Harnessing pattern-by-pattern linear classifiers for prediction with missing data

Angel D Reyero Lobo, Alexis Ayme, Claire Boyer, Erwan Scornet

Harnessing pattern-by-pattern linear classifiers for prediction with missing data

Angel D. Reyero Lobo
Paris-Saclay University
Gif-sur-Yvette, France
angel.reyero-lobo@universite-paris-saclay.fr

Alexis Ayme
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université and Université Paris Cité, CNRS, F-75005
Paris, France
alexis.ayme@sorbonne-universite.fr

Claire Boyer
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université and Université Paris Cité, CNRS, F-75005
Paris, France
claire.boyer@sorbonne-universite.fr

Erwan Scornet
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université and Université Paris Cité, CNRS, F-75005
Paris, France
erwan.scornet@sorbonne-universite.fr

Abstract

Missing values have been thoroughly analyzed in the context of linear models, where the final aim is to build coefficient estimates. However, estimating coefficients does not directly solve the problem of prediction with missing entries: a manner to address empty components must be designed. Major approaches to deal with prediction with missing values are empirically driven and can be decomposed into two families: imputation (filling in empty fields) and pattern-by-pattern prediction, where a predictor is built on each missing pattern. Unfortunately, most simple imputation techniques used in practice (as constant imputation) are not consistent when combined with linear models. In this paper, we focus on the more flexible pattern-by-pattern approaches and study their predictive performances on Missing Completely At Random (MCAR) data. We first show that a pattern-by-pattern logistic regression model is intrinsically ill-defined, implying that even classical logistic regression is impossible to apply to missing data. We then analyze the perceptron model and show how the linear separability property extends to partially-observed inputs. Finally, we use the Linear Discriminant Analysis to prove that pattern-by-pattern LDA is consistent in a high-dimensional regime. We refine our analysis to more complex MNAR data.

Preprint. Under review.

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

1 Introduction

Due to the large size of modern data sets, and the automatization of data collection, missing values are ubiquitous in real-world applications. Missing data can arise due to various reasons, such as sensor malfunctions, survey respondents skipping questions, or integration of data from diverse sources, collected using different methods. In his seminal work, Rubin (1976) categorizes missing value scenarios into three types: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR), depending on relationships between observed variables, missing variables, and the missing data pattern.

Much of the focus in missing value literature is on parameter estimation. Regarding linear models, closed-form coefficient estimators have been derived (Little, 1992; Jones, 1996; Robins et al., 1994), including sparsity constraints (Rosenbaum and Tsybakov, 2010; Loh and Wainwright, 2012) or the study of the optimization procedure (Sportisse et al., 2020). Regarding logistic regression models, no closed-form solutions are available and one may resort to the Expectation-Maximization algorithm (Consentino and Claeskens, 2011). Using the EM for parameter estimation in generalized linear models was introduced by Ibrahim (1990) for MAR data (with asymptotic theoretical guarantees) and later extended to some MNAR settings by modelling the missing indicators (Ibrahim et al., 1999). Methods for estimating the parameters in a high-dimensional LDA framework with MCAR missing data have also been proposed (see, e.g., Tony Cai and Zhang, 2019).

Prediction tasks with missing values differ from model estimation: estimated model parameters alone cannot directly predict outcomes on test samples containing missing values. A first strategy commonly encountered in practice is to impute the training dataset, before applying standard algorithms. Josse et al. (2019); Bertsimas et al. (2024) prove the consistency of constant imputation strategies preceding non-parametric learning methods, a result later extended for almost all imputation functions by Le Morvan et al. (2021). While these results are asymptotic and strongly rely on non-parametric estimators, Ayme et al. (2023, 2024) provides a finite-sample analysis of imputation in linear models.

An 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. Agarwal et al. (2019) examined the Principal Component Regression (PCR) strategy for handling missing values in high-dimensional settings. Le Morvan et al. (2020a,b) and Ayme et al. (2022) analyze pattern-by-pattern linear predictors, in finite-sample settings. Regarding classification, in fact, few analyses exist on predicting on missing data. Pelckmans et al. (2005) adapted Support Vector Machine (SVM) classifiers to accommodate missing values. Sell et al. (2023) establish minimax rate for prediction with missing values and propose HAM, an algorithm based on a sequential fit of k nearest neighbors on each missing pattern, which is minimax. Jiang et al. (2020) is one of the few methods able to estimate parameter and predict in presence of missing values.

From a practical perspective, many methods have been proposed to deal with missing values. For example, García-Laencina et al. (2009) propose to use K nearest neighbors to impute and predict with missing data, a work later refined by Choudhury and Kosorok (2020). Besides, MissForest Stekhoven and Bühlmann (2012) is one of the most versatile supervised learning algorithm, able to deal with discrete and continuous features. The interested reader may refer to Emmanuel et al. (2021) for a review of methods able to perform classification with missing values. However, most of these methods are not theoretically grounded.

Contributions Prediction with missing values requires to either use imputation or dedicated pattern-by-pattern strategies. Most previous works focus on the first approach, trying to derive Bayes optimality for generic imputation function in a non-parametric setting, or rate of convergence for the specific high-dimensional linear regression. Surprisingly, few results are available for pattern-by-pattern strategies. This paper aims at filling this gap. After formalizing the problem of missing inputs for prediction purposes (Section 2), we study the validity of pattern-by-pattern linear predictors for MCAR data. First, we show that the widely-used logistic regression is ill-specified to handle missing inputs (Section 3). More particularly, both pattern-by-pattern and imputation strategies are shown to be invalid, due to the modelling of the outcome probability, too rigid to be adapted to missing data scenarios. We then choose to break free from the straitjacket of this model by considering linearly separable data and the pattern-by-pattern perceptron algorithm (Section 4). We quantify the probability of maintaining linear separability despite missing values. Our results highlight that preserving linear separability across all missing patterns is restrictive and strongly depends on the geometry of the problem, but holds in specific high-dimensional sparse settings. To conclude our analysis of linear predictors, we use the linear discriminant analysis (LDA) framework (Section 5). Under MCAR data, we quantify the difference between the risk of an empirical classifier with missing data and that of the complete Bayes predictor. Such an error converges to zero as the number of samples and the number of inputs grow to infinity. Thus, the LDA is a sound theoretical procedure to handle missing values. Our analysis also highlights the difficulty of prediction in general MNAR settings. However, a simple thresholded pattern-by-pattern LDA predictor is shown to be efficient in MNAR situations, even when all missing patterns are admissible.

2 Preliminaries on supervised statistical learning with missing values

Supervised learning The main objective of binary classification tasks is to predict a target $Y \in \{-1, 1\}$ given some observation $X \in \mathbb{R}^d$. A canonical way of quantifying the performance of a classifier $h : \mathbb{R}^d \to \{-1, 1\}$ is given by the probability of misclassification

$$\mathcal{R}_{\mathrm{comp}}(h) = \mathbb{P}(Y \neq h(X)),$$

where the index "comp" stands for complete data. The Bayes predictor, minimizing the risk $\mathcal{R}_{\mathrm{comp}}$, takes the form $h^\star_{\mathrm{comp}}(X) = \mathrm{sign}(\mathbb{E}[Y|X])$. As the data distribution is unknown, learning consists in estimating $h^\star_{\mathrm{comp}}$ given a training sample $\mathcal{D}_n := \{(X_i, Y_i), i = 1, \ldots, n\}$.

Missing data in learning In the context of supervised learning with missing values, we assume that the input observation $X \in \mathbb{R}^d$ is only partially observed, with $M \in \{0, 1\}^d$ the associated missing pattern: each coordinate $M_j = 1$ indicates that the jth component of the input vector X_j is missing (and $M_j = 0$ if X_j is observed). Given a specific missing pattern $m \in \{0, 1\}^d$, we define $\mathrm{obs}(m)$ (resp. $\mathrm{mis}(m)$) as the set of indices where m is 0 (resp. 1), representing the observed (resp. missing) variables. Subsequently, $X_{\mathrm{obs}(M)}$ (resp. $X_{\mathrm{mis}(M)}$) refers to the subvector of X containing the observed (resp. missing) entries of X. Our aim is to predict the output Y from a pair consisting of the masked observation and the missing pattern, denoted as $Z := (X_{\mathrm{obs}(M)}, M)$ belonging to $\mathcal{Z}$. In presence of missing data, the performance of a classifier $h : \mathcal{Z} \to \{-1, 1\}$ is evaluated via

$$\mathcal{R}_{\mathrm{mis}}(h) = \mathbb{P}(Y \neq h(Z)),$$

and the Bayes predictor $h^\star_{\mathrm{mis}} : \mathcal{Z} \to \{-1, 1\}$ that minimizes $\mathcal{R}_{\mathrm{mis}}$ is defined as $h^\star_{\mathrm{mis}}(Z) = \mathrm{sign}(\mathbb{E}[Y|Z])$. Our analysis is based on the fact that the Bayes predictor $h^\star_{\mathrm{mis}}$ can be decomposed with respect to the missing patterns (see Lemma A.1), that is

$$h^\star_{\mathrm{mis}}(Z) = \sum_{m \in \mathcal{M}} h^\star_m(X_{\mathrm{obs}(m)}) \mathbb{1}_{M=m} \quad \text{with} \quad h^\star_m(X_{\mathrm{obs}(m)}) = \mathrm{sign}(\mathbb{E}\left[Y | X_{\mathrm{obs}(m)}, M = m\right]), \quad (1)$$

where $\mathcal{M} \subset \{1, \ldots, d\}$ is the set of admissible missing patterns. Learning with missing values can be seen as estimating $h^\star_m$ for all $m \in \{0, 1\}^d$, given an incomplete i.i.d. training sample $\mathcal{D}_n := \big\{(X_{i,\mathrm{obs}(M_i)}, M_i, Y_i), i = 1, \ldots, n\big\}$.

3 Logistic Regression

One of the most popular parametric methods for binary classification (with complete data) relies on the following logistic model for the distribution of $Y|X$.

Assumption 1 (Logistic model)**.** *Let $\sigma(t) = 1/(1 + e^{-t})$. There exist $\beta^\star_0, \ldots, \beta^\star_d \in \mathbb{R}$ such that the distribution of the output $Y \in \{-1, 1\}$ given the complete input X satisfies $\mathbb{P}(Y = 1|X) = \sigma(\beta^\star_0 + \sum_{j=1}^d \beta^\star_j X_j)$.*

In presence of missing data, one could be tempted to learn the parameters of the logistic model on complete data and use a logistic model with these estimators in order to predict on an incomplete vector. Proposition 3.1 below shows that such a strategy is doomed to fail when missing data are uninformative, regardless of the estimation procedure used on the complete data.

Assumption 2 (Missing Completely At Random (MCAR))**.** *M is independent of (X, Y).*

Proposition 3.1. *Under the logistic model specified by Assumption 1 for complete data, assume that the components $X_1, \ldots, X_d$ are independent, each one with an unbounded support, satisfying $\mathbb{E}\left[\exp(\beta_j^\star X_j)\right] < \infty$. Assume also Assumption 2. Let $m \in \{0,1\}^d$ and assume that the logistic model holds on the missing pattern $M = m$, that is there exist $\beta_{0,m}^\star, \ldots, \beta_{d,m}^\star \in \mathbb{R}$ such that*

$$\mathbb{P}\left(Y = 1 | X_{\mathrm{obs}(M)}, M = m\right) = \sigma\Big(\beta_{0,m}^\star + \sum_{j \in \mathrm{obs}(M)} \beta_{j,m}^\star X_j\Big).$$

Then, for all $j \in \mathrm{mis}(m)$, $\beta_j^\star = 0$.

Proposition 3.1 emphasizes that under MCAR missing data, the logistic model cannot be valid on the complete input vector and on any incomplete vector simultaneously, unless the unobserved components are not involved in the original logistic regression model. Using logistic models for all missing patterns is thus an ill-specified strategy, which will lead to inconsistent estimators. Note that such a result highlights that constant imputation is also an ill-specified strategy, even in the most simple case of independent entries. Interestingly, this result holds for each missing pattern separately. In particular, the logistic model should not be used even if only two missing patterns are possible.

Contrary to linear regression for which the prediction structure is preserved when the inputs are partially observed (see, e.g., Le Morvan et al., 2020c; Ayme et al., 2022), logistic models are not suited for missing data, assuming in both settings independent input variables with MCAR missingness. Due to the nonlinearity relation between the probability of success and the input vector, we do not have that the conditional expectation of the full model output is equal to the link function applied to the conditional expectation of the inputs, that is

$$\mathbb{P}[Y = 1 | X_{\mathrm{obs}(M)}, M = m] = \mathbb{E}\Big[\mathbb{P}[Y = 1 | X] | X_{\mathrm{obs}(M)}\Big] \neq \sigma\Big(\mathbb{E}[\beta_0^\star + \sum_{j=1}^{d} \beta_j^\star X_j | X_{\mathrm{obs}(M)}]\Big). \tag{2}$$

Therefore, the logistic model is not preserved on missing patterns and resulting imputation strategies will inevitably fail. To circumvent this issue, one may resort to traditional likelihood approaches at the price of additional assumptions on the input distribution (see, e.g., Jiang et al., 2020). Although this approach reduces the applicability and appeal of traditional logistic regression, it brings estimation and prediction down to the same problem. As modelling the output probability in each missing pattern by a logistic model is too restrictive, we opt in the next section for a deterministic approach and analyze how linear separability is preserved in presence of missing data.

4 Perceptron

We explore in this section how missing values impact geometry-based predictors such as the perceptron. The principle of the perceptron algorithm (Rosenblatt, 1958) is to iteratively find a hyperplane separating the data. The convergence of the method is ensured under the separability of the observations (Novikoff, 1962). In order to capture the influence of missing data, the goal is therefore to quantify the probability of maintaining linear separability in the presence of missing values, thus ensuring the validity of a *pattern-by-pattern perceptron*.

4.1 Setting

When dealing with complete observations, we say that the points $(X_i, Y_i)_{i=1,\ldots,n} \in \mathbb{R}^d \times \{-1, +1\}$ are linearly separable if there exists a hyperplane, parameterized by $(w^\star, b^\star) \in \mathbb{R}^d \times \mathbb{R}$, such that for all $i \in \{1, ..., n\}$, $Y_i\left(X_i^\top w^\star + b^\star\right) > 0$.

When dealing with missing inputs, the training data $(X_i \odot (1 - M_i), M_i, Y_i)_{i=1,\ldots,n} \in (\mathbb{R}^d \times \{0,1\}^d \times \{-1, +1\})^n$ is said linearly separable if $\forall m \in \{0,1\}^d$, $\exists (w_{(m)}^\star, b_{(m)}^\star) \in \mathbb{R}^d \times \mathbb{R}$ such that

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

Remark 4.1 (Related work: the rare eclipse problem). In Bandeira et al. (2014), the authors investigate the preservation of linear separability between two convex sets under random Gaussian projections. This particular problem is referred to as the *rare eclipse problem*. Unlike the Gaussian projections covered in Bandeira et al. (2014), the case of missing values involves random projections aligned with canonical axes.

Lemma 4.2. *Linear separability of complete data does not imply that of incomplete data.*

In all generality, the perceptron model cannot be transferred from complete to missing data patterns. In the following, we make additional assumptions on the input distribution (adapted to the perceptron model), to highlight favourable cases of predictor adaptability to missing inputs.

