NeuMiss network classifiers: deep learning for classifying with missing values

Alexandre Perez-Lebel[1], Marine Le Morvan[1], and Gaël Varoquaux[1]

Inria, Saclay, France first.last@inria.fr

Abstract. Many application settings, such as health or business, call for prediction on data with missing values. The optimal predictor then builds upon correlations between observed and missing covariates. In regression settings, NeuMiss networks have been introduced to approximate such Bayes predictor, providing the corresponding inductive bias in deep learning. Here, we extend these results to specific classification settings: we show that the same functional forms are Bayes consistent in the binary classification setting assuming Gaussian data and a probit link, we adapt the neural architecture to this new setting, and we benchmark the resulting predictive pipeline showing that it predicts favorably compared to classical missing-values pipelines and is more robust to the missingness mechanism, including missing not at random settings.

Keywords: Missing values · Differentiable programming · Classification

1 Introduction: classification with missing values

Statistics on data with missing values has long been studied. A central result is that in "missing at random" (MAR) settings, i.e. when the probability that an entry is missing only depends on observed values, valid maximum-likelihood estimates can be obtained while ignoring the details of the missing-value mechanism [10, 6]. This result justifies expectation maximization and imputation based approaches [6, 12, 7]. However, building predictive models in the presence of missing values in a supervised learning framework can lead to different tradeoffs [3].

Predictive models can be directly optimized to account for missing values, as in popular tree-based approaches [11, 2]. The challenge is that for d variables, there are 2^d possible missing-value patterns. Even in the simple case of linear data-generating mechanisms, statistically and computationally efficient learning must capture the links between the optimal predictors across all possible missing data patterns [8, 4]. For this, NeuMiss networks [4] encode the appropriate inductive bias to approximate simultaneously the optimal predictors across all missing-data patterns in regression settings. They are more *robust to the missingness mechanism* than typical missing-values approaches based on imputation or expectation maximization. In particular, they are suited for missing not at random (MNAR) settings.

Here we adapt NeuMiss networks to specific classification settings. Section 3 exposes the original NeuMiss networks. Section 4 shows how they can be adapted

to binary classification with a probit model. Finally, section 5 studies the performance of NeuMiss networks on a simple classification task with missing values.

2 Problem setting: Notations and assumptions

Notations We consider a pair of random variables (X, Y), where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ for regression or $Y \in \{-1, 1\}$ for binary classification. We consider a random vector $M \in \{0, 1\}^d$ acting as a mask on X: for all $1 \leq j \leq d$, $M_j = 1 \Leftrightarrow X_j$ not observed. The incomplete feature vector $\tilde{X} \in \mathcal{X} := (\mathbb{R} \cup \{\texttt{NA}\})^d$ is defined as $\tilde{X}_j = \texttt{NA}$ if $M_j = 1$ and $\tilde{X}_j = X_j$ if $M_j \neq 1$.

We note $obs(M)$ (resp. $mis(M)$) the zero (resp. non-zero) entries of M. We denote by X_{obs} (resp. X_{mis}) the observed (resp. missing) values of X.

Assumption 1 (Gaussian data)**.** *The complete data is distributed as a multivariate Gaussian:* $X \sim \mathcal{N}(\mu, \Sigma)$.

Assumption 2 (Missing At Random (MAR))**.** *For all* $m \in \{0, 1\}$, $\mathbb{P}[M = m \mid X] = \mathbb{P}[M = m \mid X_{obs}]$.

Missing Completely At Random (MCAR) is a special case of MAR where the probability of M does not depend on the covariates.

Assumption 3 (Gaussian self-masking)**.** *The probability that a variable is missing depends on its own value through a Gaussian function:*

$$\mathbb{P}[M \mid X] = \prod_{k=1}^{d} \mathbb{P}[M_k \mid X_k] \tag{1}$$

$$\forall\ 1 \leq k \leq d, \quad \mathbb{P}[M_k = 1 \mid X_k] = K_k \exp\left(-\frac{(X_k - \tilde{\mu}_k)^2}{2\,\tilde{\sigma}_k^2}\right) \tag{2}$$

where $0 < K_k < 1$.

