sum_{s\neq i,j}^{n}\mathbb{1}_{Y_s=k}\mathbb{1}_{M_{s,l}=0}}\right)\right.\\
&\qquad\left.|Y_i=k,M_{i,r}=0,Y_j=k,M_{j,l}=0\right]\mathbb{P}\left(Y_i=k,M_{i,r}=0,Y_j=k,M_{j,l}=0|\mathcal{B}_{k,l,r}\right)\\
&\quad=\mathbb{E}\left[\frac{1}{\left(1+\mathbb{1}_{M_{j,r}=0}+\sum_{s\neq i,j}^{n}\mathbb{1}_{Y_s=k}\mathbb{1}_{M_{s,r}=0}\right)\left(1+\mathbb{1}_{M_{i,l}=0}+\sum_{s\neq i,j}^{n}\mathbb{1}_{Y_s=k}\mathbb{1}_{M_{s,l}=0}\right)}\right]\\
&\qquad\mathbb{E}\left[(X_{i,r}-\mu_{k,r})|Y_i=k\right]\mathbb{E}\left[(X_{j,l}-\mu_{k,l})|Y_j=k\right]\mathbb{P}\left(Y_i=k,M_{i,r}=0,Y_j=k,M_{j,l}=0|\mathcal{B}_{k,l,r}\right)\\
&\qquad\qquad\text{(using Assumption 2 and independence)}\\
&\quad=0.
\end{aligned}$$

Then, retaking the expression,

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{\mu}_{k,r}-\mu_{k,r})\left(\widehat{\mu}_{k,l}-\mu_{k,l}\right)|\mathcal{B}_{k,l,r}\right]\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{(X_{i,r}-\mu_{k,r})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}\right)\left(\frac{(X_{i,l}-\mu_{k,l})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}\right)\middle|\mathcal{B}_{k,l,r}\right]\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\left[\frac{(X_{i,r}-\mu_{k,r})}{1+\sum_{j\neq i}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,r}=0}}\frac{(X_{i,l}-\mu_{k,l})}{1+\sum_{j\neq i}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}}\middle|Y_i=k,M_{i,r}=0,M_{i,l}=0\right]\\
&\qquad\mathbb{P}\left(Y_i=k,M_{i,r}=0,M_{i,l}=0|\mathcal{B}_{k,l,r}\right)\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\left[(X_{i,r}-\mu_{k,r})(X_{i,l}-\mu_{k,l})|Y_i=k\right]\mathbb{E}\left[\frac{1}{1+\sum_{j\neq i}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,r}=0}}\frac{1}{1+\sum_{j\neq i}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}}\right]\\
&\qquad\mathbb{P}\left(Y_i=k,M_{i,r}=0,M_{i,l}=0|\mathcal{B}_{k,l,r}\right)\\
&\quad=n\Sigma_{r,l}B(n,\eta)\frac{(1-\eta)^2}{2\mathbb{P}\left(\mathcal{B}_{k,l,r}\right)}, \qquad (63)
\end{aligned}$$

where $B(n,\eta):=\mathbb{E}\left[\frac{1}{1+\sum_{j=2}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,r}=0}}\frac{1}{1+\sum_{j=2}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}}\right]$.

– The ultimate case is where only one mean can be estimated $\mathcal{C}(k,m)_{k,l,r}$. We can compute the probability of this case as

$$\mathbb{P}(\mathcal{C}(k,m)_{k,l,r}) = \mathbb{P}\left((\mathcal{B}_{k,l,r} \cup \mathcal{A}_{k,l,r})^c\right) = 2\left(\frac{1+\eta}{2}\right)^n - 2\left(\frac{\eta^2+1}{2}\right)^n.$$

Nevertheless, note that in this case $\mathbb{E}\left[\left(\widehat{\mu}_{k,r} - \mu_{k,r}\right)\left(\widehat{\mu}_{k,l} - \mu_{k,l}\right)|\mathcal{C}(k,m)_{k,l,r}\right] = 0$. Effectively, decompose $\mathcal{C}(k,m)_{k,l,r} = \mathcal{C}(k,m)^1_{k,l,r} \coprod \mathcal{C}(k,m)^2_{k,l,r}$ without loss of generality, where $\mathcal{C}(k,m)^1_{k,l,r}$ is the event where the one that can be estimated is $\widehat{\mu}_{k,r}$ and there are no samples of class k where the l-th coordinate is observed. Then, note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\widehat{\mu}_{k,r} - \mu_{k,r}\right)\left(\widehat{\mu}_{k,l} - \mu_{k,l}\right)\middle|\mathcal{C}(k,m)^1_{k,l,r}\right] \\
&\quad = -\mu_{k,l}\mathbb{E}\left[\widehat{\mu}_{k,r} - \mu_{k,r}\middle|\mathcal{C}(k,m)^1_{k,l,r}\right] \\
&\quad = -\mu_{k,l}\mathbb{E}\left[\frac{\sum_{i=1}^n (X_{i,r} - \mu_{k,r})\mathbb{1}_{M_{i,r}=0}\mathbb{1}_{Y_i=k}}{\sum_{i=1}^n \mathbb{1}_{M_{i,r}=0}\mathbb{1}_{Y_i=k}}\middle|\mathcal{C}(k,m)^1_{k,l,r}\right] \\
&\quad = -\mu_{k,l}n\mathbb{E}\left[\frac{(X_{1,r} - \mu_{k,r})}{1+\sum_{i=2}^n \mathbb{1}_{M_{i,r}=0}\mathbb{1}_{Y_i=k}}\middle|M_{1,r}=0, Y_1=k\right]\mathbb{P}(M_{1,r}=0, Y_1=k|\mathcal{C}(k,m)^1_{k,l,r}) \\
&\quad = -\mu_{k,l}n\mathbb{E}\left[(X_{1,r} - \mu_{k,r})|Y_1=k\right]\mathbb{E}\left[\frac{1}{1+\sum_{i=2}^n \mathbb{1}_{M_{i,r}=0}\mathbb{1}_{Y_i=k}}\right]\mathbb{P}(M_{1,r}=0, Y_1=k|\mathcal{C}(k,m)^1_{k,l,r}) \\
&\qquad\qquad \text{(using MCAR and independence)} \\
&\quad = 0. \qquad (64)
\end{aligned}$$

It is completely symmetric for the case $\mathcal{C}(k,m)^2_{k,l,r}$.

Based on the previously discussed details, we are able to decompose $\mathcal{C}(k,m)_{r,l}$ as follows

$$\begin{aligned}
\mathcal{C}(k,m)_{r,l} &= \mathbb{E}\left[\left(\widehat{\mu}_{k,r} - \mu_{k,r}\right)\left(\widehat{\mu}_{k,l} - \mu_{k,l}\right)\right] \\
&= \mathbb{E}\left[\left(\widehat{\mu}_{k,r} - \mu_{k,r}\right)\left(\widehat{\mu}_{k,l} - \mu_{k,l}|\mathcal{A}_{k,l,r}\right)\right]\mathbb{P}\left(\mathcal{A}_{k,l,r}\right) \\
&\quad + \mathbb{E}\left[\left(\widehat{\mu}_{k,r} - \mu_{k,r}\right)\left(\widehat{\mu}_{k,l} - \mu_{k,l}|\mathcal{B}_{k,l,r}\right)\right]\mathbb{P}\left(\mathcal{B}_{k,l,r}\right) \\
&\quad + \mathbb{E}\left[\left(\widehat{\mu}_{k,r} - \mu_{k,r}\right)\left(\widehat{\mu}_{k,l} - \mu_{k,l}|\mathcal{C}(k,m)_{k,l,r}\right)\right]\mathbb{P}\left(\mathcal{C}(k,m)_{k,l,r}\right) \\
&= \mu_{k,r}\mu_{k,l}\left(\frac{\eta^2+1}{2}\right)^n \qquad \text{(using (60) and (61))} \\
&\quad + n\Sigma_{r,l}B(n,\eta)\frac{(1-\eta)^2}{2\mathbb{P}\left(\mathcal{B}_{k,l,r}\right)}\mathbb{P}\left(\mathcal{B}_{k,l,r}\right) \qquad \text{(using (63))} \\
&\quad + 0\mathbb{P}\left(\mathcal{C}(k,m)_{k,l,r}\right) \qquad \text{(using (64))} \\
&= \mu_{k,r}\mu_{k,l}\left(\frac{\eta^2+1}{2}\right)^n + n\Sigma_{r,l}B(n,\eta)\frac{(1-\eta)^2}{2} \qquad (65)
\end{aligned}$$

Then, from (59) and (65), we have that

$$\mathcal{C}(k,m) = F \odot \mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^\top + n\frac{1-\eta}{2}G \odot \Sigma_{\mathrm{obs}(m)},$$

where F is defined with a diagonal of $\left(\frac{1+\eta}{2}\right)^n$ and the remaining elements in the matrix as $\left(\frac{1+\eta^2}{2}\right)^n$, and G is defined with a diagonal of $A(n,\eta)$ and the remaining elements as $(1-\eta)B(n,\eta)$.

Then, if we take the first inequality given by (57), we have that

$$\mathbb{E}\left[\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)} - \mu_{k,\mathrm{obs}(m)})\right\|\right] \leq \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mathcal{C}(k,m)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)^{\frac{1}{2}}$$

$$= \left(\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(F \odot \mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) + n\frac{1-\eta}{2}\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(G \odot \Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\right)^{\frac{1}{2}} \tag{66}$$

- To solve $\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(F \odot \mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)$:

  We decompose F as:

$$F = \begin{bmatrix} \left(\frac{1+\eta}{2}\right)^n & \left(\frac{1+\eta^2}{2}\right)^n & \cdots & \left(\frac{1+\eta^2}{2}\right)^n \\ \left(\frac{1+\eta^2}{2}\right)^n & \left(\frac{1+\eta}{2}\right)^n & \cdots & \left(\frac{1+\eta^2}{2}\right)^n \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1+\eta^2}{2}\right)^n & \left(\frac{1+\eta^2}{2}\right)^n & \cdots & \left(\frac{1+\eta}{2}\right)^n \end{bmatrix}$$

$$= \left(\frac{1+\eta^2}{2}\right)^n \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{\mathbf{1}} + \left(\left(\frac{1+\eta}{2}\right)^n - \left(\frac{1+\eta^2}{2}\right)^n\right) I_{d-\|m\|_0}.$$

Then, we have that

$$
\begin{aligned}
&\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(F \odot \mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\quad= \left(\frac{1+\eta^2}{2}\right)^{n} \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mathbf{1} \odot \mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\qquad\quad + \left(\left(\frac{1+\eta}{2}\right)^{n} - \left(\frac{1+\eta^2}{2}\right)^{n}\right) \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(I_{d-\|m\|_0} \odot \mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\quad= \left(\frac{1+\eta^2}{2}\right)^{n} \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\qquad\quad + \left(\left(\frac{1+\eta}{2}\right)^{n} - \left(\frac{1+\eta^2}{2}\right)^{n}\right) \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\operatorname{diag}\left(\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\quad\leq \left(\frac{1+\eta^2}{2}\right)^{n} \operatorname{tr}\left(\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\right)\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\right)^{\top}\right)\\
&\qquad\quad + \left(\left(\frac{1+\eta}{2}\right)^{n} - \left(\frac{1+\eta^2}{2}\right)^{n}\right) \|\mu\|_{\infty}^{2} \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-1}\right) \qquad \text{(using Lemma A.5)}\\
&\quad\leq \left(\frac{1+\eta^2}{2}\right)^{n} \operatorname{tr}\left(\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\right)^{\top}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\right)\right)\\
&\qquad\quad + \left(\left(\frac{1+\eta}{2}\right)^{n} - \left(\frac{1+\eta^2}{2}\right)^{n}\right) \|\mu\|_{\infty}^{2} \frac{1}{\lambda_{\min}\left(\Sigma_{\mathrm{obs}(m)}\right)}(d-\|m\|_0)\\
&\quad= \left(\frac{1+\eta^2}{2}\right)^{n} \left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\right\|^{2} + \left(\left(\frac{1+\eta}{2}\right)^{n} - \left(\frac{1+\eta^2}{2}\right)^{n}\right) \frac{\|\mu\|_{\infty}^{2}(d-\|m\|_0)}{\lambda_{\min}(\Sigma)}\\
&\quad\leq \left(\frac{1+\eta^2}{2}\right)^{n} \frac{\left\|\mu_{k,\mathrm{obs}(m)}\right\|^{2}}{\lambda_{\min}(\Sigma)} + \left(\left(\frac{1+\eta}{2}\right)^{n} - \left(\frac{1+\eta^2}{2}\right)^{n}\right) \frac{\|\mu\|_{\infty}^{2}(d-\|m\|_0)}{\lambda_{\min}(\Sigma)}\\
&\quad\leq \left(\frac{1+\eta^2}{2}\right)^{n} \frac{\|\mu\|_{\infty}^{2}(d-\|m\|_0)}{\lambda_{\min}(\Sigma)} + \left(\left(\frac{1+\eta}{2}\right)^{n} - \left(\frac{1+\eta^2}{2}\right)^{n}\right) \frac{\|\mu\|_{\infty}^{2}(d-\|m\|_0)}{\lambda_{\min}(\Sigma)}.
\end{aligned}
$$

Thus, conclude that

$$
\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(F \odot \mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \leq \left(\frac{1+\eta}{2}\right)^{n} \frac{\|\mu\|_{\infty}^{2}(d-\|m\|_0)}{\lambda_{\min}(\Sigma)} \qquad (67)
$$

- To solve $\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(G \odot \Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)$:

We decompose G as

$$\begin{aligned}G &= \begin{bmatrix} A(n,\eta) & (1-\eta)B(n,\eta) & \cdots & (1-\eta)B(n,\eta) \\ (1-\eta)B(n,\eta) & A(n,\eta) & \cdots & (1-\eta)B(n,\eta) \\ \vdots & \vdots & \ddots & \vdots \\ (1-\eta)B(n,\eta) & (1-\eta)B(n,\eta) & \cdots & A(n,\eta) \end{bmatrix} \\ &= (1-\eta)B(n,\eta) \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{\mathbf{1}} + (A(n,\eta) - (1-\eta)B(n,\eta)) I_{d-\|m\|_0}.\end{aligned}$$

Note that $A(n,\eta) \geq B(n,\eta)$, then $A(n,\eta) - (1-\eta)B(n,\eta) \geq 0$. To prove this, remark that

$$\begin{aligned}B(n,\eta) &:= \mathbb{E}\left[\frac{1}{1+\sum_{j=1}^{n-1} \mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,r}=0}} \frac{1}{1+\sum_{j=1}^{n-1} \mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{1+\sum_{j=1}^{Z} \mathbb{1}_{M_{j,r}=0}} \frac{1}{1+\sum_{j=1}^{Z} \mathbb{1}_{M_{j,l}=0}} \middle| Z\right]\right]\end{aligned}$$

where $Z := \sum_{i=1}^{n-1} \mathbb{1}_{Y_i=k} \sim \mathcal{B}(n-1, 1/2)$ using the exchangeability as the sample is i.i.d. By leveraging the independence between the missingness at coordinate r and coordinate l, as well as the independence of each sample from the rest, we can conclude that

