



**HAL**  
open science

## A primer on linear classification with missing data

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

► **To cite this version:**

Angel D Reyero Lobo, Alexis Ayme, Claire Boyer, Erwan Scornet. A primer on linear classification with missing data. 2024. hal-04575204v2

**HAL Id: hal-04575204**

**<https://hal.science/hal-04575204v2>**

Preprint submitted on 9 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---

# A primer on linear classification with missing data

---

**Angel Reyero Lobo**  
Université Paul Sabatier

**Alexis Ayme**  
Sorbonne Université

**Claire Boyer**  
Université Paris-Saclay

**Erwan Scornet**  
Sorbonne Université

## Abstract

Supervised learning with missing data aims at building the best prediction of a target output based on partially-observed inputs. Major approaches to address this problem can be decomposed into (i) impute-then-predict strategies, which first fill in the empty input components and then apply a unique predictor and (ii) Pattern-by-Pattern (P-b-P) approaches, where a predictor is built on each missing pattern. In this paper, we theoretically analyze how three classical linear classifiers, namely perceptron, logistic regression and linear discriminant analysis (LDA), behave with Missing Completely At Random (MCAR) data, depending on the strategy (imputation or P-b-P) used to handle missing values. We prove that both imputation and P-b-P approaches are ill-specified in a logistic regression framework, thus questioning the relevance of such approaches to handle missing data. The most favorable auspices to perform classification with missing data concern P-b-P LDA methods. We provide finite-sample bounds for the excess risk in this framework, even for high-dimensional or MNAR settings. Experiments illustrate our theoretical findings.

**Keywords:** Linear discriminant analysis, Logistic regression, Bayes risk, 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.

**Estimation** 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 high-dimensional LDA frameworks with MCAR data have also been proposed (see, e.g., [Tony Cai and Zhang, 2019](#)).

**Regression** Prediction tasks with missing values differ from model estimation: estimated model parameters alone cannot directly predict outcomes on test samples containing missing values. Impute-then-predict strategies, often encountered in practice, consists in imputing 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 anal-

ysis 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 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.

**Classification** Regarding classification, in fact, few theoretical analyses exist on prediction with missing data. Pelckmans et al. (2005) adapted Support Vector Machine (SVM) classifiers to accommodate missing values. Sell et al. (2023) establish minimax rate for non-parametric prediction with missing values and propose a minimax algorithm called HAM.

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. Jiang et al. (2020) is one of the few methods able to estimate parameter and predict in presence of missing values. Recently, a strong interest has been put on comparing the empirical performances of various imputation strategies (Bertsimas et al., 2018; Poulos and Valle, 2018; Jäger et al., 2021). 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** While predictions in linear regression with missing data has been extensively studied (see paragraph *regression* above), to the best of our knowledge, there exists no theoretical work analyzing the benefits of Pattern-by-Pattern (P-b-P) and impute-then-predict strategies on linear *classifiers*. In this paper, we theoretically study the validity of these two approaches with linear classifiers for MCAR missing data. First, we analyze the perceptron model (Section 3) and prove that both strategies to handle missing values are likely to fail, due to the intrinsic difficulty of maintaining the linear separability on each missing pattern. Since linear separability is a strong assumption, we turn to the widely-used logistic regression model. We prove that this model is ill-specified to handle MCAR data (Section 4), whether imputation or P-b-P strategies are used. As modelling the joint distribution of input-output seems to be unavoidable, we turn into the linear discriminant analysis (LDA)

framework (Section 5). We prove that imputation is not optimal for LDA but P-b-P is. We establish finite-sample bounds on the excess risk of LDA, in high-dimensional regimes, also allowing missing data to be more complex (MNAR). Experiments (Section 6) illustrate our findings.

## 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  $h : \mathbb{R}^d \rightarrow \{-1, 1\}$  is given by the probability of misclassification

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

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

**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  $j$ th 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  $\text{obs}(m)$  (resp.  $\text{mis}(m)$ ) as the set of indices where  $m$  is 0 (resp. 1), representing the observed (resp. missing) variables. Subsequently,  $X_{\text{obs}(M)}$  (resp.  $X_{\text{mis}(M)}$ ) 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  $Z := (X_{\text{obs}(M)}, M)$  belonging to  $\mathcal{Z}$ . In presence of missing data, the performance of a classifier  $h : \mathcal{Z} \rightarrow \{-1, 1\}$  is evaluated via

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

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

$$h_{\text{mis}}^*(Z) = \sum_{m \in \mathcal{M}} h_m^*(X_{\text{obs}(m)}) \mathbf{1}_{M=m}, \quad (1)$$

with  $h_m^*(X_{\text{obs}(m)}) = \text{sign}(\mathbb{E}[Y|X_{\text{obs}(m)}, M = m])$ , where  $\mathcal{M} \subset \{1, \dots, d\}$  is the set of admissible missing patterns. Learning with missing values can be seen as estimating  $h_m^*$  for all  $m \in \{0, 1\}^d$ , given an incomplete i.i.d. training sample  $\mathcal{D}_n := \{(X_{i,\text{obs}(M_i)}, M_i, Y_i), i = 1, \dots, n\}$ .

### 3 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 the perceptron algorithm in the presence of missing values.

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

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

$$Y_i \left( (w_{(m)}^*)^\top (1 - M_i) \odot X_i + b_{(m)}^* \right) > 0.$$

*Remark 3.1* (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 3.2.** *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. Thus, both imputation and P-b-P approaches fail in all generality, as the separability of incomplete data is required for the convergence of the perceptron algorithm. In the following, we make additional assumptions on the input distribution (adapted to the perceptron model), to highlight favorable cases of predictor adaptability to missing inputs.

**Assumption 1** (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  $R_1, R_2 \sim \mathcal{U}(0, \frac{1}{2} \|c_1 - c_2\|_2)^{\otimes 2}$ .*

**Assumption 2** (MCAR).  $M \perp\!\!\!\perp (X, Y)$ .

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.

**Proposition 3.3** (Separability of two balls with different radius). *Grant Assumption 1 and Assumption 2 (MCAR). For all  $1 \leq j \leq d$ , let  $\eta_j := \mathbb{P}(M_j = 1)$ . Then*

$$\alpha \leq \mathbb{P}(B_{1,\text{obs}(M)} \cap B_{2,\text{obs}(M)} = \emptyset) \leq \sqrt{\alpha}, \quad (2)$$

with  $\alpha = \langle 1 - \eta, (c_1 - c_2)^2 \rangle / \|c_1 - c_2\|_2^2 \leq 1$ .

The lower bound in (2) is informative when the probability of missing values on each coordinate remains low. When for any coordinate  $j$ ,  $\eta_j = \eta$ , the bounds (2) become independent of the centroids:

$$(1 - \eta) \leq \mathbb{P}(B_{1,\text{obs}(M)} \cap B_{2,\text{obs}(M)} = \emptyset) \leq \sqrt{1 - \eta}.$$

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

$$1 - \frac{(c_{1,j_0} - c_{2,j_0})^2}{\|c_1 - c_2\|_2^2} \leq \mathbb{P}(B_{1,\text{obs}(M)} \cap B_{2,\text{obs}(M)} = \emptyset). \quad (3)$$

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.

While some positive results can be obtained with high probability in some specific settings (with few missing data, see above, or in high-dimensional settings, see Appendix B.4), linear separability for each missing pattern is impossible to obtain without additional assumptions. Thus, assuming only the linear separability of complete data, a P-b-P approach is not Bayes optimal and nor is the constant imputation, as shown below.

**Lemma 3.4.** *If a P-b-P approach with linear classifiers is not Bayes optimal, then constant imputation with linear classifiers is not Bayes optimal.*

### 4 Logistic Regression

Since linear separability is difficult to maintain for each missing pattern, we now relax this assumption by modelling the distribution of  $Y|X$ , following the popular logistic regression framework.

**Assumption 3** (Logistic model). *Let  $\sigma(t) = 1/(1 + e^{-t})$ . There exist  $\beta_0^*, \dots, \beta_d^* \in \mathbb{R}$  such that the distribution of the output  $Y \in \{-1, 1\}$  given the complete input  $X$  satisfies  $\mathbb{P}(Y = 1|X) = \sigma(\beta_0^* + \sum_{j=1}^d \beta_j^* X_j)$ .*

Assuming that the logistic model holds on complete data, one could be tempted to use a P-b-P approach using a different logistic regression model for each missing pattern. Proposition 4.1 below shows that such a strategy is doomed to fail when input variables are independent and missing data uninformative (MCAR).

**Proposition 4.1.** *Grant Assumption 2 (MCAR) and Assumption 3 for complete data. Furthermore, assume that the components  $X_1, \dots, X_d$  are independent, each one with an unbounded support, satisfying  $\mathbb{E}[\exp(\beta_j^* X_j)] < \infty$ . Let  $m \in \{0, 1\}^d$  and assume that the logistic model holds on the missing pattern  $M = m$ , that is there exist a vector  $\beta_m^* \in \mathbb{R}^{|\text{obs}(m)|+1}$  such that*

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

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

Proposition 4.1 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 probability estimators. Note that such a result highlights that constant imputation is also an ill-specified strategy, 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 can be preserved when the inputs are partially observed (see, e.g., Le Morvan et al., 2020b; Ayme et al., 2022), logistic models are not suited for missing data, assuming in both settings independent input variables with MCAR missingness. Indeed, due to the non-linearity,

$$\mathbb{E}[Y|X_{\text{obs}(M)}, M = m] \neq \sigma\left(\mathbb{E}[\beta_0^*] + \sum_{j=1}^d \beta_j^* \mathbb{E}[X_j|X_{\text{obs}(M)}]\right).$$

Therefore, the logistic model is not preserved on missing patterns and resulting P-b-P/imputation strategies are ill-posed, as shown in Proposition 4.1 for independent input. Thus, modelling the conditional distribution  $Y|X$  without additional constraints on the distribution of  $X$  appears to be insufficient to ensure the soundness of P-b-P/imputation strategies. This is therefore the direction we will explore in the next section.

## 5 Linear Discriminant Analysis with missing data

In this section, we analyze the finite-sample property of pattern-by-pattern LDA.

### 5.1 Bayes optimality

Linear discriminant analysis (LDA) relies on Gaussian assumptions of the distributions of  $X|Y = k$  for each class  $k$ .

**Assumption 4** (Balanced LDA). *Let  $\Sigma$  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  $\mu_k \in \mathbb{R}^d$ .*

LDA can be seen as a specific case of logistic regression, with well-chosen coefficients and a specific distribution on  $X$  (mixture of two Gaussian components) as detailed in Lemma D.1, thus restricting the setting of Section 4. Notably, in the complete case of LDA, the Bayes predictor reads

$$h_{\text{comp}}^*(x) := \text{sign}\left(\left(\mu_1 - \mu_{-1}\right)^\top \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_{-1}}{2}\right)\right), \quad (4)$$

minimizing the misclassification probability  $\mathcal{R}_{\text{comp}}$  (see Section D.1 for details). When MCAR data occurs, by denoting  $\Sigma_{\text{obs}(M)} := \Sigma_{\text{obs}(M) \times \text{obs}(M)}$  (and  $\Sigma_{\text{obs}(M)}^{-1} = (\Sigma_{\text{obs}(M)})^{-1}$ ), 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 4 (LDA), the pattern-by-pattern Bayes classifier is given by*

$$h_m^*(x_{\text{obs}(m)}) = \text{sign}\left(\left(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}\right)^\top \Sigma_{\text{obs}(m)}^{-1} \times \left(x_{\text{obs}(m)} - \frac{\mu_{1,\text{obs}(m)} + \mu_{-1,\text{obs}(m)}}{2}\right)\right).$$

The decomposition provided in Proposition 5.1 relies on the fact that, under MCAR assumption, the distribution of  $X_{\text{obs}(M)}|Y, M = m$  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. Therefore, in the following, we first consider MCAR missing data.

*Example 5.2* (LDA+MAR is not pattern-by-pattern LDA). Let  $X \in \mathbb{R}^2$  be a random variable satisfying Assumption 4, 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  $X_{\text{obs}(M)}|Y=k, M=m$ , for the pattern  $m = (0, 1)$ , is not Gaussian, as its first component is always positive.

According to Proposition 5.1, Pattern-by-Pattern LDA is Bayes optimal for any choice of the covariance matrix  $\Sigma$  and any signal  $\mu_1 - \mu_{-1}$ . This contrasts with perceptron and logistic framework for which the P-b-P strategy was not Bayes optimal or well-defined.

On the other hand, Proposition 5.3 below shows that constant imputation strategies are not Bayes optimal in all LDA frameworks. Therefore, we focus solely on the pattern-by-pattern strategy in the sequel.

**Proposition 5.3.** *Under Assumptions 2 (MCAR) and 4 (LDA) with  $d \geq 3$ , constant imputation is Bayes optimal for all values of  $\mu_{-1}$  and  $\mu_1$  if and only if  $\Sigma$  is diagonal. In this case, the optimal imputation is given by  $\alpha_j = (\mu_{1,j} - \mu_{-1,j})/2$  for all  $1 \leq j \leq d$ .*

## 5.2 Finite-sample bounds

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  $\mathcal{R}_{\text{mis}}(h^*) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*)$ .

**Assumption 5** (Constant  $\mathbb{P}(M_j = 1)$ ). *The random variables  $M_1, \dots, M_d$  are independent, and follow a Bernoulli distribution with parameter  $\eta$ .*

**Assumption 6** (Constant  $(\mu_1 - \mu_{-1})_j$ ).  $\forall j \in \{1, \dots, d\}, (\mu_1 - \mu_{-1})_j = \pm\mu$ , with  $\mu > 0$ .