Gaussian self-masking is an instance of a Missing Not At Random (MNAR) setting.

3 NeuMiss: regression with missing values

For regression settings, Le Morvan et al. [4] introduced a theoretically-grounded architecture, NeuMiss networks, designed to approximate optimal predictors with missing values. Therefore, we recall below the expression of the Bayes predictors for a linear regression setting.

Assumption 4 (Linear model)**.** *The response* $Y \in \mathbb{R}$ *is linked to the complete data* X *through a linear model:*

$$Y = \beta_0^\star + \langle X, \beta^\star \rangle + \epsilon, \quad \textit{where } \beta_0^\star \in \mathbb{R},\ \beta^\star \in \mathbb{R}^d,\ \textit{and } \epsilon \sim \mathcal{N}\left(0, \sigma^2\right),\ \sigma \in \mathbb{R}. \tag{3}$$

Bayes predictor for least-squares regression For a least-squares loss, the Bayes predictor $f^\star_{\tilde{X}}$ is defined as a solution of the following optimization problem:

$$f^\star_{\tilde{X}} \in \underset{f:\tilde{\mathcal{X}}\to\mathcal{Y}}{\operatorname{argmin}}\, \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right] \tag{4}$$

The solution verifies $f^\star_{\tilde{X}} = \mathbb{E}\left[Y \mid \tilde{X}\right]$ (see *e.g.* sec 1.5.5 of [1]). Le Morvan et al. [4] showed that under Assumptions 4 (linear model) and 1 (Gaussian data), and in MAR settings (Assumption 2), the Bayes predictor reads:

$$f^\star(X_{obs}, M) = \beta^\star_0 + \langle \beta^\star_{obs}, X_{obs}\rangle + \left\langle \beta^\star_{mis}, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs})\right\rangle \tag{5}$$

They also provide an expression under the Gaussian self masking assumption, but in the sake of brevity we do not introduce its expression here.

Approximation with differentiable programming Le Morvan et al. [4] introduce a Neural network to approximate the above Bayes predictors. The key difficulty to approximate the Bayes predictor (5) lies in the approximation of $(\Sigma_{obs})^{-1}$ for any possible set of observed values. [4] solve this problem with a NeuMiss block: it consists of a number of layers that compute, based on an algorithm unfolding strategy, a truncated Neumann series for $(\Sigma_{obs})^{-1}$. Practically, this strategy can be implemented as follows: each layer i applies a $x \mapsto W^{(i)}x + x$ transformation followed by a new type of non-linearity, the elementwise multiplication by the mask. The NeuMiss block takes as input $(X - \mu)$ where the missing values have been imputed by zero, and μ is a parameter learned by the network, so that the first block approximates the quantity $(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs})$. The second block approximates the remaining operations, and notably the dot product with $\beta^\star$. Both blocks combined approximate the Bayes predictor in a MAR setting, with an error that decays exponentially fast with the depth of the NeuMiss block ([4]). A diagram summarizing the operations is given in Figure 1. Note that this architecture avoids explicitely estimating the covariance matrix and inverting it for each missing-value pattern. An important benefit from this is that contrary to methods such as EM, it was shown to perform well even under the Gaussian self-masking MNAR Bayes predictor.

4 Optimal predictors in classification

In this section, we derive the expression of the Bayes predictors in a classification setting. Based on these expressions we then show that NeuMiss networks can be adapted to classification settings with a simple modification of the architecture.