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[\frac{1}{1+\sum_{j=1}^{Z} \mathbb{1}_{M_{j,r}=0}} \frac{1}{1+\sum_{j=1}^{Z} \mathbb{1}_{M_{j,l}=0}} \middle| Z\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{1+\sum_{j=1}^{Z} \mathbb{1}_{M_{j,r}=0}} \middle| Z\right] \mathbb{E}\left[\frac{1}{1+\sum_{j=1}^{Z} \mathbb{1}_{M_{j,l}=0}} \middle| Z\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{1+\sum_{j=1}^{Z} \mathbb{1}_{M_{j,r}=0}} \middle| Z\right]^2\right] && \text{(using that } M_{j,r} \sim M_{j,l}\text{)} \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\left(1+\sum_{j=1}^{Z} \mathbb{1}_{M_{j,r}=0}\right)^2} \middle| Z\right]\right] && \text{(using Jensen Inequality)} \\ &= A(n,\eta).\end{aligned}$$

Thus, we have that

$$
\begin{aligned}
&\mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(G\odot\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\quad=(1-\eta)B(n,\eta)\mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mathbf{1}\odot\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\qquad+(A(n,\eta)-(1-\eta)B(n,\eta))\mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(I_{d-\|m\|_0}\odot\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\quad=(1-\eta)B(n,\eta)\mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\Sigma_{\mathrm{obs}(m)}\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\qquad+(A(n,\eta)-(1-\eta)B(n,\eta))\mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mathrm{diag}\left(\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\quad\leq(1-\eta)B(n,\eta)(d-\|m\|_0)\\
&\qquad+(A(n,\eta)-(1-\eta)B(n,\eta))\max_{i\in[d]}\left(\Sigma_{i,i}\right)\mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-1}\right) &&\text{(using Lemma A.5)}\\
&\quad\leq(1-\eta)B(n,\eta)(d-\|m\|_0)\\
&\qquad+(A(n,\eta)-(1-\eta)B(n,\eta))\frac{\max_{i\in[d]}\left(\Sigma_{i,i}\right)}{\lambda_{\min}\left(\Sigma\right)}(d-\|m\|_0)\\
&\quad\leq\rho(1-\eta)B(n,\eta)(d-\|m\|_0)+(A(n,\eta)-(1-\eta)B(n,\eta))\rho(d-\|m\|_0)\\
&&&\text{(using that } \rho:=\tfrac{\max_{i\in[d]}(\Sigma_{i,i})}{\lambda_{\min}(\Sigma)}\geq 1)\\
&\quad=A(n,\eta)\rho(d-\|m\|_0).
\end{aligned}
$$

Furthermore, recall that

$$
A(n,\eta):=\mathbb{E}\left[\frac{1}{(1+B)^2}\right],
$$

where $B\sim\mathcal{B}(n-1,(1-\eta)/2)$. Then, using Lemma A.4, conclude that

$$
\mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(G\odot\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\leq A(n,\eta)\rho(d-\|m\|_0)\leq\frac{2\rho(d-\|m\|_0)}{n(n+1)\left(\frac{1-\eta}{2}\right)^2} \tag{68}
$$

Finally, combining all the previous results in (66), conclude that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right\|\right]\leq\\
&\quad=\left(\mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(F\odot\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)+n\frac{1-\eta}{2}\mathrm{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(G\odot\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\right)^{\frac{1}{2}}\\
&\quad\leq\left(\left(\frac{1+\eta}{2}\right)^n\frac{\|\mu\|_\infty^2\left(d-\|m\|_0\right)}{\lambda_{\min}\left(\Sigma\right)}+n\frac{1-\eta}{2}\frac{2\rho(d-\|m\|_0)}{n(n+1)\left(\frac{1-\eta}{2}\right)^2}\right)^{\frac{1}{2}}\\
&&&\text{(using Equation (67) and Equation (68))}\\
&\quad\leq\left(\left(\frac{1+\eta}{2}\right)^n\frac{\|\mu\|_\infty^2\left(d-\|m\|_0\right)}{\lambda_{\min}\left(\Sigma\right)}+\frac{4\rho(d-\|m\|_0)}{(n+1)\left(1-\eta\right)}\right)^{\frac{1}{2}}
\end{aligned}
$$

□

C.3.7 Proof of Theorem 3.12

Proof. By Lemma C.4,

$$
\begin{aligned}
&L(\widehat{h}) - L(h^\star)\\
&\quad\le \sum_{m\in\mathcal{M}} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(-\widehat{\mu}_{1,\text{obs}(m)} + \mu_{1,\text{obs}(m)})\right\| + \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|\right]\right) p_m\\
&\quad\le \frac{2}{\sqrt{2\pi}} \sum_{m\in\mathcal{M}} \left(\left(\frac{1+\eta}{2}\right)^n \frac{\|\mu\|_\infty^2 (d - \|m\|_0)}{\lambda_{\min}(\Sigma)} + \frac{4\rho(d-\|m\|_0)}{(n+1)(1-\eta)}\right)^{\frac{1}{2}} p_m. \qquad \text{(using Lemma C.6)}
\end{aligned}
$$

Now, using Assumption 10, we have that $\|M\|_0 \sim \mathcal{B}(d,\eta)$, so that

$$
\begin{aligned}
&\frac{2}{\sqrt{2\pi}} \sum_{m\in\mathcal{M}} \left(\left(\frac{1+\eta}{2}\right)^n \frac{\|\mu\|_\infty^2 (d - \|m\|_0)}{\lambda_{\min}(\Sigma)} + \frac{4\rho(d-\|m\|_0)}{(n+1)(1-\eta)}\right)^{\frac{1}{2}} p_m\\
&\quad= \frac{2}{\sqrt{2\pi}}\mathbb{E}\left[\left(\left(\frac{1+\eta}{2}\right)^n \frac{\|\mu\|_\infty^2 (d - B)}{\lambda_{\min}(\Sigma)} + \frac{4\rho(d-B)}{(n+1)(1-\eta)}\right)^{\frac{1}{2}}\right] \qquad \text{(where } B\sim\mathcal{B}(d,\eta)\text{)}\\
&\quad\le \frac{2}{\sqrt{2\pi}}\mathbb{E}\left[\left(\left(\frac{1+\eta}{2}\right)^n \frac{\|\mu\|_\infty^2 (d - B)}{\lambda_{\min}(\Sigma)} + \frac{4\rho(d-B)}{(n+1)(1-\eta)}\right)\right]^{\frac{1}{2}} \qquad \text{(using Jensen Inequality)}\\
&\quad\le \frac{2}{\sqrt{2\pi}}\left(\left(\frac{1+\eta}{2}\right)^n \frac{\|\mu\|_\infty^2 d(1-\eta)}{\lambda_{\min}(\Sigma)} + \frac{4\rho d}{n}\right)^{\frac{1}{2}}.
\end{aligned}
$$

□

C.4 Proofs of Section 3.2.4

C.4.1 Lemma for Theorem 3.13

Lemma C.7. *Given a missing pattern $m\in\{0,1\}^d$ and dataset $\mathcal{D}_n$ that satisfies Assumptions 2 and 9, with a covariance matrix $\Sigma = \sigma^2 I_d$, then we can state the following:*

$$
\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}\right\|^2\right] \le &\sum_{j\notin\text{supp}(\mu)\cap\text{obs}(m)} \frac{\sqrt{6}\sigma^2}{d}\frac{4}{(1-\eta)(n+1)}\\
&+ \sum_{j\in\text{supp}(\mu)\cap\text{obs}(m)} \left(\sigma^2\left(1+8e^{-1}+8\log(d)\right)\frac{4}{(1-\eta)(n+1)} + \|\mu\|_\infty^2\left(\frac{1+\eta}{2}\right)^n\right),
\end{aligned}
$$

where the estimate $\widetilde{\mu}_k$ was defined in (21) and $\text{supp}(\mu)$ *is the set of indices of discriminating coordinates.*

Proof. First, recall that $N_{k,j} := \sum_{i=1}^n \mathbb{1}_{M_{i,j}=0}\mathbb{1}_{Y_i=k}$ represents the number of observations for the j-th coordinate of class k. Note that $N_{k,j}\sim\mathcal{B}(n,\frac{1-\eta}{2})$. Second, lets denote $\overline{\mu}_{k,j}$ the Estimate (18) knowing $N_{k,j}$. Then,

$$
\overline{\mu}_{k,j} := \widehat{\mu}_{k,j}|N_{k,j} = \frac{\sum_{i=1}^{N_{k,j}} X_{i,j}}{N_{k,j}} \sim \mathcal{N}\left(\mu_{k,j}, \frac{\sigma^2}{N_{k,j}}\right),
$$

where we have used that the sample is i.i.d. Lets separate the sum depending on the coordinates that are discriminating with the ones that are not:

$$\mathbb{E}\left[\left\|\widetilde{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)}\right\|^2\right]=\sum_{j\notin\mathrm{supp}(\mu)\cap\mathrm{obs}(m)}\mathbb{E}\left[\widetilde{\mu}_{k,j}^2\right]+\sum_{j\in\mathrm{supp}(\mu)\cap\mathrm{obs}(m)}\mathbb{E}\left[\left(\widetilde{\mu}_{k,j}-\mu_{k,j}\right)^2\right]. \tag{69}$$

Hence, it is possible to decompose the proof into two parts: the first part calculates the first term, and the second part calculates the other term.

- Computation of $\mathbb{E}\left[\widetilde{\mu}_{k,j}^2\right]$ knowing that $j\notin\mathrm{supp}(\mu)$, i.e. $\mu_{k,j}=0$.

  Remember that when the estimate (18) was defined, the convention of $0/0=0$ was adopted. Therefore, if there were no observations available to estimate the j-th coordinate of class k, the estimate was set to 0. Then, note that

$$\begin{aligned}\mathbb{E}\left[\widetilde{\mu}_{k,j}^2\right]&=\mathbb{E}\left[\widetilde{\mu}_{k,j}^2\mathbb{1}_{N_{k,j}>0}\right]\\&=\mathbb{E}\left[\mathbb{E}\left[\overline{\mu}_{k,j}^2\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}}\middle|N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right]\\&\leq\mathbb{E}\left[\mathbb{E}\left[\overline{\mu}_{k,j}^4\middle|N_{k,j}\right]^{\frac{1}{2}}\mathbb{E}\left[\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}}\middle|N_{k,j}\right]^{\frac{1}{2}}\mathbb{1}_{N_{k,j}>0}\right]&&\text{(using Cauchy-Schwarz)}\\&=\mathbb{E}\left[\mathbb{E}\left[\overline{\mu}_{k,j}^4\middle|N_{k,j}\right]^{\frac{1}{2}}\mathbb{P}\left(\left|\overline{\mu}_{k,j}\right|>\tau_{k,j}\middle|N_{k,j}\right)^{\frac{1}{2}}\mathbb{1}_{N_{k,j}>0}\right].&&(70)\end{aligned}$$

  On the one hand, remark that as $\overline{\mu}_{k,j}\sim\mathbb{N}\left(0,\sigma^2/N_{k,j}\right)$, then $\frac{\overline{\mu}_{k,j}^2}{\sigma^2/N_{k,j}}\sim\chi^2(1)$. Thus, the first term in the previous expectation is simplified as

$$\begin{aligned}\mathbb{E}\left[\overline{\mu}_{k,j}^4\middle|N_{k,j}\right]&=\mathbb{E}\left[\left(\frac{\overline{\mu}_{k,j}^2}{\sigma^2/N_{k,j}}\right)^2\middle|N_{k,j}\right]\frac{\sigma^4}{N_{k,j}^2}\\&=\left(\mathrm{var}\left(\frac{\overline{\mu}_{k,j}^2}{\sigma^2/N_{k,j}}\right)+\mathbb{E}\left[\frac{\overline{\mu}_{k,j}^2}{\sigma^2/N_{k,j}}\right]^2\right)\frac{\sigma^4}{N_{k,j}^2}\\&=3\frac{\sigma^4}{N_{k,j}^2}.&&\text{(using that }\tfrac{\overline{\mu}_{k,j}^2}{\sigma^2/N_{k,j}}\sim\chi^2(1)\text{)}\end{aligned}$$

  On the other hand, using Lemma A.1, as $\overline{\mu}_{k,j}\sim\mathcal{N}(0,\sigma^2/N_{k,j})$, we have that

$$\mathbb{P}\left(\left|\overline{\mu}_{k,j}\right|>\tau_{k,j}\middle|N_{k,j}\right)\leq 2e^{-\frac{\tau_{k,j}^2N_{k,j}}{2\sigma^2}}.$$

  Then, combining both results in Inequality (70), we have that

$$\begin{aligned}\mathbb{E}\left[\widetilde{\mu}_{k,j}^2\right]&\leq\mathbb{E}\left[\mathbb{E}\left[\overline{\mu}_{k,j}^4\middle|N_{k,j}\right]^{\frac{1}{2}}\mathbb{P}\left(\left|\overline{\mu}_{k,j}\right|>\tau_{k,j}\middle|N_{k,j}\right)^{\frac{1}{2}}\mathbb{1}_{N_{k,j}>0}\right]\\&\leq\sqrt{6}\mathbb{E}\left[\frac{\sigma^2}{N_{k,j}}e^{-\frac{\tau_{k,j}^2N_{k,j}}{4\sigma^2}}\mathbb{1}_{N_{k,j}>0}\right]\\&=\sqrt{6}\mathbb{E}\left[\frac{\sigma^2}{N_{k,j}}e^{-\log(d)}\mathbb{1}_{N_{k,j}>0}\right]&&\text{(using Definition (22))}\\&=\frac{\sqrt{6}\sigma^2}{d}\mathbb{E}\left[\frac{\mathbb{1}_{N_{k,j}>0}}{N_{k,j}}\right]\end{aligned}$$

Using Lemma A.3 we obtain that

$$\mathbb{E}\left[\widetilde{\mu}_{k,j}^2\right] \leq \frac{\sqrt{6}\sigma^2}{d}\frac{4}{(1-\eta)(n+1)} \tag{71}$$

- Computation of $\mathbb{E}\left[\left(\widetilde{\mu}_{k,j}-\mu_{k,j}\right)^2\right]$ knowing that $j \in \mathrm{supp}(\mu)$, i.e. $\mu_{k,j} \neq 0$.