Assumption 5 ensures that the missingness probability is the same for each input coordinate. Assumption 6 can be achieved up to a change of coordinates. In the sequel, we refer to  $\text{SNR} := \mu/\sqrt{\lambda_{\max}(\Sigma)}$  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 5.4.** *Under Assumptions 2, 4, 5 and 6,*

$$\begin{aligned} & \mathcal{R}_{\text{mis}}(h^*) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) \\ & \leq \frac{\eta^d}{2} + \frac{\mu\eta}{2\sqrt{2\pi}} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left( \epsilon(\eta, \text{SNR})^{d-1} - \eta^{d-1} \right), \end{aligned}$$

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

The bound provided in Proposition 5.4 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 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 5.5.** *Under Assumptions 2, 4, 5, 6,*

$$\lim_{\text{SNR} \rightarrow \infty} \sqrt{\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}} \frac{\text{SNR}}{e^{\text{SNR}^2/8}} = 0$$

implies  $\lim_{\text{SNR} \rightarrow \infty} \mathcal{R}_{\text{mis}}(h^*) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) = \frac{\eta^d}{2}$ .

The limit established in Corollary 5.5 matches that of the limit of the bound of Proposition 5.4 when the SNR tends to infinity. It is important to note that the assumption on the structure of  $\Sigma$  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  $\Sigma$  is arbitrary but constant, with increasingly separated classes.

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

$$\hat{\mu}_{k,j} = \frac{\sum_{i=1}^n X_{i,j} \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}} \quad (5)$$

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

$$\begin{aligned} \hat{h}_m(x_{\text{obs}(m)}) &= \text{sign} \left( \left( \hat{\mu}_{1,\text{obs}(m)} - \hat{\mu}_{-1,\text{obs}(m)} \right)^\top \Sigma^{-1} \right. \\ & \quad \left. \times \left( x_{\text{obs}(m)} - \frac{\hat{\mu}_{1,\text{obs}(m)} + \hat{\mu}_{-1,\text{obs}(m)}}{2} \right) \right). \end{aligned} \quad (6)$$

Remark that under MCAR assumption, the estimates  $\hat{\mu}_{k,j}$  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  $\Sigma$ .

**Theorem 5.6** (Bound on P-b-P LDA with known  $\Sigma$ ). *Grant Assumptions 2, 4 and 5. Then the excess risk of the classifier  $\hat{h}$ , defined in (6), satisfies*

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

Then, for  $n$  large enough, we have

$$\mathcal{R}_{\text{mis}}(\hat{h}) - \mathcal{R}_{\text{mis}}(h^*) \lesssim \sqrt{\kappa d/n}. \quad (7)$$

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  $\Sigma$  is mild, since the respective term decreases exponentially (corresponding to the case where all data are missing).

The upper bound presented in (7) is independent of the missingness probability  $\eta$ . 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 5.7.** *Grant Assumptions 2, 4, 5, 6. Then the classifier  $\hat{h}$ , defined in (6) satisfies*

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

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

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

$$n \ll \frac{1}{(\eta \cdot \text{SNR})^2} \frac{1}{\epsilon(\eta, \text{SNR})^d}, \quad (8)$$

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,  $\mathcal{R}_{\text{mis}}(\hat{h}) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) = O(\sqrt{d/n})$ , 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 5.8* (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  $\hat{h}_{\text{LPD}}$  on complete data which satisfies  $\mathcal{R}_{\text{comp}}(\hat{h}_{\text{LPD}}) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) = O((s \log(d)/n)^{1/2})$ . Although Tony Cai and Zhang (2019) follow a completely different approach, their estimator applied on complete data reaches the same rate of convergence.

### 5.3 Extension to MNAR settings

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

Under Assumption 7, the Bayes predictor can be decomposed pattern by pattern as follows.

**Proposition 5.9** (MNAR P-b-P LDA). *Under Assumption 7, the pattern-by-pattern Bayes classifier is*

$$\begin{aligned} h_m^*(x_{\text{obs}(m)}) &= \text{sign} \left( (\mu_{m,1} - \mu_{m,-1})^\top \Sigma_m^{-1} \right. \\ & \quad \left. \times \left( x_{\text{obs}(m)} - \frac{\mu_{m,1} + \mu_{m,-1}}{2} \right) - \log \left( \frac{\pi_{m,-1}}{\pi_{m,1}} \right) \right). \end{aligned}$$

Given the expression of the Bayes predictor in Proposition 5.9, we build a plug-in estimate based on the estimation  $\hat{\mu}_{m,k}$  of the mean  $\mu_{m,k}$  on pattern  $m \in \{0, 1\}^d$  and class  $k$ , defined as

$$\hat{\mu}_{m,k} := \frac{\sum_{i=1}^n X_i \mathbf{1}_{Y_i=k} \mathbf{1}_{M_i=m}}{\mathbf{1}_{Y_i=k} \mathbf{1}_{M_i=m}}. \quad (9)$$

Due to the potential exponential number of missing patterns, it may be difficult to estimate the  $2^{d+1}$  estimates  $\hat{\mu}_{m,k}$ . 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

$$\tilde{\mu}_{m,k} := \hat{\mu}_{m,k} \mathbf{1}_{\frac{N_{m,k}}{n} > \tau}, \quad (10)$$

with  $\tau := \sqrt{d/n}$  and  $N_{m,k} := \sum_{i=1}^n \mathbf{1}_{M_i=m} \mathbf{1}_{Y_i=k}$  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  $\tilde{h}$  defined as

$$\begin{aligned} \tilde{h}_m(x_{\text{obs}(m)}) &= \text{sign}\left(\left(\tilde{\mu}_{1,\text{obs}(m)} - \tilde{\mu}_{-1,\text{obs}(m)}\right)^\top \Sigma_m^{-1} \right. \\ &\quad \left. \times \left(x_{\text{obs}(m)} - \frac{\tilde{\mu}_{1,\text{obs}(m)} + \tilde{\mu}_{-1,\text{obs}(m)}}{2}\right)\right). \end{aligned} \quad (11)$$

**Theorem 5.10** (MNAR P-b-P LDA estimation). *Grant Assumption 7 and assume that the classes are balanced on each missing pattern. Let  $\tau \geq \sqrt{d/n}$ . Then, the plug-in classifier based on (11) satisfies*

$$\begin{aligned} &\mathcal{R}_{\text{mis}}(\tilde{h}) - \mathcal{R}_{\text{mis}}(h^*) \\ &\leq \sum_{m \in \{0,1\}^d} \left( \frac{4}{\sqrt{2\pi}} + \frac{8}{\sqrt{\pi}} \frac{\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}} \right) \tau \wedge p_m \\ &\quad + \sum_{\substack{m \in \{0,1\}^d, \\ p_m \geq \tau}} \frac{\sqrt{2}\|\mu_m\|}{\sqrt{\pi\lambda_{\min}(\Sigma_m)}} p_m (1 - p_m)^{n/2}. \end{aligned} \quad (12)$$

Theorem 5.10 holds for various types of missingness. Indeed, Assumption 7 is very generic and may correspond to very difficult MNAR settings in which there is no relation between any covariances matrices  $\Sigma_m$  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  $\mu_{-1}, \mu_1, \Sigma$  such that  $\mu_{\pm 1, m} = \mu_{\pm 1, \text{obs}(m)}$  and  $\Sigma_m = \Sigma_{\text{obs}(m)}$  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 5.10 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 variable with parameter  $\eta \leq d/n$ . In such a setting, even if each missing pattern is admissible,

$$\mathcal{R}_{\text{mis}}(\tilde{h}) - \mathcal{R}_{\text{mis}}(h^*) \lesssim \frac{d^2}{n} + (1 - \min_{p_m > 0} p_m)^{n/2}, \quad (13)$$

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 5.7, we do not compare  $\mathcal{R}_{\text{mis}}(\tilde{h})$  to  $\mathcal{R}_{\text{comp}}(h_{\text{comp}}^*)$  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 7, the distribution of the complete pattern (corresponding to  $m = 0$ ) may be chosen independently of the other distributions ( $m \neq 0$ ). Thus, the difference  $\mathcal{R}_{\text{mis}}(\tilde{h}) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*)$  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 Experiments

**Simulated data.** We run experiments with an input dimension  $d = 5$ . We simulate data according to the following models:

1. the LDA framework (Assumption 4) with  $\Sigma$  equal to  $I_d$  or Toeplitz of the form  $(0.6^{|i-j|})_{ij}$ , and where  $\mu_1$  and  $\mu_{-1}$  are such that  $\mu_1 - \mu_{-1}$  is a vector with equal non-zero components. More precisely, we first sample  $\mu_0 \sim \mathcal{N}(0, \sigma^2 I_d)$  with  $\sigma^2 = 25$ , and then  $\mu_1 | \mu_0 \sim \mu_0 + 1.5\epsilon$ , with each coordinate of  $\epsilon$  being an independent Rademacher.
2. a logistic setting in which  $X \sim \mathcal{N}(0, \Sigma)$  where  $\Sigma$  is as above and  $Y$  follows a logistic model defined in Assumption 3 with  $\beta^* \sim \mathcal{N}(0, I_d)$ .

Missing values are either MCAR with the same probability of missingness  $\eta = 0.5$  on each component or self-masking MNAR (the missingness probability only depends on its value) that is  $\mathbb{P}(M_j = 1 | X) = \frac{1}{1 + \exp(-X^j)}$ , where coefficients have been chosen to ensure a probability of missingness close to  $\eta$ .

**Algorithms.** For prediction purposes, we assess the performances of the following algorithms on simulated data:

- Logistic regression with Pattern-by-pattern strategy (pbp LogReg), imputed-by-0 data (0-imp+LogReg) or ICE-imputed data (ICE-imp+LogReg); ICE imputation refers to single imputation by chained equations ([Van Buuren and Groothuis-Oudshoorn, 2011](#)) implemented as `IterativeImputer` in scikit-learn ([Pedregosa et al., 2011](#)).
- Perceptron with Pattern-by-pattern strategy (pbp Perceptron), imputed-by-0 data (0-imp+Perceptron) or ICE-imputed data (ICE-imp+Perceptron);
- LDA with Pattern-by-pattern strategy (pbp LDA), imputed-by-0 data (0-imp+LDA) or (ICE-imp+LDA);
- LDA described by Equation (7) and denoted by LDA-MCAR.



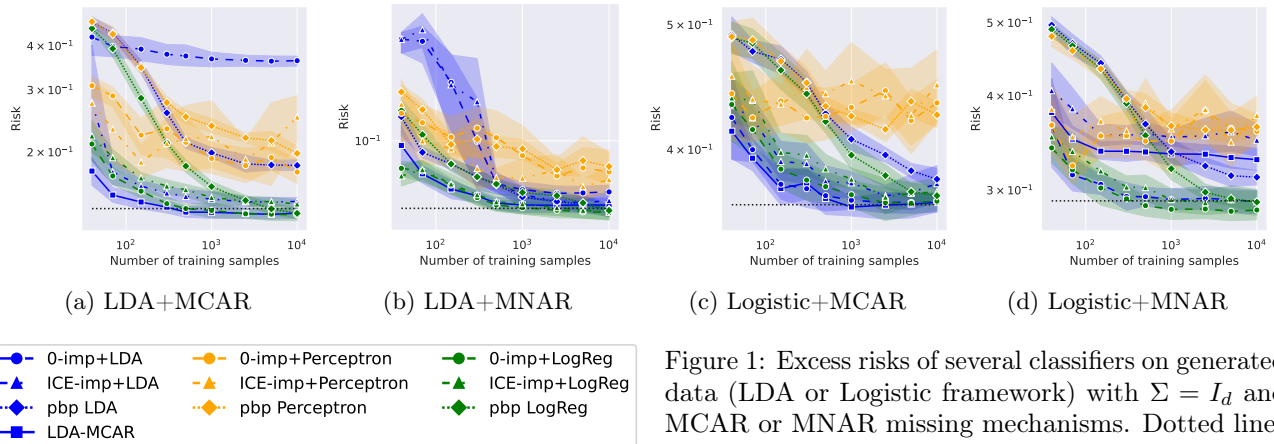


Figure 1: Excess risks of several classifiers on generated data (LDA or Logistic framework) with  $\Sigma = I_d$  and MCAR or MNAR missing mechanisms. Dotted lines stand for the Bayes risk  $\mathcal{R}_{\text{mis}}(h_{\text{mis}}^*)$  with missing data.

## 6.1 Results

Results are displayed in Figure 1 for data generated via LDA or Logistic framework ( $\Sigma = I_d$ ) and MCAR or MNAR missingness. Experiments with a Toeplitz covariance matrix can be found in Appendix G.

In Figure 1a and Figure 1b, we observe that P-b-P strategies show bad performances for small sample sizes. Indeed, in this case, few samples are available in each missing pattern, which makes it difficult (or even impossible) to train each P-b-P classifier. Note also that Perceptron-based methods are not very accurate, since such methods do not converge for non-linearly separable data.

In Figure 1a, we see that LDA-MCAR is among the best methods, in terms of predictive accuracy. Notably, P-b-P LDA requires a significantly larger training sample compared to LDA-MCAR, as it does not use information from other missing patterns. Surprisingly, imputation by zero followed by a logistic classifier is on par with LDA-MCAR with MCAR data and outperforms this method for MNAR data (Figure 1b).

In Figure 1c and Figure 1d, we first note that all methods have poor performances, with an excess risk around 0.3 with  $n = 10,000$  samples. Such bad performances highlight that classification with logistic data and missing values is a challenging task, despite its simplistic nature. Trying to distinguish methods in this difficult scenario, we observe that zero-imputation with logistic regression and LDA-MCAR are among the best classifiers both in MCAR and MNAR scenario (even if data do not satisfy LDA assumptions).