From now on we consider a binary random variable $Y \in \{-1, 1\}$. In this binary classification setting the goal is to predict the class Y from $\tilde{X}$. With the 0-1 loss, the Bayes predictor $f^\star_{\tilde{X}}$ is a solution of the optimization problem:

$$f^\star_{\tilde{X}} \in \underset{f:\tilde{\mathcal{X}}\to\mathcal{Y}}{\operatorname{argmin}}\, \mathbb{E}\left[\mathbb{1}_{f(\tilde{X})\neq Y}\right], \tag{6}$$

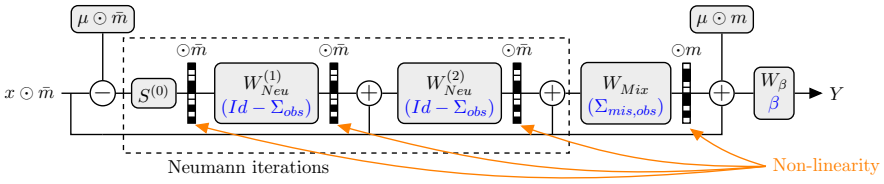


Fig. 1: **NeuMiss network architecture with a depth of 4** — $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}$ corresponds to a simple transformation of the covariance matrix indicated in blue. Taken from [4].

which is minimal for:

$$f^\star_{\tilde{X}}(\tilde{X}) = \begin{cases} +1 & \text{if} \quad \eta(\tilde{X}) \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}, \quad \text{with } \eta(\tilde{x}) := \mathbb{P}[Y = 1 \mid \tilde{X} = \tilde{x}]. \tag{7}$$

Intuitively, it means that in classification, the Bayes predictor predicts the class with highest posterior class probability for a given point $\tilde{x}$.

Assumption 5 (Probit model)**.** *The response $Y \in \{-1, 1\}$ is linked to the complete data X through the following relation:*

$$\mathbb{P}\left[Y = 1 \mid X, \beta_0^\star, \beta^\star\right] = \Phi(\langle X, \beta^\star \rangle + \beta_0^\star), \tag{8}$$

where $\Phi : \mathbb{R} \to [0, 1]$ is the cumulative distribution function of the standard normal distribution.

Proposition 1 (M(C)AR Bayes predictor). *For data generated via the probit model (Assumption 5), and under Assumption 1 (Gaussian data) and Assumption 2 (MAR), then:*

$$\eta(\tilde{X}) = \Phi\left(\frac{\nu}{(1 + \sigma^2)^{\frac{1}{2}}}\right), \tag{9}$$

with:

$$\nu := \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \left\langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \right\rangle \tag{10}$$

$$\sigma^2 := \beta_{mis}^{\star T} \Sigma_{mis} \beta_{mis}^\star - \beta_{mis}^{\star T} \Sigma_{mis,obs}(\Sigma_{obs})^{-1} \Sigma_{mis,obs}^T \beta_{mis}^\star \tag{11}$$

and the Bayes predictor can be written $f^\star_{\tilde{X}}(\tilde{X}) = sign(\nu)$.

Proposition 1 shows that the Bayes classifier under the probit model is related to the Bayes predictor in regression (5). Indeed, the quantity ν exactly corresponds to (5). Similar results hold under the Gaussian self-masking mechanism:

Proposition 2 (Gaussian self-masking Bayes predictor). *For data generated via the probit model, and under Assumption 1 and Assumption 3, then:*

$$\eta(\tilde{X}) = \Phi\left(\frac{\nu}{(1+\sigma^2)^{\frac{1}{2}}}\right), \tag{12}$$

where:

$$\begin{aligned}\nu = \; & \beta_0^\star + \langle \beta_{obs}^\star, X_{obs}\rangle + \langle \beta_{mis}^\star, (Id + D_{mis}\Sigma^{-1}_{mis\,|\,obs}) \\ & \times (\tilde{\mu}_{mis} + D_{mis}\Sigma^{-1}_{mis\,|\,obs}(\mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs})))\rangle\end{aligned} \tag{13}$$

$$\sigma^2 = \; \beta_{mis}^{\star T}\left(D_{mis}^{-1} + \Sigma^{-1}_{mis|obs}\right)^{-1}\beta_{mis}^\star \tag{14}$$

with $\Sigma_{mis\,|\,obs} := \Sigma_{mis} - \Sigma_{mis,obs}(\Sigma_{obs})^{-1}\Sigma_{obs,mis}$ *and* D *diagonal of* $(\tilde{\sigma}_1^2, ..., \tilde{\sigma}_d^2)$.