  As in the previous computation, separate the case where there are samples to estimate the mean from the case where we are only able to impute by 0:

$$\begin{aligned}\mathbb{E}\left[\left(\widetilde{\mu}_{k,j}-\mu_{k,j}\right)^2\right] &= \mathbb{E}\left[\left(\widetilde{\mu}_{k,j}-\mu_{k,j}\right)^2 \mathbb{1}_{N_{k,j}>0}\right] + \mathbb{E}\left[\left(0-\mu_{k,j}\right)^2 \mathbb{1}_{N_{k,j}=0}\right] \\ &= \mathbb{E}\left[\left(\widetilde{\mu}_{k,j}-\mu_{k,j}\right)^2 \mathbb{1}_{N_{k,j}>0}\right] + \mu_{k,j}^2\mathbb{P}(N_{k,j}=0).\end{aligned}$$

  Note the event of not having observed any j-th coordinate of the class k was the event defined in (58). Then, this probability was already computed in (60), so we have that

$$\mathbb{E}\left[\left(\widetilde{\mu}_{k,j}-\mu_{k,j}\right)^2\right] = \mathbb{E}\left[\left(\widetilde{\mu}_{k,j}-\mu_{k,j}\right)^2 \mathbb{1}_{N_{k,j}>0}\right] + \mu_{k,j}^2\left(\frac{1+\eta}{2}\right)^n. \tag{72}$$

  Then, lets bound the first term:

$$\begin{aligned}&\mathbb{E}\left[\left(\widetilde{\mu}_{k,j}-\mu_{k,j}\right)^2 \mathbb{1}_{N_{k,j}>0}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\widetilde{\mu}_{k,j}-\mu_{k,j}\right)^2 \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\overline{\mu}_{k,j}\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}}-\mu_{k,j}\right)^2 \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\left(\overline{\mu}_{k,j}-\mu_{k,j}\right)\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}}+\mu_{k,j}\left(\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}}-1\right)\right)^2 \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\overline{\mu}_{k,j}-\mu_{k,j}\right)^2\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}} \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &\quad + 2\mu_{k,j}\mathbb{E}\left[\mathbb{E}\left[\left(\overline{\mu}_{k,j}-\mu_{k,j}\right)\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}}\left(\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}}-1\right) \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &\quad + \mu_{k,j}^2\mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}}-1\right)^2 \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\overline{\mu}_{k,j}-\mu_{k,j}\right)^2\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}} \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &\quad + 2\mu_{k,j}\mathbb{E}\left[\mathbb{E}\left[\left(\overline{\mu}_{k,j}-\mu_{k,j}\right)\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}}\mathbb{1}_{|\overline{\mu}_{k,j}|\leq\tau_{k,j}} \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &\quad + \mu_{k,j}^2\mathbb{E}\left[\mathbb{E}\left[1-\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}} \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\overline{\mu}_{k,j}-\mu_{k,j}\right)^2\mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}} \middle| N_{k,j}\right]\mathbb{1}_{N_{k,j}>0}\right] \\ &\quad + \mu_{k,j}^2\mathbb{E}\left[\mathbb{P}\left(|\overline{\mu}_{k,j}|\leq\tau_{k,j}\middle|N_{k,j}\right)\mathbb{1}_{N_{k,j}>0}\right].\end{aligned}$$

  To do so, compute the terms separately as follows:

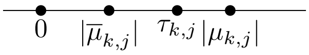


Figure 9: Assuming that $|\mu_{k,j}| \geq \tau_{k,j}$, then the event $|\overline{\mu}_{k,j}| \leq \tau_{k,j}$ implies $|\overline{\mu}_{k,j} - \mu_{k,j}| \geq |\mu_{k,j}| - \tau_{k,j}$.

– Lets compute $\mathbb{E}\left[\mathbb{E}\left[\left(\overline{\mu}_{k,j} - \mu_{k,j}\right)^2 \mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}} \middle| N_{k,j}\right] \mathbb{1}_{N_{k,j}>0}\right]$:
As we have that

$$\mathbb{E}\left[\left(\overline{\mu}_{k,j} - \mu_{k,j}\right)^2 \mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}} \middle| N_{k,j}\right] \leq \mathbb{E}\left[\left(\overline{\mu}_{k,j} - \mu_{k,j}\right)^2 \middle| N_{k,j}\right] = \frac{\sigma^2}{N_{k,j}},$$

then we have that

$$\mathbb{E}\left[\mathbb{E}\left[\left(\overline{\mu}_{k,j} - \mu_{k,j}\right)^2 \mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}} \middle| N_{k,j}\right] \mathbb{1}_{N_{k,j}>0}\right] \leq \sigma^2 \mathbb{E}\left[\frac{\mathbb{1}_{N_{k,j}>0}}{N_{k,j}}\right]. \tag{73}$$

– Lets compute $\mu_{k,j}^2 \mathbb{E}\left[\mathbb{P}\left(|\overline{\mu}_{k,j}| \leq \tau_{k,j} \middle| N_{k,j}\right) \mathbb{1}_{N_{k,j}>0}\right]$:
Begin by computing $\mu_{k,j}^2 \mathbb{P}\left(|\overline{\mu}_{k,j}| \leq \tau_{k,j} \middle| N_{k,j}\right)$. Now, we differentiate between the cases where $|\mu_{k,j}|$ is smaller than $\tau_{k,j}$ and the other cases. The first case is really simple as we are able to bound $\mu_{k,j}^2 \mathbb{P}\left(|\overline{\mu}_{k,j}| \leq \tau_{k,j} \middle| N_{k,j}\right)$ by $\tau_{k,j}^2$. Then, in the sequel suppose that $|\mu_{k,j}| \geq \tau_{k,j}$. Thus, as shown in Figure 9, the event of having $|\overline{\mu}_{k,j}| \leq \tau_{k,j}$ is equivalent to $|\mu_{k,j}| - |\overline{\mu}_{k,j}| \geq |\mu_{k,j}| - \tau_{k,j}$, and using the triangle inequality, it is included in $|\overline{\mu}_{k,j} - \mu_{k,j}| \geq |\mu_{k,j}| - \tau_{k,j}$. Then, we have that

$$\begin{aligned}
\mu_{k,j}^2 \mathbb{P}\left(|\overline{\mu}_{k,j}| \leq \tau_{k,j} \middle| N_{k,j}\right) &\leq \mu_{k,j}^2 \mathbb{P}\left(|\overline{\mu}_{k,j} - \mu_{k,j}| \geq |\mu_{k,j}| - \tau_{k,j} \middle| N_{k,j}\right) \\
&= \left(|\mu_{k,j}| - \tau_{k,j} + \tau_{k,j}\right)^2 \mathbb{P}\left(|\overline{\mu}_{k,j} - \mu_{k,j}| \geq |\mu_{k,j}| - \tau_{k,j} \middle| N_{k,j}\right) \\
&\leq 2\left(|\mu_{k,j}| - \tau_{k,j}\right)^2 \mathbb{P}\left(|\overline{\mu}_{k,j} - \mu_{k,j}| \geq |\mu_{k,j}| - \tau_{k,j} \middle| N_{k,j}\right) \\
&\quad + 2\tau_{k,j}^2 \mathbb{P}\left(|\overline{\mu}_{k,j} - \mu_{k,j}| \geq |\mu_{k,j}| - \tau_{k,j} \middle| N_{k,j}\right) \\
&\leq 4\left(|\mu_{k,j}| - \tau_{k,j}\right)^2 \exp\left(-\frac{\left(|\mu_{k,j}| - \tau_{k,j}\right)^2 N_{k,j}}{2\sigma^2}\right) + 2\tau_{k,j}^2. \\
&\hfill \text{(using Lemma A.1)}
\end{aligned}$$

To establish an upper bound for the first term, let's analyze the maximum value of the function:

$$f(x) := x \exp\left(-\frac{x N_{k,j}}{2\sigma^2}\right)$$

where $x \geq 0$. By computing the derivative, we find:

$$f'(x) = \left(1 - \frac{x N_{k,j}}{2\sigma^2}\right) \exp\left(-\frac{x N_{k,j}}{2\sigma^2}\right),$$

The maximum value occurs at $x = \frac{2\sigma^2}{N_{k,j}}$. Therefore, substituting the value of $x = \frac{2\sigma^2}{N_{k,j}}$ into the previous equation, we obtain:

$$\begin{aligned}
\mu_{k,j}^2 \mathbb{P}\left(|\overline{\mu}_{k,j}| \leq \tau_{k,j} \middle| N_{k,j}\right) &\leq 4\frac{2\sigma^2}{N_{k,j}} \exp\left(-\frac{\frac{2\sigma^2}{N_{k,j}} N_{k,j}}{2\sigma^2}\right) + 2\tau_{k,j}^2 \\
&= \frac{8\sigma^2}{N_{k,j}} \exp(-1) + 2\tau_{k,j}^2
\end{aligned}$$

Then, combining the case of $|\mu_{k,j}| \geq \tau_{k,j}$ with $|\mu_{k,j}| < \tau_{k,j}$, we have that

$$\begin{aligned}
\mu_{k,j}^2 \mathbb{P}\left(|\overline{\mu}_{k,j}| \leq \tau_{k,j} \middle| N_{k,j}\right) &\leq \max\left(\tau_{k,j}^2, \frac{8\sigma^2}{N_{k,j}} \exp\left(-1\right) + 2\tau_{k,j}^2\right) \\
&= \frac{8\sigma^2}{N_{k,j}} e^{-1} + 2\tau_{k,j}^2 \\
&= \frac{8\sigma^2}{N_{k,j}} e^{-1} + 2\frac{4\sigma^2 \log(d)}{N_{k,j}} \\
&= \frac{8\sigma^2}{N_{k,j}} \left(e^{-1} + \log(d)\right).
\end{aligned}$$

Hence, conclude that

$$\mu_{k,j}^2 \mathbb{E}\left[\mathbb{P}\left(|\overline{\mu}_{k,j}| \leq \tau_{k,j} \middle| N_{k,j}\right) \mathbb{1}_{N_{k,j}>0}\right] \leq 8\sigma^2 \left(e^{-1} + \log(d)\right) \mathbb{E}\left[\frac{\mathbb{1}_{N_{k,j}>0}}{N_{k,j}}\right]. \tag{74}$$

Retaking the decomposition and applying inequalities (73) and (74), this results in

$$\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{\mu}_{k,j} - \mu_{k,j}\right)^2 \mathbb{1}_{N_{k,j}>0}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\overline{\mu}_{k,j} - \mu_{k,j}\right)^2 \mathbb{1}_{|\overline{\mu}_{k,j}|>\tau_{k,j}} \middle| N_{k,j}\right] \mathbb{1}_{N_{k,j}>0}\right] \\
&\quad + \mu_{k,j}^2 \mathbb{E}\left[\mathbb{P}\left(|\overline{\mu}_{k,j}| \leq \tau_{k,j} \middle| N_{k,j}\right) \mathbb{1}_{N_{k,j}>0}\right] \\
&\leq \sigma^2 \mathbb{E}\left[\frac{\mathbb{1}_{N_{k,j}>0}}{N_{k,j}}\right] + 8\sigma^2 \left(e^{-1} + \log(d)\right) \mathbb{E}\left[\frac{\mathbb{1}_{N_{k,j}>0}}{N_{k,j}}\right] \\
&\leq \sigma^2 \left(1 + 8e^{-1} + 8\log(d)\right) \frac{4}{(1-\eta)(n+1)}. && \text{(using Lemma A.3)}
\end{aligned}$$

Then, from Equation (72), we have that

$$\begin{aligned}
\mathbb{E}\left[\left(\widetilde{\mu}_{k,j} - \mu_{k,j}\right)^2\right] &= \mathbb{E}\left[\left(\widetilde{\mu}_{k,j} - \mu_{k,j}\right)^2 \mathbb{1}_{N_{k,j}>0}\right] + \mu_{k,j}^2 \left(\frac{1+\eta}{2}\right)^n \\
&\leq \sigma^2 \left(1 + 8e^{-1} + 8\log(d)\right) \frac{4}{(1-\eta)(n+1)} + \|\mu\|_\infty^2 \left(\frac{1+\eta}{2}\right)^n.
\end{aligned} \tag{75}$$

Finally, conclude that

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{\mu}_{k,\mathrm{obs}(m)} - \mu_{k,\mathrm{obs}(m)}\right\|^2\right] &= \sum_{j \notin \mathrm{supp}(\mu) \cap \mathrm{obs}(m)} \mathbb{E}\left[\widetilde{\mu}_{k,j}^2\right] + \sum_{j \in \mathrm{supp}(\mu) \cap \mathrm{obs}(m)} \mathbb{E}\left[\left(\widetilde{\mu}_{k,j} - \mu_{k,j}\right)^2\right] \\
& && \text{(using Equation (69))} \\
&\leq \sum_{j \notin \mathrm{supp}(\mu) \cap \mathrm{obs}(m)} \frac{\sqrt{6}\sigma^2}{d} \frac{4}{(1-\eta)(n+1)} && \text{(using Inequality (71))} \\
&\quad + \sum_{j \in \mathrm{supp}(\mu) \cap \mathrm{obs}(m)} \left(\sigma^2 \left(1 + 8e^{-1} + 8\log(d)\right) \frac{4}{(1-\eta)(n+1)} + \|\mu\|_\infty^2 \left(\frac{1+\eta}{2}\right)^n\right) \\
& && \text{(using Inequality (75))}
\end{aligned}$$

□

C.4.2 Proof of Theorem 3.13

Proof. By Lemma C.4,

$$
\begin{aligned}
&L(\widetilde{h}) - L(h^\star) \\
&\quad\le \sum_{m\in\mathcal{M}} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\frac{1}{\sigma} \left\| -\widetilde{\mu}_{1,\mathrm{obs}(m)} + \mu_{1,\mathrm{obs}(m)} \right\| + \frac{1}{\sigma} \left\| \widetilde{\mu}_{-1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)} \right\| \right] \right) p_m \\
&\quad\le \sum_{m\in\mathcal{M}} \frac{1}{\sigma\sqrt{2\pi}} \left(\sum_{k=\pm 1} \mathbb{E}\left[\left\| \widetilde{\mu}_{k,\mathrm{obs}(m)} - \mu_{k,\mathrm{obs}(m)} \right\| \right] \right) p_m \\
&\quad\le \sum_{m\in\mathcal{M}} \frac{1}{\sigma\sqrt{2\pi}} \left(\sum_{k=\pm 1} \mathbb{E}\left[\left\| \widetilde{\mu}_{k,\mathrm{obs}(m)} - \mu_{k,\mathrm{obs}(m)} \right\|^2 \right]^{\frac{1}{2}} \right) p_m \qquad \text{(using Jensen inequality)} \\
&\quad\le \sum_{m\in\mathcal{M}} \frac{2}{\sigma\sqrt{2\pi}} \left(\sum_{j\notin \mathrm{supp}(\mu)\cap \mathrm{obs}(m)} \frac{\sqrt{6}\sigma^2}{d} \frac{4}{(1-\eta)(n+1)} \right. \\
&\qquad\quad \left. + \sum_{j\in \mathrm{supp}(\mu)\cap \mathrm{obs}(m)} \left(\sigma^2 \left(1 + 8e^{-1} + 8\log(d)\right) \frac{4}{(1-\eta)(n+1)} + \|\mu\|_\infty^2 \left(\frac{1+\eta}{2}\right)^n \right) \right)^{\frac{1}{2}} p_m \\
&\qquad\qquad \text{(using Lemma C.7)} \\
&\quad= \sum_{m\in\mathcal{M}} \frac{2}{\sigma\sqrt{2\pi}} \left(\frac{\sqrt{6}\sigma^2}{d} \frac{4}{(1-\eta)(n+1)} \mathrm{card}\left(\mathrm{supp}(\mu)^c \cap \mathrm{obs}(m)\right) \right. \\
&\qquad\quad \left. + \left(\sigma^2 \left(1 + 8e^{-1} + 8\log(d)\right) \frac{4}{(1-\eta)(n+1)} + \|\mu\|_\infty^2 \left(\frac{1+\eta}{2}\right)^n \right) \mathrm{card}\left(\mathrm{supp}(\mu) \cap \mathrm{obs}(m)\right) \right)^{\frac{1}{2}} p_m.
\end{aligned}
$$

Observe that without loss of generality thanks to Assumption 10, we can suppose that up to a reordering of the terms, the first s covariates are the support and the rest are the non-discriminant ones. Then,card $(\mathrm{supp}(\mu) \cap \mathrm{obs}(m)) = \sum_{i=1}^{s} \mathbb{1}_{m_i=0}$ and card $(\mathrm{supp}(\mu)^c \cap \mathrm{obs}(m)) = \sum_{i=s+1}^{d} \mathbb{1}_{m_i=0}$.