4.2 Fixed centroids

Assumption 3 (Fixed centroids and random radii)**.** *For given centroids c_1 and c_2, both classes are arbitrarily distributed in disjoint Euclidean balls B_1 and B_2, of radii R_1 and R_2, centered around the centroids. Radii R_1 and R_2 are uniformly distributed as $R_1, R_2 \sim \mathcal{U}(0, \frac{1}{2}\left\|c_1 - c_2\right\|_2)^{\otimes 2}$.*

Under MCAR assumption, remark that preserving linear separability despite missing values means that the Euclidean balls used to generate data remains disjoint when restricted to the support of observed entries, and that

$$\begin{aligned}\mathbb{P}\left(B_{1,\mathrm{obs}(M)} \cap B_{2,\mathrm{obs}(M)} = \emptyset\right) &= \mathbb{P}\left(R_1 + R_2 < \left\|c_{1,\mathrm{obs}(M)} - c_{2,\mathrm{obs}(M)}\right\|_2\right) \\ &= \mathbb{P}\left(R_1 + R_2 < \left\|(1-M)\odot(c_1 - c_2)\right\|_2\right). \end{aligned} \tag{3}$$

Proposition 4.3 (Separability of two balls with different radius)**.** *Given two fixed centroids c_1 and c_2, assume that complete data is generated as in Assumption 3. Under MCAR Assumption 2, with $\eta_j := \mathbb{P}(M_j = 1)$ for any coordinate $j \in \{1, \dots, d\}$, then*

$$\frac{\sum_{j=1}^d (1-\eta_j)(c_{1j} - c_{2j})^2}{\sum_{j=1}^d (c_{1j} - c_{2j})^2} \le \mathbb{P}\left(B_{1,\mathrm{obs}(M)} \cap B_{2,\mathrm{obs}(M)} = \emptyset\right) \le \sqrt{\frac{\sum_{j=1}^d (1-\eta_j)(c_{1j} - c_{2j})^2}{\sum_{j=1}^d (c_{1j} - c_{2j})^2}}.$$

This lower bound is informative when the probability of missing values on each coordinate remains low. When the centroids differ only from one coordinate j_0, note that the balls $B_{1,\mathrm{obs}(M)}$ and $B_{2,\mathrm{obs}(M)}$ do not overlap if and only if $j_0 \in \mathrm{obs}(M)$, i.e., $m_{j_0} = 0$, which happens with probability $1 - \eta_{j_0}$. When for any coordinate j, $\eta_j = \eta$, the bounds obtained in Proposition 4.3 become independent of the centroids:

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

On the contrary, when there is only one coordinate j_0 always missing ($\eta_{j_0} = 1$ and $\eta_j = 0$ for $j \neq j_0$), the bounds reveal that

$$1 - \frac{(c_{1,j_0} - c_{2,j_0})^2}{\|c_1 - c_2\|_2^2} \le \mathbb{P}\left(B_{1,\mathrm{obs}(M)} \cap B_{2,\mathrm{obs}(M)} = \emptyset\right) \le \sqrt{1 - \frac{(c_{1,j_0} - c_{2,j_0})^2}{\|c_1 - c_2\|_2^2}}.$$

This highlights that for high proportions of missing values that are very localized at certain coordinates, the linear separation will be all the more preserved if the quantity $\|c_1 - c_2\|_2^2$ is carried uniformly across the coordinates, i.e., when the vector $c_1 - c_2$ is anti-sparse.

4.3 Random centroids

The bounds derived in the previous section strongly depends on the geometry of the problem, via the centroid coordinates. To establish more general result, we consider random centroids C_1 and $C_2 \in \mathbb{R}^d$ and work with disjoint ℓ^p-balls (of same radius for simplicity). The former point is particularly suited to preserve the data geometry after random projections induced by missing entries.

Assumption 4. *We assume that (i) the coordinates of $C_1 - C_2$ are i.i.d., (ii) for all $j \in \{1, ..., d\}$, $\mathbb{E}\left[(C_1 - C_2)_j^p\right] < \infty$ and (iii) conditional to the centers C_1 and C_2, the radii R_1 is uniformly distributed as $R_1 | (C_1, C_2) \sim \mathcal{U}(0, \frac{1}{2}\left\|C_1 - C_2\right\|_p)$, with $R_2 = R_1$.*

Assumption 4 trivially includes the cases where $(C_1, C_2) \sim \mathcal{N}(\mu_1, \lambda_1 I_d) \otimes \mathcal{N}(\mu_2, \lambda_2 I_d)$, or where $(C_1, C_2) \sim \mathcal{U}(a_1, b_1)^{\otimes d} \otimes \mathcal{U}(a_2, b_2)^{\otimes d}$.

Assumption 5 (Uniform s-missing patterns)**.** *The missing pattern M is sampled uniformly at random among missing patterns admitting s missing values in total, i.e., $M \sim \mathcal{U}(\{m \in \{0,1\}^d, \|m\|_0 = s\})$.*

In the next proposition, we characterize the probability of preserving linear separability despite missing values, when the dimension d tends to ∞.

Proposition 4.4 (Asymptotic separability of two balls with the same radius)**.** *Under Assumption 4 and Assumption 5, let $\rho := \lim_{d\to\infty} \frac{s}{d}$. Then,*

$$\lim_{d\to+\infty} \mathbb{P}\left(B_{1,\mathrm{obs}(M)} \cap B_{2,\mathrm{obs}(M)} = \emptyset\right) = \sqrt[p]{1-\rho}. \tag{4}$$

Therefore, in high-dimensional regimes, pattern-by-pattern perceptron is a valid procedure with a probability converging to $\sqrt[p]{1-\rho}$, where ρ is the asymptotic ratio of missing values. Note that when s/d tends to zero, as s and d tend to infinity, the separability of the balls is ensured with probability 1. Besides this asymptotic separability probability $\sqrt[p]{1-\rho}$ increases when p increases. This is due to the fact that when p increases, the radius $R_1|(C_1, C_2) \sim \mathcal{U}(0, \frac{1}{2}\|C_1 - C_2\|_p)$ is shrinked ($p \mapsto \|x\|_p$ is non-increasing) and the balls are more and more separated. Beyond this restrictive sparse high-dimensional setting, the linear separability strongly depends on the geometry of the inputs. To be more conclusive on the efficiency of pattern-by-pattern classifiers, modeling both the distribution of $Y|X$ and the distribution of X seems to be unavoidable.

5 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 $h^\star(X) = \mathrm{sign}(\mathbb{E}[Y|X])$ when working with complete data. In this section, we analyze the finite-sample property of pattern-by-pattern LDA.

5.1 Setting

Assumption 6 (Balanced 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 $\pi_1 = \pi_{-1}$. For each class $k \in \{-1, 1\}$, $X|Y=k \sim \mathcal{N}(\mu_k, \Sigma)$, with $\mu_k \in \mathbb{R}^d$.*

In the complete case of LDA, the Bayes predictor reads as

$$h^\star_{\mathrm{comp}}(x) := \mathrm{sign}\left((\mu_1 - \mu_{-1})^\top \Sigma^{-1}\left(x - \frac{\mu_1 + \mu_{-1}}{2}\right)\right), \tag{5}$$

minimizing the misclassification probability $\mathcal{R}_{\mathrm{comp}}$ (see Section D.1 for details). When MCAR data occurs, by denoting $\Sigma_{\mathrm{obs}(M)} := \Sigma_{\mathrm{obs}(M)\times\mathrm{obs}(M)}$ (and $\Sigma^{-1}_{\mathrm{obs}(M)} = (\Sigma_{\mathrm{obs}(M)})^{-1}$), the pattern-by-pattern Bayes predictor (1) can be written as follows.

Proposition 5.1 (Pattern-by-pattern Bayes predictor for LDA with MCAR data)**.** *Under Assumptions 2 (MCAR) and 6 (LDA), the pattern-by-pattern Bayes classifier is given by*

$$h^\star_m(x_{\mathrm{obs}(m)}) = \mathrm{sign}\Big(\big(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\big)^\top \Sigma^{-1}_{\mathrm{obs}(m)}\Big(x_{\mathrm{obs}(m)} - \frac{\mu_{1,\mathrm{obs}(m)} + \mu_{-1,\mathrm{obs}(m)}}{2}\Big)\Big).$$

The decomposition provided in Proposition 5.1 relies on the fact that, under MCAR assumption, the distribution of $X_{\mathrm{obs}(M)}|Y, M=m$ is Gaussian for all $m \in \mathcal{M}$ (see Lemma F.6), similarly to the complete case. This does not hold anymore with a MAR missing mechanism, as shown below.

Example 5.2 (LDA+MAR is not pattern-by-pattern LDA). Let $X \in \mathbb{R}^2$ be a random variable satisfying Assumption 6, 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. In this case, the input distribution of $X_{\mathrm{obs}(M)}|Y=k, M=m$, for the pattern $m=(0,1)$, is not Gaussian, as its first component is positive.

Our goal is to study whether the Bayes risk with missing values converges to the Bayes risk with complete data as the dimension d increases. To do so, we scrutinize the error $\mathcal{R}_{\text{mis}}(h^\star) - \mathcal{R}_{\text{comp}}(h^\star_{\text{comp}})$.

Assumption 7 (Constant $\mathbb{P}(M_j = 1)$)**.** *The random variables $M_1, \ldots, M_d$ are independent, and follow a Bernoulli distribution with parameter η.*

Assumption 8 (Constant $(\mu_1 - \mu_{-1})_j$)**.** $\forall j \in \{1, ...d\}, (\mu_1 - \mu_{-1})_j = \pm\mu$*, with* $\mu > 0$.

Assumption 7 ensures that the missingness probability is the same for each input coordinate. Assumption 8 can be achieved up to a change of coordinates. In the sequel, we refer to $\mathrm{SNR} := \mu/\sqrt{\lambda_{\max}(\Sigma)}$ as the signal-to-noise ratio, where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of the input covariance matrix. This quantity describes the overlapping of the classes, and thus the difficulty of the classification task.

Proposition 5.3. *Under Assumptions 2, 6, 7 and 8, we have that*

$$\mathcal{R}_{\text{mis}}(h^\star) - \mathcal{R}_{\text{comp}}(h^\star_{\text{comp}}) \leq \frac{\eta^d}{2} + \frac{\mu\eta}{2\sqrt{2\pi}} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left(\epsilon\left(\eta, \mathrm{SNR}\right)^{d-1} - \eta^{d-1} \right),$$

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

The bound provided in Proposition 5.3 outlines that the difference between the Bayes risk with missing and complete data decreases exponentially fast with the input dimension d, assuming that the minimum eigenvalue of the covariance matrix is lower bounded or decreases at most polynomially with d (an assumption already considered in high-dimensional statistics, see e.g., Tony Cai and Zhang, 2019; Cai and Liu, 2011). This is the first analysis of the bias term due to learning with missing data in a classification context. When the signal-to-noise ratio SNR goes to infinity, one should expect the classification rate to be improved.

Corollary 5.4. *Under Assumptions 2, 6, 7, 8,*

$$\lim_{\mathrm{SNR}\to\infty} \sqrt{\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}} \frac{\mathrm{SNR}}{e^{\mathrm{SNR}^2/8}} = 0 \qquad \Longrightarrow \qquad \mathcal{R}_{\text{mis}}(h^\star) - \mathcal{R}_{\text{comp}}(h^\star_{\text{comp}}) \xrightarrow[\lambda\to\infty]{} \frac{\eta^d}{2}.$$

The limit established in Corollary 5.4 matches that of the limit of the bound of Proposition 5.3 when the SNR tends to infinity. It is important to note that the assumption on the structure of Σ is mild (as $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ may increase exponentially) and encompasses various scenarios, for example when $\Sigma = \sigma^2 I_d$ or when Σ is arbitrary but constant, with increasingly separated classes.

5.2 LDA estimation with missing values

Based on Proposition 5.1, we consider the pattern-by-pattern plug-in predictor $\widehat{h}$, in which

$$\widehat{\mu}_{k,j} = \frac{\sum_{i=1}^n X_{i,j} \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}} = \frac{\sum_{i=1}^n (X_i \odot (1-M_i))_j \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}, \tag{6}$$

estimates $\mu_{k,j}$, with the convention $0/0 = 0$, where the covariance matrix Σ is assumed to be known. More precisely,

$$\widehat{h}_m(x_{\text{obs}(m)}) = \mathrm{sign}\left(\left(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)} \right)^\top \Sigma^{-1} \left(x_{\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2} \right) \right). \tag{7}$$

Remark that under MCAR assumption, the estimates $\widehat{\mu}_{k,j}$ are built with all the observed inputs, independently of their missing patterns. This departs from a pattern-by-pattern estimation strategy where each mean is computed pattern-wise, using each observation once. We define $\kappa := \max_{i\in[d]}(\Sigma_{i,i})/\lambda_{\min}(\Sigma)$ as the largest value of the diagonal of the covariance over its smallest eigenvalue, which can be regarded as a non-standard condition number of Σ.

Theorem 5.5 (Bound on p-b-p LDA with known Σ)**.** *Grant Assumptions 2, 6 and 7. Then the excess risk of the classifier $\widehat{h}$, defined in* (7)*, satisfies*

$$\mathcal{R}_{\text{mis}}(\widehat{h}) - \mathcal{R}_{\text{mis}}(h^\star) \leq \frac{2}{\sqrt{2\pi}} \left(\left(\frac{1+\eta}{2}\right)^n \frac{\|\mu\|_\infty^2 d(1-\eta)}{\lambda_{\min}(\Sigma)} + \frac{4\kappa d}{n} \right)^{\frac{1}{2}}.$$

Then, for n large enough, we have

$$\mathcal{R}_{\mathrm{mis}}(\widehat{h}) - \mathcal{R}_{\mathrm{mis}}(h^\star) \lesssim \sqrt{\kappa d/n}. \tag{8}$$

The convergence rate of the LDA classifier in presence of missing values (and with a known covariance) is of the order of $(d/n)^{1/2}$. Moreover, the dependence of the upper bound on the covariance matrix Σ is mild, since the respective term decreases exponentially (corresponding to the case where all data are missing).

The upper bound presented in (8) is independent of the missingness probability η. If this could be surprising at first sight, it is important to note that the quantity of interest here is the difference between the misclassification probabilities of the estimated LDA predictor and the pattern-by-pattern LDA Bayes predictor given in Proposition 5.1. Both risks are integrated w.r.t. the distribution of missing inputs, so that both risks include the same missing data scenario. However, the influence of the probability of missingness should be expected when comparing predictors dealing with incomplete data on the one hand and the complete case on the other, as shown in the following corollary.

Corollary 5.6. *Grant Assumptions 2, 6, 7, 8. Then the classifier $\widehat{h}$, defined in* (7) *satisfies*

$$\begin{aligned}\mathcal{R}_{\mathrm{mis}}(\widehat{h}) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) \leq \frac{2}{\sqrt{2\pi}} &\left(\left(\frac{1+\eta}{2}\right)^n \frac{\|\mu\|_\infty^2\, d(1-\eta)}{\lambda_{\min}(\Sigma)} + \frac{4\kappa d}{n}\right)^{\frac{1}{2}} \\ &+ \frac{\eta^d}{2} + \frac{\mu\eta}{2\sqrt{2\pi}}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\left(\epsilon\left(\eta,\mathrm{SNR}\right)^{d-1} - \eta^{d-1}\right)\end{aligned}$$

with $\epsilon(\eta,\mathrm{SNR}) := \eta + e^{-\frac{\mathrm{SNR}^2}{8}}(1-\eta) < 1$ *and* $\mathrm{SNR} := \mu/\sqrt{\lambda_{\max}(\Sigma)}$.

In the previous bound, the first term is the learning error $\mathcal{R}_{\mathrm{mis}}(\widehat{h}) - \mathcal{R}_{\mathrm{mis}}(h^\star)$ and scales as $\sqrt{d/n}$; the second term is the bias $\mathcal{R}_{\mathrm{mis}}(h^\star) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}})$ due to missing values. When

$$n \ll \frac{1}{(\eta\cdot\mathrm{SNR})^2}\frac{1}{\epsilon(\eta,\mathrm{SNR})^d}, \tag{9}$$

the learning error inherent to the estimation procedure prevails over the approximation error due to missing values. Then, the impact of missing values on the predictive performances is negligible, and, $\mathcal{R}_{\mathrm{mis}}(\widehat{h}) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) = O(\sqrt{d/n})$, which corresponds to classical rates (see, e.g. Anderson, 2003). Assuming that $d = o(n)$, the misclassification risk of the estimated LDA with missing values converges to the Bayes risk with complete data.

Remark 5.7 (Related work on LDA with missing data). Previous work on LDA with missing values (Cai and Liu, 2011; Tony Cai and Zhang, 2019) focus on parameter estimation, which is not sufficient to design a procedure to predict with missing values. More precisely, Cai and Liu (2011) assume the s-sparsity of the so-called discriminant direction $\beta := \Sigma^{-1}(\mu_1 - \mu_{-1})$ and prove that, estimating this vector via linear programming discriminant (LPD) leads to a predictor $\widehat{h}_{\mathrm{LPD}}$ on complete data which satisfies $\mathcal{R}_{\mathrm{comp}}(\widehat{h}_{\mathrm{LPD}}) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) = O\left((s\log(d)/n)^{1/2}\right)$. Although Tony Cai and Zhang (2019) follow a completely different approach, their estimator applied on complete data reaches the same rate of convergence.

5.2.1 LDA under MNAR assumption

Extending LDA predictors to handle more general missing values is challenging. Indeed, as shown in Example 5.2, even under a MAR assumption, a pattern-wise approach for LDA is not valid. In this section, we exhibit a MNAR setting compatible with pattern-by-pattern LDA as follows.