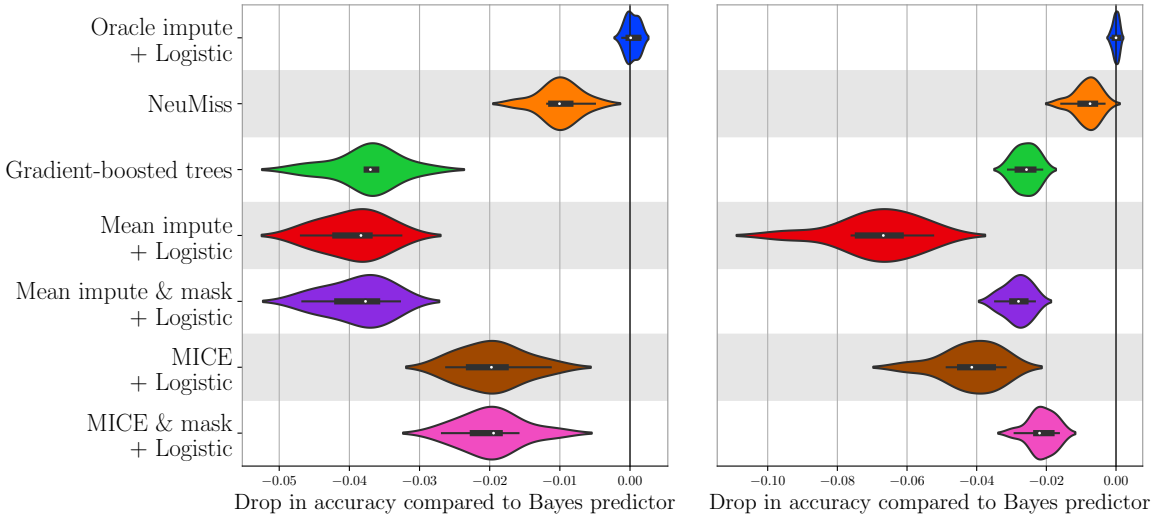
The Bayes predictor can be written: $f^\star_{\tilde{X}}(\tilde{X}) = sign(\nu)$

Again here, the quantity ν exactly corresponds to the expression of the Bayes predictor in a regression setting with Gaussian self-masking. Thus, Propositions 1 and 2 show that the Bayes classifier only depends on the sign of the same expressions as in regression settings. Hence, the NeuMiss architecture can be readily adapted to classification settings, as it approximates this specific expression. In practice, it suffices to add a sigmoid non-linearity at the output of the NeuMiss architecture and optimize a binary cross-entropy loss instead of a least-squares loss.

5 Empirical study of classification with missing values

We simulated two datasets of 500 000 samples and 50 numerical features drawn from a multivariate gaussian $\mathcal{N}(\mu, \Sigma)$. One has MCAR missing-values and the other has Gaussian self-masking missing values, both having about 50% missing values. Entries of mean μ are sampled from a normal distribution and the ones of covariance Σ are computed from $\Sigma = B^T B + \text{diag}(\epsilon)$ with $B \in \mathbb{R}^{\frac{d}{2}\times d}$ having its entries sampled from a normal distribution and ϵ having its entries drawn uniformly from $[0.01, 0.1]$. The outcome Y is generated under the probit model. We drew 10 simulations for each of the two settings. Classes are balanced with around 54% positive samples on average. Although the label generation model is probit, using logistic regression is justified by the proximity between the logit and probit functions. We chose arbitrarily a depth of 25 for the first block of NeuMiss without cross-validating since weights are shared in this block, and thus the deeper the network, the better is the approximation without the risk of overfitting. For the training, we used the Adam optimizer, an adaptive learning rate and early stopping.

