A primer on linear classification with missing data

Angel Reyero Lobo[1,2] Alexis Ayme[3] Claire Boyer[4] Erwan Scornet[5]

[1]Université de Toulouse [2]Inria Paris-Saclay [3] École Normale Supérieure [4] Université Paris-Saclay [5] Sorbonne Université

Abstract

Supervised learning with missing data aims to build the best possible prediction of a target output based on partially observed inputs. The main approaches to address this problem can be divided into: (i) **impute-then-predict** strategies, which first fill in the missing input components and then apply a single predictor; and (ii) **pattern-by-pattern** approaches, where a separate predictor is trained for each missing data pattern.

It is essential to study, from a theoretical perspective, how standard linear classifiers—namely the perceptron, logistic regression, and linear discriminant analysis (LDA)—can be adapted to effectively handle missing values.

Missing values framework

Notation: Let an observation with missing values $(X_{\mathrm{obs}(M)}, M, Y)$ be:

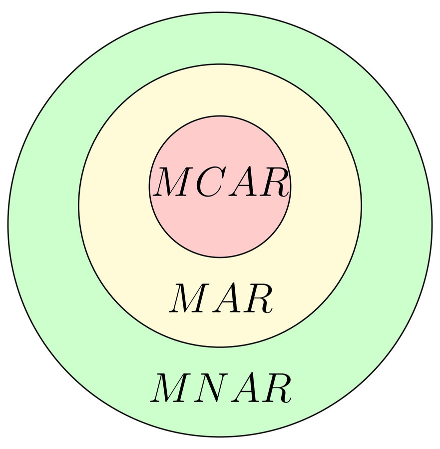
- Missing value pattern $M \in \{0,1\}^d$ such that
$$M_j = 1 \iff X_j \text{ is missing.}$$
- $\mathrm{obs}(M) := \{j \in \{1, ...d\} | M_j = 0\}$.
- $X_{\mathrm{obs}(M)}$ observed covariates.
- $Y \in \{-1, 1\}$ the label (*always* observed).

Example:

$$\begin{aligned} X &= (6, 3, \mathtt{NA}, 3, \mathtt{NA}),\\ M &= (0, 0, 1, 0, 1),\\ \mathrm{obs}(M) &= (1, 2, \quad 4 \quad),\\ X_{\mathrm{obs}(M)} &= (6, 3, \quad 3 \quad). \end{aligned}$$

Missing values mechanism: Assumptions on $M \mid X, Y$ categorized into

- **MCAR** (Missing completely at random). $M \perp\!\!\!\perp X, Y$.
- **MAR** (Missing at random). $\forall m \in \{0,1\}^d$, $\mathbb{P}(M = m \mid X, Y) = \mathbb{P}(M = m \mid X_{\mathrm{obs}(m)})$.
- **MNAR** (Missing not at random). M depends on X and Y.



Classification with missing values

- **Complete data case**
  - Dataset: $\mathcal{D}_n = \{(X_i, Y_i), i \in \{1, ...n\}\}$
  - Misclassification probability:
$$\mathcal{L}_{\mathrm{comp}}\left(\widehat{h}_{\mathrm{comp}}\right) := \mathbb{P}\left(\widehat{h}_{\mathrm{comp}}(X) \neq Y\right).$$
  - Bayes classifier:
$$h^\star_{\mathrm{comp}}(X) = \mathrm{sign}(\mathbb{E}[Y \mid X]).$$
- **Incomplete data case**
  - Dataset: $\mathcal{D}^\star_n = \{(X_{i,\mathrm{obs}(M_i)}, M_i, Y_i), i \in \{1, ...n\}\}$
  - Misclassification probability:
$$\mathcal{L}\left(\widehat{h}\right) := \mathbb{P}\left(\widehat{h}(X_{\mathrm{obs}(M)}, M) \neq Y\right).$$
  - Bayes classifier:
$$h^\star(X_{\mathrm{obs}(M)}, M) := \mathrm{sign}(\mathbb{E}[Y \mid X_{\mathrm{obs}(M)}, M]) = \sum_{m \in \{0,1\}^d} h^\star_m(X_{\mathrm{obs}(m)}) \mathbb{1}_{M=m},$$

  with $h^\star_m(X_{\mathrm{obs}(m)}) := \mathrm{sign}(\mathbb{E}[Y \mid X_{\mathrm{obs}(m)}, M = m])$. It can be decomposed pattern-by-pattern!

Prediction VS Model Inference

Model estimation can be done via MLE using EM algorithm.

