Linear predictor on linearly-generated data with missing values: non consistency and solutions

Marine Le Morvan

INRIA (Parietal), CNRS (IJCLab)

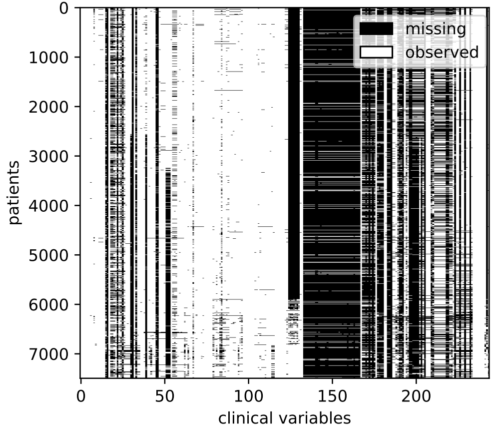
N. Prost

J. Josse

E. Scornet

G. Varoquaux

Missing values are ubiquitous in various fields



Traumabase clinical health records.

Most off-the-shelf supervised learning methods cannot be applied with missing values.

What to do:

- Complete-case analysis?
- Imputation prior to learning?
- Expectation Maximization?

We will study the case of linear regression with missing values, which has surprisingly received little attention up to now.

Content

Outline

Notation

- $\mathbf{x}_n \in \mathbb{R}^{n \times d}$: complete data (unavailable).
- $\mathbf{z}_n \in \{R \times \mathtt{na}\}^{\mathtt{n \times d}}$: incomplete data (available).
- $\mathbf{m}_n \in \{0, 1\}^{n \times d}$: mask. 0s (1s) indicate the observed (missing) values.
- $\mathbf{y}_n \in \mathbb{R}^n$: the response vector.

$$\mathbf{z}_n = \begin{pmatrix} 9.1 & 8.5 \\ 2.1 & \mathtt{na} \\ \mathtt{na} & 9.6 \\ \mathtt{na} & \mathtt{na} \end{pmatrix}, \ \mathbf{x}_n = \begin{pmatrix} 9.1 & 8.5 \\ 2.1 & 3.5 \\ 6.7 & 9.6 \\ 4.2 & 5.5 \end{pmatrix}, \ \mathbf{m}_n = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}, \ \mathbf{y}_n = \begin{pmatrix} 4.6 \\ 7.9 \\ 8.3 \\ 4.6 \end{pmatrix}$$

Each row of $\mathbf{x}_n, \mathbf{z}_n, \mathbf{m}_n, \mathbf{y}_n$ are realization of the generic random variable X, Z, M, Y.

The incomplete vector is related to X and M by:

$$Z = X \odot (1 - M) + \mathtt{na} \odot M.$$

Problem setting

- **Working hypothesis:**

  In this work, we assume that the response is linearly generated:

Assumption (Linear model)

$$Y = \beta_0 + \langle X, \beta \rangle + \varepsilon, \quad X \in \mathbb{R}^d, \ \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

- **Problem formulation:**

  We wish to solve a least squares regression problem with missing values:

$$\min_{f:\ \{\mathbb{R} \times \mathtt{na}\}^{\mathtt{d}} \to \mathbb{R}} \mathbb{E}\left[(Y - f(Z))^2 \right],$$

Outline

Characterizing optimal regressors: the Bayes predictor

- A **Bayes predictor** f^* is the a minimizer of the loss (in our case least squares),

$$f^\star \in \underset{f:\ \{\mathbb{R}\times\texttt{na}\}^{\texttt{d}}\to\mathbb{R}}{\operatorname{argmin}} \mathbb{E}\left[\left(Y - f(Z)\right)^2\right].$$

- For the least squares loss, we know it is the **conditional expectation of the response given the input**:

  ✓ In the complete case: $f^\star = \mathbb{E}\left[Y|X\right] = \langle \beta, X\rangle + \beta_0$.

  ✓ In the incomplete case: $f^\star = \mathbb{E}\left[Y|Z\right] = \mathbb{E}\left[Y|M, X_{obs(M)}\right]$

- In the incomplete case, the Bayes predictor need not be linear.