Assumption 9 (GPMM-LDA)**.** *For all $m \in \mathcal{M}$ and $k \in \{-1,+1\}$, $X_{\mathrm{obs}(m)}|(M{=}m, Y{=}k) \sim \mathcal{N}(\mu_{m,k},\Sigma_m)$ with $\pi_{m,1} = \pi_{m,-1}$ where $\pi_{m,k} := \mathbb{P}(Y=k, M=m)$.*

Under Assumption 9, the Bayes predictor can be decomposed pattern by pattern as follows.

Proposition 5.8 (MNAR p-b-p LDA)**.** *Under Assumption 9, the pattern-by-pattern Bayes classifier is*

$$h^\star_m(x_{\mathrm{obs}(m)}) = \quad \mathrm{sign}\left(\left(\mu_{m,1} - \mu_{m,-1}\right)^\top \Sigma_m^{-1}\left(x_{\mathrm{obs}(m)} - \frac{\mu_{m,1}+\mu_{m,-1}}{2}\right) - \log\left(\frac{\pi_{m,-1}}{\pi_{m,1}}\right)\right).$$

Given the expression of the Bayes predictor in Proposition 5.8, we build a plug-in estimate based on the estimation $\widehat{\mu}_{m,k}$ of the mean $\mu_{m,k}$ on pattern $m \in \{0,1\}^d$ and class k, defined as

$$\widehat{\mu}_{m,k} := \frac{\sum_{i=1}^{n} X_i \mathbb{1}_{Y_i=k} \mathbb{1}_{M_i=m}}{\mathbb{1}_{Y_i=k} \mathbb{1}_{M_i=m}}. \tag{10}$$

Due to the potential exponential number of missing patterns, it may be difficult to estimate the 2^{d+1} estimates $\widehat{\mu}_{m,k}$. In line with Ayme et al. (2022), we employ a thresholded estimate, which boils down to estimating only the mean over the most frequent missing patterns, that is

$$\widetilde{\mu}_{m,k} := \widehat{\mu}_{m,k} \mathbb{1}_{\frac{N_{m,k}}{n} > \tau}, \tag{11}$$

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 $d < n$. Assuming that the covariance matrix for each missing pattern is known, we construct the pattern-by-pattern predictor $\widetilde{h}$ defined as

$$\widetilde{h}_m(x_{\mathrm{obs}(m)}) = \mathrm{sign}\left(\left(\widetilde{\mu}_{1,\mathrm{obs}(m)} - \widetilde{\mu}_{-1,\mathrm{obs}(m)} \right)^\top \Sigma_m^{-1} \left(x_{\mathrm{obs}(m)} - \frac{\widetilde{\mu}_{1,\mathrm{obs}(m)} + \widetilde{\mu}_{-1,\mathrm{obs}(m)}}{2} \right) \right). \tag{12}$$

Theorem 5.9 (MNAR p-b-p LDA estimation)**.** *Grant Assumption 9 and assume that the classes are balanced on each missing pattern. Let $\tau \geq \sqrt{d/n}$. Then, the plug-in classifier based on* (12) *satisfies*

$$\begin{aligned} &\mathcal{R}_{\mathrm{mis}}(\widetilde{h}) - \mathcal{R}_{\mathrm{mis}}(h^\star) \qquad (13) \\ &\leq \sum_{m \in \{0,1\}^d} \left(\frac{4}{\sqrt{2\pi}} + \frac{8}{\sqrt{\pi}} \frac{\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}} \right) \tau \wedge p_m + \sum_{\substack{m \in \{0,1\}^d, \\ p_m \geq \tau}} \frac{\sqrt{2}\,\|\mu_m\|}{\sqrt{\pi \lambda_{\min}(\Sigma_m)}} p_m (1-p_m)^{n/2}. \end{aligned}$$

Theorem 5.9 holds for various types of missingness. Indeed, Assumption 9 is very generic and may correspond to very difficult MNAR settings in which there is no relation between any covariances matrices Σ_m or any mean vector $\mu_{m,k}$. In this setting, building consistent predictions requires to build 2^d estimates of covariances matrices and 2^{d+1} mean estimates, an exponentially difficult task. On the other hand, assuming that there exists unique μ_{-1}, μ_1, Σ such that $\mu_{\pm 1,m} = \mu_{\pm 1,\mathrm{obs}(m)}$ and $\Sigma_m = \Sigma_{\mathrm{obs}(m)}$ allows us to study a MCAR setting in which proportion of missing values are different across coordinates, a generalization of Section 5.2.

The upper bound established in Theorem 5.9 is low when few missing patterns are admissible, but it appears to be very large when all 2^d missing patterns may occur. However, when the missing distribution is concentrated enough, one can control this upper bound. To see this, let us introduce the missing pattern distribution complexity $\mathfrak{C}_p(\tau) := \sum_{m \in \{0,1\}^d} \tau \wedge p_m$ used in Ayme et al. (2022), and assume that the missingness indicators $M_1, \dots, M_d$ are independent, distributed as a Bernoulli with parameter $\eta \leq d/n$. In such a setting, even if each missing pattern is admissible,

$$\mathcal{R}_{\mathrm{mis}}(\widetilde{h}) - \mathcal{R}_{\mathrm{mis}}(h^\star) \lesssim \frac{d^2}{n} + (1 - \min_{p_m > 0} p_m)^{n/2}, \tag{14}$$

which is much better than the initial upper bound, scaling as $d2^d/n$. This upper bound benefits from the concentration of the missing patterns, as a high number of missing components is unlikely to occur for independent Bernoulli distribution, with a small parameter $\eta \leq d/n$.

Contrary to Corollary 5.6, we do not compare $\mathcal{R}_{\mathrm{mis}}(\widehat{h})$ to $\mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}})$ as, in a MNAR setting, the distribution of the fully observed pattern may not be identifiable from the distribution of all missing patterns. Indeed, note that, in Assumption 9, the distribution of the complete pattern (corresponding to $m = 0$) may be chosen independently of the other distributions ($m \neq 0$). Thus, the difference $\mathcal{R}_{\mathrm{mis}}(\widehat{h}) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}})$ may be arbitrary large. This highlights the fact that all strategies that first estimate parameters from the complete distribution and then predict on each missing pattern by using these estimations are doomed to fail.

6 Conclusion

Coefficient estimation in parametric models is different from prediction, as one needs to specify a way to handle missing fields. Imputation and pattern-by-pattern approaches are the two most common strategies. The former are difficult to combine with linear classifiers, as the consistency may be lost. In this paper, we focus on the latter for linear classification purposes. We prove that pattern-by-pattern logistic regression (and even constant imputation in conjunction with logistic regression) leads to inconsistent probability estimates. We then study how the linear separability of complete data may extend to incomplete data, which, if true, underpins the suitability of a pattern-by-pattern perceptron. Under strong constraints, we prove that such a separability holds in a high-dimensional sparse setting. Finally, we turn to the LDA framework and propose a finite-sample analysis, highlighting that in MCAR scenarios, a pattern-by-pattern LDA approach is consistent in high dimensions. We extend our analysis to more complex types of missing data MNAR and provide a generic upper bound, which appears to be self-explanatory and legible when missing patterns are modeled as independent Bernoulli variables.

Our work provides a first analysis on how pattern-by-pattern classifiers may help to handle missing data in a predictive framework. If probabilistic models can be undermined by missing data (as is the case for logistic regression), one can expect that their decision frontier remains valid. Indeed, even if our result shows that the probability of classification cannot be properly estimated for any missing pattern, it may be possible that the decision frontier is close to the correct one, which should deserve further study. Regarding the discriminant analysis, the Gaussian assumption of the (conditional) distribution of the inputs helps to be theoretically conclusive when copping with missing data. Adapting this study framework to manage categorical inputs would confirm the applicability and relevance of LDA-type predictors in the presence of missing data.

References

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Naive imputation implicitly regularizes high-dimensional linear models. In *International Conference on Machine Learning*, pages 1320–1340. PMLR, 2023.

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Random features models: a way to study the success of naive imputation. *to appear in International Conference on Machine Learning proceedings*, 2024.

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

Dimitris Bertsimas, Arthur Delarue, and Jean Pauphilet. Simple imputation rules for prediction with missing data: Contrasting theoretical guarantees with empirical performance, 2024.

Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association*, 106(496):1566–1577, 2011.

Arkopal Choudhury and Michael R Kosorok. Missing data imputation for classification problems. *arXiv preprint arXiv:2002.10709*, 2020.

Fabrizio Consentino and Gerda Claeskens. Missing covariates in logistic regression, estimation and distribution selection. *Statistical Modelling*, 11(2):159–183, 2011.

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

Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big data*, 8:1–37, 2021.

Pedro J García-Laencina, José-Luis Sancho-Gómez, Aníbal R Figueiras-Vidal, and Michel Verleysen. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, 72(7-9):1483–1493, 2009.

Joseph G Ibrahim. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769, 1990.

Joseph G Ibrahim, Stuart R Lipsitz, and M-H Chen. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):173–190, 1999.

Wei Jiang, Julie Josse, Marc Lavielle, TraumaBase Group, et al. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907, 2020.

Michael P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91:222–230, 1996.

Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

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.

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, 2020c.

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What'sa good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.

Roderick JA Little. Regression with missing x's: a review. *Journal of the American statistical association*, 87(420):1227–1237, 1992.

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):1637 – 1664, 2012. doi: 10.1214/12-AOS1018. URL `https://doi.org/10.1214/12-AOS1018`.

Albert BJ Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622. New York, NY, 1962.

Kristiaan Pelckmans, Jos De Brabanter, Johan AK Suykens, and Bart De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5-6):684–692, 2005.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

Mathieu Rosenbaum and Alexandre B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620 – 2651, 2010. doi: 10.1214/10-AOS793. URL `https://doi.org/10.1214/10-AOS793`.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

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`.

Torben Sell, Thomas B. Berrett, and Timothy I. Cannings. Nonparametric classification with missing data, 2023.

Aude Sportisse, Claire Boyer, Aymeric Dieuleveut, and Julie Josses. Debiasing averaged stochastic gradient descent to handle missing values. *Advances in Neural Information Processing Systems*, 33, 2020.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

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`.

Notations. For $n \in \mathbb{N}$, we denote $[n] = \{1, \dots, n\}$. 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 values $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ respectively designate the largest and the smallest eigenvalues of any matrix A. We denote $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

A Proofs of Section 2

Lemma A.1. *Let $h^\star$ be a minimizer of $\mathcal{R}_{\text{mis}}(h) := \mathbb{P}(Y \neq h(Z))$, where $Z = (X_{\text{obs}(M)}, M)$. Then,*

$$h^\star(Z) = \sum_{m \in \mathcal{M}} h^\star_m(X_{\text{obs}(m)}) \mathbb{1}_{M=m},$$

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

Proof of Lemma A.1. Recall that we quantify the accuracy of a classifier using the probability of misclassification given by

$$\mathcal{R}_{\text{mis}}(h) := \mathbb{P}(Y \neq h(Z)) \tag{15}$$

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

$$\mathcal{R}_{\text{mis}}(h) = \frac{1}{4}\mathbb{E}\left[(Y - h(Z))^2\right] = \frac{1}{4}\mathbb{E}\left[(Y - \mathbb{E}\left[Y|Z\right])^2\right] + \frac{1}{4}\mathbb{E}\left[(\mathbb{E}\left[Y|Z\right] - h(Z))^2\right]. \tag{16}$$

Thus, the Bayes predictor is

$$h^\star(Z) := \text{sign}(\mathbb{E}\left[Y|Z\right]) = \text{sign}(\mathbb{E}\left[Y|X_{\text{obs}(M)}, M\right]) \text{ where } \text{sign}(x) = \mathbb{1}_{x \geq 0} - \mathbb{1}_{x<0}. \tag{17}$$

As we have that

$$\mathbb{E}\left[Y|X_{\text{obs}(M)}, M\right] = \sum_{m \in \mathcal{M}} \mathbb{E}\left[Y|X_{\text{obs}(m)}, M = m\right] \mathbb{1}_{M=m}, \tag{18}$$

then, the Bayes predictor can be written as

$$
\begin{aligned}
h^\star(Z) &= \mathrm{sign}(\mathbb{E}\left[Y|Z\right]) \\
&= \mathrm{sign}\left(\sum_{m\in\mathcal{M}} \mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M=m\right] \mathbb{1}_{M=m}\right) \\
&= \sum_{m\in\mathcal{M}} \mathrm{sign}(\mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M=m\right])\mathbb{1}_{M=m} \\
&= \sum_{m\in\mathcal{M}} h^\star_m(X_{\mathrm{obs}(m)})\mathbb{1}_{M=m} \qquad (19)
\end{aligned}
$$

with

$$
h^\star_m(X_{\mathrm{obs}(m)}) := \mathrm{sign}(\mathbb{E}\left[Y|X_{\mathrm{obs}(m)}, M=m\right]). \qquad (20)
$$

□

B (Logistic Model) Proof of Proposition 3.1

Proof. Let $m \in \{0,1\}^d$,

$$
\begin{aligned}
\mathbb{P}\left(Y=1|X_{\mathrm{obs}(m)}, M=m\right) &= \mathbb{P}\left(Y=1|X_{\mathrm{obs}(m)}\right) && \text{(using Assumption 2)} \\
&= \mathbb{E}\left[\mathbb{P}\left(Y=1|X\right)|X_{\mathrm{obs}(m)}\right] && (21) \\
&= \mathbb{E}\left[\frac{1}{1+\exp(-\beta^\star_0 - \sum_{j=1}^d \beta^\star_j X_j)}|X_{\mathrm{obs}(m)}\right]. && (22)
\end{aligned}
$$

Now, assume that there exists $\beta^\star_m \in \mathbb{R}^{d-\|m\|_0}$ such that

$$
\mathbb{P}\left(Y=1|X_{\mathrm{obs}(m)}, M=m\right) = \frac{1}{1+\exp(-\beta^\star_{0,m} - \sum_{j\in\mathrm{obs}(m)} \beta^\star_{j,m} X_j)}. \qquad (23)
$$

Combining the two previous equations leads to

$$
\begin{aligned}
&\frac{1}{1+\exp(-\beta^\star_{0,m} - \sum_{j\in\mathrm{obs}(m)} \beta^\star_{j,m} X_j)} \\
&= \mathbb{E}\left[\frac{1}{1+\exp(-\beta^\star_0 - \sum_{j=1}^d \beta^\star_j X_j)}|X_{\mathrm{obs}(m)}\right] && (24) \\
&\geq \frac{1}{\mathbb{E}\left[1+\exp(-\beta^\star_0 - \sum_{j=1}^d \beta^\star_j X_j)|X_{\mathrm{obs}(m)}\right]} && \text{(using Jensen Inequality)} \\
&= \frac{1}{1+\mathbb{E}\left[\exp\left(-\beta^\star_0 - \sum_{j\in\mathrm{obs}(m)} \beta^\star_j X_j - \sum_{j\in\mathrm{mis}(m)} \beta^\star_j X_j\right) \mid X_{\mathrm{obs}(m)}\right]} && (25) \\
&= \frac{1}{1+\exp\left(-\beta^\star_0 - \sum_{j\in\mathrm{obs}(m)} \beta^\star_j X_j\right)\mathbb{E}\left[\exp\left(-\sum_{j\in\mathrm{mis}(m)} \beta^\star_j X_j\right) \mid X_{\mathrm{obs}(m)}\right]}, && (26) \\
& && (27)
\end{aligned}
$$

which is equivalent to

$$
\exp\left(-(\beta^\star_{0,m} - \beta^\star_0) - \sum_{j\in\mathrm{obs}(m)} (\beta^\star_{j,m} - \beta^\star_j) X_j\right) \leq \mathbb{E}\left[\exp\left(-\sum_{j\in\mathrm{mis}(m)} \beta^\star_j X_j\right) \mid X_{\mathrm{obs}(m)}\right]. \qquad (28)
$$

Now, assuming that variables $X_1, \ldots, X_d$ are independent, we have

$$\exp\left(-(\beta^\star_{0,m} - \beta^\star_0) - \sum_{j \in \mathrm{obs}(m)} (\beta^\star_{j,m} - \beta^\star_j) X_j\right) \leq \mathbb{E}\left[\exp\left(-\sum_{j \in \mathrm{mis}(m)} \beta^\star_j X_j\right)\right]. \tag{29}$$