## 7 Conclusion

In this paper, we analyze the two main classes of methods able to handle missing values for classification:

Pattern-by-Pattern (P-b-P) and imputation. We prove that both imputation and P-b-P strategies are theoretically deficient for perceptron or logistic classifiers. In the LDA framework, P-b-P is Bayes optimal in all generality, whereas imputation is for diagonal covariance matrices only. We establish finite sample bounds for the excess risk of LDA classifier with MCAR data, which vanishes as  $n$  and  $d$  grows at some prescribed rates. We extend our analysis to specific MNAR scenarios. Experiments illustrate our theoretical findings, showing that P-b-P LDA outperforms other classifiers for LDA-generated data with MCAR missingness. Surprisingly, despite its ill-posedness, imputation with logistic regression exhibits good performances in this setting. A proper analysis of this phenomenon, by identifying favorable underlying behaviors during training, remains to be conducted. In this regard, 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 coping 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 Do-kyoon 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, Colin Pawlowski, and Ying Daisy Zhuo. From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Research*, 18(196):1–39, 2018.
- 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.
- Tlameo 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.
- Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. A benchmark for data imputation methods. *Frontiers in big Data*, 4:693674, 2021.
- 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, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What’s a 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.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- 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.
- Jason Poulos and Rafael Valle. Missing data imputation for supervised learning. *Applied Artificial Intelligence*, 32(2):186–196, 2018.
- 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. Canning. 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>.
- Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

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

$$h^*(Z) = \sum_{m \in \mathcal{M}} h_m^*(X_{\text{obs}(m)}) \mathbf{1}_{M=m},$$

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

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

$$\mathcal{R}_{\text{mis}}(h) := \mathbb{P}(Y \neq h(Z)) \tag{14}$$

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

$$\mathcal{R}_{\text{mis}}(h) = \frac{1}{4} \mathbb{E}[(Y - h(Z))^2] = \frac{1}{4} \mathbb{E}[(Y - \mathbb{E}[Y|Z])^2] + \frac{1}{4} \mathbb{E}[(\mathbb{E}[Y|Z] - h(Z))^2]. \tag{15}$$

Thus, the Bayes predictor is

$$h^*(Z) := \text{sign}(\mathbb{E}[Y|Z]) = \text{sign}(\mathbb{E}[Y|X_{\text{obs}(M)}, M]) \text{ where } \text{sign}(x) = \mathbf{1}_{x \geq 0} - \mathbf{1}_{x < 0}. \tag{16}$$

As we have that

$$\mathbb{E}[Y|X_{\text{obs}(M)}, M] = \sum_{m \in \mathcal{M}} \mathbb{E}[Y|X_{\text{obs}(m)}, M = m] \mathbf{1}_{M=m}, \tag{17}$$

then, the Bayes predictor can be written as

$$\begin{aligned} h^*(Z) &= \text{sign}(\mathbb{E}[Y|Z]) \\ &= \text{sign} \left( \sum_{m \in \mathcal{M}} \mathbb{E}[Y|X_{\text{obs}(m)}, M = m] \mathbf{1}_{M=m} \right) \\ &= \sum_{m \in \mathcal{M}} \text{sign}(\mathbb{E}[Y|X_{\text{obs}(m)}, M = m]) \mathbf{1}_{M=m} \\ &= \sum_{m \in \mathcal{M}} h_m^*(X_{\text{obs}(m)}) \mathbf{1}_{M=m} \end{aligned} \tag{18}$$

with

$$h_m^*(X_{\text{obs}(m)}) := \text{sign}(\mathbb{E}[Y|X_{\text{obs}(m)}, M = m]). \tag{19}$$

□

## B (Perceptron) Proofs of Section 3

### B.1 Proof of Lemma 3.2

*Proof.* Suppose that we only have two points  $X_1, X_2 \in \mathbb{R}^d$  where  $x_2 = (x_{1,1}, \dots, x_{1,(k-1)}, x_{2,k}, x_{1,(k+1)}, \dots, x_{1,d})$  with  $x_{1,k} \neq x_{2,k}$ . We have  $y_2 = -y_1$ . We also suppose that  $m_{1,k} = m_{2,k} = 1$  and  $m_1 = m_2$ . Then,  $\mathcal{W}$  is not empty, but  $\mathcal{W}_{\text{mis}}$  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. □

---

## B.2 Separability characterization

**Lemma B.1** (Separability characterization). *Consider the  $\ell^p$ -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,

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

□

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

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

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

## B.3 Proof of Proposition 3.3

*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,

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

as shown in Lemma B.1. We have that

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

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

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

Moreover, note that

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

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

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

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

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

□

#### B.4 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  $\ell^p$ -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 8.** We assume that (i) the coordinates of  $C_1 - C_2$  are i.i.d., (ii) for all  $j \in \{1, \dots, d\}$ ,  $\mathbb{E}[(C_1 - C_2)_j^p] < \infty$  and (iii) conditional to the centers  $C_1$  and  $C_2$ , the radii  $R_1$  is uniformly distributed as  $R_1 | (C_1, C_2) \sim \mathcal{U}(0, \frac{1}{2} \|C_1 - C_2\|_p)$ , with  $R_2 = R_1$ .

Assumption 8 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 9** (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  $\infty$ .

**Proposition B.2** (Asymptotic separability of two balls with the same radius). *Under Assumption 8 and Assumption 9, let  $\rho := \lim_{d \rightarrow \infty} \frac{s}{d}$ . Then,*

$$\lim_{d \rightarrow +\infty} \mathbb{P}(B_{1, \text{obs}(M)} \cap B_{2, \text{obs}(M)} = \emptyset) = \sqrt[p]{1 - \rho}. \quad (20)$$

Therefore, in high-dimensional regimes, pattern-by-pattern perceptron is a valid procedure with a probability converging to  $\sqrt[p]{1 - \rho}$ , where  $\rho$  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 shrunk ( $p \mapsto \|x\|_p$  is non-increasing) and the balls are more and more separated.

## B.5 Proof of Proposition B.2

*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 B.1. Since  $R := R_1 = R_2$ , this probability corresponds to

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

Using Assumption 8, we have that

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

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

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

after having reordered the terms using the exchangeability of the  $(C_1 - C_2)_j$  (Assumption 8(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  $\rho = \lim_{d \rightarrow \infty} \frac{s}{d} = 0$  and  $\frac{s}{d-s} \frac{1}{s} \sum_{j=d-s+1}^d (C_{1j} - C_{2j})^p \xrightarrow{d \rightarrow \infty} 0$ , so we would have the result using that

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

Then, combining Assumption 8 and the law of large numbers, we get

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

Using Slutsky's theorem,

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

Re-using Slutsky's theorem,

$$\frac{\frac{s}{d-s} \frac{1}{s} \sum_{j=d-s+1}^d (C_{1j} - C_{2j})^p}{\frac{1}{d-s} \sum_{j=1}^{d-s} (C_{1j} - C_{2j})^p} \xrightarrow{d \rightarrow \infty} \frac{\rho}{1 - \rho}.$$

Finally, using the continuous mapping theorem, we have that

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

□

## B.6 Link between non-optimality of P-b-P and constant imputation approaches

*Proof of Lemma 3.4.* Recall that  $Z := (X_{\text{obs}(M)}, M)$ . A Pattern-by-Pattern (P-b-P) approach with linear classifier  $h_{\text{PbP}}$  is defined as,

$$h_{\text{PbP}}(Z) = \sum_{m \in \mathcal{M}} h_m(X_{\text{obs}(m)}) \mathbf{1}_{M=m},$$

where, for all  $m \in \{0, 1\}^d$ , there exists  $\beta_{j,m}$ , such that

$$h_m(X_{\text{obs}(m)}) = \beta_{0,m} + \sum_{j \in \text{obs}(m)} \beta_{j,m} X_j \tag{21}$$

$$= \beta_{0,m} + \sum_{j=1}^d \beta_{j,m} X_j \mathbf{1}_{M_j=0}. \tag{22}$$



On the contrary, an imputation strategy that consists in replacing missing values in  $X_j$  by  $\alpha_j$  for all  $1 \leq j \leq d$  can be written as

$$h_{\text{imp}}(Z) = \beta_0 + \sum_{j=1}^d \beta_j (X_j \mathbb{1}_{M_j=0} + \alpha_j \mathbb{1}_{M_j=1}) \quad (23)$$

$$= \beta_0 + \sum_{j=1}^d \beta_j \alpha_j \mathbb{1}_{M_j=1} + \sum_{j=1}^d \beta_j X_j \mathbb{1}_{M_j=0}. \quad (24)$$

Thus, comparing (22) and (24), we see that a constant imputation strategy with linear classifiers can always be written as a P-b-P approach with linear classifiers. Thus, assuming that P-b-P approach with linear classifiers are not Bayes optimal, we conclude that constant imputation strategies with linear classifiers are not Bayes optimal either.  $\square$

### C (Logistic Model) Proof of Proposition 4.1

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

$$\begin{aligned} \mathbb{P}(Y = 1 | X_{\text{obs}(m)}, M = m) &= \mathbb{P}(Y = 1 | X_{\text{obs}(m)}) && \text{(using Assumption 2)} \\ &= \mathbb{E} \left[ \mathbb{P}(Y = 1 | X) | X_{\text{obs}(m)} \right] && (25) \end{aligned}$$

$$= \mathbb{E} \left[ \frac{1}{1 + \exp(-\beta_0^* - \sum_{j=1}^d \beta_j^* X_j)} | X_{\text{obs}(m)} \right]. \quad (26)$$

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

$$\mathbb{P}(Y = 1 | X_{\text{obs}(m)}, M = m) = \frac{1}{1 + \exp(-\beta_{0,m}^* - \sum_{j \in \text{obs}(m)} \beta_{j,m}^* X_j)}. \quad (27)$$

Combining the two previous equations leads to

$$\begin{aligned} &\frac{1}{1 + \exp(-\beta_{0,m}^* - \sum_{j \in \text{obs}(m)} \beta_{j,m}^* X_j)} \\ &= \mathbb{E} \left[ \frac{1}{1 + \exp(-\beta_0^* - \sum_{j=1}^d \beta_j^* X_j)} | X_{\text{obs}(m)} \right] && (28) \end{aligned}$$

$$\geq \frac{1}{\mathbb{E} \left[ 1 + \exp(-\beta_0^* - \sum_{j=1}^d \beta_j^* X_j) | X_{\text{obs}(m)} \right]} \quad \text{(using Jensen Inequality)}$$

$$= \frac{1}{1 + \mathbb{E} \left[ \exp \left( -\beta_0^* - \sum_{j \in \text{obs}(m)} \beta_j^* X_j - \sum_{j \in \text{mis}(m)} \beta_j^* X_j \right) | X_{\text{obs}(m)} \right]} \quad (29)$$

$$= \frac{1}{1 + \exp \left( -\beta_0^* - \sum_{j \in \text{obs}(m)} \beta_j^* X_j \right) \mathbb{E} \left[ \exp \left( - \sum_{j \in \text{mis}(m)} \beta_j^* X_j \right) | X_{\text{obs}(m)} \right]} \quad (30)$$

$$(31)$$

which is equivalent to

$$\exp \left( -(\beta_{0,m}^* - \beta_0^*) - \sum_{j \in \text{obs}(m)} (\beta_{j,m}^* - \beta_j^*) X_j \right) \leq \mathbb{E} \left[ \exp \left( - \sum_{j \in \text{mis}(m)} \beta_j^* X_j \right) | X_{\text{obs}(m)} \right]. \quad (32)$$

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

$$\exp\left(-(\beta_{0,m}^* - \beta_0^*) - \sum_{j \in \text{obs}(m)} (\beta_{j,m}^* - \beta_j^*) X_j\right) \leq \mathbb{E} \left[ \exp\left(- \sum_{j \in \text{mis}(m)} \beta_j^* X_j\right) \right]. \quad (33)$$

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

$$\exp\left(-(\beta_{0,m}^* - \beta_0^*) - (\beta_{k,m}^* - \beta_k^*) X_k\right) \leq \mathbb{E} \left[ \exp\left(- \sum_{j \in \text{mis}(m)} \beta_j^* X_j\right) \right]. \quad (34)$$

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

$$\beta_{k,m}^* = \beta_k^*. \quad (35)$$

Injecting this into (28) leads to

$$\frac{1}{1 + \exp(-\beta_{0,m}^* - \sum_{j \in \text{obs}(m)} \beta_j^* X_j)} = \mathbb{E} \left[ \frac{1}{1 + \exp(-\beta_0^* - \sum_{j=1}^d \beta_j^* X_j)} \middle| X_{\text{obs}(m)} \right], \quad (36)$$

that is

$$\mathbb{E} \left[ \frac{1 + \exp(-\beta_{0,m}^* - \sum_{j \in \text{obs}(m)} \beta_j^* X_j)}{1 + \exp(-\beta_0^* - \sum_{j=1}^d \beta_j^* X_j)} \middle| X_{\text{obs}(m)} \right] = 1. \quad (37)$$

Let

$$u = \exp\left(- \sum_{j \in \text{obs}(m)} \beta_j^* X_j\right) \quad \text{and} \quad Z_{\text{mis}(m)} = \exp\left(- \sum_{j \in \text{mis}(m)} \beta_j^* X_j\right). \quad (38)$$

According to (37), for all  $u \in (0, \infty)$ ,

$$\mathbb{E} \left[ \frac{1 + u \exp(-\beta_{0,m}^*)}{1 + u Z_{\text{mis}(m)} \exp(-\beta_0^*)} \right] = 1. \quad (39)$$

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

$$\lim_{u \rightarrow \infty} \mathbb{E} \left[ \frac{1 + u \exp(-\beta_{0,m}^*)}{1 + u Z_{\text{mis}(m)} \exp(-\beta_0^*)} \right] = \mathbb{E} \left[ \lim_{u \rightarrow \infty} \frac{1 + u \exp(-\beta_{0,m}^*)}{1 + u Z_{\text{mis}(m)} \exp(-\beta_0^*)} \right] \quad (40)$$

$$= \mathbb{E} \left[ \frac{\exp(-\beta_{0,m}^*)}{Z_{\text{mis}(m)} \exp(-\beta_0^*)} \right]. \quad (41)$$

Thus,

$$\mathbb{E} \left[ \frac{1}{Z_{\text{mis}(m)}} \right] = \exp(\beta_{0,m}^* - \beta_0^*). \quad (42)$$

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

$$\exp(\beta_{0,m}^* - \beta_0^*) = \mathbb{E} \left[ \frac{1}{\prod_{j \in \text{mis}(m)} \exp(-\beta_j^* X_j)} \right] \quad (43)$$

$$= \mathbb{E} \left[ \prod_{j \in \text{mis}(m)} \exp(\beta_j^* X_j) \right] \quad (44)$$

$$= \prod_{j \in \text{mis}(m)} \mathbb{E} [\exp(\beta_j^* X_j)]. \quad (45)$$

Thus,

$$\exp(-\beta_{0,m}^*) = \frac{\exp(-\beta_0^*)}{\prod_{j \in \text{mis}(m)} \mathbb{E}[\exp(\beta_j^* X_j)]} \quad (46)$$

$$= \frac{\exp(-\beta_0^*)}{\mathbb{E}[Z'_{\text{mis}(m)}]}, \quad (47)$$

where

$$Z'_{\text{mis}(m)} = 1/Z_{\text{mis}(m)} = \exp\left(\sum_{j \in \text{mis}(m)} \beta_j^* X_j\right). \quad (48)$$