Therefore,

$$
\begin{aligned}
&L(\widetilde{h}) - L(h^\star) \\
&\quad \le \mathbb{E}\left[\frac{2}{\sigma\sqrt{2\pi}}\left(\frac{\sqrt{6}\sigma^2}{d}\frac{4}{(1-\eta)(n+1)}\sum_{i=s+1}^{d}\mathbb{1}_{M_i=0}\right.\right. \\
&\qquad \left.\left. + \left(\sigma^2\left(1+8e^{-1}+8\log(d)\right)\frac{4}{(1-\eta)(n+1)} + \|\mu\|_\infty^2\left(\frac{1+\eta}{2}\right)^n\right)\sum_{i=1}^{s}\mathbb{1}_{M_i=0}\right)^{\frac{1}{2}}\right] \\
&\quad \le \frac{2}{\sigma\sqrt{2\pi}}\mathbb{E}\left[\frac{\sqrt{6}\sigma^2}{d}\frac{4}{(1-\eta)(n+1)}\sum_{i=s+1}^{d}\mathbb{1}_{M_i=0}\right. \\
&\qquad \left. + \left(\sigma^2\left(1+8e^{-1}+8\log(d)\right)\frac{4}{(1-\eta)(n+1)} + \|\mu\|_\infty^2\left(\frac{1+\eta}{2}\right)^n\right)\sum_{i=1}^{s}\mathbb{1}_{M_i=0}\right]^{\frac{1}{2}} \\
&\qquad \text{(using Jensen Inequality)} \\
&\quad = \frac{2}{\sigma\sqrt{2\pi}}\left(\frac{\sqrt{6}\sigma^2}{d}\frac{4}{(1-\eta)(n+1)}(d-s)(1-\eta)\right. \\
&\qquad \left. + \left(\sigma^2\left(1+8e^{-1}+8\log(d)\right)\frac{4}{(1-\eta)(n+1)} + \|\mu\|_\infty^2\left(\frac{1+\eta}{2}\right)^n\right)s(1-\eta)\right)^{\frac{1}{2}} \\
&\quad = \frac{2}{\sqrt{2\pi}}\left(\frac{4\sqrt{6}}{(n+1)}(1-\frac{s}{d})\right. \\
&\qquad \left. + \left(1+8e^{-1}+8\log(d)\right)\frac{4s}{n+1} + \frac{\|\mu\|_\infty^2}{\sigma^2}\left(\frac{1+\eta}{2}\right)^n s(1-\eta)\right)^{\frac{1}{2}} \\
&\quad \lesssim \left(\frac{s\log(d)}{n} + \frac{\|\mu\|_\infty^2}{\sigma^2}\left(\frac{1+\eta}{2}\right)^n s(1-\eta)\right)^{\frac{1}{2}}.
\end{aligned}
$$

To conclude the proof, remark the exponential decay of the last term. Then, for n large enough, this term is negligible compared to the first one. □

C.5 Proofs of Section 3.2.5

C.5.1 Proof of Proposition 3.14

Proof. Expanding (5), we have that

$$
\begin{aligned}
&h_m^\star(X_{\mathrm{obs}(m)}) \\
&\quad = \mathrm{sign}(\mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M=m\right]) \\
&\quad = \mathrm{sign}\left(\mathbb{P}\left(Y=1\middle|X_{\mathrm{obs}(m)}, M=m\right) - \mathbb{P}\left(Y=-1\middle|X_{\mathrm{obs}(m)}, M=m\right)\right) \\
&\quad = \mathrm{sign}\left(\frac{\mathbb{P}\left(Y=1, X_{\mathrm{obs}(m)}, M=m\right)}{\mathbb{P}(X_{\mathrm{obs}(m)}, M=m)} - \frac{\mathbb{P}\left(Y=-1, X_{\mathrm{obs}(m)}, M=m\right)}{\mathbb{P}(X_{\mathrm{obs}(m)}, M=m)}\right) \\
&\quad = \mathrm{sign}\left(\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|M=m, Y=1\right)\pi_{m,1} - \mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|M=m, Y=-1\right)\pi_{m,-1}\right),
\end{aligned}
$$

with $\pi_{m,k} := \mathbb{P}(M = m, Y = k)$. Remark that the objective is to study whether

$$\mathbb{P}\left(X_{\text{obs}(m)}\middle|M = m, Y = 1\right)\pi_{m,1} > \mathbb{P}\left(X_{\text{obs}(m)}\middle|M = m, Y = -1\right)\pi_{m,-1},$$

or equivalently,

$$\frac{\mathbb{P}\left(X_{\text{obs}(m)}\middle|M = m, Y = 1\right)}{\mathbb{P}\left(X_{\text{obs}(m)}\middle|M = m, Y = -1\right)} > \frac{\pi_{m,-1}}{\pi_{m,1}}$$

or again

$$\log\left(\frac{\mathbb{P}\left(X_{\text{obs}(m)}\middle|M = m, Y = 1\right)}{\mathbb{P}\left(X_{\text{obs}(m)}\middle|M = m, Y = -1\right)}\right) > \log\left(\frac{\pi_{m,-1}}{\pi_{m,1}}\right).$$

Note that by using Assumption 14, we have $X_{\text{obs}(m)}|M = m, Y = k \sim \mathcal{N}(\mu_{m,k}, \Sigma_m)$. To conclude the proof, follow the same strategy employed in Appendix C.1.1, but use this new assumption instead of $X_{\text{obs}(m)}|Y = k \sim \mathcal{N}(\mu_{k,\text{obs}(m)}, \Sigma_{\text{obs}(m)})$ as previously done.

□

C.5.2 Proof of Proposition 3.15

Proof. W.l.o.g., we focus only on $\mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = 1\middle|Y = -1, M = m\right)$, and cover the other case by symmetry. Using Proposition 3.14, we have that $\mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = 1\middle|Y = -1, M = m\right)$ is equal to

$$\mathbb{P}\left(\left(\mu_{m,1} - \mu_{m,-1}\right)^\top \Sigma_m^{-1}\left(x_{\text{obs}(m)} - \frac{\mu_{m,1} + \mu_{m,-1}}{2}\right) - \log\left(\frac{\pi_{m,-1}}{\pi_{m,1}}\right) > 0\middle|Y = -1, M = m\right)$$

Call $N := \Sigma_m^{-\frac{1}{2}}(X_{\text{obs}(m)} - \mu_{m,-1})$ and remark that $N|Y = -1, M = m \sim \mathcal{N}(0, Id_{d-\|m\|_0})$ using Assumption 14. Fix $\gamma := \Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \mu_{m,-1})$, so that

$$\begin{aligned}\mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = 1\middle|Y = -1, M = m\right) &= \mathbb{P}\left(\gamma^\top N - \frac{1}{2}\|\gamma\|^2 > \log\left(\frac{\pi_{m,-1}}{\pi_{m,1}}\right)\middle|Y = -1, M = m\right)\\ &= \mathbb{P}\left(\frac{\gamma^\top N}{\|\gamma\|} > \frac{1}{2}\|\gamma\| + \frac{1}{\|\gamma\|}\log\left(\frac{\pi_{m,-1}}{\pi_{m,1}}\right)\middle|Y = -1, M = m\right)\\ &= \Phi\left(-\frac{1}{2}\|\gamma\| - \frac{1}{\|\gamma\|}\log\left(\frac{\pi_{m,-1}}{\pi_{m,1}}\right)\right).\end{aligned}$$

□

C.5.3 General lemmas for LDA misclassification control under Assumption 14.

Lemma C.8 ($\widehat{\mu}_m$ misclassification probability)**.** *Given a sample satisfying Assumption 14, with balanced classes over all the missing patterns, then*

$$\begin{aligned}&\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) = 1\middle|Y = -1, M = m, \mathcal{D}_n\right)\\ &= \Phi\left(\frac{\left(\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1} - \widehat{\mu}_{m,-1})\right)^\top \Sigma_m^{-\frac{1}{2}}\left(\mu_{m,-1} - \frac{\widehat{\mu}_{m,1}+\widehat{\mu}_{m,-1}}{2}\right)}{\left\|\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1} - \widehat{\mu}_{m,-1})\right\|}\right) \qquad (76)\end{aligned}$$

and symmetrically,

$$\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m, \mathcal{D}_n\right)$$
$$= \Phi\left(-\frac{\left(\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1} - \widehat{\mu}_{m,-1})\right)^\top \Sigma_m^{-\frac{1}{2}}\left(\mu_{m,1} - \frac{\widehat{\mu}_{m,1}+\widehat{\mu}_{m,-1}}{2}\right)}{\left\|\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1} - \widehat{\mu}_{m,-1})\right\|}\right) \tag{77}$$

with Φ *the standard Gaussian cumulative function.*

Proof. Given the similarity to the proof of Lemma C.2, this proof is omitted. □

Lemma C.9. *Given a dataset* $\mathcal{D}_n$ *satisfying Assumptions 14, with which we can build estimates of the mean for each class, denoted as* $\widehat{\mu}_1$ *and* $\widehat{\mu}_{-1}$*, we can state that for the classifier* $\widehat{h}_m$ *defined in Equation (20),*

$$\left|\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m\right)\right|$$
$$\leq \frac{3}{2\sqrt{2\pi}}\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,-1} - \widehat{\mu}_{m,-1})\right\| + \frac{1}{2\sqrt{2\pi}}\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \widehat{\mu}_{m,1})\right\| \tag{78}$$

and symmetrically,

$$\left|\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m\right)\right|$$
$$\leq \frac{3}{2\sqrt{2\pi}}\left\|\Sigma_m^{-\frac{1}{2}}(-\widehat{\mu}_{m,1} + \mu_{m,1})\right\| + \frac{1}{2\sqrt{2\pi}}\left\|\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,-1} - \mu_{m,-1})\right\| \tag{79}$$

Proof. To proof Inequality (78) it is sufficient to notice that

$$\left|\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m\right)\right|$$
$$= \left|\Phi\left(\frac{\left(\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1} - \widehat{\mu}_{m,-1})\right)^\top \Sigma_m^{-\frac{1}{2}}\left(\mu_{m,-1} - \frac{\widehat{\mu}_{m,1}+\widehat{\mu}_{m,-1}}{2}\right)}{\left\|\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1}) - \widehat{\mu}_{m,-1})\right\|}\right)\right.$$
$$\left. - \Phi\left(-\frac{\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \mu_{m,-1})\right\|}{2}\right)\right|, \qquad \text{(using Proposition 3.15 and Lemma C.8)}$$

and then to follow the exact same steps as those followed in the proof of Lemma C.3. We can proof similarly Inequality (79). □

Lemma C.10. *Given a dataset* $\mathcal{D}_n$ *satisfying Assumptions 14,and an estimate of the mean for each class, then the classifier* $\widehat{h}_m$ *defined in Equation (19) verifies that*

$$L(\widehat{h}) - L(h^\star)$$
$$\leq \sum_{m \in \mathcal{M}} \frac{1}{\sqrt{2\pi}}\left(\mathbb{E}\left[\left\|\Sigma_m^{-\frac{1}{2}}(-\widehat{\mu}_{m,1} + \mu_{m,1})\right\| + \left\|\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,-1} - \mu_{m,-1})\right\|\right]\right) p_m.$$

Proof.

$$
\begin{aligned}
&L(\widehat{h}) - L(h^\star)\\
&\quad= \mathbb{P}\left(\widehat{h}(X_{\mathrm{obs}(M)}, M) \neq Y\right) - \mathbb{P}\left(h^\star(X_{\mathrm{obs}(M)}, M) \neq Y\right)\\
&\quad= \sum_{m\in\mathcal{M}} \left(\mathbb{P}\left(\widehat{h}(X_{\mathrm{obs}(M)}, M) \neq Y \middle| M = m\right) - \mathbb{P}\left(h^\star(X_{\mathrm{obs}(M)}, M) \neq Y \middle| M = m\right)\right) p_m\\
&\quad= \sum_{m\in\mathcal{M}} \left(\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) \neq Y \middle| M = m\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) \neq Y \middle| M = m\right)\right) p_m \qquad \text{(using (19))}\\
&\quad= \sum_{m\in\mathcal{M}} \pi_{m,-1} \left(\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m\right)\right)\\
&\qquad\quad + \sum_{m\in\mathcal{M}} \pi_{m,1} \left(\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m\right)\right)\\
&\quad= \sum_{m\in\mathcal{M}} \pi_{m,-1} \left(\mathbb{E}\left[\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \middle| Y = -1, M = m\right)\right]\right)\\
&\qquad\quad + \sum_{m\in\mathcal{M}} \pi_{m,1} \left(\mathbb{E}\left[\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \middle| Y = 1, M = m\right)\right]\right)\\
&\quad\leq \sum_{m\in\mathcal{M}} \frac{\pi_{m,-1}}{2\sqrt{2\pi}} \left(\mathbb{E}\left[3\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,-1} - \widehat{\mu}_{m,-1})\right\| + \left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \widehat{\mu}_{m,1})\right\|\right]\right)\\
&\qquad\quad + \sum_{m\in\mathcal{M}} \frac{\pi_{m,-1}}{2\sqrt{2\pi}} \left(\mathbb{E}\left[3\left\|\Sigma_m^{-\frac{1}{2}}(-\widehat{\mu}_{m,1} + \mu_{m,1})\right\| + \left\|\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,-1} - \mu_{m,-1})\right\|\right]\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(using Lemma C.9)}\\
&\quad= \sum_{m\in\mathcal{M}} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\left\|\Sigma_m^{-\frac{1}{2}}(-\widehat{\mu}_{m,1} + \mu_{m,1})\right\| + \left\|\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,-1} - \mu_{m,-1})\right\|\right]\right) p_m.\\
&\qquad\qquad\qquad\qquad\qquad \text{(using that } \pi_{m,k} = p_m\mathbb{P}(Y = k|M = m) = \tfrac{p_m}{2}\text{)}
\end{aligned}
$$

□

C.5.4 Lemmas for Theorem 3.16

Lemma C.11. *Under Assumption 14 we have that*

$$
\mathbb{E}\left[(\widetilde{\mu}_{m,k} - \mu_{m,k})(\widetilde{\mu}_{m,k} - \mu_{m,k})^\top\right] = \mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{N_{m,k}}\right] \Sigma_m + \mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau\right) \mu_{m,k}\mu_{m,k}^\top
$$

where $\widetilde{\mu}_{m,k}$ is the estimate defined at (26).

Proof. We start by decomposing similarly as done with the sparsity in the proof of Lemma C.7:

$$\begin{aligned}
&\mathbb{E}\left[(\widetilde{\mu}_{m,k} - \mu_{m,k})(\widetilde{\mu}_{m,k} - \mu_{m,k})^\top\right] \\
&\quad = \mathbb{E}\Big[(\widehat{\mu}_{m,k}\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} - \mu_{m,k}\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} + \mu_{m,k}\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} - \mu_{m,k}) \\
&\qquad\quad (\widehat{\mu}_{m,k}\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} - \mu_{m,k}\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} + \mu_{m,k}\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} - \mu_{m,k})^\top\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}(\widehat{\mu}_{m,k} - \mu_{m,k})(\widehat{\mu}_{m,k} - \mu_{m,k})^\top + \mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\left(\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} - 1\right)(\widehat{\mu}_{m,k} - \mu_{m,k})\mu_{m,k}^\top \\
&\qquad\quad + \mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\left(\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} - 1\right)\mu_{m,k}(\widehat{\mu}_{m,k} - \mu_{m,k})^\top + \left(\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} - 1\right)^2 \mu_{m,k}\mu_{m,k}^\top\Big].
\end{aligned}$$

Note that $\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\left(\mathbb{1}_{\frac{N_{m,k}}{n}>\tau} - 1\right) = 0$. Then, we simplify the previous expression to

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}(\widehat{\mu}_{m,k} - \mu_{m,k})(\widehat{\mu}_{m,k} - \mu_{m,k})^\top + \left(1 - \mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\right)\mu_{m,k}\mu_{m,k}^\top\right] \\
&\quad = \mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}(\widehat{\mu}_{m,k} - \mu_{m,k})(\widehat{\mu}_{m,k} - \mu_{m,k})^\top\right] + \mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau\right)\mu_{m,k}\mu_{m,k}^\top.
\end{aligned}$$

Finally, remark that $\widehat{\mu}_{m,k} - \mu_{m,k}|N_{m,k} \sim \mathcal{N}(0, \Sigma_m/N_{m,k})$. Thus, we conclude noticing that

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}(\widehat{\mu}_{m,k} - \mu_{m,k})(\widehat{\mu}_{m,k} - \mu_{m,k})^\top\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}(\widehat{\mu}_{m,k} - \mu_{m,k})(\widehat{\mu}_{m,k} - \mu_{m,k})^\top\middle|N_{m,k}\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}_{N_{m,k}>n\tau}\mathbb{E}\left[(\widehat{\mu}_{m,k} - \mu_{m,k})(\widehat{\mu}_{m,k} - \mu_{m,k})^\top\middle|N_{m,k}\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right]\Sigma_m.
\end{aligned}$$

□

Lemma C.12. *Under Assumption 14 we have that*

$$\mathbb{E}\left[\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k} - \mu_{m,k})\right\|\right] \leq \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right](d - \|m\|_0) + \mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau\right)\left\|\Sigma_m^{-\frac{1}{2}}\mu_{m,k}\right\|^2\right)^{\frac{1}{2}},$$

where $\widetilde{\mu}_{m,k}$ *is the estimate defined at* (26).