Let $1 \leq k \leq d$. Letting $X_j = 0$ for all $j \in \mathrm{obs}(m)$ with $j \neq k$, we have

$$\exp\left(-(\beta^\star_{0,m} - \beta^\star_0) - (\beta^\star_{k,m} - \beta^\star_k) X_k\right) \leq \mathbb{E}\left[\exp\left(-\sum_{j \in \mathrm{mis}(m)} \beta^\star_j X_j\right)\right]. \tag{30}$$

By assumption, the support of X_k is $\mathbb{R}$. Thus, letting X_k tending to $\pm\infty$, we deduce that

$$\beta^\star_{k,m} = \beta^\star_k. \tag{31}$$

Injecting this into (24) leads to

$$\frac{1}{1 + \exp(-\beta^\star_{0,m} - \sum\limits_{j \in \mathrm{obs}(m)} \beta^\star_j X_j)} = \mathbb{E}\left[\frac{1}{1 + \exp(-\beta^\star_0 - \sum_{j=1}^d \beta^\star_j X_j)} | X_{\mathrm{obs}(m)}\right], \tag{32}$$

that is

$$\mathbb{E}\left[\frac{1 + \exp(-\beta^\star_{0,m} - \sum\limits_{j \in \mathrm{obs}(m)} \beta^\star_j X_j)}{1 + \exp(-\beta^\star_0 - \sum_{j=1}^d \beta^\star_j X_j)} | X_{\mathrm{obs}(m)}\right] = 1. \tag{33}$$

Let

$$u = \exp\left(-\sum_{j \in \mathrm{obs}(m)} \beta^\star_j X_j\right) \quad \text{and} \quad Z_{\mathrm{mis}(m)} = \exp\left(-\sum_{j \in \mathrm{mis}(m)} \beta^\star_j X_j\right). \tag{34}$$

According to (33), for all $u \in (0, \infty)$,

$$\mathbb{E}\left[\frac{1 + u \exp(-\beta^\star_{0,m})}{1 + u Z_{\mathrm{mis}(m)} \exp(-\beta^\star_0)}\right] = 1. \tag{35}$$

Assume that $\mathbb{E}\left[1/Z_{\mathrm{mis}(m)}\right]$ exists. Take the limit when u tends to infinity. According to Lebesgue dominated convergence theorem, we have

$$\begin{aligned}
\lim_{u \to \infty} \mathbb{E}\left[\frac{1 + u \exp(-\beta^\star_{0,m})}{1 + u Z_{\mathrm{mis}(m)} \exp(-\beta^\star_0)}\right] &= \mathbb{E}\left[\lim_{u \to \infty} \frac{1 + u \exp(-\beta^\star_{0,m})}{1 + u Z_{\mathrm{mis}(m)} \exp(-\beta^\star_0)}\right] &(36)\\
&= \mathbb{E}\left[\frac{\exp(-\beta^\star_{0,m})}{Z_{\mathrm{mis}(m)} \exp(-\beta^\star_0)}\right]. &(37)
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\frac{1}{Z_{\mathrm{mis}(m)}}\right] = \exp(\beta^\star_{0,m} - \beta^\star_0). \tag{38}$$

By definition of $Z_{\mathrm{mis}(m)}$, we have

$$\begin{aligned}
\exp(\beta^\star_{0,m} - \beta^\star_0) &= \mathbb{E}\left[\frac{1}{\prod\limits_{j \in \mathrm{mis}(m)} \exp\left(-\beta^\star_j X_j\right)}\right] &(39)\\
&= \mathbb{E}\left[\prod_{j \in \mathrm{mis}(m)} \exp\left(\beta^\star_j X_j\right)\right] &(40)\\
&= \prod_{j \in \mathrm{mis}(m)} \mathbb{E}\left[\exp\left(\beta^\star_j X_j\right)\right]. &(41)
\end{aligned}$$

Thus,

$$\exp(-\beta^\star_{0,m}) = \frac{\exp(-\beta^\star_0)}{\prod\limits_{j\in \mathrm{mis}(m)} \mathbb{E}\left[\exp\left(\beta^\star_j X_j\right)\right]} \tag{42}$$

$$= \frac{\exp(-\beta^\star_0)}{\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right]}, \tag{43}$$

where

$$Z'_{\mathrm{mis}(m)} = 1/Z_{\mathrm{mis}(m)} = \exp\left(\sum_{j\in \mathrm{mis}(m)} \beta^\star_j X_j\right). \tag{44}$$

Injecting this equality into (35) leads to, for all $u \in (0, \infty)$,

$$\mathbb{E}\left[\frac{1 + u\exp(-\beta^\star_0)/\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right]}{1 + u\exp(-\beta^\star_0)/Z'_{\mathrm{mis}(m)}}\right] = 1 \tag{45}$$

$$\iff \mathbb{E}\left[\frac{\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right] + u\exp(-\beta^\star_0)}{\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right] + u\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right]\exp(-\beta^\star_0)/Z'_{\mathrm{mis}(m)}}\right] = 1 \tag{46}$$

$$\iff \mathbb{E}\left[\frac{\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right] + v}{\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right] + v\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right]/Z'_{\mathrm{mis}(m)}}\right] = 1. \tag{47}$$

where $v = u\exp(-\beta^\star_0)$. As this holds for all $v \in (0, \infty)$, taking the derivative of the expectation leads to, for all $v \in (0, \infty)$,

$$\mathbb{E}\left[\frac{\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right]\left(1 - \frac{\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right]}{Z'_{\mathrm{mis}(m)}}\right)}{\left(\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right] + \frac{\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right]}{Z'_{\mathrm{mis}(m)}}v\right)^2}\right] = 0. \tag{48}$$

Letting v tend to zero leads to

$$\mathbb{E}\left[\frac{1}{Z'_{\mathrm{mis}(m)}}\right] = \frac{1}{\mathbb{E}\left[Z'_{\mathrm{mis}(m)}\right]}, \tag{49}$$

which holds only if the random variable $Z'_{\mathrm{mis}(m)}$ is degenerated. By definition of $Z'_{\mathrm{mis}(m)}$, we deduce that for all $j \in \mathrm{mis}(m)$, X_j is degenerated or $\beta^\star_j = 0$. Since the support of X_j is $\mathbb{R}$, we have that $\beta^\star_j = 0$.

□

C (Perceptron) Proofs of Section 4

C.1 Proof of Lemma 4.2

Proof. Suppose that we only have two points $X_1, X_2 \in \mathbb{R}^d$ where $x_2 = (x_{1,1}, ..., x_{1,(k-1)}, x_{2,k}, x_{1,(k+1)}, ..., x_{1,d})$ 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. □

C.2 Separability characterization

Lemma C.1 (Separability characterization)**.** *Consider the ℓ^p-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. 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,

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

$\square$

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

$$\begin{aligned}
\mathbb{P}\left(B_{1,\text{obs}(M)} \cap B_{2,\text{obs}(M)} = \emptyset\right) &= \mathbb{P}\left(R_1 + R_2 < \left\|c_{1,\text{obs}(M)} - c_{2,\text{obs}(M)}\right\|_p\right) && (50)\\
&= \mathbb{P}\left(R_1 + R_2 < \|\Pi_M(c_1) - \Pi_M(c_2)\|_p\right) && (51)\\
&= \mathbb{P}\left(R_1 + R_2 < \|(1 - M) \odot (c_1 - c_2)\|_p\right). && (52)
\end{aligned}$$

In the remainder, we fix $p = 2$ (the Euclidean norm).

C.3 Proof of Proposition 4.3

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 (52). We have that

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

By Assumption 3, $(R_1, R_2) \sim U(0, \frac{1}{2}\|c_1 - c_2\|_2)^{\otimes 2}$ and assuming MCAR data ($R_1 \perp\!\!\!\perp M$),

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

Moreover, note that

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

Therefore, the lower bound is obtained using Jensen's inequality as follows

$$\begin{aligned}
\mathbb{P}\left(R_1+R_2 < \|(1-M)\odot(c_1-c_2)\|_2\right) &\geq \left(\mathbb{E}\left[\sqrt{\frac{\|(1-M)\odot(c_1-c_2)\|_2^2}{\|(c_1-c_2)\|_2^2}}\right]\right)^2 \\
&\geq \mathbb{E}\left[\frac{\|(1-M)\odot(c_1-c_2)\|_2^2}{\|(c_1-c_2)\|_2^2}\right] \\
&= \frac{\sum_{j=1}^d (1-\eta_j)(c_{1j}-c_{2j})^2}{\sum_{j=1}^d (c_{1j}-c_{2j})^2}.
\end{aligned}$$

To obtain the upper bound, one can proceed similarly, by using Jensen's inequality,

$$\begin{aligned}
\mathbb{P}\left(R_1+R_2 < \|(1-M)\odot(c_1-c_2)\|_2\right) &\leq \mathbb{P}\left(R_1 < \|(1-M)\odot(c_1-c_2)\|_2\right) \\
&= \mathbb{E}\left[\frac{\|(1-M)\odot(c_1-c_2)\|_2}{\|(c_1-c_2)\|_2}\right] \\
&= \sqrt{\left(\mathbb{E}\left[\frac{\|(1-M)\odot(c_1-c_2)\|_2}{\|(c_1-c_2)\|_2}\right]\right)^2} \\
&\leq \sqrt{\mathbb{E}\left[\frac{\|(1-M)\odot(c_1-c_2)\|_2^2}{\|(c_1-c_2)\|_2^2}\right]} \\
&= \sqrt{\frac{\sum_{j=1}^d (1-\eta_j)(c_{1j}-c_{2j})^2}{\sum_{j=1}^d (c_{1j}-c_{2j})^2}}.
\end{aligned}$$

□

C.4 Proof of Proposition 4.4

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 radii is still smaller than the distance between the two centers after projection as shown in Lemma C.1. As seen in (52), since $R := R_1 = R_2$, this probability corresponds to

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

Using Assumption 4, we have that

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

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

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

after having reordered the terms using the exchangeability of the $(C_1 - C_2)_j$ (Assumption 4(i)). One has

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

As d goes to infinity, we assume that the number of missing values s goes to infinity. Otherwise, if s is bounded, then $\rho = \lim_{d\to\infty}\frac{s}{d} = 0$ 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

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

Then, combining Assumption 4 and the law of large numbers, we get

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

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

Using Slutsky's theorem,

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

Re-using Slutsky's theorem,

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

Finally, using the continuous mapping theorem, we have that

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

$\square$

D (LDA + MCAR) Proofs of Section 5.1

D.1 Preliminary

The Bayes predictor $h^\star_{\text{comp}}$ satisfies

$$\mathcal{R}_{\text{comp}}(h^\star_{\text{comp}}) = \Phi\left(-a_m - b_m\right)\pi_{-1} + \Phi\left(a_m - b_m\right), \pi_1 \tag{53}$$

where $\Phi(x) = \mathbb{P}[\mathcal{N}(0,1) \leq x]$ is the c.d.f. of a standard Gaussian random variable, $a_m = \log\left(\frac{\pi_{-1}}{\pi_1}\right) / \|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1})\|$ and $b_m = \|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1})\|/2$.

Corollary D.1 (Bayes Risk of p-b-p LDA)**.** *Under Assumptions 2 and 6, the Bayes risk is given by*

$$\mathcal{R}_{\text{mis}}(h^\star) = \sum_{m \in \{0,1\}^d} \Phi\left(-a_m - b_m\right)\pi_{-1}p_m + \Phi\left(a_m - b_m\right)\pi_1 p_m, \tag{54}$$

where, for all $m \in \mathcal{M}$,

$$a_m = \frac{\log\left(\frac{\pi_{-1}}{\pi_1}\right)}{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|} \quad \textit{and} \quad b_m = \frac{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|}{2} \tag{55}$$

The proof can be found in Appendix D.3. Note that, from Corollary D.1 (using that $\pi_1 = \pi_{-1}$) and Equation (53), we have that

$$L(h^\star) - \mathcal{R}_{\text{comp}}(h^\star_{\text{comp}}) = \sum_{m \in \{0,1\}^d} \left(\Phi\left(-\frac{\left\|\Sigma^{-\frac{1}{2}}_{\text{obs}(m)}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|}{2}\right) - \Phi\left(-\frac{\left\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1})\right\|}{2}\right)\right) p_m, \tag{56}$$

with Φ the c.d.f. of a standard Gaussian variable.

D.2 Proof of Proposition 5.1

Proof. Expanding (20),

$$\begin{aligned} h^\star_m(X_{\text{obs}(m)}) &= \text{sign}(\mathbb{E}\left[Y | X_{\text{obs}(m)}, M = m\right]) \\ &= \text{sign}\left(\mathbb{P}\left(Y = 1 \mid X_{\text{obs}(m)}, M = m\right) - \mathbb{P}\left(Y = -1 \mid X_{\text{obs}(m)}, M = m\right)\right). \end{aligned} \tag{57}$$

Note that, for any Borelian $B \subset \mathbb{R}^{|\text{obs}(m)|}$,

$$\begin{aligned} \mathbb{P}\left(Y = k \mid X_{\text{obs}(m)} \in B, M = m\right) &= \frac{\mathbb{P}\left(Y = k, X_{\text{obs}(m)} \in B \mid M = m\right)}{\mathbb{P}\left(X_{\text{obs}(m)} \in B \mid M = m\right)} \\ &= \frac{\mathbb{P}\left(Y = k, X_{\text{obs}(m)} \in B\right)}{\mathbb{P}\left(X_{\text{obs}(m)} \in B\right)} \qquad \text{(using Assumption 2)} \\ &= \frac{\mathbb{P}\left(X_{\text{obs}(m)} \in B \mid Y = k\right)\pi_k}{\mathbb{P}\left(X_{\text{obs}(m)} \in B\right)}. \end{aligned}$$

Thus,

$$\mathbb{P}\left(Y = 1 \mid X_{\text{obs}(m)} \in B, M = m\right) > \mathbb{P}\left(Y = -1 \mid X_{\text{obs}(m)} \in B, M = m\right) \tag{58}$$

$$\iff \quad \mathbb{P}\left(X_{\text{obs}(m)} \in B \mid Y = 1\right)\pi_1 > \mathbb{P}\left(X_{\text{obs}(m)} \in B \mid Y = -1\right)\pi_{-1}. \tag{59}$$

As this holds for any Borelian $B \subset \mathbb{R}^{|\text{obs}(m)|}$, $h^\star_m$ can be rewritten as

$$h^\star_m(x) = \text{sign}\left(\pi_1 f_{X_{\text{obs}(m)}|Y=1}(x) - \pi_{-1} f_{X_{\text{obs}(m)}|Y=1}(x)\right) \tag{60}$$

$$= \text{sign}\left(\log\left(\frac{f_{X_{\text{obs}(m)}|Y=1}(x)}{f_{X_{\text{obs}(m)}|Y=-1}(x)}\right) - \log\left(\frac{\pi_{-1}}{\pi_1}\right)\right), \tag{61}$$

where $f_{X_{\mathrm{obs}(m)}|Y=k}$ is the density of $X_{\mathrm{obs}(m)} \mid Y = k$ for all $k \in \{-1, 1\}$. Under LDA model (Assumption 6), 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 F.6 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,

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

Consequently,

$$
h^\star_m(x) = \mathrm{sign}\left((\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})^\top\Sigma^{-1}_{\mathrm{obs}(m)}\left(x-\frac{\mu_{1,\mathrm{obs}(m)}+\mu_{-1,\mathrm{obs}(m)}}{2}\right) - \log\left(\frac{\pi_{-1}}{\pi_1}\right)\right), \tag{62}
$$

which concludes the proof. □

D.3 Proof of Corollary D.1

Proof. Let $N = \Sigma^{-\frac{1}{2}}_{\mathrm{obs}(m)}(X_{\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})$. Using Proposition 5.1, we have

$$
\begin{aligned}
&\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \mid Y = -1\right)\\
=&\mathbb{P}\Big(\left(\mu_{1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)}\right)^\top\Sigma^{-1}_{\mathrm{obs}(m)}\left(X_{\mathrm{obs}(m)}-\frac{\mu_{1,\mathrm{obs}(m)}+\mu_{-1,\mathrm{obs}(m)}}{2}\right)\\
&\qquad - \log\left(\frac{\pi_{-1}}{\pi_1}\right) > 0 \mid Y = -1\Big)\\
=&\mathbb{P}\left(\gamma^\top N - \frac{1}{2}\|\gamma\|^2 > \log\left(\frac{\pi_{-1}}{\pi_1}\right) \mid Y = -1\right),
\end{aligned}
$$

where $\gamma = \Sigma^{-\frac{1}{2}}_{\mathrm{obs}(m)}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})$. By Lemma F.6, $N|Y = -1 \sim \mathcal{N}(0, Id_{d-\|m\|_0})$. Thus,