⚠ Missing values in the training set and in the test set

✗ Estimating the underlying model does not help for prediction:
$$\mathbb{E}[Y|X] = f_\beta(X) \quad \Rightarrow \quad \widehat{Y} \neq f_\beta(X_{\mathrm{obs}(M)}).$$

We need to design predictors handling missing entries:

- **Impute-then-predict** (Morvan et al. [2021]).
- **Pattern-by-pattern decomposition** (Ayme et al. [2022]).

Lemma: Bayes optimality for linear classifiers

If a p-b-p approach with linear classifiers is not Bayes optimal, then constant imputation with linear classifiers is not Bayes optimal.

Main question

Does the pattern-by-pattern Bayes classifier conserve the model structure on the observed covariates as happens with the linear model (Morvan et al. [2020])?
$$\mathbb{E}[Y|X] = f_\beta(X) \quad \overset{?}{\Rightarrow} \quad \mathbb{E}[Y|X_{\mathrm{obs}}, M] = f_{\beta'}(X_{\mathrm{obs}(M)}, M).$$

Perceptron

To ensure the convergence of the perceptron, we need the linear separability.

Lemma: p-b-p perceptron

Linear separability of complete data does not imply that of incomplete data.

✗ The p-b-p and constant imputation are not Bayes optimal.

Logistic regression

We make the following assumption about the distribution of complete data.

Assumption: Logistic model

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

Proposition: p-b-p logistic regression

Assume $M \perp\!\!\!\perp X, Y$ (MCAR) and logistic model for complete data. 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 $\beta^\star_m \in \mathbb{R}^{|\mathrm{obs}(m)|+1}$ such that
$$\mathbb{P}\left(Y = 1 \mid X_{\mathrm{obs}(m)}, M = m\right) = \sigma\Big(\beta^\star_{0,m} + \sum_{j \in \mathrm{obs}(m)} \beta^\star_{j,m} X_j\Big).$$
Then, for all $j \in \mathrm{mis}(m)$, $\beta^\star_j = 0$.

✗ The p-b-p and constant imputation are not Bayes optimal.

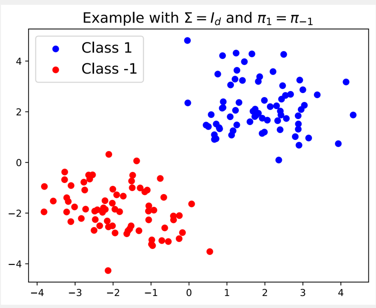
Linear discriminant analysis (LDA)

We make the following assumption about the distribution of complete data.

Assumption: Balanced LDA

Denoting $\pi_1 := \mathbb{P}(Y = 1)$ and $\pi_{-1} := \mathbb{P}(Y = -1)$, then we assume that:

- $X \mid Y = k \sim \mathcal{N}(\mu_k, \Sigma)$,
- $\pi_1 = \pi_{-1}$.



This assumption yields a closed-form Bayes classifier:

Proposition: Bayes classifier with complete data

Under the **LDA** model, the Bayes classifier is
$$h^\star(x) = \mathrm{sign}\left((\mu_1 - \mu_{-1})^\top \Sigma^{-1}\left(x - \frac{\mu_1 + \mu_{-1}}{2}\right)\right).$$

Proposition: p-b-p LDA

Under the **LDA** model with **MCAR** inputs, the p-b-p Bayes classifier is
$$h^\star_m(x_{\mathrm{obs}(m)}) = \mathrm{sign}\left(\left(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right)^\top \Sigma^{-1}_{\mathrm{obs}(m)}\left(x_{\mathrm{obs}(m)} - \frac{\mu_{1,\mathrm{obs}(m)} + \mu_{-1,\mathrm{obs}(m)}}{2}\right)\right)$$

✓ P-b-p is Bayes optimal!

✓ They are the projected parameters!

Proposition: constant imputation LDA

Under the **LDA** model with **MCAR** inputs, constant imputation is optimal only if Σ diagonal.

