A primer on linear classification with missing data

Angel REYERO LOBO

Journées de Statistique
IMT & Inria Paris-Saclay
Joint work with:
Alexis Ayme, Claire Boyer & Erwan Scornet.

June 5, 2025

Contents

Index

Motivation

Trauma.center	Heart rate	Death	Anticoagulant. therapy	Glascow score
Pitie-Salpêtrière	88	0	No	3
Beaujon	103	0	NA	5
Bicêtre	NA	0	Yes	6
Bicêtre	NA	0	No	NA
Lille	62	0	Yes	6
Lille	NA	0	No	NA
	⋮	⋮	⋮	⋮

Different sources of missing values (`Not Available(NA)`):

- Bugs/ Sensors failures
- Costs
- Sensitive data
- Data merging
- No time to measure in an emergency situation

Notation

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

- Missing value pattern $M \in \{0,1\}^d$ such that

$$M_j = 1 \quad \iff \quad X_j \text{ is missing.}$$

- $\mathrm{obs}(M) := \{j \in \{1, \dots d\} | M_j = 0\}$.
- $X_{\mathrm{obs}(M)}$ observed covariates.
- $Y \in \{-1, 1\}$ the label (*always* observed).

Example:

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

Supervised learning with missing values: Classification

- Complete data case
  - Dataset: $\mathscr{D}_n = \{(X_i, Y_i), i \in \{1, ...n\}\}$
  - Misclassification probability:

$$\mathscr{L}_{\text{comp}}\left(\widehat{h}_{\text{comp}}\right) := \mathbb{P}\left(\widehat{h}_{\text{comp}}(X) \neq Y\right).$$

- Incomplete data case
  - Dataset: $\mathscr{D}_n^\star = \{(X_{i,\text{obs}(M_i)}, M_i, Y_i), i \in \{1, ...n\}\}$
  - Misclassification probability:

$$\mathscr{L}\left(\widehat{h}\right) := \mathbb{P}\left(\widehat{h}(X_{\text{obs}(M)}, M) \neq Y\right).$$

Missing values mechanism

Assumptions on $M \mid X, Y$ Rubin [1976]:

- **MCAR** (Missing completely at random).

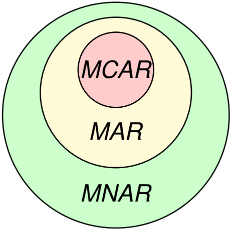
$$M \perp\!\!\!\perp X, Y.$$

- **MAR** (Missing at random).

$$\forall m \in \{0,1\}^d,$$

$$\mathbb{P}(M = m \mid X, Y) = \mathbb{P}\left(M = m \mid X_{obs(m)}\right).$$

- **MNAR** (Missing not at random). M depends on the full vectors X and Y.



Index

Pattern-by-pattern Bayes predictor

Proposition

The Bayes predictor is given by

$$h^\star(X_{\mathrm{obs}(M)}, M) := \mathrm{sign}(\mathbb{E}\left[Y \mid X_{\mathrm{obs}(M)}, M\right]).$$

Pattern-by-pattern decomposition:

$$h^\star(Z) = h^\star(X_{\mathrm{obs}(M)}, M) = \sum_{m \in \{0,1\}^d} h^\star_m(X_{\mathrm{obs}(m)}) \mathbb{1}_{M=m}$$

with

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

Problematic: prediction vs model inference

- Estimation via MLE using EM algorithm (Dempster et al. [1977]).
- (!) Missing values in the training set and in the test set

✗ Estimating the underlying model does not help for prediction:

$$\mathbb{E}[Y \mid X] = f_{\beta}(X) \quad \not\Rightarrow \quad \widehat{Y} \neq f_{\widehat{\beta}}(X_{\mathrm{obs}(M)}).$$

- We need to design predictors handling missing entries:
  - **Impute-then-predict** (Josse et al. [2019], Le Morvan et al. [2021]).
  - **Pattern-by-pattern decomposition** (Ayme et al. [2022]).
- (!) The pattern-by-pattern Bayes classifier may not conserve the model structure on the observed covariates.

$$\mathbb{E}[Y \mid X] = f_{\beta}(X) \quad \overset{?}{\Rightarrow} \quad \mathbb{E}[Y \mid X_{\mathrm{obs}(M)}, M] = f_{\beta'}(X_{\mathrm{obs}(M)}, M)$$

✓ (Linear model) Morvan et al. [2020]

✗ (Logistic model) This work.

Index

Perceptron

Lemma

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

- To ensure the convergence of the *p-b-p perceptron*, we need the linear separability (Novikoff [1962]).

Lemma

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

Assumption (Logistic model)

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

Proposition

Assume $M \perp\!\!\!\perp X, Y$ (MCAR) and logistic model for complete data. If the logistic model holds on the missing pattern $M = m$ for $m \in \{0,1\}^d$, i.e. there exist a vector $\beta_m^\star \in \mathbb{R}^{|\mathrm{obs}(m)|+1}$ such that

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

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

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

LDA in the complete data case

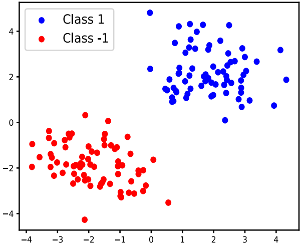
Assumption (Balanced LDA): $X \mid Y = k \sim \mathcal{N}(\mu_k, \Sigma)$, $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1)$.