We compare a NeuMiss architecture trained with a binary cross entropy loss to gradient boosted trees with Missing Incorporated in Attributes (`HistGradientBoostingClassifier` in scikit-learn [9]) and to logistic regression after imputation: either mean imputation or MICE [12] (using `IterativeImputer` in scikit-learn). For comparison purposes, we also implemented an oracle imputation:



(a) MCAR (b) Gaussian self-masking

Fig. 2: **Prediction benchmarks.** Accuracy of the benchmarked methods relative to the accuracy of the Bayes predictor, on 10 simulated datasets of Gaussian data of 500 000 samples and 50 numerical features for MCAR and GSM settings. About 50% values are missing in each.

missing values are replaced with the analytic expression of $\mathbb{E}[X_{mis} \mid X_{obs}, M]$ knowing the parameters of the data generation (μ, Σ, and the parameters for the Gaussian self-masking mechanism when relevant). As in common in MNAR settings, we also test concatenating the missingness "mask" –an indicator variable of whether an entry has been imputed or not– to the imputed values. The score obtained by the Bayes predictor is used as reference.

As expected, Oracle Impute + Logistic performs best since it only differs from the Bayes predictor by the fact that $\beta^\star$ is learned. However, in practice, the ground truth parameters of the distribution are unknown so this predictor cannot be used. NeuMiss networks obtained higher relative score than every other non-oracle methods (Fig. 2). Importantly, NeuMiss networks are also robust to the missing-value mechanism: they are the best performers both in MAR and in MNAR settings. This is unlike MICE imputation which needs the mask in MNAR settings.

Conclusion NeuMiss networks were derived to approximate Bayes predictors in regression settings with gaussian data. Our theoretical results show that they can also approximate the Bayes predictors in binary classification assuming a probit link. Here also, they are robust to the missingness mechanism, fitting Gaussian self masking MNAR beyond MAR. Further work remains to evaluate NeuMiss in classification on real-world non-Gaussian data and to consider multiclass.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer International Publishing (2006). https://doi.org/10.5555/1162264
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. KDD (2016)
3. Josse, J., Prost, N., Scornet, E., Varoquaux, G.: On the consistency of supervised learning with missing values. arXiv preprint arXiv:1902.06931 (2019)
4. Le Morvan, M., Josse, J., Moreau, T., Scornet, E., Varoquaux, G.: NeuMiss networks: differentiable programming for supervised learning with missing values. Advances in Neural Information Processing Systems **33** (2020)
5. Le Morvan, M., Josse, J., Moreau, T., Scornet, E., Varoquaux, G.: NeuMiss networks: differentiable programming for supervised learning with missing values. Advances in Neural Information Processing Systems **33**, 5980–5990 (2020)
6. Little, R.J., Rubin, D.B.: Statistical analysis with missing data, vol. 793. John Wiley & Sons (2019)
7. Mattei, P.A., Frellsen, J.: Miwae: Deep generative modelling and imputation of incomplete data sets. International Conference on Machine Learning p. 4413 (2019)
8. Morvan, M.L., Prost, N., Josse, J., Scornet, E., Varoquaux, G.: Linear predictor on linearly-generated data with missing values: non consistency and solutions http://arxiv.org/abs/2002.00658
9. Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python . Journal of Machine Learning Research **12**, 2825–2830 (2011)
10. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
11. Twala, B.E.T.H., Jones, M.C., Hand, D.J.: Good methods for coping with missing data in decision trees. Pattern Recogn. Lett. **29**, 950–956 (May 2008)
12. Van Buuren, S.: Flexible imputation of missing data. CRC press (2018)

6 Appendix

6.1 Proof of Proposition 1

Proof. As proved in section A.3 of [5], for all $j \in mis(M)$, we have $\mathbb{P}\left[X_j|M, X_{obs}\right] = \mathbb{P}\left[X_j|X_{obs}\right]$ and thus:

$$\mathbb{P}\left[X_{mis}|M, X_{obs}\right] = \mathbb{P}\left[X_{mis}|X_{obs}\right] \tag{15}$$

both under the MCAR and MAR assumptions.