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

Similarly, letting $N' = \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(X_{\mathrm{obs}(m)} - \mu_{1,\mathrm{obs}(m)})$,

$$
\begin{aligned}
&\mathbb{P}\left(h_m^\star(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1\right) \\
=&\mathbb{P}\Big(\left(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right)^\top \Sigma_{\mathrm{obs}(m)}^{-1} \left(X_{\mathrm{obs}(m)} - \frac{\mu_{1,\mathrm{obs}(m)} + \mu_{-1,\mathrm{obs}(m)}}{2}\right) \\
&\qquad - \log\left(\frac{\pi_{-1}}{\pi_1}\right) < 0 \mid Y = 1\Big) \\
=&\mathbb{P}\left(\gamma^\top N + \frac{1}{2}\left\|\gamma\right\|^2 < \log\left(\frac{\pi_{-1}}{\pi_1}\right) \mid Y = 1\right) \\
=&\mathbb{P}\left(\frac{\gamma^\top N}{\|\gamma\|} > \frac{1}{2}\left\|\gamma\right\| - \frac{1}{\|\gamma\|}\log\left(\frac{\pi_{-1}}{\pi_1}\right) \mid Y = -1\right) \\
=&\Phi\left(-\frac{1}{2}\left\|\gamma\right\| + \frac{1}{\|\gamma\|}\log\left(\frac{\pi_{-1}}{\pi_1}\right)\right).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&\mathcal{R}_{\mathrm{mis}}(h^\star) \\
&= \mathbb{P}\left(h^\star(X_{\mathrm{obs}(M)}, M) \neq Y\right) \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h^\star(X_{\mathrm{obs}(m)}, M) \neq Y \mid M = m\right) p_m \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h_m^\star(X_{\mathrm{obs}(m)}) \neq Y\right) p_m \qquad \text{(using Assumption 2)} \\
&= \sum_{m\in\{0,1\}^d} \mathbb{P}\left(h_m^\star(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1\right) \pi_1 p_m + \mathbb{P}\left(h_m^\star(X_{\mathrm{obs}(m)}) = 1 \mid Y = -1\right) \pi_{-1} p_m \\
&= \sum_{m\in\{0,1\}^d} \Phi(a_m - b_m)\pi_1 p_m + \Phi\left(-a_m - b_m\right)\pi_{-1} p_m,
\end{aligned}
$$

where, for all $m \in \mathcal{M}$,

$$
a_m = \frac{\log\left(\frac{\pi_{-1}}{\pi_1}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|} \quad \text{and} \quad b_m = \frac{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\|}{2}. \tag{63}
$$

□

D.4 Proof of Proposition 5.3

Proof. Using Assumption 8, we have that

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

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

Recall that Φ is the c.d.f. of a standard Gaussian random variable, according to Equation (56), we have

$$
\begin{aligned}
&\mathcal{R}_{\mathrm{mis}}(h^\star) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) \\
&= \sum_{m\in\{0,1\}^d} \left(\Phi\left(- \frac{\left\| \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}} (\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}) \right\|}{2} \right) - \Phi\left(- \frac{\left\| \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_{-1}) \right\|}{2} \right) \right) p_m \qquad (64) \\
&\leq \sum_{m\in\{0,1\}^d} \left(\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - \|m\|_0}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) p_m \\
&= \sum_{i=0}^{d} \sum_{\substack{m\in\{0,1\}^d \\ \text{s.t. } \|m\|_0 = i}} \left(\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - i}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) p_m \\
&= \sum_{i=0}^{d} \left(\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - i}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) \binom{d}{i} \eta^i (1-\eta)^{d-i} \\
&\hspace{25em} \text{(using Assumption 7)} \\
&= \mathbb{E}\left[\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - B}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right] \qquad (65)
\end{aligned}
$$

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

$$
\begin{aligned}
&L(h^\star) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) \\
&\leq \mathbb{E}\left[\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - B}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \mid B = d \right] \mathbb{P}(B = d) \\
&\quad + \mathbb{E}\left[\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - B}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \mid B \neq d \right] \mathbb{P}(B \neq d) \\
&= \left(\frac{1}{2} - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) \eta^d \qquad (66) \\
&\quad + \mathbb{E}\left[\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - B}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \mid B \neq d \right] (1 - \eta^d) \qquad (67)
\end{aligned}
$$

Now, we study the second term in (67). Letting $Q(x) = \int_x^\infty e^{-\frac{t^2}{2}} dt$, we have $\Phi(x) = \frac{1}{\sqrt{2\pi}} Q(-x)$, which leads to

$$
\begin{aligned}
&\mathbb{E}\left[\Phi\left(-\frac{\mu}{2} \sqrt{\frac{d - B}{\lambda_{\max}(\Sigma)}} \right) - \Phi\left(-\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \mid B \neq d \right] \\
&\quad = \mathbb{E}\left[\frac{1}{\sqrt{2\pi}} \left(Q\left(T_B\right) - Q\left(t\right) \right) \mid B \neq d \right],
\end{aligned}
$$

where $T_B := \frac{\mu}{2}\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}$ and $t = \frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}$. Applying the the mean-value inequality to the function Q on the interval $[T_B, t]$ leads to

$$
Q(T_B) - Q(t) \leq e^{-\frac{T_B^2}{2}} (t - T_B). \qquad (68)
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}\left[\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\mid B\neq d\right] \\
&\leq \frac{1}{\sqrt{2\pi}}\mathbb{E}\left[e^{-\frac{t_B^2}{2}}(t-T_B)\mid B\neq d\right] \\
&= \frac{\mu}{2\sqrt{2\pi}}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\mid B\neq d\right]. \qquad (69)
\end{aligned}
$$

Besides, since

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

we have

$$
\begin{aligned}
&\mathbb{E}\left[\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\mid B\neq d\right] \\
&= \frac{\mu}{2\sqrt{2\pi}}\frac{1}{\mathbb{P}(B\neq d)}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right] \qquad (70) \\
&\qquad -\frac{\mu}{2\sqrt{2\pi}}\frac{\mathbb{P}(B=d)}{\mathbb{P}(B\neq d)}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}. \qquad (71)
\end{aligned}
$$

Looking at the expectation in (71), we obtain

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

since

$$
\sqrt{d}-\sqrt{d-B}=\frac{d-d+B}{\sqrt{d}+\sqrt{d-B}}\leq\frac{B}{\sqrt{d}}. \qquad (72)
$$

Simple calculation shows that

$$
\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\right]=\left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^d.
$$

Besides,

$$\begin{aligned}
\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}B\right] &= \sum_{i=0}^{d}\binom{d}{i}e^{-\frac{\mu^2(d-i)}{8\lambda_{\max}(\Sigma)}}i\eta^i(1-\eta)^{d-i}\\
&= \eta d\sum_{i=1}^{d}\frac{(d-1)!}{(i-1)!(d-1-(i-1))!}\eta^{i-1}\left(e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^{d-1-(i-1)}\\
&= \eta d\left(\eta+e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)\right)^{d-1}. \qquad (73)
\end{aligned}$$

Therefore, letting $A=e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}}(1-\eta)$, we have that

$$\begin{aligned}
&\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right]\\
&\le \sqrt{d}\left(\frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}-\frac{1}{\sqrt{\lambda_{\max}(\Sigma)}}\right)(\eta+A)^d+\frac{1}{\sqrt{\lambda_{\max}(\Sigma)d}}\eta d\,(\eta+A)^{d-1}\\
&= \frac{\sqrt{d}}{\sqrt{\lambda_{\min}(\Sigma)}}(\eta+A)^d-\sqrt{\frac{d}{\lambda_{\max}(\Sigma)}}(\eta+A)^{d-1}A. \qquad (74)
\end{aligned}$$

Gathering equations (67), (71) and (74), we obtain

$$\begin{aligned}
&L(h^\star)-\mathcal{R}_{\text{comp}}(h^\star_{\text{comp}})\\
&= \left(\frac{1}{2}-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d\\
&\quad+\mathbb{E}\left[\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)-\Phi\left(\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\mid B\neq d\right](1-\eta^d)\\
&\le \left(\frac{1}{2}-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d+\frac{\mu}{2\sqrt{2\pi}}(1-\eta^d)\\
&\quad\times\left(\frac{1}{\mathbb{P}(B\neq d)}\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right]-\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\frac{\mathbb{P}(B=d)}{\mathbb{P}(B\neq d)}\right)\\
&= \left(\frac{1}{2}-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d\\
&\quad+\frac{\mu}{2\sqrt{2\pi}}\left(\mathbb{E}\left[e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}-\sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}}\right)\right]-\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\eta^d\right)\\
&\le \left(\frac{1}{2}-\Phi\left(-\frac{\mu}{2}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\right)\right)\eta^d\\
&\quad+\frac{\mu}{2\sqrt{2\pi}}\left(\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}(\eta+A)^d-\sqrt{\frac{d}{\lambda_{\max}(\Sigma)}}(\eta+A)^{d-1}A-\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\eta^d\right).
\end{aligned}$$

An upper bound of this inequality is given by

$$L(h^\star)-\mathcal{R}_{\text{comp}}(h^\star_{\text{comp}})\le\frac{\eta^d}{2}+\frac{\mu\eta}{2\sqrt{2\pi}}\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\left((\eta+A)^{d-1}-\eta^{d-1}\right). \qquad (75)$$

□

D.5 Proof of Corollary 5.4

Proof. Recall that, by Equation (56),

$$
\begin{aligned}
&L(h^\star) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) \\
&= \sum_{m \in \{0,1\}^d} \left(\Phi\left(- \frac{\left\| \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}} (\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}) \right\|}{2} \right) - \Phi\left(- \frac{\left\| \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_{-1}) \right\|}{2} \right) \right) p_m \\
&\geq \left(\Phi(0) - \Phi\left(- \frac{\left\| \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_{-1}) \right\|}{2} \right) \right) \eta^d,
\end{aligned}
$$

using only $m = \mathbf{1}$, since all terms in the above sum are positive. By Assumption 8, $\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1})\| \geq d\mu/\sqrt{\lambda_{\max}(\Sigma)}$. Hence

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

Consequently,

$$
L(h^\star) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) \geq \left(\frac{1}{2} - \Phi\left(- \frac{d\lambda}{2} \right) \right) \eta^d \xrightarrow[\lambda \to \infty]{} \frac{\eta^d}{2}. \tag{76}
$$

On the other hand, by Proposition 5.3, we have

$$
L(h^\star) - \mathcal{R}_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) \leq \frac{\eta^d}{2} + \frac{\mu}{2\sqrt{2\pi}} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left((\eta + A)^d - \eta^d \right). \tag{77}
$$

Note that

$$
\begin{aligned}
&\left| \mu \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left(\left(\eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}} (1 - \eta) \right)^d - \eta^d \right) \right| \\
&= \mu \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left(\sum_{i=0}^{d} \binom{d}{i} \eta^{d-i} e^{-\frac{i\lambda^2}{8}} (1-\eta)^i - \eta^d \right) \\
&= \frac{\mu}{\sqrt{\lambda_{\max}(\Sigma)}} \sqrt{\frac{d\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}} \left(\sum_{i=1}^{d} \binom{d}{i} \eta^{d-i} e^{-\frac{i\lambda^2}{8}} (1-\eta)^i \right) \\
&= \lambda \sqrt{\frac{d\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}} \left(\sum_{i=1}^{d} \binom{d}{i} \eta^{d-i} e^{-\frac{i\lambda^2}{8}} (1-\eta)^i \right),
\end{aligned}
$$

which tends to zero by assumption. This concludes the proof. □

D.6 Proofs of Section 5.2

D.6.1 General lemmas for LDA misclassification control.

Lemma D.2 ($\widehat{\mu}$ misclassification probability)**.** *Given a sample satisfying Assumptions 2 and 6, with balanced classes, then*

$$
\begin{aligned}
&\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = 1 \mid Y = -1, \mathcal{D}_n \right) \\
&= \Phi\left(\frac{\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)}) \right)^\top \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}} \left(\mu_{-1,\mathrm{obs}(m)} - \frac{\widehat{\mu}_{1,\mathrm{obs}(m)} + \widehat{\mu}_{-1,\mathrm{obs}(m)}}{2} \right)}{\left\| \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)}) \right\|} \right)
\end{aligned}
\tag{78}
$$

and symmetrically,

$$\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1, \mathcal{D}_n\right)$$
$$= \Phi\left(-\frac{\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right)^\top \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mu_{1,\mathrm{obs}(m)} - \frac{\widehat{\mu}_{1,\mathrm{obs}(m)}+\widehat{\mu}_{-1,\mathrm{obs}(m)}}{2}\right)}{\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right\|}\right) \tag{79}$$

with Φ *the standard Gaussian cumulative function.*

Proof. We follow the same strategy as in the proof of Corollary D.1. We have

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

Let $N = \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(X_{\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})$. By Lemma F.6, $N|Y = -1 \sim \mathcal{N}(0, Id_{d-\|m\|_0})$. Since $(X_{\mathrm{obs}(m)}, Y)$ and $\mathcal{D}_n$ are independent

$$N|Y = -1, \mathcal{D}_n \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \tag{80}$$

Letting $\widehat{\gamma} = \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})$, we have

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

Now we prove the second statement. According to the proof of Corollary D.1, we have

$$\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1, \mathcal{D}_n\right)$$
$$= \mathbb{P}\left(\left(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)}\right)^\top \Sigma_{\mathrm{obs}(m)}^{-1}\left(X_{\mathrm{obs}(m)} - \frac{\widehat{\mu}_{1,\mathrm{obs}(m)} + \widehat{\mu}_{-1,\mathrm{obs}(m)}}{2}\right) < 0 \mid Y = 1\right).$$

Let $N = \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(X_{\mathrm{obs}(m)} - \mu_{1,\mathrm{obs}(m)})$. By Lemma F.6, and since $(X_{\mathrm{obs}(m)}, Y)$ and $\mathcal{D}_n$ are independent,

$$N|Y = 1, \mathcal{D}_n \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \tag{81}$$

Letting $\widehat{\gamma} = \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})$, we have

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

$\square$

Lemma D.3. *Grant Assumptions 2 and 6. Assume that we are given two estimators $\widehat{\mu}_1$ and $\widehat{\mu}_{-1}$ of μ_1 and μ_{-1}. Then, the classifier $\widehat{h}_m$ defined in Equation* (7) *satisfies*

$$\left|\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1\right)\right| \\ \leq \frac{3}{2\sqrt{2\pi}}\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{-1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right\| + \frac{1}{2\sqrt{2\pi}}\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\mathrm{obs}(m)} - \widehat{\mu}_{1,\mathrm{obs}(m)})\right\| \tag{82}$$

and symmetrically,

$$\left|\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y=1, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \mid Y=1\right)\right| \\ \leq \frac{3}{2\sqrt{2\pi}}\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(-\widehat{\mu}_{1,\mathrm{obs}(m)} + \mu_{1,\mathrm{obs}(m)})\right\| + \frac{1}{2\sqrt{2\pi}}\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{-1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)})\right\| \tag{83}$$

Proof. We only prove the first inequality, the other one can be handled in the same manner. According to using Corollary D.1 and Lemma D.2,

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

since Φ is $(1/\sqrt{2\pi})$-Lipschitz. Note that, by injecting $\pm\widehat{\mu}_{-1,\mathrm{obs}(m)}$, the numerator of the first term can be rewritten as

$$\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right)^\top \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mu_{-1,\mathrm{obs}(m)} - \frac{\widehat{\mu}_{1,\mathrm{obs}(m)}+\widehat{\mu}_{-1,\mathrm{obs}(m)}}{2}\right) \tag{84}$$

$$= \left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right)^\top \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mu_{-1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)}\right) \tag{85}$$

$$+ \frac{1}{2}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right)^\top \Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\widehat{\mu}_{-1,\mathrm{obs}(m)} - \widehat{\mu}_{1,\mathrm{obs}(m)}\right) \tag{86}$$

$$\leq \left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right\|\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\mu_{-1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right\| \tag{87}$$

$$- \frac{1}{2}\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\mathrm{obs}(m)} - \widehat{\mu}_{-1,\mathrm{obs}(m)})\right\|^2, \tag{88}$$

where the last line results from Cauchy-Schwarz inequality. Thus, by the Triangle inequality, followed by the reverse triangle inequality, we obtain

$$\left|\mathbb{P}\left(\widehat{h}_m(X_{\text{obs}(m)}) = 1 \mid Y = -1, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\text{obs}(m)}) = 1 \mid Y = -1\right)\right| \tag{89}$$

$$\leq \frac{1}{\sqrt{2\pi}} \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\| \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\|}{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\|} \tag{90}$$

$$+ \frac{1}{\sqrt{2\pi}} \left| - \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\|}{2} + \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|}{2} \right| \tag{91}$$

$$\leq \frac{1}{\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\| \tag{92}$$

$$+ \frac{1}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(-\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)} + \mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\| \tag{93}$$

$$\leq \frac{1}{\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\| + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)})\right\| \tag{94}$$

$$+ \frac{1}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\| \tag{95}$$

$$\leq \frac{3}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)})\right\| + \frac{1}{2\sqrt{2\pi}} \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)})\right\|. \tag{96}$$

The second statement of the Lemma can be proven in the same way. □

Lemma D.4. *Grant Assumptions 2, 6 and assume the classes are balanced. Assume that we are given two estimators $\widehat{\mu}_1$ and $\widehat{\mu}_{-1}$ of μ_1 and μ_{-1}. Then, the classifier $\widehat{h}$ defined in Equation (7) satisfies*

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

Proof. We have

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

by Lemma D.3. Thus,

$$\mathcal{R}_{\text{mis}}(\widehat{h}) - \mathcal{R}_{\text{mis}}(h^\star)$$
$$= \sum_{m\in\mathcal{M}} \frac{p_m}{\sqrt{2\pi}} \left(\mathbb{E}\left[\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(-\widehat{\mu}_{1,\text{obs}(m)} + \mu_{1,\text{obs}(m)})\right\| + \left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|\right]\right).$$

□

It is worth noting that, at this juncture, neither the structure of the estimate nor the structure of the covariance matrix have been incorporated.