Injecting this equality into (39) leads to, for all  $u \in (0, \infty)$ ,

$$\mathbb{E}\left[\frac{1 + u \exp(-\beta_0^*)/\mathbb{E}[Z'_{\text{mis}(m)}]}{1 + u \exp(-\beta_0^*)/Z'_{\text{mis}(m)}}\right] = 1 \quad (49)$$

$$\Leftrightarrow \mathbb{E}\left[\frac{\mathbb{E}[Z'_{\text{mis}(m)}] + u \exp(-\beta_0^*)}{\mathbb{E}[Z'_{\text{mis}(m)}] + u \mathbb{E}[Z'_{\text{mis}(m)}] \exp(-\beta_0^*)/Z'_{\text{mis}(m)}}\right] = 1 \quad (50)$$

$$\Leftrightarrow \mathbb{E}\left[\frac{\mathbb{E}[Z'_{\text{mis}(m)}] + v}{\mathbb{E}[Z'_{\text{mis}(m)}] + v \mathbb{E}[Z'_{\text{mis}(m)}]/Z'_{\text{mis}(m)}}\right] = 1. \quad (51)$$

where  $v = u \exp(-\beta_0^*)$ . As this holds for all  $v \in (0, \infty)$ , taking the derivative of the expectation leads to, for all  $v \in (0, \infty)$ ,

$$\mathbb{E}\left[\frac{\mathbb{E}[Z'_{\text{mis}(m)}] \left(1 - \frac{\mathbb{E}[Z'_{\text{mis}(m)}]}{Z'_{\text{mis}(m)}}\right)}{\left(\mathbb{E}[Z'_{\text{mis}(m)}] + \frac{\mathbb{E}[Z'_{\text{mis}(m)}]}{Z'_{\text{mis}(m)}} v\right)^2}\right] = 0. \quad (52)$$

Letting  $v$  tend to zero leads to

$$\mathbb{E}\left[\frac{1}{Z'_{\text{mis}(m)}}\right] = \frac{1}{\mathbb{E}[Z'_{\text{mis}(m)}]}, \quad (53)$$

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

## D (LDA + MCAR) Proofs of Section 5.1

### D.1 Preliminary

**Lemma D.1.** *Assumption 4 (LDA) is equivalent to*

(i) *The following logistic regression model*

$$\mathbb{P}[Y = 1|X] = \sigma(\log(\pi C) + (\mu_1 - \mu_{-1})^\top \Sigma^{-1} x), \quad (54)$$

with  $C = \exp\left(\frac{1}{2} \mu_{-1}^\top \Sigma^{-1} \mu_{-1} - \frac{1}{2} \mu_1^\top \Sigma^{-1} \mu_1\right)$  and  $\pi = \mathbb{P}[Y = 1]/\mathbb{P}[Y = -1]$ ,

(ii) and  $X$  is distributed as a Gaussian mixture  $X = \pi_{-1}Z_0 + (1 - \pi_0)Z_1$ , where  $Z_k \sim \mathcal{N}(\mu_k, \Sigma)$  for all  $k \in \{0, 1\}$ .

*Proof of Lemma D.1.* Grant Assumption 4. Let  $f_k$  be the density of  $X|Y = k$ . Let  $\pi = \mathbb{P}[Y = 1]/\mathbb{P}[Y = 0]$ . We have

$$\frac{\mathbb{P}[Y = 1|X = x]}{\mathbb{P}[Y = 0|X = x]} = \pi \frac{f_1(x)}{f_0(x)} \quad (55)$$

$$= \pi \exp\left(-\frac{1}{2}(x - \mu_1)^\top \Sigma^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_0)^\top \Sigma^{-1}(x - \mu_0)\right) \quad (56)$$

$$= \pi C \exp\left((\mu_1 - \mu_0)^\top \Sigma^{-1}x\right), \quad (57)$$

with  $C = \exp\left(\frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1\right)$ . Consequently,

$$\log\left(\frac{\mathbb{P}[Y = 1|X = x]}{\mathbb{P}[Y = 0|X = x]}\right) = \log(\pi C) + (\mu_1 - \mu_0)^\top \Sigma^{-1}x, \quad (58)$$

which concludes the first part of the proof. Now, assume that (i) and (ii) of Lemma D.1 hold. Recall that  $f(x), f_0(x), f_1(x)$  are respectively the density of  $X, X|Y = 0$  and  $X|Y = 1$ . We have

$$f_0(x) = \frac{\mathbb{P}[Y = 0|X = x]f(x)}{\mathbb{P}[Y = 0]}. \quad (59)$$

Since  $X$  is a mixture of Gaussian, we have

$$f(x) = C' \exp\left(-\frac{1}{2}x^\top \Sigma^{-1}x\right) (\pi_0 c_0 \exp(x^\top \Sigma^{-1}\mu_0) + (1 - \pi_0)c_1 \exp(x^\top \Sigma^{-1}\mu_1)) \quad (60)$$

$$= C' \pi_0 c_0 \exp\left(-\frac{1}{2}x^\top \Sigma^{-1}x\right) \exp(x^\top \Sigma^{-1}\mu_0) \left(1 + \frac{(1 - \pi_0)c_1}{\pi_0 c_0} \exp(x^\top \Sigma^{-1}(\mu_1 - \mu_0))\right), \quad (61)$$

with  $C' = (2\pi)^{-d/2}(\det(\Sigma))^{-1/2}$ ,  $c_0 = \exp(-\frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0)$  and  $c_1 = \exp(-\frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1)$ . Besides, by assumption, we have

$$\mathbb{P}[Y = 0|X = x] = \frac{1}{1 + \exp(\log(\pi C) + (\mu_1 - \mu_0)^\top \Sigma^{-1}x)} \quad (62)$$

$$= \frac{1}{1 + \pi C \exp((\mu_1 - \mu_0)^\top \Sigma^{-1}x)}. \quad (63)$$

Gathering (61) and (63) in (59), we obtain

$$f_0(x) = C' c_0 \exp\left(-\frac{1}{2}x^\top \Sigma^{-1}x\right) \exp(x^\top \Sigma^{-1}\mu_0) \frac{1 + \frac{(1 - \pi_0)c_1}{\pi_0 c_0} \exp(x^\top \Sigma^{-1}(\mu_1 - \mu_0))}{1 + \pi C \exp((\mu_1 - \mu_0)^\top \Sigma^{-1}x)} \quad (64)$$

$$= C' c_0 \exp\left(-\frac{1}{2}x^\top \Sigma^{-1}x\right) \exp(x^\top \Sigma^{-1}\mu_0) \quad (65)$$

$$= (2\pi)^{-d/2}(\det(\Sigma))^{-1/2} \exp\left(-\frac{1}{2}(x - \mu_0)^\top \Sigma^{-1}(x - \mu_0)\right), \quad (66)$$

since  $C = c_1/c_0$  and  $\pi = (1 - \pi_0)/\pi_0$ . Consequently,  $X|Y = 0$  follows a multivariate Gaussian distribution  $\mathcal{N}(\mu_0, \Sigma)$ . The same calculus can be carried out for the distribution of  $X|Y = 1$ , which concludes the proof.  $\square$

The Bayes predictor  $h_{\text{comp}}^*$  satisfies

$$\mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) = \Phi(-a_m - b_m) \pi_{-1} + \Phi(a_m - b_m) \pi_1 \quad (67)$$

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.2** (Bayes Risk of P-b-P LDA). *Under Assumptions 2 and 4, the Bayes risk is given by*

$$\mathcal{R}_{\text{mis}}(h^*) = \sum_{m \in \{0,1\}^d} \Phi(-a_m - b_m) \pi_{-1} p_m + \Phi(a_m - b_m) \pi_1 p_m, \quad (68)$$

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

$$a_m = \frac{\log\left(\frac{\pi_{-1}}{\pi_1}\right)}{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|} \quad \text{and} \quad b_m = \frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|}{2} \quad (69)$$

The proof can be found in Appendix D.3. Note that, from Corollary D.2 (using that  $\pi_1 = \pi_{-1}$ ) and Equation (67), we have that

$$\begin{aligned} & L(h^*) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) \\ &= \sum_{m \in \{0,1\}^d} \left( \Phi\left(-\frac{\left\|\Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})\right\|}{2}\right) - \Phi\left(-\frac{\left\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1})\right\|}{2}\right) \right) p_m, \end{aligned} \quad (70)$$

with  $\Phi$  the c.d.f. of a standard Gaussian variable.

## D.2 Proof of Proposition 5.1

*Proof.* Expanding (19),

$$\begin{aligned} h_m^*(X_{\text{obs}(m)}) &= \text{sign}(\mathbb{E}[Y | X_{\text{obs}(m)}, M = m]) \\ &= \text{sign}(\mathbb{P}(Y = 1 | X_{\text{obs}(m)}, M = m) - \mathbb{P}(Y = -1 | X_{\text{obs}(m)}, M = m)). \end{aligned} \quad (71)$$

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

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

Thus,

$$\mathbb{P}(Y = 1 | X_{\text{obs}(m)} \in B, M = m) > \mathbb{P}(Y = -1 | X_{\text{obs}(m)} \in B, M = m) \quad (72)$$

$$\iff \mathbb{P}(X_{\text{obs}(m)} \in B | Y = 1) \pi_1 > \mathbb{P}(X_{\text{obs}(m)} \in B | Y = -1) \pi_{-1}. \quad (73)$$

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

$$h_m^*(x) = \text{sign}\left(\pi_1 f_{X_{\text{obs}(m)} | Y=1}(x) - \pi_{-1} f_{X_{\text{obs}(m)} | Y=-1}(x)\right) \quad (74)$$

$$= \text{sign}\left(\log\left(\frac{f_{X_{\text{obs}(m)} | Y=1}(x)}{f_{X_{\text{obs}(m)} | Y=-1}(x)}\right) - \log\left(\frac{\pi_{-1}}{\pi_1}\right)\right), \quad (75)$$

where  $f_{X_{\text{obs}(m)} | Y=k}$  is the density of  $X_{\text{obs}(m)} | Y = k$  for all  $k \in \{-1, 1\}$ . Under LDA model (Assumption 4), the objective is to determine the distribution of  $X_{\text{obs}(m)} | Y = k$  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,  $X_{\text{obs}(m)}|Y = k \sim \mathcal{N}(\mu_{k,\text{obs}(m)}, \Sigma_{\text{obs}(m)})$  and therefore,

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

Consequently,

$$h_m^*(x) = \text{sign} \left( (\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})^\top \Sigma_{\text{obs}(m)}^{-1} \left( x - \frac{\mu_{1,\text{obs}(m)} + \mu_{-1,\text{obs}(m)}}{2} \right) - \log \left( \frac{\pi_{-1}}{\pi_1} \right) \right), \quad (76)$$

which concludes the proof.  $\square$

### D.3 Proof of Corollary D.2

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

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

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

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

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

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

Finally,

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

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

$$a_m = \frac{\log\left(\frac{\pi_{-1}}{\pi_1}\right)}{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\|} \quad \text{and} \quad b_m = \frac{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\|}{2}. \quad (77)$$

□

#### D.4 Proof of Proposition 5.3

*Proof.* By Proposition 5.1, we know that the Bayes classifier takes the form

$$\begin{aligned}
h_m^*(x_{\text{obs}(m)}) &= \text{sign}\left(c_m (\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})^\top \Sigma_{\text{obs}(m)}^{-1} \left(x_{\text{obs}(m)} - \frac{\mu_{1,\text{obs}(m)} + \mu_{-1,\text{obs}(m)}}{2}\right)\right) \\
&= \text{sign}\left(c_m \sum_{j=1}^d (\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j x_j \mathbb{1}_{m_j=0} - c_m \sum_{j=1}^d (\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j v_j \mathbb{1}_{m_j=0}\right), \quad (78)
\end{aligned}$$

for some  $c_m > 0$ , where, for all  $1 \leq j \leq d$ ,  $\beta_j = \mu_{1,j} - \mu_{-1,j}$  and  $v_j = (\mu_{1,j} + \mu_{-1,j})/2$ . The predictor associated with constant imputation  $\alpha \in \mathbb{R}^d$  takes the form

$$h_{\text{imp}}(z) = \text{sign}\left(\sum_{j=1}^d \gamma_j x_j \mathbb{1}_{m_j=0} + \gamma_0 + \sum_{j=1}^d \gamma_j \alpha_j \mathbb{1}_{m_j=0}\right). \quad (79)$$

The Bayes predictor takes the form of an imputation strategy if and only if (78) equals (79), that is  $\gamma_0 = 0$  and

$$\sum_{j=1}^d (c_m (\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j - \gamma_j) x_j \mathbb{1}_{m_j=0} = \sum_{j=1}^d (c_m (\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j v_j - \gamma_j \alpha_j) \mathbb{1}_{m_j=0}, \quad (80)$$

which is equivalent to, for all  $m \in \{0,1\}^d$  and all  $j \in \text{obs}(m)$ ,

$$\left\{ \begin{array}{l} c_m (\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j = \gamma_j \\ c_m (\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j v_j = \gamma_j \alpha_j \end{array} \right\} \iff \left\{ \begin{array}{l} \gamma_j = c_m (\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j \\ (\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j v_j = \alpha_j (\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j \end{array} \right. \quad (81)$$

Choosing  $m = e_j$ , that is the missing pattern in which only the  $j$ th component is observed, we have

$$\frac{\beta_j}{\sigma_{jj}} v_j = \alpha_j \frac{\beta_j}{\sigma_{jj}}. \quad (82)$$