Proof.

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right\|\right]\\
&\le \mathbb{E}\left[\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right\|^2\right]^{\frac{1}{2}} && \text{(using Jensen Inequality)}\\
&= \mathbb{E}\left[\operatorname{tr}\left(\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right\|^2\right)\right]^{\frac{1}{2}}\\
&= \mathbb{E}\left[\operatorname{tr}\left(\left(\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right)^\top \Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right)\right]^{\frac{1}{2}}\\
&= \mathbb{E}\left[\operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\left(\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right)^\top\right)\right]^{\frac{1}{2}}\\
&= \mathbb{E}\left[\operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})(\widetilde{\mu}_{m,k}-\mu_{m,k})^\top\Sigma_m^{-\frac{1}{2}}\right)\right]^{\frac{1}{2}}\\
&= \operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}\mathbb{E}\left[(\widetilde{\mu}_{m,k}-\mu_{m,k})(\widetilde{\mu}_{m,k}-\mu_{m,k})^\top\right]\Sigma_m^{-\frac{1}{2}}\right)^{\frac{1}{2}}\\
&= \operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}\left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right]\Sigma_m+\mathbb{P}\left(\frac{N_{m,k}}{n}\le\tau\right)\mu_{m,k}\mu_{m,k}^\top\right)\Sigma_m^{-\frac{1}{2}}\right)^{\frac{1}{2}} && \text{(using Lemma C.11)}\\
&= \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right](d-\|m\|_0)+\mathbb{P}\left(\frac{N_{m,k}}{n}\le\tau\right)\operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}\mu_{m,k}\mu_{m,k}^\top\Sigma_m^{-\frac{1}{2}}\right)\right)^{\frac{1}{2}}\\
&= \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right](d-\|m\|_0)+\mathbb{P}\left(\frac{N_{m,k}}{n}\le\tau\right)\left\|\Sigma_m^{-\frac{1}{2}}\mu_{m,k}\right\|^2\right)^{\frac{1}{2}}.
\end{aligned}
$$

□

C.5.5 Proof of Theorem 3.16

Proof. Let $A_\tau := \{m \in \{0,1\}^d | p_m < \tau\}$ be the set of missing pattern with occurrence probability smaller than τ. We are going to separate the set of missing patterns using this set.

$$
\begin{aligned}
&L(\widetilde{h}) - L(h^\star)\\
&\le \sum_{m\in\mathcal{M}}\frac{1}{\sqrt{2\pi}}\left(\mathbb{E}\left[\left\|\Sigma_m^{-\frac{1}{2}}(-\widetilde{\mu}_{m,1}+\mu_{m,1})\right\|+\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,-1}-\mu_{m,-1})\right\|\right]\right)p_m\\
&\qquad\qquad \text{(using Lemma C.10)}\\
&\le \sum_{m\in\mathcal{M}}\sum_{k=\pm1}\frac{1}{\sqrt{2\pi}}\left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right](d-\|m\|_0)+\mathbb{P}\left(\frac{N_{m,k}}{n}\le\tau\right)\left\|\Sigma_m^{-\frac{1}{2}}\mu_{m,k}\right\|^2\right)^{\frac{1}{2}}p_m\\
&\qquad\qquad \text{(using Lemma C.12)}\\
&= \sum_{m\in A_\tau}\sum_{k=\pm1}\frac{1}{\sqrt{2\pi}}\left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right](d-\|m\|_0)+\mathbb{P}\left(\frac{N_{m,k}}{n}\le\tau\right)\left\|\Sigma_m^{-\frac{1}{2}}\mu_{m,k}\right\|^2\right)^{\frac{1}{2}}p_m\mathbb{1}_{p_m<\tau}\\
&\quad+\sum_{m\notin A_\tau}\sum_{k=\pm1}\frac{1}{\sqrt{2\pi}}\left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right](d-\|m\|_0)+\mathbb{P}\left(\frac{N_{m,k}}{n}\le\tau\right)\left\|\Sigma_m^{-\frac{1}{2}}\mu_{m,k}\right\|^2\right)^{\frac{1}{2}}p_m\mathbb{1}_{p_m\ge\tau}.
\end{aligned}
$$

Lets separate into cases in order to compute this quantity:

- $\underline{m \in A_\tau}$ We have that

$$\begin{aligned}
&\sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{N_{m,k}} \right] (d - \|m\|_0) + \mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau \right) \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} \\
&\quad \leq \sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{n\tau} \right] (d - \|m\|_0) + \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} \\
&\quad \leq \sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \right] \tau + \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} \\
&\quad \leq \frac{2}{\sqrt{2\pi}} \left(1 + \frac{\|\mu_m\|^2}{\lambda_{\min}(\Sigma_m)} \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau}.
\end{aligned}$$

- $\underline{m \notin A_\tau}$ First, note that

$$\begin{aligned}
\mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau \right) &= \mathbb{P}\left(N_{m,k} \leq n\tau \mathbb{1}_{N_{m,k}>0} \right) \\
&= \mathbb{P}\left(N_{m,k}^2 \leq n^2\tau^2 \mathbb{1}_{N_{m,k}>0} \right) \\
&= \mathbb{P}\left(\frac{\mathbb{1}_{N_{m,k}>0}}{N_{m,k}^2} \geq \frac{1}{n^2\tau^2} \right) \\
&\leq \frac{\mathbb{E}\left[\frac{\mathbb{1}_{N_{m,k}>0}}{N_{m,k}^2} \right]}{\frac{1}{n^2\tau^2}} && \text{(using Markov Inequality)} \\
&\leq n^2\tau^2 \frac{8}{\frac{p_m^2}{4}(n+1)(n+2)} && \text{(using Inequality (40))} \\
&\leq \tau^2 \frac{32}{p_m^2}
\end{aligned}$$

Then, we have that

$$
\begin{aligned}
&\sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{N_{m,k}} \right] (d - \|m\|_0) + \mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau \right) \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m \geq \tau} \\
&\quad \leq \sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{N_{m,k}} \right] (d - \|m\|_0) \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m \geq \tau} \\
&\qquad + \frac{1}{\sqrt{2\pi}} \left(\mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau \right) \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m \geq \tau} \\
&\quad \leq \sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\frac{2}{\frac{p_m}{2}(n+1)} (d - \|m\|_0) \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m \geq \tau} \qquad \text{(using Inequality (38))} \\
&\qquad + \frac{1}{\sqrt{2\pi}} \left(\tau^2 \frac{32}{p_m^2} \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m \geq \tau} \\
&\quad \leq \sum_{k=\pm 1} \frac{2}{\sqrt{2\pi}} \tau \sqrt{p_m} \mathbb{1}_{p_m \geq \tau} + \frac{4}{\sqrt{\pi}} \tau \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\| \mathbb{1}_{p_m \geq \tau} \\
&\quad \leq \sum_{k=\pm 1} \frac{2}{\sqrt{2\pi}} \tau \mathbb{1}_{p_m \geq \tau} + \frac{4}{\sqrt{\pi}} \tau \frac{\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}} \mathbb{1}_{p_m \geq \tau}.
\end{aligned}
$$

To reach the conclusion, we merge both results, resulting in

$$
\begin{aligned}
&L(\widetilde{h}) - L(h^\star) \\
&\quad \leq \sum_{m \in \{0,1\}^d} \frac{2}{\sqrt{2\pi}} \left(1 + \frac{\|\mu_m\|^2}{\lambda_{\min}(\Sigma_m)} \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} + \left(\frac{4}{\sqrt{2\pi}} + \frac{8}{\sqrt{\pi}} \frac{\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}} \right) \tau \mathbb{1}_{p_m \geq \tau} \\
&\quad \leq \sum_{m \in \{0,1\}^d} \max \left(\frac{2}{\sqrt{2\pi}} \left(1 + \frac{\|\mu_m\|^2}{\lambda_{\min}(\Sigma_m)} \right)^{\frac{1}{2}}, \frac{4}{\sqrt{2\pi}} + \frac{8}{\sqrt{\pi}} \frac{\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}} \right) \tau \wedge p_m,
\end{aligned}
$$

where

$$
\frac{2}{\sqrt{2\pi}} \left(1 + \frac{\|\mu_m\|^2}{\lambda_{\min}(\Sigma_m)} \right)^{\frac{1}{2}} \leq \frac{2}{\sqrt{2\pi}} + \frac{2}{\sqrt{2\pi}} \frac{\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}} < \frac{4}{\sqrt{2\pi}} + \frac{8}{\sqrt{\pi}} \frac{\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}}.
$$

□

C.6 Proof of Section 3.2.6

C.6.1 Proof of Proposition 3.17

Proof. Expanding (5), we have that

$$\begin{aligned}
&h^\star_m(X_{\mathrm{obs}(m)})\\
&\quad= \mathrm{sign}(\mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M=m\right])\\
&\quad= \mathrm{sign}\left(\mathbb{P}\left(Y=1\middle|X_{\mathrm{obs}(m)}, M=m\right) - \mathbb{P}\left(Y=-1\middle|X_{\mathrm{obs}(m)}, M=m\right)\right)\\
&\quad= \mathrm{sign}\left(\frac{\mathbb{P}\left(Y=1, X_{\mathrm{obs}(m)}, M=m\right)}{\mathbb{P}(X_{\mathrm{obs}(m)}, M=m)} - \frac{\mathbb{P}\left(Y=-1, X_{\mathrm{obs}(m)}, M=m\right)}{\mathbb{P}(X_{\mathrm{obs}(m)}, M=m)}\right)\\
&\quad= \mathrm{sign}\left(\mathbb{P}\left(Y=1, X_{\mathrm{obs}(m)}, M=m\right) - \mathbb{P}\left(Y=-1, X_{\mathrm{obs}(m)}, M=m\right)\right)\\
&\quad= \mathrm{sign}\left(\mathbb{P}\left(M=m\middle|Y=1, X_{\mathrm{obs}(m)}\right)\mathbb{P}\left(Y=1, X_{\mathrm{obs}(m)}\right)\right.\\
&\qquad\left. -\mathbb{P}\left(M=m\middle|Y=-1, X_{\mathrm{obs}(m)}\right)\mathbb{P}\left(Y=-1, X_{\mathrm{obs}(m)}\right)\right)\\
&\quad= \mathrm{sign}\left(\mathbb{P}\left(M=m\middle|Y=1, X_{\mathrm{obs}(m)}\right)\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=1\right)\mathbb{P}(Y=1)\right.\\
&\qquad\left. -\mathbb{P}\left(M=m\middle|Y=-1, X_{\mathrm{obs}(m)}\right)\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=-1\right)\mathbb{P}(Y=-1)\right)
\end{aligned}$$

Therefore, we have to study if

$$\begin{aligned}
&\mathbb{P}\left(M=m\middle|Y=1, X_{\mathrm{obs}(m)}\right)\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=1\right)\mathbb{P}(Y=1)\\
&\quad> \mathbb{P}\left(M=m\middle|Y=-1, X_{\mathrm{obs}(m)}\right)\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=-1\right)\mathbb{P}(Y=-1),
\end{aligned}$$

which is equivalent to examining whether

$$\frac{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=1\right)}{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=-1\right)} > \frac{\mathbb{P}\left(M=m\middle|Y=-1, X_{\mathrm{obs}(m)}\right)\mathbb{P}(Y=-1)}{\mathbb{P}\left(M=m\middle|Y=1, X_{\mathrm{obs}(m)}\right)\mathbb{P}(Y=1)},$$

or again

$$\log\left(\frac{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=1\right)}{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=-1\right)}\right) > \log\left(\frac{\mathbb{P}\left(M=m\middle|Y=-1, X_{\mathrm{obs}(m)}\right)\mathbb{P}(Y=-1)}{\mathbb{P}\left(M=m\middle|Y=1, X_{\mathrm{obs}(m)}\right)\mathbb{P}(Y=1)}\right).$$

Note that by employing Lemma B.1, the distribution of $X_{\mathrm{obs}(m)}|Y=k \sim \mathcal{N}(\mu_{k,\mathrm{obs}(m)}, \Sigma_{\mathrm{obs}(m)})$. Consequently, by applying a similar approach to the one used for the MCAR case in Proposition 3.1, we can deduce that

$$\log\left(\frac{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=1\right)}{\mathbb{P}\left(X_{\mathrm{obs}(m)}\middle|Y=-1\right)}\right) = \left(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right)^\top \Sigma^{-1}_{\mathrm{obs}(m)}\left(x_{\mathrm{obs}(m)} - \frac{\mu_{1,\mathrm{obs}(m)} + \mu_{-1,\mathrm{obs}(m)}}{2}\right).$$

□

C.6.2 Proof of Corollary 3.18

Proof. Observe from Proposition 3.17 that in order to express the pattern-by-pattern Bayes classifier, we need to develop $\mathbb{P}\left(M=m\middle|Y=k, X_{\mathrm{obs}(m)}\right)$ for $k \in \{-1, 1\}$. Then, we have that

$$
\begin{aligned}
&\mathbb{P}\left(M=m \middle| Y=k, X_{\mathrm{obs}(m)}\right)\\
&\quad=\mathbb{E}\left[\mathbb{P}\left(M=m|Y=k,X\right)\middle|Y=k,X_{\mathrm{obs}(m)}\right] && \text{(using Tower property)}\\
&\quad=\mathbb{E}\left[\prod_{j=1}^{d}\mathbb{P}\left(M_j=m_j|X_j,Y=k\right)\middle|Y=k,X_{\mathrm{obs}(m)}\right] && \text{(using Assumption 15)}\\
&\quad=\prod_{j\in\mathrm{obs}(m)}\mathbb{P}\left(M_j=m_j|X_j,Y=k\right)\mathbb{E}\left[\prod_{j\notin\mathrm{obs}(m)}\mathbb{P}\left(M_j=m_j|X_j,Y=k\right)\middle|Y=k\right]\\
&&& \text{(using } X_{\mathrm{obs}(m)}\perp\!\!\!\perp X_{\mathrm{mis}(m)}\text{)}\\
&\quad=\prod_{j\in\mathrm{obs}(m)}\mathbb{P}\left(M_j=m_j|X_j,Y=k\right)\prod_{j\notin\mathrm{obs}(m)}\mathbb{E}\left[\mathbb{P}\left(M_j=m_j|X_j,Y=k\right)|Y=k\right]\\
&\quad=\prod_{j\in\mathrm{obs}(m)}\mathbb{P}\left(M_j=m_j|X_j,Y=k\right)\prod_{j\notin\mathrm{obs}(m)}\mathbb{P}\left(M_j=m_j|Y=k\right)\\
&\quad=\prod_{j\in\mathrm{obs}(m)}\mathbb{P}\left(M_j=0|X_j,Y=k\right)\prod_{j\notin\mathrm{obs}(m)}\mathbb{P}\left(M_j=1|Y=k\right). && \text{(if } j\in\mathrm{obs}(m)\text{ then } m_j=0\text{)}
\end{aligned}
$$