D.6.2 Lemma for Theorem 5.5

Lemma D.5. *For all $m \in \mathcal{M}$ and all $k \in \{-1, 1\}$,*

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

with $\widehat{\mu}_{k,\text{obs}(m)}$ defined in (6) and $\kappa := \max_{i\in[n]} \Sigma_{i,i}/\lambda_{\min}(\Sigma)$ the greatest value of the diagonal of the covariance matrix divided by its smallest eigenvalue.

Proof. First, by Jensen's inequality,

$$\mathbb{E}\left[\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right\|\right] \leq \mathbb{E}\left[\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right\|^2\right]^{\frac{1}{2}} \tag{97}$$

$$= \mathbb{E}\left[\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right)^\top \left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right)\right]^{\frac{1}{2}} \tag{98}$$

$$= \mathbb{E}\left[\text{tr}\left(\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right)^\top \left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right)\right)\right]^{\frac{1}{2}} \tag{99}$$

$$= \mathbb{E}\left[\text{tr}\left(\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right) \left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right)^\top\right)\right]^{\frac{1}{2}} \tag{100}$$

$$= \text{tr}\left(\mathbb{E}\left[\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right) \left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})\right)^\top\right]\right)^{\frac{1}{2}} \tag{101}$$

$$= \text{tr}\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\mathbb{E}\left[\left(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}\right)\left(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}\right)^\top\right]\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\right)^{\frac{1}{2}} \tag{102}$$

$$= \text{tr}\left(\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\mathcal{C}(k,m)\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}\right)^{\frac{1}{2}}, \tag{103}$$

where

$$\mathcal{C}(k,m) := \mathbb{E}\left[\left(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}\right)\left(\widehat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}\right)^\top\right]. \tag{104}$$

Now, we compute the elements $\mathcal{C}(k,m)_{r,l} =$, for all $r, l \in \text{obs}(m)$.

First case. We start by computing $\mathcal{C}(k,m)_{l,l}$ for all l. Note that

$$\mathcal{C}(k,m)_{l,l} = \mathbb{E}\left[(\widehat{\mu}_{k,l} - \mu_{k,l})^2\right]. \tag{105}$$

The estimator $\widehat{\mu}_{k,l}$ equals zero if all samples of class k have a missing l-th coordinate, which corresponds to the event

$$\mathcal{A}_{k,l} := \{\forall i \in \{1, ...n\}, \quad Y_i = -k \text{ or } M_{i,l} = 1\}, \tag{106}$$

where

$$\mathbb{P}(\mathcal{A}_{k,l}) = \prod_{i=1}^n P(Y_i = -k \text{ or } M_{i,l} = 1) = \left(\frac{1+\eta}{2}\right)^n. \tag{107}$$

Thus,

$$
\begin{aligned}
&\mathbb{E}\left[(\widehat{\mu}_{k,l}-\mu_{k,l})^2\right]\\
&= \mathbb{E}\left[(\widehat{\mu}_{k,l}-\mu_{k,l})^2 \mid \mathcal{A}_{k,l}\right]\mathbb{P}\left(\mathcal{A}_{k,l}\right)+\mathbb{E}\left[(\widehat{\mu}_{k,l}-\mu_{k,l})^2 \mid \mathcal{A}_{k,l}^c\right]\mathbb{P}\left(\mathcal{A}_{k,l}^c\right)\\
&= \mu_{k,l}^2\left(\frac{1+\eta}{2}\right)^n+\mathbb{E}\left[\left(\frac{\sum_{i=1}^n (X_{i,l}-\mu_{k,l})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}\right)^2 \mid \mathcal{A}_{k,l}^c\right]\left(1-\left(\frac{1+\eta}{2}\right)^n\right)
\end{aligned}
$$

The second term can be rewritten as

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

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

$$
\mathcal{C}(k,m)_{l,l} = \mu_{k,l}^2\left(\frac{1+\eta}{2}\right)^n + n\left(\frac{1-\eta}{2}\right)\Sigma_{l,l}A(n,\eta). \tag{108}
$$

Second case. Now, we want to compute, for all $r \neq l$,

$$
\mathcal{C}(k,m)_{r,l} = \mathbb{E}\left[\left(\widehat{\mu}_{k,r}-\mu_{k,r}\right)\left(\widehat{\mu}_{k,l}-\mu_{k,l}\right)\right]. \tag{109}
$$

To this aim, we distinguish three cases, depending on the presence of available samples to compute $\widehat{\mu}_{k,r}$ and $\widehat{\mu}_{k,r}$. First, let us denote by

$$
\mathcal{A}_{k,l,r} := \{\forall i \in \{1,...n\}, (Y_i=-k \text{ or } (M_{i,r}=1 \text{ and } M_{i,l}=1))\}, \tag{110}
$$

the event in which there is no available samples to estimate any of the means $\widehat{\mu}_{k,r}$ and $\widehat{\mu}_{k,r}$, that is each sample either belongs to the other class or is missing at both coordinates. We have

$$\begin{aligned}
&\mathbb{P}(\mathcal{A}_{k,l,r}) \\
&= \mathbb{P}\left(\forall i \in \{1,...n\}, Y_i = -k \text{ or } (M_{i,r} = 1 \text{ and } M_{i,l} = 1)\right) && (111) \\
&= \left(\mathbb{P}(Y_i = -k) + \mathbb{P}(M_{i,r} = 1 \text{ and } M_{i,l} = 1) - \mathbb{P}(Y_i = -k \text{ and } M_{i,r} = 1 \text{ and } M_{i,l} = 1)\right)^n \\
&= \left(\frac{\eta^2+1}{2}\right)^n. && (112)
\end{aligned}$$

Besides, on the event $\mathcal{A}_{k,l,r}$, we have

$$\mathbb{E}\left[\left(\widehat{\mu}_{k,r} - \mu_{k,r}\right)\left(\widehat{\mu}_{k,l} - \mu_{k,l}\right)|\mathcal{A}_{k,l,r}\right] = \mu_{k,r}\mu_{k,l}. \tag{113}$$

We now consider the second case and denote by

$$\mathcal{B}_{k,l,r} := \{\exists i \in \{1,...,n\}, (Y_i = k \wedge M_{i,l} = 0) = 1\} \cap \{\exists i \in \{1,...n\}, (Y_i = k \wedge M_{i,r} = 0) = 1\}, \tag{114}$$

the event in which the there is at least one available sample to estimate both means $\widehat{\mu}_{k,r}$ and $\widehat{\mu}_{k,r}$. Observe that

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

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

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

Then, we have that

$$\mathbb{P}\left(\mathcal{B}_{k,l,r}\right) = 1 - 2\left(\frac{1+\eta}{2}\right)^n + \left(\frac{\eta^2+1}{2}\right)^n. \tag{115}$$

Besides,

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

Observe that this second sum is null. Indeed,

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

Then,

$$
\begin{aligned}
&\mathbb{E}\left[(\widehat{\mu}_{k,r}-\mu_{k,r})(\widehat{\mu}_{k,l}-\mu_{k,l})\mid \mathcal{B}_{k,l,r}\right]\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{(X_{i,r}-\mu_{k,r})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,r}=0}}\right)\left(\frac{(X_{i,l}-\mu_{k,l})\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}{\sum_{i=1}^{n}\mathbb{1}_{Y_i=k}\mathbb{1}_{M_{i,l}=0}}\right)\mid \mathcal{B}_{k,l,r}\right]\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\left[\frac{(X_{i,r}-\mu_{k,r})}{1+\sum_{j\neq i}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,r}=0}}\frac{(X_{i,l}-\mu_{k,l})}{1+\sum_{j\neq i}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}}\mid Y_i=k, M_{i,r}=0, M_{i,l}=0\right]\\
&\qquad\mathbb{P}\left(Y_i=k, M_{i,r}=0, M_{i,l}=0\mid \mathcal{B}_{k,l,r}\right)\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\left[(X_{i,r}-\mu_{k,r})(X_{i,l}-\mu_{k,l})\mid Y_i=k\right]\\
&\qquad\times\mathbb{E}\left[\frac{1}{1+\sum_{j\neq i}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,r}=0}}\frac{1}{1+\sum_{j\neq i}^{n}\mathbb{1}_{Y_j=k}\mathbb{1}_{M_{j,l}=0}}\right]\\
&\qquad\times\mathbb{P}\left(Y_i=k, M_{i,r}=0, M_{i,l}=0\mid \mathcal{B}_{k,l,r}\right)\\
&\quad=n\Sigma_{r,l}B(n,\eta)\frac{(1-\eta)^2}{2\mathbb{P}\left(\mathcal{B}_{k,l,r}\right)},
\end{aligned}
\tag{116}
$$

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

Now, we consider the last case, and denote by

$$
\mathcal{C}_{k,l,r}=(\mathcal{B}_{k,l,r}\cup\mathcal{A}_{k,l,r})^c
\tag{117}
$$

the event in which only one mean can be estimated. We have

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

Let $\mathcal{C}_{k,l,r}=\mathcal{C}_{1,k,l,r}\cup\mathcal{C}_{2,k,l,r}$, where $\mathcal{C}_{1,k,l,r}$ is the event where the one that can be estimated is $\widehat{\mu}_{k,r}$. Then,

$$
\begin{aligned}
&\mathbb{E}\left[(\widehat{\mu}_{k,r}-\mu_{k,r})(\widehat{\mu}_{k,l}-\mu_{k,l})\mid \mathcal{C}_{1,k,l,r}\right]\\
&\quad=-\mu_{k,l}\mathbb{E}\left[\frac{\sum_{i=1}^{n}(X_{i,r}-\mu_{k,r})\mathbb{1}_{M_{i,r}=0}\mathbb{1}_{Y_i=k}}{\sum_{i=1}^{n}\mathbb{1}_{M_{i,r}=0}\mathbb{1}_{Y_i=k}}\mid \mathcal{C}_{1,k,l,r}\right]\\
&\quad=-\mu_{k,l}n\mathbb{E}\left[\frac{(X_{1,r}-\mu_{k,r})}{1+\sum_{i=2}^{n}\mathbb{1}_{M_{i,r}=0}\mathbb{1}_{Y_i=k}}\mid M_{1,r}=0, Y_1=k\right]\mathbb{P}(M_{1,r}=0, Y_1=k|\mathcal{C}_{1,k,l,r})\\
&\quad=-\mu_{k,l}n\mathbb{E}\left[(X_{1,r}-\mu_{k,r})\mid Y_1=k\right]\mathbb{E}\left[\frac{1}{1+\sum_{i=2}^{n}\mathbb{1}_{M_{i,r}=0}\mathbb{1}_{Y_i=k}}\right]\mathbb{P}(M_{1,r}=0, Y_1=k|\mathcal{C}_{1,k,l,r})\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(using MCAR and independence)}\\
&\quad=0.
\end{aligned}
\tag{118}
$$

By symmetry, we also have

$$\mathbb{E}\left[\left(\widehat{\mu}_{k,r}-\mu_{k,r}\right)\left(\widehat{\mu}_{k,l}-\mu_{k,l}\right) \mid \mathcal{C}_{2,k,l,r}\right]=0. \tag{119}$$

Thus,

$$\mathbb{E}\left[\left(\widehat{\mu}_{k,r}-\mu_{k,r}\right)\left(\widehat{\mu}_{k,l}-\mu_{k,l}\right) \mid \mathcal{C}_{k,l,r}\right]=0. \tag{120}$$

Gathering (113), (116) and (120), we are able to compute $\mathcal{C}(k,m)_{r,l}$ as follows

$$\begin{aligned}
\mathcal{C}(k,m)_{r,l} &= \mathbb{E}\left[\left(\widehat{\mu}_{k,r}-\mu_{k,r}\right)\left(\widehat{\mu}_{k,l}-\mu_{k,l} \mid \mathcal{A}_{k,l,r}\right)\right]\mathbb{P}\left(\mathcal{A}_{k,l,r}\right) \\
&\quad + \mathbb{E}\left[\left(\widehat{\mu}_{k,r}-\mu_{k,r}\right)\left(\widehat{\mu}_{k,l}-\mu_{k,l} \mid \mathcal{B}_{k,l,r}\right)\right]\mathbb{P}\left(\mathcal{B}_{k,l,r}\right) \\
&\quad + \mathbb{E}\left[\left(\widehat{\mu}_{k,r}-\mu_{k,r}\right)\left(\widehat{\mu}_{k,l}-\mu_{k,l} \mid \mathcal{C}(k,m)_{k,l,r}\right)\right]\mathbb{P}\left(\mathcal{C}(k,m)_{k,l,r}\right) \\
&= \mu_{k,r}\mu_{k,l}\left(\frac{\eta^2+1}{2}\right)^n + n\Sigma_{r,l}B(n,\eta)\frac{(1-\eta)^2}{2},
\end{aligned}$$

using (112) and (115). From (108), recall that

$$\mathcal{C}(k,m)_{l,l} = \mu_{k,l}^2\left(\frac{1+\eta}{2}\right)^n + n\left(\frac{1-\eta}{2}\right)\Sigma_{l,l}A(n,\eta). \tag{121}$$

Let J be the matrix composed of 1 in each entry, and let

$$F = \left(\left(\frac{1+\eta}{2}\right)^n - \left(\frac{1+\eta^2}{2}\right)^n\right)I + \left(\frac{1+\eta^2}{2}\right)^n J \tag{122}$$

$$G = \left(A(n,\eta) - (1-\eta)B(n,\eta)\right)I + (1-\eta)B(n,\eta)J. \tag{123}$$

Thus,

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

Then, according to inequality (103), we have that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)}\right)\right\|\right] \\
&\quad\le \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mathcal{C}(k,m)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)^{\frac{1}{2}} \qquad (124)\\
&\quad= \left(\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(F\odot\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^\top\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) + n\frac{1-\eta}{2}\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(G\odot\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\right)^{\frac{1}{2}} \qquad (125)
\end{aligned}$$

The first term equals

$$\begin{aligned}
&\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(F\odot\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^\top\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \qquad (126)\\
&\quad= \left(\frac{1+\eta^2}{2}\right)^n \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(J\odot\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^\top\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \qquad (127)\\
&\qquad + \left(\left(\frac{1+\eta}{2}\right)^n - \left(\frac{1+\eta^2}{2}\right)^n\right)\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(I_{d-\|m\|_0}\odot\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^\top\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \qquad (128)\\
&\quad= \left(\frac{1+\eta^2}{2}\right)^n \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^\top\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \qquad (129)\\
&\qquad + \left(\left(\frac{1+\eta}{2}\right)^n - \left(\frac{1+\eta^2}{2}\right)^n\right)\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\operatorname{diag}\left(\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^\top\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right). \qquad (130)
\end{aligned}$$