Since this equality must hold for all  $\beta$ , choosing  $\beta \neq 0$  leads to

$$\alpha_j = v_j. \quad (83)$$

Besides, for all  $m \in \{0, 1\}^d$  and all  $j \in \text{obs}(m)$ ,

$$(\Sigma_{\text{obs}(m)}^{-1} \beta_{\text{obs}(m)})_j = \frac{\gamma_j}{c_m}, \quad (84)$$

since  $c_m > 0$  by assumption. Considering  $m = e_j$ , the missing pattern in which only the  $j$ th component is observed, we have

$$\frac{\beta_j}{\sigma_{jj}} = \frac{\gamma_j}{c_m}. \quad (85)$$

Since  $\beta_j \neq 0$ , we have  $\gamma_j \neq 0$ . Now, letting  $m = \mathbf{0}$  leads to, for all  $j \in \{1, \dots, d\}$ ,

$$\gamma_j = c_0 (\Sigma^{-1} \beta)_j. \quad (86)$$

Consequently,

$$\frac{\gamma_j}{\gamma_i} = \frac{(\Sigma^{-1} \beta)_j}{(\Sigma^{-1} \beta)_i} \quad (87)$$

Let  $i \neq j \in \{1, \dots, d\}$ . Consider  $m = 1 - e_j - e_i$ , that is the missing pattern in which only the  $j$ th or the  $i$ th components are observed. Then,

$$\gamma_j = c_m (\Sigma_{\{i,j\}}^{-1} \beta_{\{i,j\}})_j, \quad (88)$$

and similarly for  $\gamma_i$ , where  $\Sigma_{\{i,j\}}$  is the submatrix composed of the  $i$ th and  $j$ th rows and columns of  $\Sigma$ . Simple calculations show that

$$\Sigma_{\{i,j\}}^{-1} = \frac{1}{\sigma_{ii}\sigma_{jj} - \sigma_{ij}^2} \begin{bmatrix} \sigma_{jj} & -\sigma_{ij} \\ -\sigma_{ij} & \sigma_{ii} \end{bmatrix}. \quad (89)$$

Thus,

$$\Sigma_{\{i,j\}}^{-1} \beta_{\{i,j\}} = \frac{1}{\sigma_{ii}\sigma_{jj} - \sigma_{ij}^2} \begin{bmatrix} \sigma_{jj}\beta_i - \sigma_{ij}\beta_j \\ -\sigma_{ij}\beta_i + \sigma_{ii}\beta_j \end{bmatrix}. \quad (90)$$

Hence,

$$\frac{\gamma_j}{\gamma_i} = \frac{-\sigma_{ij}\beta_i + \sigma_{ii}\beta_j}{\sigma_{jj}\beta_i - \sigma_{ij}\beta_j} \quad (91)$$

Gathering (87) and (91),

$$\frac{(\Sigma^{-1} \beta)_j}{(\Sigma^{-1} \beta)_i} = \frac{-\sigma_{ij}\beta_i + \sigma_{ii}\beta_j}{\sigma_{jj}\beta_i - \sigma_{ij}\beta_j}. \quad (92)$$

Letting  $\Sigma^{-1} = (s_{ij})_{1 \leq i, j \leq d}$ , we obtain

$$\left( \sum_{k=1}^d s_{jk} \beta_k \right) (\sigma_{jj} \beta_i - \sigma_{ij} \beta_j) = \left( \sum_{k=1}^d s_{ik} \beta_k \right) (-\sigma_{ij} \beta_i + \sigma_{ii} \beta_j) \quad (93)$$

$$\iff \sum_{k \neq i, j} (s_{jk} \sigma_{jj} + \sigma_{ij} s_{ik}) \beta_k \beta_i + \sum_{k \neq i, j} (-s_{jk} \sigma_{ij} + \sigma_{ii} s_{ik}) \beta_k \beta_j + (s_{ji} \sigma_{jj} + s_{ii} \sigma_{ij}) \beta_i^2 \quad (94)$$

$$+ (-s_{jj} \sigma_{ij} - s_{ij} \sigma_{ii}) \beta_j^2 + (s_{jj} \sigma_{jj} - s_{ii} \sigma_{ii}) \beta_i \beta_j = 0 \quad (95)$$

As this equality must hold for all  $\beta \neq 0$ , we have

$$\begin{cases} s_{jk} \sigma_{jj} + \sigma_{ij} s_{ik} = 0 \\ -s_{jk} \sigma_{ij} + \sigma_{ii} s_{ik} = 0 \end{cases}. \quad (96)$$



Multiplying the first line by  $\sigma_{ij}/\sigma_{jj}$  and adding it to the second one leads to

$$\sigma_{jj}s_{ik}(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}) = 0. \quad (97)$$

Since  $\Sigma_{i,j}$  is invertible by assumption,  $\sigma_{ij}^2 \neq -\sigma_{ii}\sigma_{jj}$  and  $\sigma_{jj} \neq 0$ . Thus, for all  $k \neq i, j$ ,  $s_{jk} = s_{ik} = 0$ . As the matrix  $\Sigma^{-1}$  is symmetric, we deduce that for  $d \geq 3$ , for all  $k \neq k'$ ,  $s_{kk'} = 0$ . Consequently, the matrix  $\Sigma^{-1}$  is diagonal, and so is  $\Sigma$ . Thus, the matrix  $\Sigma$  takes the form  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_d)$ . Using (80), and taking  $\text{obs}(m) = \{j\}$ , we have

$$\sum_{j=1}^d (c_m(\Sigma_{\text{obs}(m)}^{-1}\beta_{\text{obs}(m)})_j - \gamma_j)x_j \mathbf{1}_{m_j=0} = \sum_{j=1}^d (c_m(\Sigma_{\text{obs}(m)}^{-1}\beta_{\text{obs}(m)})_j v_j - \gamma_j \alpha_j) \mathbf{1}_{m_j=0}, \quad (98)$$

$$\sum_{j=1}^d (c_m \sigma_j^{-1} \beta_j - \gamma_j) x_j \mathbf{1}_{M_j=0} = \sum_{j=1}^d (c_m \beta_j v_j \sigma_j^{-1} - \gamma_j \alpha_j) \mathbf{1}_{M_j=0}, \quad (99)$$

which is equivalent to, for all  $j \in \text{obs}(m)$ ,

$$\begin{cases} c_m \sigma_j^{-1} \beta_j = \gamma_j \\ c_m \beta_j \sigma_j^{-1} v_j = \gamma_j \alpha_j \end{cases} \iff \begin{cases} \gamma_j = c_m \sigma_j^{-1} \beta_j \\ \alpha_j = v_j \end{cases}, \quad (100)$$

which concludes the proof. □

## D.5 Proof of Proposition 5.4

*Proof.* Using Assumption 6, we have that

$$\begin{aligned} \left\| \Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1}) \right\| &\leq \frac{\|\mu_1 - \mu_{-1}\|}{\sqrt{\lambda_{\min}(\Sigma)}} = \mu \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \\ \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\| &\geq \frac{\|\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}\|}{\sqrt{\lambda_{\max}(\Sigma)}} = \mu \sqrt{\frac{d - \|m\|_0}{\lambda_{\max}(\Sigma)}} \end{aligned}$$

Recall that  $\Phi$  is the c.d.f. of a standard Gaussian random variable, according to Equation (70), we have

$$\begin{aligned} &\mathcal{R}_{\text{mis}}(h^*) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) \\ &= \sum_{m \in \{0,1\}^d} \left( \Phi \left( -\frac{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\|}{2} \right) - \Phi \left( -\frac{\left\| \Sigma^{-\frac{1}{2}}(\mu_1 - \mu_{-1}) \right\|}{2} \right) \right) p_m \end{aligned} \quad (101)$$

$$\begin{aligned} &\leq \sum_{m \in \{0,1\}^d} \left( \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d - \|m\|_0}{\lambda_{\max}(\Sigma)}} \right) - \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) p_m \\ &= \sum_{i=0}^d \sum_{\substack{m \in \{0,1\}^d \\ \text{s.t. } \|m\|_0=i}} \left( \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d-i}{\lambda_{\max}(\Sigma)}} \right) - \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) p_m \\ &= \sum_{i=0}^d \left( \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d-i}{\lambda_{\max}(\Sigma)}} \right) - \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) \binom{d}{i} \eta^i (1-\eta)^{d-i} \quad (\text{using Assumption 5}) \\ &= \mathbb{E} \left[ \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}} \right) - \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right] \end{aligned} \quad (102)$$

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

$$\begin{aligned}
& L(h^*) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) \\
& \leq \mathbb{E} \left[ \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}} \right) - \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \mid B = d \right] \mathbb{P}(B = d) \\
& \quad + \mathbb{E} \left[ \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}} \right) - \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \mid B \neq d \right] \mathbb{P}(B \neq d) \\
& = \left( \frac{1}{2} - \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \right) \eta^d \tag{103}
\end{aligned}$$

$$+ \mathbb{E} \left[ \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}} \right) - \Phi \left( \frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \mid B \neq d \right] (1 - \eta^d) \tag{104}$$

Now, we study the second term in (104). 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

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

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

$$Q(T_B) - Q(t) \leq e^{-\frac{T_B^2}{2}} (t - T_B). \tag{105}$$

Thus,

$$\begin{aligned}
& \mathbb{E} \left[ \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}} \right) - \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \mid B \neq d \right] \\
& \leq \frac{1}{\sqrt{2\pi}} \mathbb{E} \left[ e^{-\frac{t^2}{2}} (t - T_B) \mid B \neq d \right] \\
& = \frac{\mu}{2\sqrt{2\pi}} \mathbb{E} \left[ e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}} \left( \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} - \sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}} \right) \mid B \neq d \right]. \tag{106}
\end{aligned}$$

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

$$\begin{aligned}
& \mathbb{E} \left[ \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}} \right) - \Phi \left( -\frac{\mu}{2} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \right) \mid B \neq d \right] \\
& = \frac{\mu}{2\sqrt{2\pi}} \frac{1}{\mathbb{P}(B \neq d)} \mathbb{E} \left[ e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}} \left( \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} - \sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}} \right) \right] \tag{107}
\end{aligned}$$

$$- \frac{\mu}{2\sqrt{2\pi}} \frac{\mathbb{P}(B = d)}{\mathbb{P}(B \neq d)} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}. \tag{108}$$

Looking at the expectation in (108), we obtain

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

since

$$\sqrt{d} - \sqrt{d-B} = \frac{d-d+B}{\sqrt{d} + \sqrt{d-B}} \leq \frac{B}{\sqrt{d}}. \tag{109}$$

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,

$$\begin{aligned}
\mathbb{E} \left[ e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}} B \right] &= \sum_{i=0}^d \binom{d}{i} e^{-\frac{\mu^2(d-i)}{8\lambda_{\max}(\Sigma)}} i \eta^i (1-\eta)^{d-i} \\
&= \eta d \sum_{i=1}^d \frac{(d-1)!}{(i-1)!(d-1-(i-1))!} \eta^{i-1} \left( e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}} (1-\eta) \right)^{d-1-(i-1)} \\
&= \eta d \left( \eta + e^{-\frac{\mu^2}{8\lambda_{\max}(\Sigma)}} (1-\eta) \right)^{d-1}.
\end{aligned} \tag{110}$$

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

$$\begin{aligned}
& \mathbb{E} \left[ e^{-\frac{\mu^2(d-B)}{8\lambda_{\max}(\Sigma)}} \left( \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} - \sqrt{\frac{d-B}{\lambda_{\max}(\Sigma)}} \right) \right] \\
&\leq \sqrt{d} \left( \frac{1}{\sqrt{\lambda_{\min}(\Sigma)}} - \frac{1}{\sqrt{\lambda_{\max}(\Sigma)}} \right) (\eta + A)^d + \frac{1}{\sqrt{\lambda_{\max}(\Sigma)}d} \eta d (\eta + A)^{d-1} \\
&= \frac{\sqrt{d}}{\sqrt{\lambda_{\min}(\Sigma)}} (\eta + A)^d - \sqrt{\frac{d}{\lambda_{\max}(\Sigma)}} (\eta + A)^{d-1} A.
\end{aligned} \tag{111}$$

Gathering equations (104), (108) and (111), we obtain

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

An upper bound of this inequality is given by

$$L(h^*) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) \leq \frac{\eta^d}{2} + \frac{\mu\eta}{2\sqrt{2\pi}} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left( (\eta + A)^{d-1} - \eta^{d-1} \right). \quad (112)$$

□

## D.6 Proof of Corollary 5.5

*Proof.* Recall that, by Equation (70),

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

using only  $m = \mathbf{1}$ , since all terms in the above sum are positive. By Assumption 6,  $\left\| \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_{-1}) \right\| \geq d\mu / \sqrt{\lambda_{\max}(\Sigma)}$ . 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,

$$L(h^*) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) \geq \left( \frac{1}{2} - \Phi \left( -\frac{d\lambda}{2} \right) \right) \eta^d \xrightarrow{\lambda \rightarrow \infty} \frac{\eta^d}{2}. \quad (113)$$

On the other hand, by Proposition 5.4, we have

$$L(h^*) - \mathcal{R}_{\text{comp}}(h_{\text{comp}}^*) \leq \frac{\eta^d}{2} + \frac{\mu}{2\sqrt{2\pi}} \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \left( (\eta + A)^d - \eta^d \right). \quad (114)$$

Note that

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

which tends to zero by assumption. This concludes the proof.  $\square$

## D.7 Proofs of Section 5.2

### D.7.1 General lemmas for LDA misclassification control.

**Lemma D.3** ( $\hat{\mu}$  misclassification probability). *Given a sample satisfying Assumptions 2 and 4, with balanced classes, then*

$$\begin{aligned} & \mathbb{P} \left( \hat{h}_m(X_{\text{obs}(m)}) = 1 \mid Y = -1, \mathcal{D}_n \right) \\ &= \Phi \left( \frac{\left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{1,\text{obs}(m)} - \hat{\mu}_{-1,\text{obs}(m)}) \right)^\top \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left( \mu_{-1,\text{obs}(m)} - \frac{\hat{\mu}_{1,\text{obs}(m)} + \hat{\mu}_{-1,\text{obs}(m)}}{2} \right)}{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{1,\text{obs}(m)} - \hat{\mu}_{-1,\text{obs}(m)}) \right\|} \right) \end{aligned} \quad (115)$$

and symmetrically,

$$\begin{aligned} & \mathbb{P} \left( \hat{h}_m(X_{\text{obs}(m)}) = -1 \mid Y = 1, \mathcal{D}_n \right) \\ &= \Phi \left( -\frac{\left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{1,\text{obs}(m)} - \hat{\mu}_{-1,\text{obs}(m)}) \right)^\top \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left( \mu_{1,\text{obs}(m)} - \frac{\hat{\mu}_{1,\text{obs}(m)} + \hat{\mu}_{-1,\text{obs}(m)}}{2} \right)}{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{1,\text{obs}(m)} - \hat{\mu}_{-1,\text{obs}(m)}) \right\|} \right) \end{aligned} \quad (116)$$

with  $\Phi$  the standard Gaussian cumulative function.