Observe that if X_j is observed, then using Assumption 16 we have that the value is in the permissible interval. Then, $\mathbb{P}\left(M_j=0|X_j,Y=k\right)=1$. We conclude taking the p-b-p Bayes classifier and combining the previous

$$
\begin{aligned}
h^\star_m(X_{\mathrm{obs}(m)})&=\mathrm{sign}\Bigg(\left(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)}\right)^\top\Sigma^{-1}_{\mathrm{obs}(m)}\left(x_{\mathrm{obs}(m)}-\frac{\mu_{1,\mathrm{obs}(m)}+\mu_{-1,\mathrm{obs}(m)}}{2}\right)\\
&\qquad-\log\left(\frac{\mathbb{P}\left(M=m\middle|Y=-1,X_{\mathrm{obs}(m)}\right)\pi_{-1}}{\mathbb{P}\left(M=m\middle|Y=1,X_{\mathrm{obs}(m)}\right)\pi_1}\right)\Bigg)\\
&=\mathrm{sign}\Bigg(\left(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)}\right)^\top\Sigma^{-1}_{\mathrm{obs}(m)}\left(x_{\mathrm{obs}(m)}-\frac{\mu_{1,\mathrm{obs}(m)}+\mu_{-1,\mathrm{obs}(m)}}{2}\right)\\
&\qquad-\log\left(\frac{\pi_{-1}}{\pi_1}\prod_{j\notin\mathrm{obs}(m)}\frac{\mathbb{P}\left(M_j=1|Y=-1\right)}{\mathbb{P}\left(M_j=1|Y=1\right)}\right)\Bigg)
\end{aligned}
$$

□

C.6.3 Proof of Theorem 3.19

This proof is split into two sections. Initially, we utilize Lemma C.13 to establish a bound on the difference between the *density* and its estimation (Indeed, it is not the density itself, but rather a shrinking moving window of the density). Subsequently, we demonstrate that the maximizer of this empirical density converges to the true one, which corresponds to the mean of the underlying distribution. To simplify and avoid overly complex notations, we remove the subscripts for the class and coordinate, and we conduct the proofs solely for an $X\sim\mathcal{N}(\mu,\sigma)$.

Lemma C.13. *Given a $\tau>0$, G_τ defined in* (27) *and $\widehat{G}_\tau$ defined in* (28)*, then*

$$
\mathbb{E}\left[\sup_x\left|\widehat{G}_\tau(x)-G_\tau(x)\right|\right]\leq\frac{4}{n}+4\sqrt{\frac{2\log(2n)}{n}}
$$

Proof. Recall that $G_\tau(x) := \mathbb{P}(x-\tau \leq X \leq x+\tau)$ and $\widehat{G}_\tau(x) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x-\tau\leq X_i \leq x+\tau}$. Then, $G_\tau(x) = \mathbb{E}\left[\widehat{G}_\tau(x)\right]$.We begin by employing the same approach as used in the symmetrization lemma (see Giraud (2021)). Taking a n-sample $\widetilde{D}_n := \{\widetilde{X}_i\}_{i\in[n]}$ i.i.d where $\widetilde{X}_i \sim \mathcal{N}(\mu,\sigma)$ and independent of $\mathcal{D}_n := \{X_i\}_{i\in[n]}$, and denoting $\widetilde{\mathbb{E}}$ the expectation over them, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_x |\widehat{G}_\tau(x) - G_\tau(x)|\right] &= \mathbb{E}\left[\sup_x \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x-\tau\leq X_i\leq x+\tau} - \mathbb{P}(x-\tau\leq X\leq x+\tau)\right|\right] \\
&= \mathbb{E}\left[\sup_x \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x-\tau\leq X_i\leq x+\tau} - \widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x-\tau\leq \widetilde{X}_i\leq x+\tau}\right]\right|\right] \\
&\leq \widetilde{\mathbb{E}}\mathbb{E}\left[\sup_x \left|\frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}_{x-\tau\leq X_i\leq x+\tau} - \mathbb{1}_{x-\tau\leq \widetilde{X}_i\leq x+\tau}\right)\right|\right]. \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(using Jensen inequality)}
\end{aligned}$$

Let $\epsilon_i \sim \mathcal{U}\{-1,1\}$ be a Rademacher variable, i.e. it is uniformly distributed on -1 and 1. Then, by symmetry, $\mathbb{1}_{x-\tau\leq X_i\leq x+\tau} - \mathbb{1}_{x-\tau\leq \widetilde{X}_i\leq x+\tau} \sim \epsilon_i\left(\mathbb{1}_{x-\tau\leq X_i\leq x+\tau} - \mathbb{1}_{x-\tau\leq \widetilde{X}_i\leq x+\tau}\right)$. Therefore, we have that

$$\begin{aligned}
\mathbb{E}\left[\sup_x |\widehat{G}_\tau(x) - G_\tau(x)|\right] &\leq \widetilde{\mathbb{E}}\mathbb{E}\left[\sup_x \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(\mathbb{1}_{x-\tau\leq X_i\leq x+\tau} - \mathbb{1}_{x-\tau\leq \widetilde{X}_i\leq x+\tau}\right)\right|\right] \\
&\leq \mathbb{E}\left[\sup_x \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{x-\tau\leq X_i\leq x+\tau}\right|\right] + \widetilde{\mathbb{E}}\left[\sup_x \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{x-\tau\leq \widetilde{X}_i\leq x+\tau}\right|\right] \\
&\qquad\qquad\qquad\qquad \text{(using the triangular inequality)} \\
&= 2\mathbb{E}\left[\sup_x \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{x-\tau\leq X_i\leq x+\tau}\right|\right].
\end{aligned}$$

Now, we notice that $\sum_{i=1}^n \epsilon_i \mathbb{1}_{x-\tau\leq X_i\leq x+\tau}$ only changes when there is an i such that $x = X_i \pm \tau$. Consequently, being the supremum over $[\beta_1+\tau, \beta_2-\tau]$ is equivalent to being the maximum over the set $\{X_i \pm \tau\}_{i\in[n]}$. Let's denote $\mathcal{D}_n^+ := \{X_i+\tau\}_{i\in[n]}$ and $\mathcal{D}_n^- := \{X_i-\tau\}_{i\in[n]}$. Then, we have that

$$\begin{aligned}
&\mathbb{E}\left[\sup_x \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{x-\tau\leq X_i\leq x+\tau}\right|\right] \\
&= \mathbb{E}\left[\max_{x\in\mathcal{D}_n^+\cup\mathcal{D}_n^-} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{x-\tau\leq X_i\leq x+\tau}\right|\right] \\
&\leq \mathbb{E}\left[\max_{x\in\mathcal{D}_n^+} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{x-\tau\leq X_i\leq x+\tau}\right|\right] + \mathbb{E}\left[\max_{x\in\mathcal{D}_n^-} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{x-\tau\leq X_i\leq x+\tau}\right|\right] \\
&\qquad\qquad\qquad\qquad \text{(using that } \max(a,b)\leq a+b \text{ for } a,b\geq 0) \\
&= \mathbb{E}\left[\max_{i_0\in[n]} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau}\right|\right] + \mathbb{E}\left[\max_{i_0\in[n]} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{X_{i_0}-2\tau\leq X_i\leq X_{i_0}}\right|\right].
\end{aligned}$$

We only treat the first term as the second one can be treated identically. We start by removing the term of the sum that is always 1:

$$\mathbb{E}\left[\max_{i_0\in[n]}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau}\right|\right]=\mathbb{E}\left[\max_{i_0\in[n]}\left|\frac{1}{n}\sum_{i\neq i_0}\epsilon_i\mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau}+\frac{\epsilon_{i_0}}{n}\right|\right]$$
$$\leq\frac{1}{n}\mathbb{E}\left[\max_{i_0\in[n]}\left|\sum_{i\neq i_0}\epsilon_i\mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau}\right|\right]+\frac{1}{n}.$$
(using triangular Inequality)

We denote X_{-i_0} as the set of all observations except X_{i_0} and equivalently for ϵ_{-i_0}. We define $\psi(X_{i_0},\epsilon_{-i_0},X_{-i_0}):=\sum_{i\neq i_0}\epsilon_i\mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau}$. Observe that

$$\max_{i_0\in[n]}\left|\sum_{i\neq i_0}\epsilon_i\mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau}\right|=\max_{i_0\in[n]}\left(\sum_{i\neq i_0}\epsilon_i\mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau},-\sum_{i\neq i_0}\epsilon_i\mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau}\right)$$
$$=\max_{i_0\in[n]}\left(\psi(X_{i_0},\epsilon_{-i_0},X_{-i_0}),\psi(X_{i_0},-\epsilon_{-i_0},X_{-i_0})\right).$$

Observe that for all $i\in[n]$, we have $-\epsilon_i\sim\epsilon_i\sim\mathcal{U}\{-1,1\}$. Therefore, $\psi(X_{i_0},\epsilon_{-i_0},X_{-i_0})\sim\psi(X_{i_0},-\epsilon_{-i_0},X_{-i_0})$.

Now, our objective is to apply Lemma A.6. To do so, note that on one hand, using symmetry, for all $\imath_0\in[n]$, we have $\mathbb{E}[\psi(X_{i_0},\epsilon_{-i_0},X_{-i_0})]=\mathbb{E}[\psi(X_{i_0},-\epsilon_{-i_0},X_{-i_0})]=0$. On the other hand, we have that, for all $\lambda\in\mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda\psi(X_{i_0},\epsilon_{-i_0},X_{-i_0})\right)\right]=\mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{i\neq i_0}\epsilon_i\mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau}\right)\middle|X_{i_0}\right]\right]$$
$$=\mathbb{E}\left[\prod_{i\neq i_0}\mathbb{E}\left[\exp\left(\lambda\epsilon_i\mathbb{1}_{X_{i_0}\leq X_i\leq X_{i_0}+2\tau}\right)\middle|X_{i_0}\right]\right]$$
$$\leq\mathbb{E}\left[\prod_{i\neq i_0}\exp\left(\frac{\lambda^2}{2}\right)\right]\qquad\text{(using Hoeffding's inequality(lemma A.2))}$$
$$=\exp\left(\frac{(n-1)\lambda^2}{2}\right).$$

Therefore, $\psi(X_{i_0},\epsilon_{-i_0},X_{-i_0})$ is a sub-Gaussian random variable with variance factor $n-1$. Then, using Lemma A.6 we have that:

$$\mathbb{E}\left[\max_{i_0\in[n]}\left(\psi(X_{i_0},\epsilon_{-i_0},X_{-i_0}),\psi(X_{i_0},-\epsilon_{-i_0},X_{-i_0})\right)\right]\leq\sqrt{2(n-1)\log(2(n-1))}.$$

By combining the previous results, we can conclude that

$$
\begin{aligned}
&\mathbb{E}\left[\sup_x |\widehat{G}_\tau(x) - G_\tau(x)|\right] \\
&\quad\le 2\mathbb{E}\left[\sup_x \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{x-\tau\le X_i\le x+\tau}\right|\right] \\
&\quad\le 2\mathbb{E}\left[\max_{i_0\in[n]} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{X_{i_0}\le X_i\le X_{i_0}+2\tau}\right|\right] + 2\mathbb{E}\left[\max_{i_0\in[n]} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{X_{i_0}-2\tau\le X_i\le X_{i_0}}\right|\right] \\
&\quad\le 2\left(\frac{1}{n}\mathbb{E}\left[\max_{i_0\in[n]} \left|\sum_{i\ne i_0} \epsilon_i \mathbb{1}_{X_{i_0}\le X_i\le X_{i_0}+2\tau}\right|\right] + \frac{1}{n}\right) + 2\left(\frac{1}{n}\mathbb{E}\left[\max_{i_0\in[n]} \left|\sum_{i\ne i_0} \epsilon_i \mathbb{1}_{X_{i_0}-2\tau\le X_i\le X_{i_0}}\right|\right] + \frac{1}{n}\right) \\
&\quad\le \frac{4}{n} + \frac{4}{n}\sqrt{2(n-1)\log(2(n-1))} \\
&\quad\le \frac{4}{n} + 4\sqrt{\frac{2\log(2n)}{n}}
\end{aligned}
$$

□

Proof of Theorem 3.19. First, observe that $G_{\tau_n}(\mu) - G_{\tau_n}(\widehat{\mu})$ is always positive as $\mu = \text{argmax} G_{\tau_n}$ for every τ_n.

We assume that $\widehat{\mu} < \mu$ (the other case is completely symmetric). We have that

$$
\begin{aligned}
G_{\tau_n}(\widehat{\mu}) &= \mathbb{P}(\widehat{\mu} - \tau_n \le X \le \widehat{\mu} + \tau_n) = \Phi\left(\frac{\widehat{\mu} + \tau_n - \mu}{\sigma}\right) - \Phi\left(\frac{\widehat{\mu} - \tau_n - \mu}{\sigma}\right), \\
G_{\tau_n}(\mu) &= \mathbb{P}(\mu - \tau_n \le X \le \mu + \tau_n) = \Phi\left(\frac{\tau_n}{\sigma}\right) - \Phi\left(\frac{-\tau_n}{\sigma}\right),
\end{aligned}
$$

where Φ is the cumulative distribution function of a standard Gaussian variable. Then, we have that

$$
G_{\tau_n}(\mu) - G_{\tau_n}(\widehat{\mu}) = \Phi\left(\frac{\tau_n}{\sigma}\right) - \Phi\left(\frac{\widehat{\mu} + \tau_n - \mu}{\sigma}\right) - \Phi\left(\frac{-\tau_n}{\sigma}\right) + \Phi\left(\frac{\widehat{\mu} - \tau_n - \mu}{\sigma}\right).
$$

In the following, for the sake of simplicity, we use τ instead of τ_n.