Then, by Lemma F.4,

$$\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(F \odot \mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^{\top}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \tag{131}$$

$$\leq \left(\frac{1+\eta^2}{2}\right)^n \operatorname{tr}\left(\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\right)\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\right)^{\top}\right) \tag{132}$$

$$+\left(\left(\frac{1+\eta}{2}\right)^n - \left(\frac{1+\eta^2}{2}\right)^n\right)\|\mu\|_\infty^2 \operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-1}\right) \tag{133}$$

$$= \left(\frac{1+\eta^2}{2}\right)^n \left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\mu_{k,\mathrm{obs}(m)}\right\|^2 + \left(\left(\frac{1+\eta}{2}\right)^n - \left(\frac{1+\eta^2}{2}\right)^n\right)\frac{\|\mu\|_\infty^2 (d-\|m\|_0)}{\lambda_{\min}(\Sigma)} \tag{134}$$

$$\leq \left(\frac{1+\eta^2}{2}\right)^n \frac{\left\|\mu_{k,\mathrm{obs}(m)}\right\|^2}{\lambda_{\min}(\Sigma)} + \left(\left(\frac{1+\eta}{2}\right)^n - \left(\frac{1+\eta^2}{2}\right)^n\right)\frac{\|\mu\|_\infty^2 (d-\|m\|_0)}{\lambda_{\min}(\Sigma)} \tag{135}$$

$$\leq \left(\frac{1+\eta^2}{2}\right)^n \frac{\|\mu\|_\infty^2 (d-\|m\|_0)}{\lambda_{\min}(\Sigma)} + \left(\left(\frac{1+\eta}{2}\right)^n - \left(\frac{1+\eta^2}{2}\right)^n\right)\frac{\|\mu\|_\infty^2 (d-\|m\|_0)}{\lambda_{\min}(\Sigma)} \tag{136}$$

$$\leq \left(\frac{1+\eta}{2}\right)^n \frac{\|\mu\|_\infty^2 (d-\|m\|_0)}{\lambda_{\min}(\Sigma)}. \tag{137}$$

Regarding the second term in (125), note that $A(n,\eta) - (1-\eta)B(n,\eta) \geq 0$. Indeed, letting $Z := \sum_{i=1}^{n-1} \mathbb{1}_{Y_i=k} \sim \mathcal{B}(n-1, 1/2)$,

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

using the exchangeability as the samples are 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

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

Thus, we have that

$$\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(G \odot \Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \tag{138}$$

$$= (1-\eta)B(n,\eta)\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(\mathbf{1} \odot \Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \tag{139}$$

$$+ (A(n,\eta) - (1-\eta)B(n,\eta))\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(I_{d-\|m\|_0} \odot \Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \tag{140}$$

$$= (1-\eta)B(n,\eta)\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\Sigma_{\mathrm{obs}(m)}\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) \tag{141}$$

$$+ (A(n,\eta) - (1-\eta)B(n,\eta))\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\operatorname{diag}\left(\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right). \tag{142}$$

Using Lemma F.4 and $A(n,\eta)-(1-\eta)B(n,\eta)\geq 0$, we have

$$\begin{aligned}
&\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(G\odot\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right) && (143)\\
&\leq (1-\eta)B(n,\eta)(d-\|m\|_0) && (144)\\
&\qquad + \left(A(n,\eta)-(1-\eta)B(n,\eta)\right)\max_{i\in[d]}\left(\Sigma_{i,i}\right)\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-1}\right) && (145)\\
&\leq (1-\eta)B(n,\eta)(d-\|m\|_0) && (146)\\
&\qquad + \left(A(n,\eta)-(1-\eta)B(n,\eta)\right)\frac{\max_{i\in[d]}\left(\Sigma_{i,i}\right)}{\lambda_{\min}\left(\Sigma\right)}(d-\|m\|_0) && (147)\\
&\leq \kappa(1-\eta)B(n,\eta)(d-\|m\|_0)+\left(A(n,\eta)-(1-\eta)B(n,\eta)\right)\kappa(d-\|m\|_0) && (148)\\
&= A(n,\eta)\kappa(d-\|m\|_0) && (149)\\
&\leq \frac{2\kappa(d-\|m\|_0)}{n(n+1)\left(\frac{1-\eta}{2}\right)^2}, && (150)
\end{aligned}$$

where $\kappa := \frac{\max_{i\in[d]}(\Sigma_{i,i})}{\lambda_{\min}(\Sigma)}\geq 1$. Finally, combining (137) and (150) in (125), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{k,\mathrm{obs}(m)}-\mu_{k,\mathrm{obs}(m)})\right\|\right]\\
&\leq \Bigg(\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(F\odot\mu_{k,\mathrm{obs}(m)}\mu_{k,\mathrm{obs}(m)}^\top\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\\
&\qquad + n\frac{1-\eta}{2}\operatorname{tr}\left(\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\left(G\odot\Sigma_{\mathrm{obs}(m)}\right)\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}\right)\Bigg)^{\frac{1}{2}}\\
&\leq \left(\left(\frac{1+\eta}{2}\right)^n\frac{\|\mu\|_\infty^2(d-\|m\|_0)}{\lambda_{\min}(\Sigma)}+n\frac{1-\eta}{2}\frac{2\kappa(d-\|m\|_0)}{n(n+1)\left(\frac{1-\eta}{2}\right)^2}\right)^{\frac{1}{2}}\\
&\leq \left(\left(\frac{1+\eta}{2}\right)^n\frac{\|\mu\|_\infty^2(d-\|m\|_0)}{\lambda_{\min}(\Sigma)}+\frac{4\kappa(d-\|m\|_0)}{(n+1)(1-\eta)}\right)^{\frac{1}{2}}.
\end{aligned}$$

□

D.6.3 Proof of Theorem 5.5

Proof. By Lemma D.4,

$$\begin{aligned}
&\mathcal{R}_{\mathrm{mis}}(\widehat{h})-\mathcal{R}_{\mathrm{mis}}(h^\star)\\
&\leq \sum_{m\in\mathcal{M}}\frac{1}{\sqrt{2\pi}}\Big(\mathbb{E}\Big[\left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(-\widehat{\mu}_{1,\mathrm{obs}(m)}+\mu_{1,\mathrm{obs}(m)})\right\|\\
&\qquad + \left\|\Sigma_{\mathrm{obs}(m)}^{-\frac{1}{2}}(\widehat{\mu}_{-1,\mathrm{obs}(m)}-\mu_{-1,\mathrm{obs}(m)})\right\|\Big]\Big)p_m\\
&\leq \frac{2}{\sqrt{2\pi}}\sum_{m\in\mathcal{M}}\left(\left(\frac{1+\eta}{2}\right)^n\frac{\|\mu\|_\infty^2(d-\|m\|_0)}{\lambda_{\min}(\Sigma)}+\frac{4\kappa(d-\|m\|_0)}{(n+1)(1-\eta)}\right)^{\frac{1}{2}}p_m.
\end{aligned}$$

(using Lemma D.5)

Now, using Assumption 7, we have that $\|M\|_0 \sim \mathcal{B}(d,\eta)$, so that

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

□

D.6.4 Proof of Corollary 5.6

Proof. From Proposition 5.3 and Theorem 5.5 we have that

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

□

E (LDA + MNAR) Proofs of Section 5.2.1

E.1 Proof of Proposition 5.8

Proof. By definition of the Bayes classifier (see (20)),

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

with $\pi_{m,k} = \mathbb{P}\left(M=m, Y=k\right)$. Thus, our objective is to study when

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

Note that by using Assumption 9, we have $X_{\mathrm{obs}(m)}|M=m, Y=k \sim \mathcal{N}(\mu_{m,k}, \Sigma_m)$. Therefore,

$$\begin{aligned}
&\log\left(\frac{f_{X_{\mathrm{obs}(m)}|M=m,Y=1}(x)}{f_{X_{\mathrm{obs}(m)}|M=m,Y=-1}(x)}\right) \\
&\quad= \log\left(\frac{(\sqrt{2\pi})^{-(d-\|m\|_0)}\sqrt{\det(\Sigma_m^{-1})}\exp\left(-\frac{1}{2}(x-\mu_{1,m})^\top\Sigma_m^{-1}(x-\mu_{1,m})\right)}{(\sqrt{2\pi})^{-(d-\|m\|_0)}\sqrt{\det(\Sigma_m^{-1})}\exp\left(-\frac{1}{2}(x-\mu_{-1,m})^\top\Sigma_m^{-1}(x-\mu_{-1,m})\right)}\right) \\
&\quad= -\frac{1}{2}(x-\mu_{1,m})^\top\Sigma_m^{-1}(x-\mu_{1,m}) + \frac{1}{2}(x-\mu_{-1,m})^\top\Sigma_m^{-1}(x-\mu_{-1,m}) \\
&\quad= (\mu_{1,m}-\mu_{-1,m})^\top\Sigma_m^{-1}\left(x - \frac{\mu_{1,m}+\mu_{-1,m}}{2}\right).
\end{aligned}$$

Consequently,

$$h^\star_m(x) = \mathrm{sign}\left((\mu_{1,m}-\mu_{-1,m})^\top\Sigma_m^{-1}\left(x - \frac{\mu_{1,m}+\mu_{-1,m}}{2}\right) - \log\left(\frac{\pi_{m,-1}}{\pi_{m,1}}\right)\right), \tag{151}$$

which concludes the proof.

□

E.2 General lemmas for LDA misclassification control under Assumption 9.

Lemma E.1 ($\widehat{\mu}_m$ misclassification probability)**.** *Grant Assumption 9. Then,*

$$\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, M=m\right) = \Phi\left(-\frac{1}{2}\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1}-\mu_{m,-1})\right\|\right), \tag{152}$$

and

$$= \Phi\left(\frac{\left(\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1}-\widehat{\mu}_{m,-1})\right)^\top\Sigma_m^{-\frac{1}{2}}\left(\mu_{m,-1} - \frac{\widehat{\mu}_{m,1}+\widehat{\mu}_{m,-1}}{2}\right)}{\left\|\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1}-\widehat{\mu}_{m,-1})\right\|}\right) \tag{153}$$

Symmetrically,

$$\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \mid Y=1, M=m\right) = \Phi\left(-\frac{1}{2}\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1}-\mu_{m,-1})\right\|\right), \tag{154}$$

and

$$\mathbb{P}\left(\widehat{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1, M = m, \mathcal{D}_n\right)$$
$$= \Phi\left(-\frac{\left(\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1} - \widehat{\mu}_{m,-1})\right)^\top \Sigma_m^{-\frac{1}{2}}\left(\mu_{m,1} - \frac{\widehat{\mu}_{m,1}+\widehat{\mu}_{m,-1}}{2}\right)}{\left\|\Sigma_m^{-\frac{1}{2}}(\widehat{\mu}_{m,1} - \widehat{\mu}_{m,-1})\right\|}\right) \tag{155}$$

with Φ *the c.d.f. of a standard Gaussian distribution.*

Proof. Using Proposition 5.8, and recalling that the classes are balanced on each missing patterns ($\pi_{m,1} = \pi_{m,-1}$),

$$\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \mid Y = -1, M = m\right)$$
$$= \mathbb{P}\left((\mu_{m,1} - \mu_{m,-1})^\top \Sigma_m^{-1}\left(X_{\mathrm{obs}(m)} - \frac{\mu_{m,1} + \mu_{m,-1}}{2}\right) > 0 \mid Y = -1, M = m\right).$$

Let $N = \Sigma_m^{-\frac{1}{2}}(X_{\mathrm{obs}(m)} - \mu_{m,-1})$. By Assumption 9,

$$N|Y = -1, M = m \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \tag{156}$$

Letting $\gamma = \Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \mu_{m,-1})$, we have

$$\begin{aligned}\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \mid Y = -1, M = m\right) &= \mathbb{P}\left(\gamma^\top N - \frac{1}{2}\|\gamma\|^2 > 0 \mid Y = -1, M = m\right)\\ &= \mathbb{P}\left(\frac{\gamma^\top N}{\|\gamma\|} > \frac{1}{2}\|\gamma\| \mid Y = -1, M = m\right)\\ &= \Phi\left(-\frac{1}{2}\|\gamma\|\right).\end{aligned}$$

Similarly, using Proposition 5.8,

$$\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1, M = m\right)$$
$$= \mathbb{P}\left((\mu_{m,1} - \mu_{m,-1})^\top \Sigma_m^{-1}\left(X_{\mathrm{obs}(m)} - \frac{\mu_{m,1} + \mu_{m,-1}}{2}\right) < 0 \mid Y = 1, M = m\right).$$

Let $N = \Sigma_m^{-\frac{1}{2}}(X_{\mathrm{obs}(m)} - \mu_{m,1})$. By Assumption 9,

$$N|Y = 1, M = m \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \tag{157}$$

Letting $\gamma = \Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \mu_{m,-1})$, we have

$$\begin{aligned}\mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1, M = m\right) &= \mathbb{P}\left(\gamma^\top N + \frac{1}{2}\|\gamma\|^2 < 0 \mid Y = 1, M = m\right)\\ &= \mathbb{P}\left(\frac{\gamma^\top N}{\|\gamma\|} < -\frac{1}{2}\|\gamma\| \mid Y = 1, M = m\right)\\ &= \Phi\left(-\frac{1}{2}\|\gamma\|\right).\end{aligned}$$

This proves the first and third statements. Regarding the second and fourth statements, following the same strategy as in the proof of Corollary D.1, we have

$$\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = 1 \mid Y = -1, \mathcal{D}_n\right)$$
$$= \mathbb{P}\left((\widetilde{\mu}_{1,m} - \widetilde{\mu}_{-1,m})^\top \Sigma_m^{-1}\left(X_{\mathrm{obs}(m)} - \frac{\widetilde{\mu}_{1,m} + \widetilde{\mu}_{-1,m}}{2}\right) > 0 \mid Y = -1, \mathcal{D}_n\right)$$

Let $N = \Sigma_m^{-\frac{1}{2}}(X_{\mathrm{obs}(m)} - \mu_{-1,m})$. By Lemma F.6, $N|Y = -1 \sim \mathcal{N}(0, Id_{d-\|m\|_0})$. Since $(X_{\mathrm{obs}(m)}, Y)$ and $\mathcal{D}_n$ are independent

$$N|Y = -1, \mathcal{D}_n \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \tag{158}$$

Letting $\widetilde{\gamma} = \Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{1,m} - \widetilde{\mu}_{-1,m})$, we have

$$\begin{aligned}
&\mathbb{P}\left(h_m^\star(X_{\mathrm{obs}(m)}) = 1 \mid Y = -1, \mathcal{D}_n\right) \\
&= \mathbb{P}\left(\widetilde{\gamma}^\top N + \widetilde{\gamma}^\top \Sigma_m^{-\frac{1}{2}}\left(\mu_{-1,m} - \frac{\widetilde{\mu}_{1,m} + \widetilde{\mu}_{-1,m}}{2}\right) > 0 \mid Y = -1, \mathcal{D}_n\right) \\
&= \mathbb{P}\left(\frac{\widetilde{\gamma}^\top N}{\|\widetilde{\gamma}\|} > -\frac{\widetilde{\gamma}^\top}{\|\widetilde{\gamma}\|}\Sigma_m^{-\frac{1}{2}}\left(\mu_{-1,m} - \frac{\widetilde{\mu}_{1,m} + \widetilde{\mu}_{-1,m}}{2}\right) \mid Y = -1, \mathcal{D}_n\right) \\
&= \Phi\left(\frac{\widetilde{\gamma}^\top}{\|\widetilde{\gamma}\|}\Sigma_m^{-\frac{1}{2}}\left(\mu_{-1,m} - \frac{\widetilde{\mu}_{1,m} + \widetilde{\mu}_{-1,m}}{2}\right)\right).
\end{aligned}$$

Regarding the fourth statement, the proof is similar. Indeed,

$$\begin{aligned}
&\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1, \mathcal{D}_n\right) \\
&= \mathbb{P}\left((\widetilde{\mu}_{1,m} - \widetilde{\mu}_{-1,m})^\top \Sigma_m^{-1}\left(X_{\mathrm{obs}(m)} - \frac{\widetilde{\mu}_{1,m} + \widetilde{\mu}_{-1,m}}{2}\right) < 0 \mid Y = 1\right).
\end{aligned}$$

Let $N = \Sigma_m^{-\frac{1}{2}}(X_{\mathrm{obs}(m)} - \mu_{1,m})$. By Lemma F.6, and since $(X_{\mathrm{obs}(m)}, Y)$ and $\mathcal{D}_n$ are independent,

$$N|Y = 1, \mathcal{D}_n \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \tag{159}$$