$$\begin{aligned}
\eta(\tilde{X}) &= \mathbb{P}\left[Y = 1|M, X_{obs(M)}\right] \\
&= \int_{X_{mis(M)}} \mathbb{P}\left[Y = 1|M, X_{obs(M)}, X_{mis(M)}\right] \mathbb{P}\left[X_{mis(M)}|M, X_{obs(M)}\right] \\
&= \int_{X_{mis(M)}} \mathbb{P}\left[Y = 1|X\right] \mathbb{P}\left[X_{mis(M)}|M, X_{obs(M)}\right] \\
&= \int_{X_{mis}} \mathbb{P}\left[Y = 1|X\right] \mathbb{P}\left[X_{mis}|X_{obs}\right] dX_{mis} \qquad \text{using (15)} \\
&= \int_{X_{mis}} \Phi\left(\langle X, \beta^\star\rangle + \beta_0^\star\right) \mathbb{P}\left[X_{mis}|X_{obs}\right] dX_{mis} \qquad \text{using (8)} \\
&= \int_{X_{mis}} \Phi\left(\langle X_{mis}, \beta_{mis}^\star\rangle + c\right) \mathbb{P}\left[X_{mis}|X_{obs}\right] dX_{mis} \qquad (16)
\end{aligned}$$

where we note $c := \langle X_{obs}, \beta^\star_{obs} \rangle + \beta^\star_0$ for clarity.

Since $X \sim \mathcal{N}(\mu, \Sigma)$, the conditional property of multivariate normal distributions gives:

$$X_{mis}|X_{obs} \sim \mathcal{N}\left(\mu_{mis|obs}, \Sigma_{mis|obs}\right) \tag{17}$$

$$\textit{with} \quad \begin{aligned} \mu_{mis|obs} &:= \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}\left(X_{obs} - \mu_{obs}\right) \\ \Sigma_{mis|obs} &:= \Sigma_{mis} - \Sigma_{mis,obs}(\Sigma_{obs})^{-1}\Sigma^T_{mis,obs} \end{aligned} \tag{18}$$

using the notations of [5].

Using (17) in (16) gives:

$$\eta(\tilde{X}) = \int \Phi\left(\langle X_{mis}, \beta^\star_{mis} \rangle + c\right) \mathcal{N}\left(X_{mis}; \mu_{mis|obs}, \Sigma_{mis|obs}\right) dX_{mis} \tag{19}$$

To get an analytic formulation of $\eta(\tilde{X})$ from (19), we use the computations by Bishop section 4.5.2 of [1]. First note that:

$$\Phi\left(\langle X_{mis}, \beta^\star_{mis} \rangle + c\right) = \int \delta\left(a - \langle X_{mis}, \beta^\star_{mis} \rangle + c\right) \Phi(a) da \tag{20}$$

We note:

$$p(a) := \int \delta\left(a - \langle X_{mis}, \beta^\star_{mis} \rangle + c\right) \mathcal{N}\left(X_{mis}; \mu_{mis|obs}, \Sigma_{mis|obs}\right) dX_{mis} \tag{21}$$

Using (20) we have:

$$\eta(\tilde{X}) = \int \Phi(a) p(a) da \tag{22}$$

[1] shows that p(a) is also Gaussian: $p(a) \sim \mathcal{N}\left(\nu, \sigma^2\right)$ and that the convolution of the probit function by a Gaussian is another probit function.
Equation (22) gives:

$$\begin{aligned} \eta(\tilde{X}) &= \int \Phi(a) \mathcal{N}\left(a; \nu, \sigma^2\right) da \\ &= \Phi\left(\frac{\nu}{(1+\sigma^2)^{\frac{1}{2}}}\right) \end{aligned} \tag{23}$$

Let us compute its mean and variance.

$$\begin{aligned} \nu &= \int_a p(a) a da \\ &= \int \left(\langle X_{mis}, \beta^\star_{mis} \rangle + c\right) \mathcal{N}\left(X_{mis}; \mu_{mis|obs}, \Sigma_{mis|obs}\right) dX_{mis} \\ &= \left\langle \mu_{mis|obs}, \beta^\star_{mis} \right\rangle + c \end{aligned} \tag{24}$$

$$\begin{aligned}\sigma^2 &= \int_a p(a)\left(a^2-\nu^2\right)da \\ &= \int \left(\left(\langle X_{mis},\beta^\star_{mis}\rangle + c\right)^2 - \left(\left\langle \mu_{mis|obs},\beta^\star_{mis}\right\rangle + c\right)^2\right)\mathcal{N}\left(X_{mis};\mu_{mis|obs},\Sigma_{mis|obs}\right)dX_{mis} \\ &= ... \\ &= \beta^{\star T}_{mis}\Sigma_{mis|obs}\beta^\star_{mis} \end{aligned} \tag{25}$$