Now, by employing the integral remainder of the Taylor expansion, we find that

$$
\Phi\left(\frac{\widehat{\mu} + \tau - \mu}{\sigma}\right) = \Phi\left(\frac{\tau}{\sigma}\right) + \left(\frac{\widehat{\mu} - \mu}{\sigma}\right)\frac{\exp\left(-\frac{\tau^2}{2\sigma^2}\right)}{\sqrt{2\pi}} + \int_{\frac{\tau}{\sigma}}^{\frac{\widehat{\mu}+\tau-\mu}{\sigma}} \frac{\widehat{\mu} - \mu}{\sigma}\frac{-t}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)\mathrm{dt},
$$

which is equal to

$$
\Phi\left(\frac{\tau}{\sigma}\right) - \Phi\left(\frac{\widehat{\mu} + \tau - \mu}{\sigma}\right) = -\left(\frac{\widehat{\mu} - \mu}{\sigma}\right)\frac{\exp\left(-\frac{\tau^2}{2\sigma^2}\right)}{\sqrt{2\pi}} + \int_{\frac{\tau}{\sigma}}^{\frac{\widehat{\mu}+\tau-\mu}{\sigma}} \frac{\widehat{\mu} - \mu}{\sigma}\frac{t}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)\mathrm{dt}.
$$

Similarly, we develop

$$
\Phi\left(\frac{\widehat{\mu} - \tau - \mu}{\sigma}\right) = \Phi\left(-\frac{\tau}{\sigma}\right) + \left(\frac{\widehat{\mu} - \mu}{\sigma}\right)\frac{\exp\left(-\frac{\tau^2}{2\sigma^2}\right)}{\sqrt{2\pi}} + \int_{-\frac{\tau}{\sigma}}^{\frac{\widehat{\mu}-\tau-\mu}{\sigma}} \frac{\widehat{\mu} - \mu}{\sigma}\frac{-t}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)\mathrm{dt},
$$

or equivalently,

$$\Phi\left(\frac{\widehat{\mu}-\tau-\mu}{\sigma}\right)-\Phi\left(-\frac{\tau}{\sigma}\right)=\left(\frac{\widehat{\mu}-\mu}{\sigma}\right)\frac{\exp\left(-\frac{\tau^2}{2\sigma^2}\right)}{\sqrt{2\pi}}+\int_{-\frac{\tau}{\sigma}}^{\frac{\widehat{\mu}-\tau-\mu}{\sigma}}\frac{\widehat{\mu}-\mu}{\sigma}\frac{-t}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)\mathrm{dt}.$$

By combining both developments, we have that

$$\begin{aligned}
G_\tau(\mu)-G_\tau(\widehat{\mu}) &= \left(\frac{\widehat{\mu}-\mu}{\sigma\sqrt{2\pi}}\right)\left(\int_{\frac{\tau}{\sigma}}^{\frac{\widehat{\mu}+\tau-\mu}{\sigma}} t\exp\left(-\frac{t^2}{2}\right)\mathrm{dt}+\int_{-\frac{\tau}{\sigma}}^{\frac{\widehat{\mu}-\tau-\mu}{\sigma}} -t\exp\left(-\frac{t^2}{2}\right)\mathrm{dt}\right)\\
&= \left(\frac{\widehat{\mu}-\mu}{\sigma\sqrt{2\pi}}\right)\left(-\exp\left(-\frac{(\widehat{\mu}+\tau-\mu)^2}{2\sigma^2}\right)+\exp\left(-\frac{\tau^2}{2\sigma^2}\right)\right.\\
&\quad\left.+\exp\left(-\frac{(\widehat{\mu}-\tau-\mu)^2}{2\sigma^2}\right)-\exp\left(-\frac{\tau^2}{2\sigma^2}\right)\right)\\
&= \left(\frac{-\widehat{\mu}+\mu}{\sigma\sqrt{2\pi}}\right)\left(\exp\left(-\frac{(\widehat{\mu}+\tau-\mu)^2}{2\sigma^2}\right)-\exp\left(-\frac{(\widehat{\mu}-\tau-\mu)^2}{2\sigma^2}\right)\right).
\end{aligned}$$

Recall that we have supposed that $\widehat{\mu}<\mu$. Then, observe that $\widehat{\mu}-\mu\in[-(\mu-\beta_1),0]\subset[-A,0]$, where $A=\max(\mu-\beta_1,\beta_2-\mu)$. Moreover, we can suppose that $\widehat{\mu}-\mu\in[-A,-\tau]$, because if $\widehat{\mu}-\mu\in[-\tau,0]$ then we have $0\leq\mu-\widehat{\mu}\leq\tau=\sqrt[4]{\log(n)/n}$, which what we are looking for.

We denote

$$\delta(x):=-\exp\left(-\frac{(x-\tau)^2}{2\sigma^2}\right)+\exp\left(-\frac{(x+\tau)^2}{2\sigma^2}\right).$$

Its derivative is given by

$$\delta'(x)=\frac{x-\tau}{\sigma^2}\exp\left(-\frac{(x-\tau)^2}{2\sigma^2}\right)-\frac{x+\tau}{\sigma^2}\exp\left(-\frac{(x+\tau)^2}{2\sigma^2}\right).$$

Note that if $x\in[-A,-\tau]$, then $(x-\tau)^2>(x+\tau)^2$, and $\exp\left(-\frac{(x-\tau)^2}{2\sigma^2}\right)<\exp\left(-\frac{(x+\tau)^2}{2\sigma^2}\right)$. We get that

$$\delta'(x)\geq\frac{x+\tau}{\sigma^2}\left(\exp\left(-\frac{(x-\tau)^2}{2\sigma^2}\right)-\exp\left(-\frac{(x+\tau)^2}{2\sigma^2}\right)\right)\geq 0.$$

Therefore, the minimum is reached at $x=-A$. Then,

$$\begin{aligned}
\delta(x) &\geq -\exp\left(-\frac{(A-\tau)^2}{2\sigma^2}\right)+\exp\left(-\frac{(A+\tau)^2}{2\sigma^2}\right)\\
&= 2\tau\frac{t^\star}{\sigma^2}\exp\left(-\frac{t^{\star 2}}{2\sigma^2}\right) && \text{(for } t^\star\in[A-\tau,A+\tau])\\
&\geq 2\tau\frac{A-\tau}{\sigma^2}\exp\left(-\frac{(A+\tau)^2}{2\sigma^2}\right)\\
&\geq 2\tau\frac{A-\frac{A}{2}}{\sigma^2}\exp\left(-\frac{(A+\frac{A}{2})^2}{2\sigma^2}\right)\\
&= 2\tau\frac{A}{2\sigma^2}\exp\left(-\frac{9A^2}{8\sigma^2}\right).
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}G_\tau(\mu) - G_\tau(\widehat{\mu}) &= \left(\frac{\widehat{\mu}-\mu}{\sigma\sqrt{2\pi}}\right)\left(-\exp\left(-\frac{(\widehat{\mu}+\tau-\mu)^2}{2\sigma^2}\right)+\exp\left(-\frac{(\widehat{\mu}-\tau-\mu)^2}{2\sigma^2}\right)\right)\\ &\geq \left(\frac{-\widehat{\mu}+\mu}{\sigma\sqrt{2\pi}}\right)2\tau\frac{A}{2\sigma^2}\exp\left(-\frac{9A^2}{8\sigma^2}\right).\end{aligned}$$

Observe that

$$\begin{aligned}0 \leq G_\tau(\mu) - G_\tau(\widehat{\mu}) &= G_\tau(\mu) - \widehat{G}_\tau(\mu) + \widehat{G}_\tau(\mu) - \widehat{G}_\tau(\widehat{\mu}) + \widehat{G}_\tau(\widehat{\mu}) - G_\tau(\widehat{\mu})\\ &\leq G_\tau(\mu) - \widehat{G}_\tau(\mu) + \widehat{G}_\tau(\widehat{\mu}) - G_\tau(\widehat{\mu}). \quad \text{(using that } \widehat{\mu} \text{ is the maximizer of } \widehat{G}_\tau)\end{aligned}$$

To conclude, we apply Lemma C.13 so that

$$\begin{aligned}2\left(\frac{4}{n}+4\sqrt{\frac{2\log(2n)}{n}}\right) &\geq 2\mathbb{E}\left[\sup_x |\widehat{G}_\tau(x) - G_\tau(x)|\right]\\ &\geq \mathbb{E}\left[G_\tau(\mu) - \widehat{G}_\tau(\mu) + \widehat{G}_\tau(\widehat{\mu}) - G_\tau(\widehat{\mu})\right]\\ &\geq \left(\frac{\mathbb{E}\left[|\widehat{\mu}-\mu|\right]}{\sigma\sqrt{2\pi}}\right)2\tau\frac{A}{2\sigma^2}\exp\left(-\frac{9A^2}{8\sigma^2}\right).\end{aligned}$$

Hence, we obtain

$$\mathbb{E}\left[|\widehat{\mu}-\mu|\right] \lesssim \sqrt[4]{\frac{\log(n)}{n}}\frac{\sigma^3}{A}\exp\left(\frac{9A^2}{8\sigma^2}\right).$$

□

D Proofs of Section 3.3 (Logistic Model)

D.1 Proof of Proposition 3.20

Proof. We show a counterexample to proof the statement. Let $m \in \{0,1\}^d$,

$$\begin{aligned}\eta_m(X_{\mathrm{obs}(m)}) &= \mathbb{P}\left(Y=1|X_{\mathrm{obs}(m)}, M=m\right)\\ &= \mathbb{P}\left(Y=1|X_{\mathrm{obs}(m)}\right) && \text{(using Assumption 2)}\\ &= \mathbb{E}\left[\mathbb{P}\left(Y=1|X\right)|X_{\mathrm{obs}(m)}\right] && \text{(using tower property)}\\ &= \mathbb{E}\left[\frac{1}{1+\exp(-X^\top\beta)}|X_{\mathrm{obs}(m)}\right] && \text{(using Assumption 18)}\end{aligned}$$

On the one hand, supposing that the logistic model remains valid on the observed data only, we have that for every $m \in \{0,1\}^d$,

$$\exists \beta^\star_m \in \mathbb{R}^{d-\|m\|_0} \text{ such that } \eta_m(X_{\mathrm{obs}(m)}) = \frac{1}{1+e^{-X_{\mathrm{obs}(m)}^\top \beta^\star_m}}.$$

Thus, joining the two expressions, we have that given an m, there exists a $\beta^\star_m$ such that

$$\begin{aligned}
\frac{1}{1+\exp(-X_{\mathrm{obs}(m)}^\top\beta^\star_m)} &= \mathbb{E}\left[\frac{1}{1+\exp(-X^\top\beta)}|X_{\mathrm{obs}(m)}\right] \\
&\geq \frac{1}{\mathbb{E}\left[1+\exp\left(-X^\top\beta\right)|X_{\mathrm{obs}(m)}\right]} && \text{(using Jensen Inequality)} \\
&= \frac{1}{1+\mathbb{E}\left[\exp\left(-X_{\mathrm{obs}(m)}^\top\beta_{\mathrm{obs}(m)} - X_{\mathrm{mis}(m)}^\top\beta_{\mathrm{mis}(m)}\right)\Big|X_{\mathrm{obs}(m)}\right]} \\
&= \frac{1}{1+\exp\left(-X_{\mathrm{obs}(m)}^\top\beta_{\mathrm{obs}(m)}\right)\mathbb{E}\left[\exp\left(-X_{\mathrm{mis}(m)}^\top\beta_{\mathrm{mis}(m)}\right)\Big|X_{\mathrm{obs}(m)}\right]},
\end{aligned}$$

which is equivalent to

$$\exp\left(-X_{\mathrm{obs}(m)}^\top\beta_{\mathrm{obs}(m)}\right)\mathbb{E}\left[\exp\left(-X_{\mathrm{mis}(m)}^\top\beta_{\mathrm{mis}(m)}\right)\Big|X_{\mathrm{obs}(m)}\right] \geq \exp(-X_{\mathrm{obs}(m)}^\top\beta^\star_m),$$

or again

$$\exp\left(-X_{\mathrm{obs}(m)}^\top\left(\beta^\star_m - \beta_{\mathrm{obs}(m)}\right)\right) \leq \mathbb{E}\left[\exp\left(-X_{\mathrm{mis}(m)}^\top\beta_{\mathrm{mis}(m)}\right)\Big|X_{\mathrm{obs}(m)}\right]. \tag{80}$$

We note that, at this stage, the only assumptions used are 2 and 18.

Now show a counterexample that satisfies the assumptions made: suppose $d=2, m=(0,1), X_2 = \exp(X_1)$ and that $\beta_2 \neq 0$, then we have that $X_{\mathrm{obs}(0,1)} = X_1$ and that $X_{\mathrm{mis}(0,1)} = X_2 = \exp(X_1)$. Hence, retaking Expression (80) we have that there exists a $\beta^\star_{(0,1)} \in \mathbb{R}$ such that

$$\begin{aligned}
\exp\left(-X_1\left(\beta^\star_{(0,1)} - \beta_1\right)\right) &\leq \mathbb{E}\left[\exp\left(-X_2\beta_2\right)|X_1\right] \\
&= \exp\left(-\exp(X_1)\beta_2\right),
\end{aligned}$$

or equivalently

$$\exp(X_1)\beta_2 \leq X_1\left(\beta^\star_{(0,1)} - \beta_1\right)$$

for any $X_1 \in \mathbb{R}$. Obviously, such $\beta^\star_{(0,1)}$ does not exist. □

D.2 Proof of Proposition 3.21

proof 1. First, prove that $\beta^\star_m = \beta_{\mathrm{obs}(m)}$. Following the same steps as in the proof of Proposition 3.20, we get the same expression as in Equation (80). Hence,

$$\begin{aligned}
\exp\left(-X_{\mathrm{obs}(m)}^\top\left(\beta^\star_m - \beta_{\mathrm{obs}(m)}\right)\right) &\leq \mathbb{E}\left[\exp\left(-X_{\mathrm{mis}(m)}^\top\beta_{\mathrm{mis}(m)}\right)\Big|X_{\mathrm{obs}(m)}\right] \\
&= \mathbb{E}\left[\exp\left(-X_{\mathrm{mis}(m)}^\top\beta_{\mathrm{mis}(m)}\right)\right]. && \text{(using Assumption 5)}
\end{aligned}$$

Then, we bound the left expression by a constant, but as $X_{\mathrm{obs}(m)}$ is a Gaussian vector, that is only possible if $\beta^\star_m = \beta_{\mathrm{obs}(m)}$.

For the sake of simplicity, we take the example of $X \sim \mathcal{N}((0,\mu_2), I_2)$ (without fixing μ_2 yet) and $\beta = (1,1)$. As shown before, $\beta^\star_{(0,1)} = 1$. Then, supposing that the logistic assumption is preserved on the observed covariates, we have that

$$\begin{aligned}
\frac{1}{1+\exp(-X_1\beta^\star_{(0,1)})} &= \frac{1}{1+\exp(-X_1)} \\
&= \mathbb{E}\left[\frac{1}{1+\exp(-X^\top\beta)}|X_1\right] \\
&\geq \frac{1}{\mathbb{E}\left[1+\exp\left(-X^\top\beta\right)|X_1\right]} && \text{(using Jensen Inequality)} \\
&= \frac{1}{1+\exp(-X_1)\mathbb{E}\left[\exp(-X_2)\right]} && \text{(using Assumption 5)} \\
&= \frac{1}{1+\exp(-X_1)\exp\left(-\mu_2+\frac{1}{2}\right)}, && \text{(using } \mathbb{E}\left[\exp(tX)\right] = \exp\left(\mu t + \tfrac{\sigma^2 t^2}{2}\right)\text{)}
\end{aligned}$$

which is satisfied if

$$\exp\left(-\mu_2 + \frac{1}{2}\right) \geq 1,$$

or again,

$$-\mu_2 + \frac{1}{2} \geq 0.$$

Therefore, assuming $\mu_2 \leq 1/2$ leads to a contradiction. □

proof 2. The proof relies on a simple counterexample. Under Assumption 18, consider the simple case of a 2-dimensional input (X_1, X_2) where the first component X_1 is always observed, and only the second component X_2 can be MCAR. Moreover, assume that $X = (X_1, X_2) \sim \mathcal{N}(0, I_2)$, and fix $(\beta_1, \beta_2) = (1,1)$ (β_1 could be chosen arbitrarily). According to the pattern-by-pattern Bayes classifier decomposition, we compute η for $m = (0,1)$, (thus $X_{\mathrm{obs}(m)} = X_1$),

$$\begin{aligned}
\eta_{(0,1)}(X_1) &= \mathbb{P}\left(Y=1|X_1, M=(0,1)\right) \\
&= \mathbb{P}\left(Y=1|X_1\right) && \text{(using Assumption 2)} \\
&= \mathbb{E}\left[\mathbb{P}\left(Y=1|X_1,X_2\right)|X_1\right] && \text{(using tower property)} \\
&= \mathbb{E}\left[\frac{1}{1+\exp(-X_2-X_1)}|X_1\right].
\end{aligned}$$

Assume that the logistic model is preserved when the first variable is only observed, i.e.,

$$\exists \beta^\star_{(0,1)}, \qquad \eta_{(0,1)}(X_1) = \frac{1}{1+\exp(-\beta^\star_{(0,1)}X_1)}.$$

That is,

$$\exists \beta^\star_{(0,1)}, \qquad \beta^\star_{(0,1)} = -\frac{1}{X_1}\ln\left(\frac{1}{\eta_{(0,1)}(X_1)} - 1\right).$$

Or equivalently $x_1 \mapsto -\frac{1}{x_1}\ln\left(\frac{1}{\eta_{(0,1)}(x_1)} - 1\right)$ is constant. We plot this in Figure 10, by doing a numerical simulation, we estimate the function $\eta_{(0,1)}(x_1)$ for different values of x_1.