Letting $\widetilde{\gamma} = \Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{1,m} - \widetilde{\mu}_{-1,m})$, we have

$$\begin{aligned}
&\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1, \mathcal{D}_n\right) \\
&= \mathbb{P}\left(\widetilde{\gamma}^\top N + \widetilde{\gamma}^\top \Sigma_m^{-\frac{1}{2}}\left(\mu_{1,m} - \frac{\widetilde{\mu}_{1,m} + \widetilde{\mu}_{-1,m}}{2}\right) < 0 \mid Y = 1, \mathcal{D}_n\right) \\
&= \mathbb{P}\left(\frac{\widetilde{\gamma}^\top N}{\|\widetilde{\gamma}\|} < -\frac{\widetilde{\gamma}^\top}{\|\widetilde{\gamma}\|}\Sigma_m^{-\frac{1}{2}}\left(\mu_{1,m} - \frac{\widetilde{\mu}_{1,m} + \widetilde{\mu}_{-1,m}}{2}\right) \mid Y = 1, \mathcal{D}_n\right) \\
&= \Phi\left(-\frac{\widetilde{\gamma}^\top}{\|\widetilde{\gamma}\|}\Sigma_m^{-\frac{1}{2}}\left(\mu_{1,m} - \frac{\widetilde{\mu}_{1,m} + \widetilde{\mu}_{-1,m}}{2}\right)\right).
\end{aligned}$$

□

Lemma E.2. *Grant Assumption 9. Assume that we are given two estimates $\widetilde{\mu}_1$ and $\widetilde{\mu}_{-1}$. Then, for all $m \in \mathcal{M}$, the classifier $\widetilde{h}_m$ defined in Equation* (12) *satisfies*

$$\begin{aligned}
&\left|\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = 1 \mid Y = -1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h_m^\star(X_{\mathrm{obs}(m)}) = 1 \mid Y = -1, M = m\right)\right| \\
&\quad \leq \frac{3}{2\sqrt{2\pi}}\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,-1} - \mu_{m,-1})\right\| + \frac{1}{2\sqrt{2\pi}}\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,1} - \mu_{m,1})\right\|
\end{aligned} \tag{160}$$

and symmetrically,

$$\begin{aligned}
&\left|\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h_m^\star(X_{\mathrm{obs}(m)}) = -1 \mid Y = 1, M = m\right)\right| \\
&\quad \leq \frac{3}{2\sqrt{2\pi}}\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,1} - \mu_{m,1})\right\| + \frac{1}{2\sqrt{2\pi}}\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,-1} - \widetilde{\mu}_{m,-1})\right\|
\end{aligned} \tag{161}$$

Proof. To prove Inequality (160), notice that, by Lemma E.1,

$$\left|\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, M=m, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, M=m\right)\right|$$

$$= \left|\Phi\left(\frac{\left(\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,1} - \widetilde{\mu}_{m,-1})\right)^\top \Sigma_m^{-\frac{1}{2}}\left(\mu_{m,-1} - \frac{\widetilde{\mu}_{m,1}+\widetilde{\mu}_{m,-1}}{2}\right)}{\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,1)} - \widetilde{\mu}_{m,-1})\right\|}\right)\right.$$
$$\left. - \Phi\left(-\frac{\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \mu_{m,-1})\right\|}{2}\right)\right|.$$

We can then apply the same steps as in the proof of Lemma D.3, and the result follows. The proof of Inequality (161) is similar. ☐

Lemma E.3. *Grant Assumption 9, with balanced classes. Assume that we are given two estimates $\widetilde{\mu}_1$ and $\widetilde{\mu}_{-1}$. Then, for all $m \in \mathcal{M}$, the classifier $\widetilde{h}_m$ defined in Equation* (7) *satisfies*

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

Proof.

$$\mathcal{R}_{\mathrm{mis}}(\widetilde{h}) - \mathcal{R}_{\mathrm{mis}}(h^\star)$$
$$= \mathbb{P}\left(\widetilde{h}(X_{\mathrm{obs}(M)}, M) \neq Y\right) - \mathbb{P}\left(h^\star(X_{\mathrm{obs}(M)}, M) \neq Y\right)$$
$$= \sum_{m\in\mathcal{M}}\left(\mathbb{P}\left(\widetilde{h}(X_{\mathrm{obs}(M)}, M) \neq Y \mid M=m\right) - \mathbb{P}\left(h^\star(X_{\mathrm{obs}(M)}, M) \neq Y \mid M=m\right)\right) p_m$$
$$= \sum_{m\in\mathcal{M}}\left(\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) \neq Y \mid M=m\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) \neq Y \mid M=m\right)\right) p_m \quad \text{(using (7))}$$
$$= \sum_{m\in\mathcal{M}} \pi_{m,-1}\left(\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, M=m\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, M=m\right)\right)$$
$$+ \sum_{m\in\mathcal{M}} \pi_{m,1}\left(\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y=1, M=m\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \mid Y=1, M=m\right)\right).$$

Note that

$$\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, M=m\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, M=m\right) \tag{162}$$
$$= \mathbb{E}\left[\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, M=m, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = 1 \mid Y=-1, M=m\right)\right] \tag{163}$$
$$\leq \frac{1}{2\sqrt{2\pi}}\mathbb{E}\left[3\left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,-1} - \widetilde{\mu}_{m,-1})\right\| + \left\|\Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \widetilde{\mu}_{m,1})\right\|\right], \tag{164}$$

according to Lemma E.2. Similarly,

$$\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y=1, M=m\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \mid Y=1, M=m\right) \tag{165}$$
$$= \mathbb{E}\left[\mathbb{P}\left(\widetilde{h}_m(X_{\mathrm{obs}(m)}) = -1 \mid Y=1, M=m, \mathcal{D}_n\right) - \mathbb{P}\left(h^\star_m(X_{\mathrm{obs}(m)}) = -1 \mid Y=1, M=m\right)\right] \tag{166}$$
$$\leq \frac{\pi_{m,-1}}{2\sqrt{2\pi}}\left(\mathbb{E}\left[3\left\|\Sigma_m^{-\frac{1}{2}}(-\widetilde{\mu}_{m,1} + \mu_{m,1})\right\| + \left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,-1} - \mu_{m,-1})\right\|\right]\right). \tag{167}$$

Consequently, since for all $m \in \mathcal{M}$, $\pi_{1,m} = \pi_{-1,m}$,

$$\begin{aligned}
&\mathcal{R}_{\mathrm{mis}}(\widetilde{h}) - \mathcal{R}_{\mathrm{mis}}(h^\star) \\
&\leq \sum_{m \in \mathcal{M}} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\left\| \Sigma_m^{-\frac{1}{2}} (-\widetilde{\mu}_{m,1} + \mu_{m,1}) \right\| + \left\| \Sigma_m^{-\frac{1}{2}} (\widetilde{\mu}_{m,-1} - \mu_{m,-1}) \right\| \right] \right) p_m.
\end{aligned}$$

□

E.3 Lemmas for Theorem 5.9

Lemma E.4. *Grant Assumption 9. Then, for all $k \in \{-1, 1\}$,*

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

where $\widetilde{\mu}_{m,k}$ is the estimate defined at (11).

Proof. We have

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

Since $\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \left(\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} - 1 \right) = 0$, we obtain

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

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

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

□

Lemma E.5. *Grant Assumption 9. Then, for all $k \in \{-1, 1\}$,*

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

where $\widetilde{\mu}_{m,k}$ is the estimate defined in (11).

Proof. By Jensen's inequality,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right\|\right] \\
&\leq \mathbb{E}\left[\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right\|^2\right]^{\frac{1}{2}} \\
&= \mathbb{E}\left[\operatorname{tr}\left(\left\|\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right\|^2\right)\right]^{\frac{1}{2}} \\
&= \mathbb{E}\left[\operatorname{tr}\left(\left(\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right)^\top \Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right)\right]^{\frac{1}{2}} \\
&= \mathbb{E}\left[\operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\left(\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})\right)^\top\right)\right]^{\frac{1}{2}} \\
&= \mathbb{E}\left[\operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}(\widetilde{\mu}_{m,k}-\mu_{m,k})(\widetilde{\mu}_{m,k}-\mu_{m,k})^\top\Sigma_m^{-\frac{1}{2}}\right)\right]^{\frac{1}{2}} \\
&= \operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}\mathbb{E}\left[(\widetilde{\mu}_{m,k}-\mu_{m,k})(\widetilde{\mu}_{m,k}-\mu_{m,k})^\top\right]\Sigma_m^{-\frac{1}{2}}\right)^{\frac{1}{2}} \\
&= \operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}\left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right]\Sigma_m+\mathbb{P}\left(\frac{N_{m,k}}{n}\leq\tau\right)\mu_{m,k}\mu_{m,k}^\top\right)\Sigma_m^{-\frac{1}{2}}\right)^{\frac{1}{2}} \\
&\hspace{20em}\text{(using Lemma E.4)} \\
&= \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right](d-\|m\|_0)+\mathbb{P}\left(\frac{N_{m,k}}{n}\leq\tau\right)\operatorname{tr}\left(\Sigma_m^{-\frac{1}{2}}\mu_{m,k}\mu_{m,k}^\top\Sigma_m^{-\frac{1}{2}}\right)\right)^{\frac{1}{2}} \\
&= \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n}>\tau}\frac{1}{N_{m,k}}\right](d-\|m\|_0)+\mathbb{P}\left(\frac{N_{m,k}}{n}\leq\tau\right)\left\|\Sigma_m^{-\frac{1}{2}}\mu_{m,k}\right\|^2\right)^{\frac{1}{2}}.
\end{aligned}
$$

□

E.4 Proof of Theorem 5.9

Proof. Let $A_\tau := \{m \in \{0,1\}^d | p_m < \tau\}$ be the set of missing pattern with occurrence probability smaller than τ. According to Lemma E.3 and Lemma E.5, we have

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

Now, for all $m \in A_\tau$, recalling that $\tau \geq \sqrt{d/n}$,

$$\sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{N_{m,k}} \right] (d - \|m\|_0) + \mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau \right) \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} \tag{168}$$

$$\leq \sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{n\tau} \right] (d - \|m\|_0) + \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} \tag{169}$$

$$\leq \sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \right] \tau + \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} \tag{170}$$

$$\leq \frac{2}{\sqrt{2\pi}} \left(1 + \frac{\|\mu_m\|^2}{\lambda_{\min}(\Sigma_m)} \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau}. \tag{171}$$

On the other hand, for all $m \notin A_\tau$,

$$\begin{aligned}
\mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau \right) &= \mathbb{P}\left(N_{m,k} \leq n\tau \right) \\
&= \mathbb{P}\left(\frac{\mathbb{1}_{N_{m,k} > 0}}{N_{m,k}^2} \geq \frac{1}{n^2 \tau^2} \right) + \mathbb{P}\left(N_{m,k} = 0 \right) \\
&\leq \frac{32 n^2 \tau^2}{p_m^2 (n+1)(n+2)} + (1 - p_m)^n \\
&\leq \frac{32 \tau^2}{p_m^2} + (1 - p_m)^n,
\end{aligned}$$

using Markov Inequality and Inequality (180). Then, for all $m \notin A_\tau$, we have

$$\sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left(\mathbb{E}\left[\mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{N_{m,k}} \right] (d - \|m\|_0) + \mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau \right) \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m \geq \tau} \tag{172}$$

$$\leq \frac{p_m \mathbb{1}_{p_m \geq \tau}}{\sqrt{2\pi}} \sum_{k=\pm 1} \left[\left(\mathbb{E}\left[\frac{\mathbb{1}_{\frac{N_{m,k}}{n} > \tau}}{N_{m,k}} \right] (d - \|m\|_0) \right)^{\frac{1}{2}} + \left(\mathbb{P}\left(\frac{N_{m,k}}{n} \leq \tau \right) \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} \right] \tag{173}$$

$$\leq \frac{1}{\sqrt{2\pi}} \sum_{k=\pm 1} \left[\left(\frac{4(d - \|m\|_0)}{p_m(n+1)} \right)^{\frac{1}{2}} + \left(\frac{32\tau^2}{p_m^2} + (1 - p_m)^n \right)^{1/2} \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\| \right] p_m \mathbb{1}_{p_m \geq \tau}$$

(using Inequality (178))

$$\leq \frac{4\tau \sqrt{p_m} \mathbb{1}_{p_m \geq \tau}}{\sqrt{2\pi}} + \left(\frac{4\tau}{\sqrt{\pi}} + \frac{1}{\sqrt{2\pi}} p_m (1 - p_m)^{n/2} \right) \mathbb{1}_{p_m \geq \tau} \sum_{k=\pm 1} \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\| \tag{174}$$

$$\leq \frac{4\tau \sqrt{p_m} \mathbb{1}_{p_m \geq \tau}}{\sqrt{2\pi}} + \left(\frac{4\tau}{\sqrt{\pi}} + \frac{1}{\sqrt{2\pi}} p_m (1 - p_m)^{n/2} \right) \frac{2\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}} \mathbb{1}_{p_m \geq \tau}. \tag{175}$$

Combining (171) and (175), we obtain

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

Since

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

we have

$$\begin{aligned}\mathcal{R}_{\text{mis}}(\widetilde{h})-\mathcal{R}_{\text{mis}}(h^{\star}) \leq & \sum_{m\in\{0,1\}^d}\left(\frac{4}{\sqrt{2\pi}}+\frac{8}{\sqrt{\pi}}\frac{\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}}\right)\tau\wedge p_m \\ & + \sum_{m\in\{0,1\}^d}\frac{\sqrt{2}\,\|\mu_m\|}{\sqrt{\pi\lambda_{\min}(\Sigma_m)}}p_m(1-p_m)^{n/2}\mathbb{1}_{p_m\geq\tau}. \end{aligned} \tag{176}$$

□

F Technical results

Lemma F.1 (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 $\lambda \in \mathbb{R}$,

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

Lemma F.2. *(Devroye et al., 2013, Lemma A2 p 587) Let $B \sim \mathcal{B}(p,n)$, we have*

$$\frac{1}{1+np} \leq \mathbb{E}\left[\frac{1}{1+B}\right] \leq \frac{1}{p(n+1)} \tag{177}$$

and

$$\mathbb{E}\left[\frac{\mathbb{1}\{B>0\}}{B}\right] \leq \frac{2}{p(n+1)}. \tag{178}$$

Proof.

- To prove the lower bound in (177), we use Jensen's inequality as follows:

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

- To prove the upper bound in (177), note that

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

using binomial formula.

- For (178), we use $1/x \leq 2/(x+1)$ for all $x \geq 1$ together with the previous result.

$\square$

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

Lemma F.3. *Given an $B \sim \mathcal{B}(n,p)$, we have that*

$$\mathbb{E}\left[\frac{1}{(1+B)^2}\right] \leq \frac{2}{(n+1)(n+2)p^2} \tag{179}$$

and

$$\mathbb{E}\left[\frac{\mathbb{1}_{B>0}}{B^2}\right] \leq \frac{8}{(n+1)(n+2)p^2} \tag{180}$$

Proof.

- In order to prove (179), note that

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

- Inequality (180) can be deduced using the fact that, for all $x \geq 1$, $1/x \leq 2/(x+1)$.

$\square$

Lemma F.4 (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. Rewrite the product of the matrices block-by-block, where $A_i \in \mathcal{M}_{n,1}(\mathbb{R})$ are the columns of A:

$$
\begin{aligned}
&\operatorname{tr}\left(\begin{bmatrix}A_1 & A_2 & A_3 & \cdots & A_n\end{bmatrix}\begin{bmatrix}b_1 & 0 & 0 & \cdots & 0\\ 0 & b_2 & 0 & \cdots & 0\\ 0 & 0 & b_3 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & b_n\end{bmatrix}\begin{bmatrix}A_1^\top\\ A_2^\top\\ A_3^\top\\ \vdots\\ A_n^\top\end{bmatrix}\right)\\
&= \operatorname{tr}\left(\begin{bmatrix}b_1A_1 & b_2A_2 & b_3A_3 & \cdots & b_nA_n\end{bmatrix}\begin{bmatrix}A_1^\top\\ A_2^\top\\ A_3^\top\\ \vdots\\ A_n^\top\end{bmatrix}\right)\\
&= \operatorname{tr}\left(\sum_{i=1}^n b_i A_i A_i^\top\right)\\
&= \sum_{i=1}^n b_i \operatorname{tr}\left(A_i A_i^\top\right)\\
&\leq C\sum_{i=1}^n \operatorname{tr}\left(A_i A_i^\top\right)\\
&= C\operatorname{tr}(A^2)
\end{aligned}
$$

□

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 F.5 (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

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

Lemma F.6 (Projection of a Gaussian vector)**.** *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,

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

□