*Proof.* We follow the same strategy as in the proof of Corollary D.2. We have

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

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

$$N|Y = -1, \mathcal{D}_n \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \quad (117)$$

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

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

Now we prove the second statement. According to the proof of Corollary D.2, we have

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

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

$$N|Y = 1, \mathcal{D}_n \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \quad (118)$$

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

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

□

**Lemma D.4.** *Grant Assumptions 2 and 4. Assume that we are given two estimators  $\hat{\mu}_1$  and  $\hat{\mu}_{-1}$  of  $\mu_1$  and  $\mu_{-1}$ . Then, the classifier  $\hat{h}_m$  defined in Equation (6) satisfies*

$$\begin{aligned} & \left| \mathbb{P}\left(\hat{h}_m(X_{\text{obs}(m)}) = 1 | Y = -1, \mathcal{D}_n\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = 1 | Y = -1\right) \right| \\ & \leq \frac{3}{2\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{-1,\text{obs}(m)} - \hat{\mu}_{-1,\text{obs}(m)}) \right\| + \frac{1}{2\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\mu_{1,\text{obs}(m)} - \hat{\mu}_{1,\text{obs}(m)}) \right\| \end{aligned} \quad (119)$$

and symmetrically,

$$\begin{aligned} & \left| \mathbb{P}\left(\hat{h}_m(X_{\text{obs}(m)}) = -1 | Y = 1, \mathcal{D}_n\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = -1 | Y = 1\right) \right| \\ & \leq \frac{3}{2\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(-\hat{\mu}_{1,\text{obs}(m)} + \mu_{1,\text{obs}(m)}) \right\| + \frac{1}{2\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}}(\hat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\| \end{aligned} \quad (120)$$

*Proof.* We only prove the first inequality, the other one can be handled in the same manner. According to using Corollary D.2 and Lemma D.3,

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

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

$$\left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right)^\top \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left( \mu_{-1,\text{obs}(m)} - \frac{\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)}}{2} \right) \quad (121)$$

$$= \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right)^\top \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \quad (122)$$

$$+ \frac{1}{2} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right)^\top \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{-1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)}) \quad (123)$$

$$\leq \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\| \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\| \quad (124)$$

$$- \frac{1}{2} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\|^2, \quad (125)$$

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

$$\left| \mathbb{P} \left( \widehat{h}_m(X_{\text{obs}(m)}) = 1 \mid Y = -1, \mathcal{D}_n \right) - \mathbb{P} \left( h_m^*(X_{\text{obs}(m)}) = 1 \mid Y = -1 \right) \right| \quad (126)$$

$$\leq \frac{1}{\sqrt{2\pi}} \frac{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\| \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\|}{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\|} \quad (127)$$

$$+ \frac{1}{\sqrt{2\pi}} \left| - \frac{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\|}{2} + \frac{\left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\|}{2} \right| \quad (128)$$

$$\leq \frac{1}{\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\| \quad (129)$$

$$+ \frac{1}{2\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (-\widehat{\mu}_{1,\text{obs}(m)} + \widehat{\mu}_{-1,\text{obs}(m)} + \mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\| \quad (130)$$

$$\leq \frac{1}{\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\| + \frac{1}{2\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)}) \right\| \quad (131)$$

$$+ \frac{1}{2\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\widehat{\mu}_{-1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)}) \right\| \quad (132)$$

$$\leq \frac{3}{2\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{-1,\text{obs}(m)} - \widehat{\mu}_{-1,\text{obs}(m)}) \right\| + \frac{1}{2\sqrt{2\pi}} \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mu_{1,\text{obs}(m)} - \widehat{\mu}_{1,\text{obs}(m)}) \right\|. \quad (133)$$

The second statement of the Lemma can be proven in the same way.  $\square$

**Lemma D.5.** *Grant Assumptions 2, 4 and assume the classes are balanced. Assume that we are given two estimators  $\hat{\mu}_1$  and  $\hat{\mu}_{-1}$  of  $\mu_1$  and  $\mu_{-1}$ . Then, the classifier  $\hat{h}$  defined in Equation (6) satisfies*

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

*Proof.* We have

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

by Lemma D.4. Thus,

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

$\square$

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.7.2 Lemma for Theorem 5.6

**Lemma D.6.** *For all  $m \in \mathcal{M}$  and all  $k \in \{-1, 1\}$ ,*

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

with  $\hat{\mu}_{k,\text{obs}(m)}$  defined in (5) 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,

$$\mathbb{E} \left[ \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right\|^2 \right] \leq \mathbb{E} \left[ \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right\|^2 \right]^{\frac{1}{2}} \quad (134)$$

$$= \mathbb{E} \left[ \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right)^\top \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right) \right]^{\frac{1}{2}} \quad (135)$$

$$= \mathbb{E} \left[ \text{tr} \left( \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right)^\top \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right) \right) \right]^{\frac{1}{2}} \quad (136)$$

$$= \mathbb{E} \left[ \text{tr} \left( \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right) \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right)^\top \right) \right]^{\frac{1}{2}} \quad (137)$$

$$= \text{tr} \left( \mathbb{E} \left[ \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right) \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) \right)^\top \right] \right)^{\frac{1}{2}} \quad (138)$$

$$= \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \mathbb{E} \left[ (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})^\top \right] \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right)^{\frac{1}{2}} \quad (139)$$

$$= \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \mathcal{C}(k, m) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right)^{\frac{1}{2}}, \quad (140)$$

where

$$\mathcal{C}(k, m) := \mathbb{E} \left[ (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)}) (\hat{\mu}_{k,\text{obs}(m)} - \mu_{k,\text{obs}(m)})^\top \right]. \quad (141)$$

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

$$\mathcal{C}(k, m)_{l,l} = \mathbb{E} \left[ (\hat{\mu}_{k,l} - \mu_{k,l})^2 \right]. \quad (142)$$

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

$$\mathcal{A}_{k,l} := \{\forall i \in \{1, \dots, n\}, \quad Y_i = -k \text{ or } M_{i,l} = 1\}, \quad (143)$$

where

$$\mathbb{P}(\mathcal{A}_{k,l}) = \prod_{i=1}^n P(Y_i = -k \text{ or } M_{i,l} = 1) = \left( \frac{1 + \eta}{2} \right)^n. \quad (144)$$

Thus,

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

The second term can be rewritten as

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

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

$$\mathcal{C}(k, m)_{l,l} = \mu_{k,l}^2 \left(\frac{1+\eta}{2}\right)^n + n \left(\frac{1-\eta}{2}\right) \Sigma_{l,l} A(n, \eta). \quad (145)$$

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

$$\mathcal{C}(k, m)_{r,l} = \mathbb{E} [(\widehat{\mu}_{k,r} - \mu_{k,r})(\widehat{\mu}_{k,l} - \mu_{k,l})]. \quad (146)$$

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

$$\mathcal{A}_{k,l,r} := \{\forall i \in \{1, \dots, n\}, (Y_i = -k \text{ or } (M_{i,r} = 1 \text{ and } M_{i,l} = 1))\}, \quad (147)$$

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

$$\begin{aligned}
&\mathbb{P}(\mathcal{A}_{k,l,r}) \\
&= \mathbb{P}(\forall i \in \{1, \dots, n\}, Y_i = -k \text{ or } (M_{i,r} = 1 \text{ and } M_{i,l} = 1)) \\
&= (\mathbb{P}(Y_i = -k) + \mathbb{P}(M_{i,r} = 1 \text{ and } M_{i,l} = 1) - \mathbb{P}(Y_i = -k \text{ and } M_{i,r} = 1 \text{ and } M_{i,l} = 1))^n \\
&= \left(\frac{\eta^2 + 1}{2}\right)^n.
\end{aligned} \quad (148)$$

$$= \left(\frac{\eta^2 + 1}{2}\right)^n. \quad (149)$$

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

$$\mathbb{E} [(\widehat{\mu}_{k,r} - \mu_{k,r})(\widehat{\mu}_{k,l} - \mu_{k,l}) \mid \mathcal{A}_{k,l,r}] = \mu_{k,r} \mu_{k,l}. \quad (150)$$

We now consider the second case and denote by

$$\mathcal{B}_{k,l,r} := \{\exists i \in \{1, \dots, n\}, (Y_i = k \wedge M_{i,l} = 0) = 1\} \cap \{\exists i \in \{1, \dots, n\}, (Y_i = k \wedge M_{i,r} = 0) = 1\}, \quad (151)$$

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

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

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

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

Then, we have that

$$\mathbb{P}(\mathcal{B}_{k,l,r}) = 1 - 2 \left(\frac{1+\eta}{2}\right)^n + \left(\frac{\eta^2+1}{2}\right)^n. \quad (152)$$

Besides,

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

Observe that this second sum is null. Indeed,

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

Then,

$$\begin{aligned}
& \mathbb{E}[(\widehat{\mu}_{k,r} - \mu_{k,r})(\widehat{\mu}_{k,l} - \mu_{k,l}) \mid \mathcal{B}_{k,l,r}] \\
&= \sum_{i=1}^n \mathbb{E} \left[ \left( \frac{(X_{i,r} - \mu_{k,r}) \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,r}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,r}=0}} \right) \left( \frac{(X_{i,l} - \mu_{k,l}) \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,l}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,l}=0}} \right) \mid \mathcal{B}_{k,l,r} \right] \\
&= \sum_{i=1}^n \mathbb{E} \left[ \frac{(X_{i,r} - \mu_{k,r})}{1 + \sum_{j \neq i}^n \mathbb{1}_{Y_j=k} \mathbb{1}_{M_{j,r}=0}} \frac{(X_{i,l} - \mu_{k,l})}{1 + \sum_{j \neq i}^n \mathbb{1}_{Y_j=k} \mathbb{1}_{M_{j,l}=0}} \mid Y_i = k, M_{i,r} = 0, M_{i,l} = 0 \right] \\
&\quad \mathbb{P}(Y_i = k, M_{i,r} = 0, M_{i,l} = 0 \mid \mathcal{B}_{k,l,r}) \\
&= \sum_{i=1}^n \mathbb{E}[(X_{i,r} - \mu_{k,r})(X_{i,l} - \mu_{k,l}) \mid Y_i = k] \\
&\quad \times \mathbb{E} \left[ \frac{1}{1 + \sum_{j \neq i}^n \mathbb{1}_{Y_j=k} \mathbb{1}_{M_{j,r}=0}} \frac{1}{1 + \sum_{j \neq i}^n \mathbb{1}_{Y_j=k} \mathbb{1}_{M_{j,l}=0}} \right] \\
&\quad \times \mathbb{P}(Y_i = k, M_{i,r} = 0, M_{i,l} = 0 \mid \mathcal{B}_{k,l,r}) \\
&= n \Sigma_{r,l} B(n, \eta) \frac{(1 - \eta)^2}{2 \mathbb{P}(\mathcal{B}_{k,l,r})}, \tag{153}
\end{aligned}$$

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

Now, we consider the last case, and denote by

$$\mathcal{C}_{k,l,r} = (\mathcal{B}_{k,l,r} \cup \mathcal{A}_{k,l,r})^c \tag{154}$$

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

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

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,

$$\begin{aligned}
& \mathbb{E}[(\widehat{\mu}_{k,r} - \mu_{k,r})(\widehat{\mu}_{k,l} - \mu_{k,l}) \mid \mathcal{C}_{1,k,l,r}] \\
&= -\mu_{k,l} \mathbb{E} \left[ \frac{\sum_{i=1}^n (X_{i,r} - \mu_{k,r}) \mathbb{1}_{M_{i,r}=0} \mathbb{1}_{Y_i=k}}{\sum_{i=1}^n \mathbb{1}_{M_{i,r}=0} \mathbb{1}_{Y_i=k}} \mid \mathcal{C}_{1,k,l,r} \right] \\
&= -\mu_{k,l} n \mathbb{E} \left[ \frac{(X_{1,r} - \mu_{k,r})}{1 + \sum_{i=2}^n \mathbb{1}_{M_{i,r}=0} \mathbb{1}_{Y_i=k}} \mid M_{1,r} = 0, Y_1 = k \right] \mathbb{P}(M_{1,r} = 0, Y_1 = k \mid \mathcal{C}_{1,k,l,r}) \\
&= -\mu_{k,l} n \mathbb{E}[(X_{1,r} - \mu_{k,r}) \mid Y_1 = k] \mathbb{E} \left[ \frac{1}{1 + \sum_{i=2}^n \mathbb{1}_{M_{i,r}=0} \mathbb{1}_{Y_i=k}} \right] \mathbb{P}(M_{1,r} = 0, Y_1 = k \mid \mathcal{C}_{1,k,l,r}) \\
&\quad \text{(using MCAR and independence)} \\
&= 0. \tag{155}
\end{aligned}$$

By symmetry, we also have

$$\mathbb{E}[(\widehat{\mu}_{k,r} - \mu_{k,r})(\widehat{\mu}_{k,l} - \mu_{k,l}) \mid \mathcal{C}_{2,k,l,r}] = 0. \tag{156}$$

Thus,

$$\mathbb{E}[(\widehat{\mu}_{k,r} - \mu_{k,r})(\widehat{\mu}_{k,l} - \mu_{k,l}) \mid \mathcal{C}_{k,l,r}] = 0. \tag{157}$$

Gathering (150), (153) and (157), we are able to compute  $\mathcal{C}(k, m)_{r,l}$  as follows