□

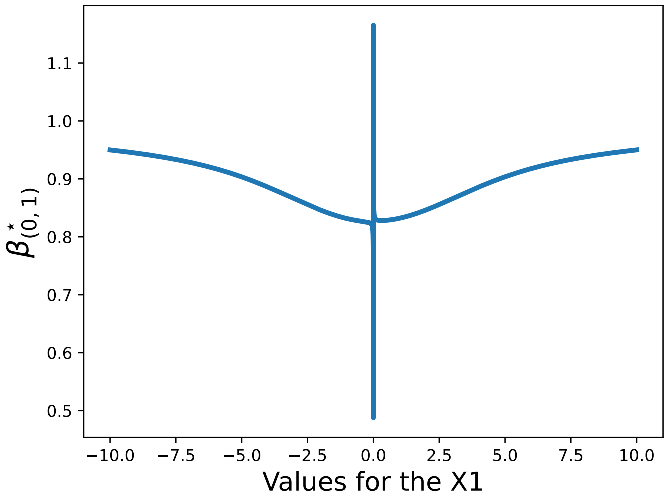


Figure 10: (MCAR with independent Gaussian covariates) Counterexample for the logistic model in pattern-by-pattern Bayes predictors: we plot the function $x_1 \mapsto -\frac{1}{x_1} \ln\left(\frac{1}{\eta_{(0,1)}(x_1)} - 1\right)$, which is not constant, showing that the logistic model cannot be preserved when particularized to a specific missing pattern.

E Proofs of Section 3.4(Perceptron)

E.1 Proof of Lemma 3.23

Proof. The goal of this proof is to find an $a \in \mathbb{R}^d, b \in \mathbb{R}$ such that for all $y \in C_1, \langle a, y\rangle + b \leq 0$ and for all $y \in C_2, \langle a, y\rangle + b \geq 0$. To do so, we start by defining the distance between two sets as $d(C_1, C_2) := \inf_{x_1 \in C_1, x_2 \in C_2} \|x_1 - x_2\|$. Using the compacity of both sets and the continuity of the norm, the existence of $x_1 \in C_1, x_2 \in C_2$ that reach the infimum is guaranteed. We notice that $x_1 \neq x_2$ because $C_1 \cap C_2 = \emptyset$. In addition, using that C_2 is convex and closed, we can take $x_2 = \Pi_{C_2}(x_1)$ because

$$\begin{aligned} \|x_1 - x_2\| &\leq \|x_1 - \Pi_{C_2}(x_1)\| && \text{(using the definition of infimum and } \Pi_{C_2}(x_1) \in C_2) \\ &\leq \|x_1 - x_2\|\,. && \text{(using the projection on a convex closed set)} \end{aligned}$$

If we use the same argument on the convex closed set C_1, we get that $x_1 = \Pi_{C_1}(\Pi_{C_2}(x_1))$, because

$$\|x_1 - \Pi_{C_2}(x_1)\| \leq \|\Pi_{C_1}(\Pi_{C_2}(x_1)) - \Pi_{C_2}(x_1)\| \leq \|x_1 - \Pi_{C_2}(x_1)\|\,.$$

Using a characterization of the projection on convex closed sets, we have that

$$\forall y \in C_2, \qquad \langle \Pi_{C_2}(x_1) - x_1, y - \Pi_{C_2}(x_1)\rangle = \langle \Pi_{C_2}(x_1) - x_1, y\rangle + \langle \Pi_{C_2}(x_1) - x_1, -\Pi_{C_2}(x_1)\rangle \geq 0.$$

Therefore, if we denote $a := \Pi_{C_2}(x_1) - x_1, b := \langle \Pi_{C_2}(x_1) - x_1, -\Pi_{C_2}(x_1)\rangle$ we have that for all $y \in C_2, \langle a, y\rangle + b \geq 0$. It remains for us to verify that for all $y \in C_1, \langle a, y\rangle + b \leq 0$. For this, using the same characterization on the projection over C_1, we have

$$\forall y \in C_1, \qquad \langle \Pi_{C_1}(\Pi_{C_2}(x_1)) - \Pi_{C_2}(x_1), y - \Pi_{C_1}(\Pi_{C_2}(x_1))\rangle = \langle x_1 - \Pi_{C_2}(x_1), y - x_1\rangle \geq 0. \quad (81)$$

Then,

$$\begin{aligned}\forall y \in C_1, \qquad \langle a, y\rangle + b &= \langle \Pi_{C_2}(x_1) - x_1, y\rangle + \langle \Pi_{C_2}(x_1) - x_1, -\Pi_{C_2}(x_1)\rangle \\ &= \langle \Pi_{C_2}(x_1) - x_1, y - \Pi_{C_2}(x_1)\rangle \\ &= \langle \Pi_{C_2}(x_1) - x_1, y - x_1\rangle + \langle \Pi_{C_2}(x_1) - x_1, x_1 - \Pi_{C_2}(x_1)\rangle \\ &= -\langle x_1 - \Pi_{C_2}(x_1), y - x_1\rangle - \|x_1 - \Pi_{C_2}(x_1)\|^2 \leq 0. \qquad \text{(using (81))}\end{aligned}$$

□

E.2 Proof of Lemma 3.25

Proof. On the one hand, if $\|C_1 - C_2\|_p \leq R_1 + R_2$, then $B_1 \cap B_2 \neq \emptyset$. For example, $x \in B_1 \cap B_2$ for $x := C_1 + \frac{R_1}{R_1+R_2}(C_2 - C_1)$ because

$$\|x - C_1\|_p = \left\|\frac{R_1}{R_1 + R_2}(C_2 - C_1)\right\|_p \leq \frac{R_1}{R_1 + R_2}(R_1 + R_2) = R_1$$

then $x \in B_1$ and

$$\|x - C_2\|_p = \left\|\frac{R_2}{R_1 + R_2}(C_2 - C_1)\right\|_p \leq \frac{R_2}{R_1 + R_2}(R_1 + R_2) = R_2$$

so $x \in B_2$.

On the other hand, if there exist an x such that $x \in B_1 \cap B_2 \neq \emptyset$, then $\|x - C_1\|_p \leq R_1$ and $\|x - C_2\|_p \leq R_2$. Using the triangle inequality,

$$\|(C_1 - C_2)\|_p \leq \|(C_1 - x)\|_p + \|(x - C_2)\|_p \leq R_1 + R_2$$

□

E.3 Proof of Proposition 3.24

Proof of Proposition 3.24. In order to study the separability of the two balls after projection through the missing pattern, we need to study the probability that the sum of radii is still smaller than the distance between the two centers after projection as shown in Lemma 3.25. As seen in (36), since $R := R_1 = R_2$, this probability corresponds to

$$\mathbb{P}\left(R < \frac{1}{2}\|(1 - M) \odot (C_1 - C_2)\|_p\right).$$

Using Assumption 21, we have that

$$\mathbb{P}\left(R < \frac{1}{2}\|(1 - M) \odot (C_1 - C_2)\|_p |M, C_1, C_2\right) = \frac{\|(1 - M) \odot (C_1 - C_2)\|_p}{\|(C_1 - C_2)\|_p}.$$

Therefore, if we define $\mathcal{M}_s = \left\{m \in \{0,1\}^d, \|m\|_0 = s\right\}$,

$$\begin{aligned}
\mathbb{P}\left(R < \frac{1}{2}\left\|(1-M)\odot(C_1 - C_2)\right\|_p\right) &= \mathbb{E}\left[\frac{\left\|(1-M)\odot(C_1-C_2)\right\|_p}{\left\|(C_1-C_2)\right\|_p}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\left\|(1-M)\odot(C_1-C_2)\right\|_p}{\left\|(C_1-C_2)\right\|_p}\middle| C_1, C_2\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sqrt[p]{\frac{\sum_{j=1}^d (1-M_j)(C_{1j}-C_{2j})^p}{\sum_{j=1}^d (C_{1j}-C_{2j})^p}}\middle| C_1, C_2\right]\right] \\
&= \mathbb{E}\left[\sum_{m\in\mathcal{M}_s}\frac{1}{\binom{d}{s}}\sqrt[p]{\frac{\sum_{j,m_j=0}(C_{1j}-C_{2j})^p}{\sum_{j=1}^d (C_{1j}-C_{2j})^p}}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(using } M\sim\mathcal{U}(\mathcal{M}_s)\text{)} \\
&= \mathbb{E}\left[\sqrt[p]{\frac{\sum_{j=1}^{d-s}(C_{1j}-C_{2j})^p}{\sum_{j=1}^d (C_{1j}-C_{2j})^p}}\right]
\end{aligned}$$

after having reordered the terms using the exchangeability of the $(C_1 - C_2)_j$ (Assumption 19). Decomposing,

$$\begin{aligned}
\frac{\sum_{j=1}^{d-s}(C_{1j}-C_{2j})^p}{\sum_{j=1}^d (C_{1j}-C_{2j})^p} &= \frac{\frac{1}{d-s}\sum_{j=1}^{d-s}(C_{1j}-C_{2j})^p}{\frac{1}{d-s}\sum_{j=1}^{d-s}(C_{1j}-C_{2j})^p + \frac{s}{d-s}\frac{1}{s}\sum_{j=d-s+1}^{d}(C_{1j}-C_{2j})^p} \\
&= \frac{1}{1+\frac{\frac{s}{d-s}\frac{1}{s}\sum_{j=d-s+1}^{d}(C_{1j}-C_{2j})^p}{\frac{1}{d-s}\sum_{j=1}^{d-s}(C_{1j}-C_{2j})^p}}.
\end{aligned}$$

As d goes to infinity, we assume that the number of missing values s goes to infinity. Otherwise, if s is bounded, then $\gamma = \lim_{d\to\infty}\frac{s}{d} = 0$ and $\frac{s}{d-s}\frac{1}{s}\sum_{j=d-s+1}^{d}(C_{1j}-C_{2j})^p \xrightarrow{d\to\infty} 0$, so we would have the result using that

$$\mathbb{P}\left(R < \frac{1}{2}\left\|(1-M)\odot(C_1-C_2)\right\|_p\right) \xrightarrow[d\to\infty]{} 1 = \sqrt[p]{1-\gamma}.$$

Then, combining Assumption 19, Assumption 20 and the law of large numbers,

$$\begin{aligned}
\frac{1}{d-s}\sum_{j=1}^{d-s}(C_{1j}-C_{2j})^p &\xrightarrow[d\to\infty]{\mathbb{P}} \mathbb{E}\left[(C_{11}-C_{21})^p\right] \\
\frac{1}{s}\sum_{j=d-s+1}^{d}(C_{1j}-C_{2j})^p &\xrightarrow[d\to\infty]{\mathbb{P}} \mathbb{E}\left[(C_{11}-C_{21})^p\right].
\end{aligned}$$

Using Slutsky's theorem,

$$\frac{s}{d-s}\frac{1}{s}\sum_{j=d-s+1}^{d}(C_{1j}-C_{2j})^p \xrightarrow[d\to\infty]{\mathbb{P}} \frac{\gamma}{1-\gamma}(\mathbb{E}\left[(C_{11}-C_{21})^p\right]).$$

Re-using Slutsky's theorem,

$$\frac{\frac{s}{d-s}\frac{1}{s}\sum_{j=d-s+1}^{d}(C_{1j}-C_{2j})^p}{\frac{1}{d-s}\sum_{j=1}^{d-s}(C_{1j}-C_{2j})^p} \xrightarrow[d\to\infty]{\mathbb{P}} \frac{\gamma}{1-\gamma}.$$

Finally, using the continuous mapping theorem, we have that

$$\mathbb{P}\left(R < \frac{1}{2}\left\|(1-M)\odot(C_1-C_2)\right\|_p\right) \xrightarrow[d\to\infty]{} \sqrt[p]{1-\gamma}.$$

□

E.4 Proof of Proposition 3.26

Proof. In order to study the separability of the two balls after projection through the missing pattern, we need to study the probability that the sum of the radii is still smaller than the distance between the two centers after projection. Equivalently,

$$\mathbb{P}\left(R_1+R_2 < \|(1-M)\odot(c_1-c_2)\|_2\right)$$

as shown in (36). We have that

$$\begin{aligned}\mathbb{P}\left(R_1+R_2 < \|(1-M)\odot(c_1-c_2)\|_2\right) &\geq \mathbb{P}\left(\max(R_1,R_2) < \frac{1}{2}\|(1-M)\odot(c_1-c_2)\|_2\right)\\ &= \prod_{i=1}^{2}\mathbb{P}\left(R_i < \frac{1}{2}\|(1-M)\odot(c_1-c_2)\|_2\right)\\ &\qquad\qquad \text{(using that } R_1 \perp\!\!\!\perp R_2\text{)}\\ &= \mathbb{P}\left(R_1 < \frac{1}{2}\|(1-M)\odot(c_1-c_2)\|_2\right)^2.\\ &\qquad\qquad \text{(using that } R_1 \sim R_2\text{)}\end{aligned}$$

Using Assumption 23, i.e. $(R_1,R_2)\sim U(0,\frac{1}{2}\|c_1-c_2\|_p)^{\otimes 2}$ and using Assumption 24 ($R_1 \perp\!\!\!\perp M$),

$$\mathbb{P}\left(R_1 < \frac{1}{2}\|(1-M)\odot(c_1-c_2)\|_2\,\middle|\,M\right) = \frac{\|(1-M)\odot(c_1-c_2)\|_2}{\|(c_1-c_2)\|_2}.$$

Moreover, note that we have that

$$\begin{aligned}\mathbb{E}\left[\frac{\|(1-M)\odot(c_1-c_2)\|_2^2}{\|(c_1-c_2)\|_2^2}\right] &= \mathbb{E}\left[\frac{\sum_{j=1}^{d}(1-M_j)(c_{1j}-c_{2j})^2}{\sum_{j=1}^{d}(c_{1j}-c_{2j})^2}\right]\\ &= \frac{\sum_{j=1}^{d}\mathbb{E}\left[(1-M_j)\right](c_{1j}-c_{2j})^2}{\sum_{j=1}^{d}(c_{1j}-c_{2j})^2}\\ &= (1-\eta)\frac{\sum_{j=1}^{d}(c_{1j}-c_{2j})^2}{\sum_{j=1}^{d}(c_{1j}-c_{2j})^2} \qquad \text{(using Assumption 10)}\\ &= 1-\eta.\end{aligned}$$

Therefore, we conclude that

$$\mathbb{P}\left(R_1 + R_2 < \|(1 - M) \odot (c_1 - c_2)\|_2\right) \geq 1 - \eta.$$

□