Experiments

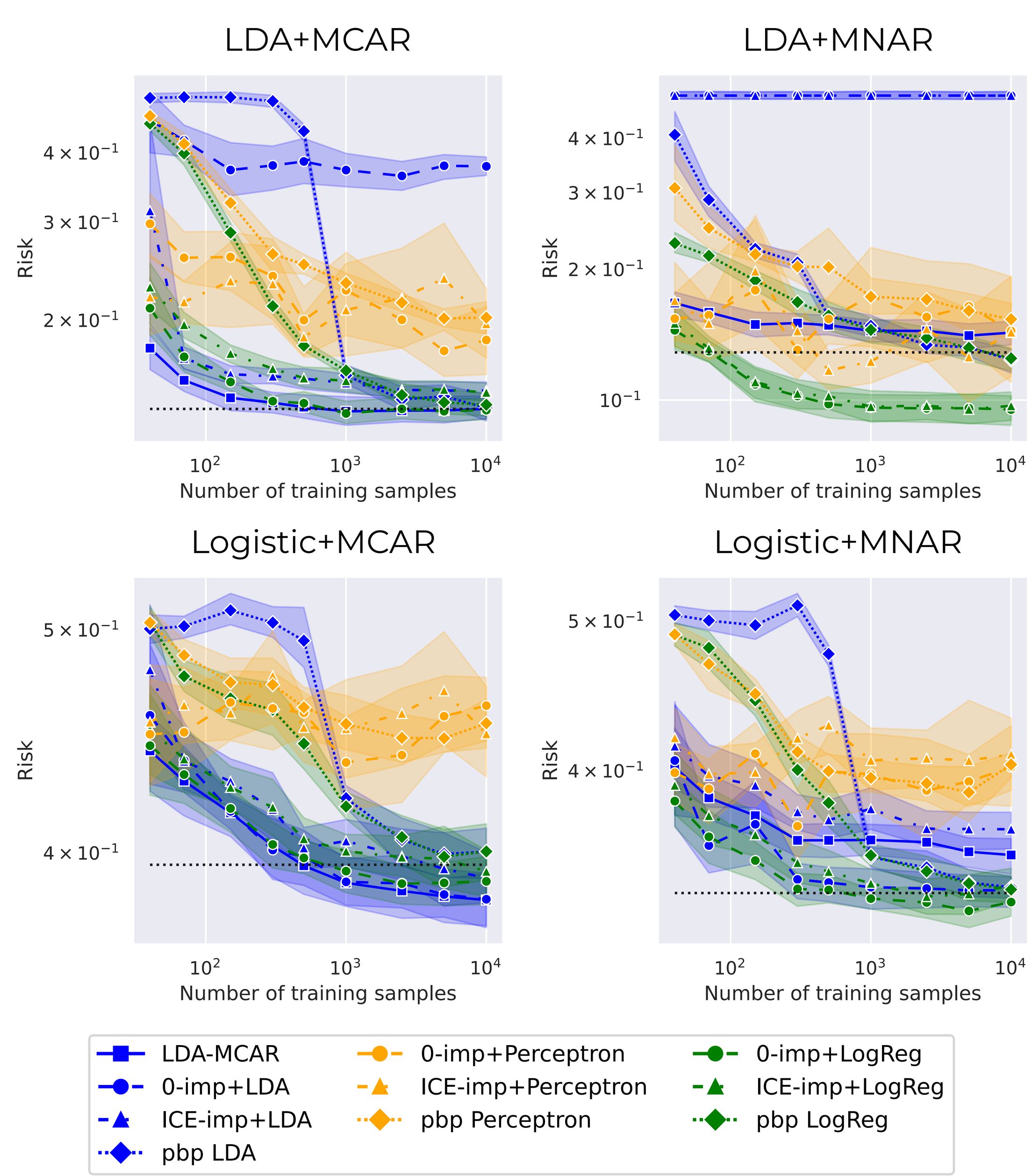


Figure: Excess risks of several classifiers on generated data (LDA or Logistic framework) with $\Sigma = I_d$ and MCAR or MNAR missing mechanisms. Dotted lines stand for the missing Bayes risk.

Take-home message

Scarcity of methods for prediction with missing values $\Rightarrow$ **p-b-p decomposition**

- On the **perceptron**:
  - P-b-p linear separability not preserved in general $\Rightarrow$ imputation and p-b-p do not work.
- On the **logistic regression**:
  - Logistic model assumption not preserved $\Rightarrow$ imputation and p-b-p do not work.
- On the **LDA**(with MCAR):
  - It accepts p-b-p decomposition!
  - Imputation only valid with Σ diagonal.
  - ✓ Other finite-sample analyses for parameter estimation and MNAR data are readily available.

References

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Near-optimal rate of consistency for linear models with missing values. In *ICML*, 2022. URL `https://proceedings.mlr.press/v162/ayme22a.html`.

Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gael Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *AISTATS*, 2020. URL `https://proceedings.mlr.press/v108/morvan20a.html`.

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gael Varoquaux. What's a good imputation to predict with missing values? In *NeurIPS*, 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/5fe8fdc79ce292c39c5f209d734b7206-Paper.pdf`.

angel.reyero-lobo@inria.fr