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

using (149) and (152). From (145), recall that

$$\mathcal{C}(k, m)_{l,l} = \mu_{k,l}^2 \left( \frac{1+\eta}{2} \right)^n + n \left( \frac{1-\eta}{2} \right) \Sigma_{l,l} A(n, \eta). \quad (158)$$

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

$$F = \left( \left( \frac{1+\eta}{2} \right)^n - \left( \frac{1+\eta^2}{2} \right)^n \right) I + \left( \frac{1+\eta^2}{2} \right)^n J \quad (159)$$

$$G = (A(n, \eta) - (1-\eta)B(n, \eta)) I + (1-\eta)B(n, \eta)J. \quad (160)$$

Thus,

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

Then, according to inequality (140), we have that

$$\begin{aligned} & \mathbb{E} \left[ \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\hat{\mu}_{k, \text{obs}(m)} - \mu_{k, \text{obs}(m)}) \right\| \right] \\ & \leq \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \mathcal{C}(k, m) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right)^{\frac{1}{2}} \end{aligned} \quad (161)$$

$$= \left( \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left( F \odot \mu_{k, \text{obs}(m)} \mu_{k, \text{obs}(m)}^\top \right) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) + n \frac{1-\eta}{2} \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (G \odot \Sigma_{\text{obs}(m)}) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \right)^{\frac{1}{2}} \quad (162)$$

The first term equals

$$\text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left( F \odot \mu_{k, \text{obs}(m)} \mu_{k, \text{obs}(m)}^\top \right) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (163)$$

$$= \left( \frac{1+\eta^2}{2} \right)^n \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left( J \odot \mu_{k, \text{obs}(m)} \mu_{k, \text{obs}(m)}^\top \right) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (164)$$

$$+ \left( \left( \frac{1+\eta}{2} \right)^n - \left( \frac{1+\eta^2}{2} \right)^n \right) \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left( I_{d-\|m\|_0} \odot \mu_{k, \text{obs}(m)} \mu_{k, \text{obs}(m)}^\top \right) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (165)$$

$$= \left( \frac{1+\eta^2}{2} \right)^n \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \mu_{k, \text{obs}(m)} \mu_{k, \text{obs}(m)}^\top \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (166)$$

$$+ \left( \left( \frac{1+\eta}{2} \right)^n - \left( \frac{1+\eta^2}{2} \right)^n \right) \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \text{diag} \left( \mu_{k, \text{obs}(m)} \mu_{k, \text{obs}(m)}^\top \right) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right). \quad (167)$$

Then, by Lemma F.4,

$$\text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \left( F \odot \mu_{k, \text{obs}(m)} \mu_{k, \text{obs}(m)}^\top \right) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (168)$$

$$\leq \left( \frac{1+\eta^2}{2} \right)^n \text{tr} \left( \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \mu_{k, \text{obs}(m)} \right) \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \mu_{k, \text{obs}(m)} \right)^\top \right) \quad (169)$$

$$+ \left( \left( \frac{1+\eta}{2} \right)^n - \left( \frac{1+\eta^2}{2} \right)^n \right) \|\mu\|_\infty^2 \text{tr} \left( \Sigma_{\text{obs}(m)}^{-1} \right) \quad (170)$$

$$= \left( \frac{1+\eta^2}{2} \right)^n \left\| \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \mu_{k, \text{obs}(m)} \right\|^2 + \left( \left( \frac{1+\eta}{2} \right)^n - \left( \frac{1+\eta^2}{2} \right)^n \right) \frac{\|\mu\|_\infty^2 (d - \|m\|_0)}{\lambda_{\min}(\Sigma)} \quad (171)$$

$$\leq \left( \frac{1+\eta^2}{2} \right)^n \frac{\|\mu_{k, \text{obs}(m)}\|^2}{\lambda_{\min}(\Sigma)} + \left( \left( \frac{1+\eta}{2} \right)^n - \left( \frac{1+\eta^2}{2} \right)^n \right) \frac{\|\mu\|_\infty^2 (d - \|m\|_0)}{\lambda_{\min}(\Sigma)} \quad (172)$$

$$\leq \left( \frac{1+\eta^2}{2} \right)^n \frac{\|\mu\|_\infty^2 (d - \|m\|_0)}{\lambda_{\min}(\Sigma)} + \left( \left( \frac{1+\eta}{2} \right)^n - \left( \frac{1+\eta^2}{2} \right)^n \right) \frac{\|\mu\|_\infty^2 (d - \|m\|_0)}{\lambda_{\min}(\Sigma)} \quad (173)$$

$$\leq \left( \frac{1+\eta}{2} \right)^n \frac{\|\mu\|_\infty^2 (d - \|m\|_0)}{\lambda_{\min}(\Sigma)}. \quad (174)$$

Regarding the second term in (162), note that  $A(n, \eta) - (1 - \eta)B(n, \eta) \geq 0$ . Indeed, letting  $Z := \sum_{i=1}^{n-1} \mathbb{1}_{Y_i=k} \sim \mathcal{B}(n-1, 1/2)$ ,

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

using the exchangeability as the samples are i.i.d. By leveraging the independence between the missingness at coordinate  $r$  and coordinate  $t$ , as well as the independence of each sample from the rest, we can conclude that

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

Thus, we have that

$$\text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (G \odot \Sigma_{\text{obs}(m)}) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (175)$$

$$= (1 - \eta)B(n, \eta) \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (\mathbf{1} \odot \Sigma_{\text{obs}(m)}) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (176)$$

$$+ (A(n, \eta) - (1 - \eta)B(n, \eta)) \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (I_{d-\|m\|_0} \odot \Sigma_{\text{obs}(m)}) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (177)$$

$$= (1 - \eta)B(n, \eta) \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \Sigma_{\text{obs}(m)} \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (178)$$

$$+ (A(n, \eta) - (1 - \eta)B(n, \eta)) \text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \text{diag}(\Sigma_{\text{obs}(m)}) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right). \quad (179)$$

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

$$\text{tr} \left( \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} (G \odot \Sigma_{\text{obs}(m)}) \Sigma_{\text{obs}(m)}^{-\frac{1}{2}} \right) \quad (180)$$

$$\leq (1 - \eta)B(n, \eta)(d - \|m\|_0) \quad (181)$$

$$+ (A(n, \eta) - (1 - \eta)B(n, \eta)) \max_{i \in [d]} (\Sigma_{i,i}) \text{tr} \left( \Sigma_{\text{obs}(m)}^{-1} \right) \quad (182)$$

$$\leq (1 - \eta)B(n, \eta)(d - \|m\|_0) \quad (183)$$

$$+ (A(n, \eta) - (1 - \eta)B(n, \eta)) \frac{\max_{i \in [d]} (\Sigma_{i,i})}{\lambda_{\min}(\Sigma)} (d - \|m\|_0) \quad (184)$$

$$\leq \kappa(1 - \eta)B(n, \eta)(d - \|m\|_0) + (A(n, \eta) - (1 - \eta)B(n, \eta))\kappa(d - \|m\|_0) \quad (185)$$

$$= A(n, \eta)\kappa(d - \|m\|_0) \quad (186)$$

$$\leq \frac{2\kappa(d - \|m\|_0)}{n(n+1) \left(\frac{1-\eta}{2}\right)^2}, \quad (187)$$

where  $\kappa := \frac{\max_{i \in [d]} (\Sigma_{i,i})}{\lambda_{\min}(\Sigma)} \geq 1$ . Finally, combining (174) and (187) in (162), we have

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

□

### D.7.3 Proof of Theorem 5.6

*Proof.* By Lemma D.5,

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

Now, using Assumption 5, we have that  $\|M\|_0 \sim \mathcal{B}(d, \eta)$ , so that

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

□

#### D.7.4 Proof of Corollary 5.7

*Proof.* From Proposition 5.4 and Theorem 5.6 we have that

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

□

## E (LDA + MNAR) Proofs of Section 5.3

### E.1 Proof of Proposition 5.9

*Proof.* By definition of the Bayes classifier (see (19)),

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

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

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



Note that by using Assumption 7, we have  $X_{\text{obs}(m)}|M = m, Y = k \sim \mathcal{N}(\mu_{m,k}, \Sigma_m)$ . Therefore,

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

Consequently,

$$h_m^*(x) = \text{sign} \left( (\mu_{1,m} - \mu_{-1,m})^\top \Sigma_m^{-1} \left( x - \frac{\mu_{1,m} + \mu_{-1,m}}{2} \right) - \log \left( \frac{\pi_{m,-1}}{\pi_{m,1}} \right) \right), \quad (188)$$

which concludes the proof.  $\square$

## E.2 General lemmas for LDA misclassification control under Assumption 7.

**Lemma E.1** ( $\hat{\mu}_m$  misclassification probability). *Grant Assumption 7. Then,*

$$\mathbb{P}(h_m^*(X_{\text{obs}(m)}) = 1 | Y = -1, M = m) = \Phi \left( -\frac{1}{2} \left\| \Sigma_m^{-\frac{1}{2}} (\mu_{m,1} - \mu_{m,-1}) \right\| \right), \quad (189)$$

and

$$\mathbb{P}(h_m^*(X_{\text{obs}(m)}) = 1 | Y = -1, \mathcal{D}_n) = \Phi \left( \frac{\left( \Sigma_m^{-\frac{1}{2}} (\hat{\mu}_{m,1} - \hat{\mu}_{m,-1}) \right)^\top \Sigma_m^{-\frac{1}{2}} \left( \mu_{m,-1} - \frac{\hat{\mu}_{m,1} + \hat{\mu}_{m,-1}}{2} \right)}{\left\| \Sigma_m^{-\frac{1}{2}} (\hat{\mu}_{m,1} - \hat{\mu}_{m,-1}) \right\|} \right) \quad (190)$$

Symmetrically,

$$\mathbb{P}(h_m^*(X_{\text{obs}(m)}) = -1 | Y = 1, M = m) = \Phi \left( -\frac{1}{2} \left\| \Sigma_m^{-\frac{1}{2}} (\mu_{m,1} - \mu_{m,-1}) \right\| \right), \quad (191)$$

and

$$\begin{aligned} & \mathbb{P}(\hat{h}_m(X_{\text{obs}(m)}) = -1 | Y = 1, M = m, \mathcal{D}_n) \\ &= \Phi \left( -\frac{\left( \Sigma_m^{-\frac{1}{2}} (\hat{\mu}_{m,1} - \hat{\mu}_{m,-1}) \right)^\top \Sigma_m^{-\frac{1}{2}} \left( \mu_{m,1} - \frac{\hat{\mu}_{m,1} + \hat{\mu}_{m,-1}}{2} \right)}{\left\| \Sigma_m^{-\frac{1}{2}} (\hat{\mu}_{m,1} - \hat{\mu}_{m,-1}) \right\|} \right) \end{aligned} \quad (192)$$

with  $\Phi$  the c.d.f. of a standard Gaussian distribution.

*Proof.* Using Proposition 5.9, and recalling that the classes are balanced on each missing patterns ( $\pi_{m,1} = \pi_{m,-1}$ ),

$$\begin{aligned} & \mathbb{P}(h_m^*(X_{\text{obs}(m)}) = 1 | Y = -1, M = m) \\ &= \mathbb{P} \left( (\mu_{m,1} - \mu_{m,-1})^\top \Sigma_m^{-1} \left( X_{\text{obs}(m)} - \frac{\mu_{m,1} + \mu_{m,-1}}{2} \right) > 0 | Y = -1, M = m \right). \end{aligned}$$

Let  $N = \Sigma_m^{-\frac{1}{2}} (X_{\text{obs}(m)} - \mu_{m,-1})$ . By Assumption 7,

$$N|Y = -1, M = m \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \quad (193)$$

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

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

Similarly, using Proposition 5.9,

$$\begin{aligned} &\mathbb{P}(h_m^*(X_{\text{obs}(m)}) = -1 \mid Y = 1, M = m) \\ &= \mathbb{P}\left((\mu_{m,1} - \mu_{m,-1})^\top \Sigma_m^{-1} \left(X_{\text{obs}(m)} - \frac{\mu_{m,1} + \mu_{m,-1}}{2}\right) < 0 \mid Y = 1, M = m\right). \end{aligned}$$

Let  $N = \Sigma_m^{-\frac{1}{2}}(X_{\text{obs}(m)} - \mu_{m,1})$ . By Assumption 7,

$$N \mid Y = 1, M = m \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \quad (194)$$

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

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

This proves the first and third statements. Regarding the second and fourth statements, following the same strategy as in the proof of Corollary D.2, we have

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

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

$$N \mid Y = -1, \mathcal{D}_n \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \quad (195)$$

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

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

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

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

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

$$N|Y = 1, \mathcal{D}_n \sim \mathcal{N}(0, Id_{d-\|m\|_0}). \quad (196)$$

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

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

□

**Lemma E.2.** *Grant Assumption 7. Assume that we are given two estimates  $\tilde{\mu}_1$  and  $\tilde{\mu}_{-1}$ . Then, for all  $m \in \mathcal{M}$ , the classifier  $\tilde{h}_m$  defined in Equation (11) satisfies*

$$\begin{aligned} & \left| \mathbb{P}\left(\tilde{h}_m(X_{\text{obs}(m)}) = 1 \mid Y = -1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = 1 \mid Y = -1, M = m\right) \right| \\ & \leq \frac{3}{2\sqrt{2\pi}} \left\| \Sigma_m^{-\frac{1}{2}}(\tilde{\mu}_{m,-1} - \mu_{m,-1}) \right\| + \frac{1}{2\sqrt{2\pi}} \left\| \Sigma_m^{-\frac{1}{2}}(\tilde{\mu}_{m,1} - \mu_{m,1}) \right\| \end{aligned} \quad (197)$$

and symmetrically,

$$\begin{aligned} & \left| \mathbb{P}\left(\tilde{h}_m(X_{\text{obs}(m)}) = -1 \mid Y = 1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = -1 \mid Y = 1, M = m\right) \right| \\ & \leq \frac{3}{2\sqrt{2\pi}} \left\| \Sigma_m^{-\frac{1}{2}}(\tilde{\mu}_{m,1} - \mu_{m,1}) \right\| + \frac{1}{2\sqrt{2\pi}} \left\| \Sigma_m^{-\frac{1}{2}}(\mu_{m,-1} - \tilde{\mu}_{m,-1}) \right\| \end{aligned} \quad (198)$$