Their full expression:

$$\begin{aligned}\nu &= \beta^{\star T}_{mis}\mu_{mis} + \beta^{\star T}_{mis}\Sigma_{mis,obs}(\Sigma_{obs})^{-1}\left(X_{obs}-\mu_{obs}\right) + \beta^{\star T}_{obs}X_{obs} + \beta^\star_0 \\ \sigma^2 &= \beta^{\star T}_{mis}\Sigma_{mis}\beta^\star_{mis} - \beta^{\star T}_{mis}\Sigma_{mis,obs}(\Sigma_{obs})^{-1}\Sigma^T_{mis,obs}\beta^\star_{mis}\end{aligned}$$

Since $\Phi(z) \geq \frac{1}{2} \Leftrightarrow z \geq 0$, using (23) we have $\eta(\tilde{X}) \geq \frac{1}{2} \Leftrightarrow \nu \geq 0$.

The Bayes estimator under the MCAR and MAR assumptions for the probit model thus writes:

$$f^\star_{\tilde{X}}(\tilde{X}) = \text{sign}(\nu)$$

The probability of the positive class being: $\eta(X) = \Phi\left(\frac{\nu}{(1+\sigma^2)^{\frac{1}{2}}}\right)$.

□

6.2 Proof of Proposition 2

Proof. As proved in section A.4 of [5], we have:

$$\mathbb{P}\left[X_{mis}|M,X_{obs}\right] = \mathcal{N}\left(X_{mis};a_M,A_M\right) \tag{26}$$

$$\text{with} \quad \begin{aligned} A_M &:= \left(D^{-1}_{mis} + \Sigma^{-1}_{mis|obs}\right)^{-1} \\ a_M &:= A_M\left(D^{-1}_{mis}\tilde{\mu}_{mis} + \Sigma^{-1}_{mis|obs}\mu_{mis|obs}\right)\end{aligned} \tag{27}$$

Using the definition of A_M we get:

$$a_M = \left(Id + D_{mis}\Sigma^{-1}_{mis|obs}\right)^{-1}\left(\tilde{\mu}_{mis} + D_{mis}\Sigma^{-1}_{mis|obs}\left(\mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}\left(X_{obs}-\mu_{obs}\right)\right)\right) \tag{28}$$

with $\Sigma_{mis|obs}$ defined in (18).

Since $\mathbb{P}\left[X_{mis}|M,X_{obs}\right]$ is a multivariate normal distribution we can use the results from the proof of Proposition 1. We have:

$$\eta(\tilde{X}) = \int \Phi\left(\langle X_{mis},\beta^\star_{mis}\rangle + c\right)\mathcal{N}\left(X_{mis};a_M,A_M\right)dX_{mis} \tag{29}$$

which is the same equation as (19) with a_M and A_M instead of $\mu_{mis|obs}$ and $\Sigma_{mis|obs}$ respectively. Thus, as in Proposition 1, the probability of the positive class reads:

$$\eta(\tilde{X}) = \Phi\left(\frac{\nu}{(1+\sigma^2)^{\frac{1}{2}}}\right)$$

with:

$$\begin{aligned}\nu &= \langle a_M, \beta^\star_{mis}\rangle + \langle X_{obs}, \beta^\star_{obs}\rangle + \beta^\star_0 \\ \sigma^2 &= \beta^{\star T}_{mis} A_M \beta^\star_{mis}\end{aligned}$$

The Bayes estimator under the Gaussian self-masking assumption for the probit model is:

$$f^\star_{\tilde{X}}(\tilde{X}) = \text{sign}(\nu) \tag{30}$$

□