Proposition

The Bayes predictor reads as

$$h^\star(x) = \operatorname{sign}\left(\left(\mu_1 - \mu_{-1}\right)^\top \Sigma^{-1}\left(x - \frac{\mu_1 + \mu_{-1}}{2}\right)\right).$$

Example with $\Sigma = I_d$ and $\pi_1 = \pi_{-1}$



LDA in the missing data case

Proposition

Under the **LDA** model with **MCAR** inputs, the p-b-p classifier is

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

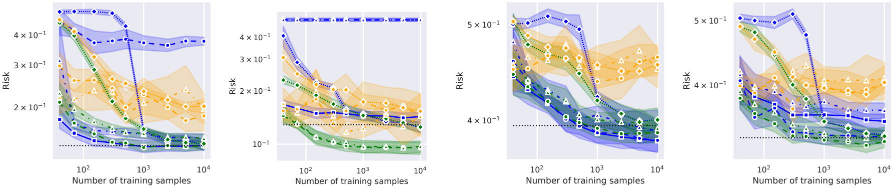
- ✓ P-b-p is Bayes optimal!
- ✓ They are the projected parameters!

Proposition

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

Index

Experiments



(a) LDA+MCAR (b) LDA+MNAR (c) Logistic+MCAR (d) Logistic+MNAR

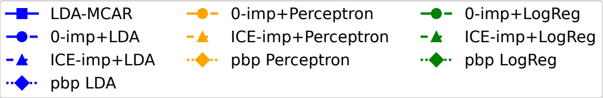


Figure 1: Excess risks of several classifiers on generated data (LDA or Logistic framework) with $\Sigma = I_d$ and MCAR or MNAR missing mechanisms. Dotted lines stand for the Bayes risk $\mathscr{R}_{\mathrm{mis}}(h^{\star}_{\mathrm{mis}})$.

Index

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 (see Reyero Lobo et al. [2025]).

Index

References

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Near-optimal rate of consistency for linear models with missing values. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1211–1243. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ayme22a.html.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL http://www.jstor.org/stable/2984875.

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

References

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

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 Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3165–3174. PMLR, 26–28 Aug 2020. URL `https://proceedings.mlr.press/v108/morvan20a.html`.

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.

References

Angel Reyero Lobo, Alexis Ayme, Claire Boyer, and Erwan Scornet. A primer on linear classification with missing data, 2025. URL https://arxiv.org/abs/2405.09196.

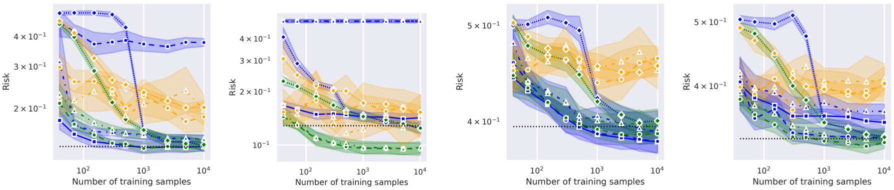
Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335739.

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.

Thank You, Questions?

Index

Experiments



(a) LDA+MCAR (b) LDA+MNAR (c) Logistic+MCAR (d) Logistic+MNAR

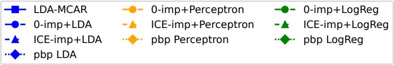


Figure 2: Excess risks of several classifiers on generated data (LDA or Logistic framework) with $\Sigma = I_d$ and MCAR or MNAR missing mechanisms. Dotted lines stand for the Bayes risk $\mathscr{R}_{\mathrm{mis}}(h^\star_{\mathrm{mis}})$.

Experiments

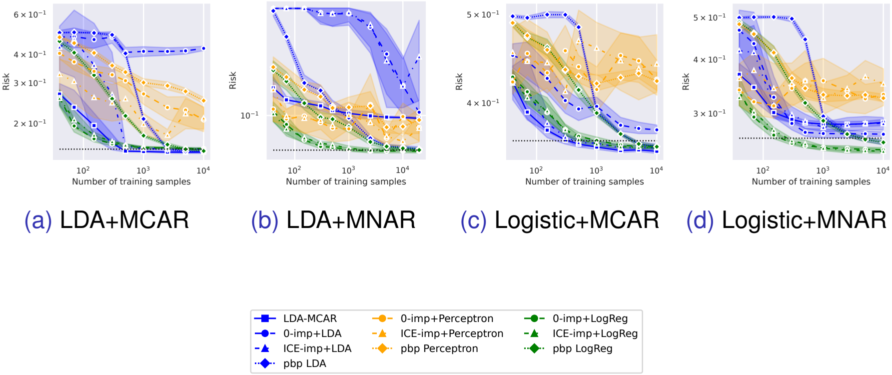


Figure 3: Excess risks of several classifiers on generated data (LDA or Logistic framework) with $\Sigma = \{0.6^{|i-j|}\}_{i,j\in\{1\dots,d\}}$ and MCAR or MNAR missing mechanisms. Dotted lines stand for the Bayes risk $\mathscr{R}_{\mathrm{mis}}(h^\star_{\mathrm{mis}})$.