*Proof.* To prove Inequality (197), notice that, by Lemma E.1,

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

We can then apply the same steps as in the proof of Lemma D.4, and the result follows. The proof of Inequality (198) is similar. □

**Lemma E.3.** *Grant Assumption 7, with balanced classes. Assume that we are given two estimates  $\tilde{\mu}_1$  and  $\tilde{\mu}_{-1}$ . Then, for all  $m \in \mathcal{M}$ , the classifier  $\tilde{h}_m$  defined in Equation (6) satisfies*

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

*Proof.*

$$\begin{aligned}
& \mathcal{R}_{\text{mis}}(\tilde{h}) - \mathcal{R}_{\text{mis}}(h^*) \\
&= \mathbb{P}\left(\tilde{h}(X_{\text{obs}(M)}, M) \neq Y\right) - \mathbb{P}\left(h^*(X_{\text{obs}(M)}, M) \neq Y\right) \\
&= \sum_{m \in \mathcal{M}} \left( \mathbb{P}\left(\tilde{h}(X_{\text{obs}(M)}, M) \neq Y \mid M = m\right) - \mathbb{P}\left(h^*(X_{\text{obs}(M)}, M) \neq Y \mid M = m\right) \right) p_m \\
&= \sum_{m \in \mathcal{M}} \left( \mathbb{P}\left(\tilde{h}_m(X_{\text{obs}(m)}) \neq Y \mid M = m\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) \neq Y \mid M = m\right) \right) p_m \quad (\text{using (6)}) \\
&= \sum_{m \in \mathcal{M}} \pi_{m,-1} \left( \mathbb{P}\left(\tilde{h}_m(X_{\text{obs}(m)}) = 1 \mid Y = -1, M = m\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = 1 \mid Y = -1, M = m\right) \right) \\
&\quad + \sum_{m \in \mathcal{M}} \pi_{m,1} \left( \mathbb{P}\left(\tilde{h}_m(X_{\text{obs}(m)}) = -1 \mid Y = 1, M = m\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = -1 \mid Y = 1, M = m\right) \right).
\end{aligned}$$

Note that

$$\mathbb{P}\left(\tilde{h}_m(X_{\text{obs}(m)}) = 1 \mid Y = -1, M = m\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = 1 \mid Y = -1, M = m\right) \quad (199)$$

$$= \mathbb{E} \left[ \mathbb{P}\left(\tilde{h}_m(X_{\text{obs}(m)}) = 1 \mid Y = -1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = 1 \mid Y = -1, M = m\right) \right] \quad (200)$$

$$\leq \frac{1}{2\sqrt{2\pi}} \mathbb{E} \left[ 3 \left\| \Sigma_m^{-\frac{1}{2}}(\mu_{m,-1} - \tilde{\mu}_{m,-1}) \right\| + \left\| \Sigma_m^{-\frac{1}{2}}(\mu_{m,1} - \tilde{\mu}_{m,1}) \right\| \right], \quad (201)$$

according to Lemma E.2. Similarly,

$$\mathbb{P}\left(\tilde{h}_m(X_{\text{obs}(m)}) = -1 \mid Y = 1, M = m\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = -1 \mid Y = 1, M = m\right) \quad (202)$$

$$= \mathbb{E} \left[ \mathbb{P}\left(\tilde{h}_m(X_{\text{obs}(m)}) = -1 \mid Y = 1, M = m, \mathcal{D}_n\right) - \mathbb{P}\left(h_m^*(X_{\text{obs}(m)}) = -1 \mid Y = 1, M = m\right) \right] \quad (203)$$

$$\leq \frac{\pi_{m,-1}}{2\sqrt{2\pi}} \left( \mathbb{E} \left[ 3 \left\| \Sigma_m^{-\frac{1}{2}}(-\tilde{\mu}_{m,1} + \mu_{m,1}) \right\| + \left\| \Sigma_m^{-\frac{1}{2}}(\tilde{\mu}_{m,-1} - \mu_{m,-1}) \right\| \right] \right). \quad (204)$$

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

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

□

### E.3 Lemmas for Theorem 5.10

**Lemma E.4.** *Grant Assumption 7. Then, for all  $k \in \{-1, 1\}$ ,*

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

where  $\tilde{\mu}_{m,k}$  is the estimate defined at (10).

*Proof.* We have

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

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

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

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

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

□

**Lemma E.5.** *Grant Assumption 7. Then, for all  $k \in \{-1, 1\}$ ,*

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

where  $\widetilde{\mu}_{m,k}$  is the estimate defined in (10).

*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[ \text{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[ \text{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[ \text{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[ \text{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}} \\ &= \text{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}} \\ &= \text{tr} \left( \Sigma_m^{-\frac{1}{2}} \left( \mathbb{E} \left[ \mathbf{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}} \quad (\text{using Lemma E.4}) \\ &= \left( \mathbb{E} \left[ \mathbf{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) \text{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[ \mathbf{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 5.10

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

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

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

$$\sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left( \mathbb{E} \left[ \mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{N_{m,k}} \right] (d - \|m\|_0) + \mathbb{P} \left( \frac{N_{m,k}}{n} \leq \tau \right) \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} \quad (205)$$

$$\leq \sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left( \mathbb{E} \left[ \mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{n\tau} \right] (d - \|m\|_0) + \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} \quad (206)$$

$$\leq \sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left( \mathbb{E} \left[ \mathbb{1}_{\frac{N_{m,k}}{n} > \tau} \right] \tau + \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau} \quad (207)$$

$$\leq \frac{2}{\sqrt{2\pi}} \left( 1 + \frac{\|\mu_m\|^2}{\lambda_{\min}(\Sigma_m)} \right)^{\frac{1}{2}} p_m \mathbb{1}_{p_m < \tau}. \quad (208)$$

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

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

using Markov Inequality and Inequality (217). Then, for all  $m \notin A_\tau$ , we have

$$\sum_{k=\pm 1} \frac{1}{\sqrt{2\pi}} \left( \mathbb{E} \left[ \mathbf{1}_{\frac{N_{m,k}}{n} > \tau} \frac{1}{N_{m,k}} \right] (d - \|m\|_0) + \mathbb{P} \left( \frac{N_{m,k}}{n} \leq \tau \right) \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} p_m \mathbf{1}_{p_m \geq \tau} \quad (209)$$

$$\leq \frac{p_m \mathbf{1}_{p_m \geq \tau}}{\sqrt{2\pi}} \sum_{k=\pm 1} \left[ \left( \mathbb{E} \left[ \frac{\mathbf{1}_{\frac{N_{m,k}}{n} > \tau}}{N_{m,k}} \right] (d - \|m\|_0) \right)^{\frac{1}{2}} + \left( \mathbb{P} \left( \frac{N_{m,k}}{n} \leq \tau \right) \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\|^2 \right)^{\frac{1}{2}} \right] \quad (210)$$

$$\leq \frac{1}{\sqrt{2\pi}} \sum_{k=\pm 1} \left[ \left( \frac{4(d - \|m\|_0)}{p_m(n+1)} \right)^{\frac{1}{2}} + \left( \frac{32\tau^2}{p_m^2} + (1-p_m)^n \right)^{1/2} \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\| \right] p_m \mathbf{1}_{p_m \geq \tau} \quad (\text{using Inequality (215)})$$

$$\leq \frac{4\tau\sqrt{p_m}\mathbf{1}_{p_m \geq \tau}}{\sqrt{2\pi}} + \left( \frac{4\tau}{\sqrt{\pi}} + \frac{1}{\sqrt{2\pi}} p_m (1-p_m)^{n/2} \right) \mathbf{1}_{p_m \geq \tau} \sum_{k=\pm 1} \left\| \Sigma_m^{-\frac{1}{2}} \mu_{m,k} \right\| \quad (211)$$

$$\leq \frac{4\tau\sqrt{p_m}\mathbf{1}_{p_m \geq \tau}}{\sqrt{2\pi}} + \left( \frac{4\tau}{\sqrt{\pi}} + \frac{1}{\sqrt{2\pi}} p_m (1-p_m)^{n/2} \right) \frac{2\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}} \mathbf{1}_{p_m \geq \tau}. \quad (212)$$

Combining (208) and (212), we obtain

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

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

$$\begin{aligned} \mathcal{R}_{\text{mis}}(\tilde{h}) - \mathcal{R}_{\text{mis}}(h^*) & \leq \sum_{m \in \{0,1\}^d} \left( \frac{4}{\sqrt{2\pi}} + \frac{8}{\sqrt{\pi}} \frac{\|\mu_m\|}{\sqrt{\lambda_{\min}(\Sigma_m)}} \right) \tau \wedge p_m \\ & \quad + \sum_{m \in \{0,1\}^d} \frac{\sqrt{2}\|\mu_m\|}{\sqrt{\pi}\lambda_{\min}(\Sigma_m)} p_m (1-p_m)^{n/2} \mathbf{1}_{p_m \geq \tau}. \end{aligned} \quad (213)$$

□

## 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  $k$ ,*

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

Let

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

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

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

---

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

$$\frac{1}{1+np} \leq \mathbb{E} \left[ \frac{1}{1+B} \right] \leq \frac{1}{p(n+1)} \quad (214)$$

and

$$\mathbb{E} \left[ \frac{\mathbb{1}\{B > 0\}}{B} \right] \leq \frac{2}{p(n+1)}. \quad (215)$$

*Proof.* • To prove the lower bound in (214), 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 (214), 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 (215), we use  $1/x \leq 2/(x+1)$  for all  $x \geq 1$  together with the previous result. □

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

$$\mathbb{E} \left[ \frac{1}{(1+B)^2} \right] \leq \frac{2}{(n+1)(n+2)p^2} \quad (216)$$

and

$$\mathbb{E} \left[ \frac{\mathbb{1}_{B>0}}{B^2} \right] \leq \frac{8}{(n+1)(n+2)p^2} \quad (217)$$



---

*Proof.* • In order to prove (216), 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 (217) can be deduced using the fact that, for all  $x \geq 1$ ,  $1/x \leq 2/(x+1)$ .

□

**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*

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

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

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

□

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  $Z_1, \dots, Z_K$  sub-Gaussian random variables with variance factor  $v$ , i.e.*

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

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  $X_{\text{obs}(m)}$  is still a Gaussian vector and  $X_{\text{obs}(m)} \sim \mathcal{N}(\mu_{\text{obs}(m)}, \Sigma_{\text{obs}(m) \times \text{obs}(m)})$ .*

*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  $\text{obs}(m)$  of coordinates is a Gaussian variable, then  $X_{\text{obs}(m)}$  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,

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

□

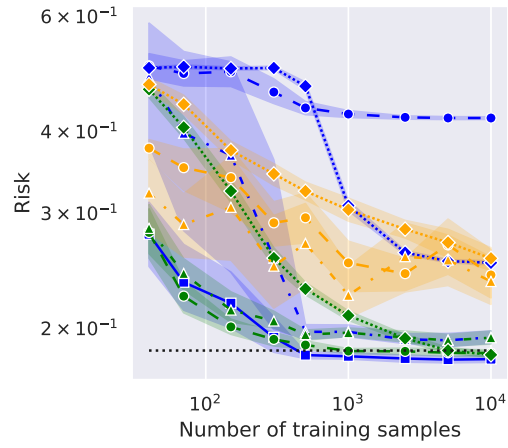


Figure 2: LDA generated data with MCAR missingness and  $\Sigma$  the Toeplitz matrix defined above.

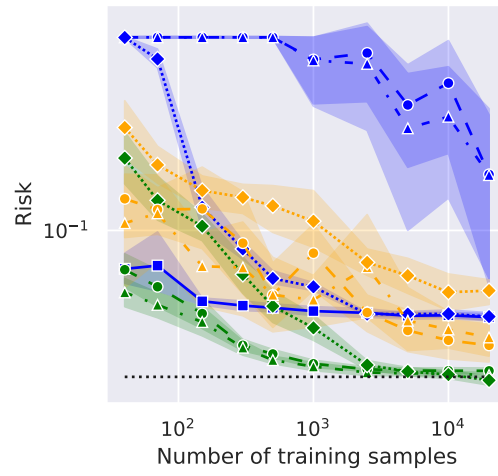


Figure 3: LDA generated data with MNAR missingness and  $\Sigma$  the Toeplitz matrix defined above.

## G Additional experiments

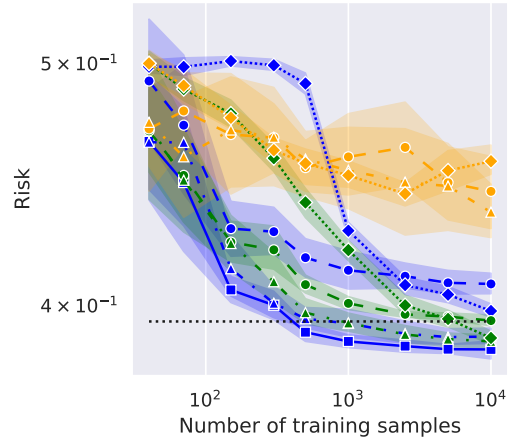


Figure 4: Logistic data with MCAR missingness and  $\Sigma$  the Toeplitz matrix defined above.

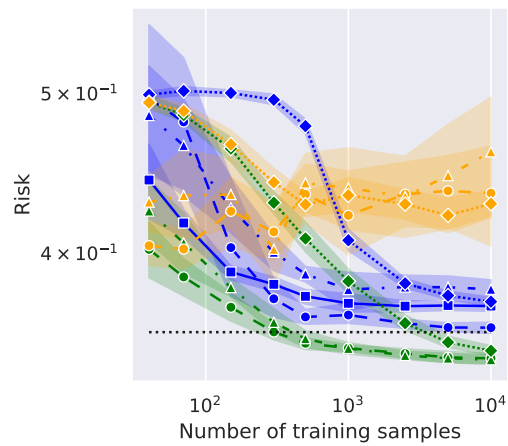


Figure 5: Logistic data with MNAR missingness and  $\Sigma$  the Toeplitz matrix defined above.