Example

Let $Y = X_1 + X_2 + \varepsilon$, where $X_2 = \exp(X_1) + \varepsilon_1$. Now, assume that only X_1 is observed. Then the Bayes predictor is:

$$f(X_1) = X_1 + \exp(X_1).$$

The Bayes predictor for incomplete data

Assumption (Gaussian pattern mixture model)

$$X \mid (M = m) \sim \mathcal{N}(\mu^m, \Sigma^m).$$

Proposition (Expanded Bayes predictor)

Under our assumptions (linear model + Gaussian pattern mixture model), the Bayes predictor takes the form

$$f^\star(Z) = \langle W, \delta \rangle,$$

where the parameter $\delta \in \mathbb{R}^p$ is a function of β, $(\mu^m)_{m \in \{0,1\}^d}$ and $(\Sigma^m)_{m \in \{0,1\}^d}$, and the random variable $W \in \mathbb{R}^p$ is the concatenation of $j = 1, \ldots, 2^d$ blocks, each one being

$$\left(\mathbb{1}_{M=m_j} \, , \; X_{obs(m_j)} \mathbb{1}_{M=m_j} \right).$$

where W is an expansion of Z.

The Bayes predictor for incomplete data

Assumption (Gaussian pattern mixture model)

$$X \mid (M = m) \sim \mathcal{N}(\mu^m, \Sigma^m).$$

Proposition (Expanded Bayes predictor)

Under our assumptions (linear model + Gaussian pattern mixture model), the Bayes predictor takes the form

$$f^\star(Z) = \langle W, \delta \rangle,$$

where (ex. d=2)

$$W = \left(\begin{array}{ccc|cc|cc|c}
1 & x_{1,1} & x_{1,2} & 0 & 0 & 0 & 0 & 0 \\
1 & x_{2,1} & x_{2,2} & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 1 & x_{3,1} & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & x_{4,1} & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 1 & x_{5,2} & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & x_{6,2} & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right)$$

Outline of the proof

Under the linear assumption we have:

$$\begin{aligned} f^\star(Z) &= \mathbb{E}[Y|Z] \\ &= \mathbb{E}[\beta_0 + \beta^\mathsf{T} X \mid Z] \\ &= \mathbb{E}[\beta_0 + \beta^\mathsf{T} X \mid M, X_{obs(M)}] \\ &= \beta_0 + \beta_{obs(M)}^\mathsf{T} X_{obs(M)} + \beta_{mis(M)}^\mathsf{T} \, \mathbb{E}[X_{mis(M)} \mid M, X_{obs(M)}] \end{aligned}$$

Moreover under the Gaussian per pattern assumption,

$$\mathbb{E}[X_{mis(M)} \mid M, X_{obs(M)}] = \theta + \Gamma^\top X_{obs(M)}$$

where θ and Γ depend on μ^M and Σ^M.

Thus,

$$f^\star(Z) = \beta_0 + \beta_{mis(M)}^\mathsf{T} \theta + \left(\beta_{obs(M)} + \Gamma\right)^\mathsf{T} X_{obs(M)}$$

i.e., the Bayes predictor is linear per pattern.

The expanded linear model

$f^*(Z) = \langle W, \delta \rangle$ where (example $d = 2$):

$$W = \left(\begin{array}{ccc|cc|cc|c}
1 & x_{1,1} & x_{1,2} & 0 & 0 & 0 & 0 & 0 \\
1 & x_{2,1} & x_{2,2} & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 1 & x_{3,1} & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & x_{4,1} & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 1 & x_{5,2} & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & x_{6,2} & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right)$$

Problem: the dimension of W is

$$p = \sum_{k=0}^{d} \binom{d}{k} \times (k+1) = 2^{d-1} \times (d+2).$$

Outline

The linear approximation model

The Bayes predictor can be expressed as a polynome of X and M, which can be truncated to a first order approximation.

Definition (Linear approximation)

We define the linear approximation of $f^\star$ as

$$f^\star_{\mathrm{approx}}(Z) = \beta^\star_{0,0} + \sum_{j=1}^{d} \beta^\star_{j,0} M_j + \sum_{j=1}^{d} \beta^\star_j X_j (1 - M_j).$$

Estimation of the linear approximation model

- $f^{\star}_{\mathrm{approx}}$ can be estimated by fitting a linear model on X imputed by 0 concatenated with the mask.
- This is equivalent to jointly fitting a linear model on X and optimizing an imputation constant for each variable.

$$\text{Given } \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{pmatrix} 1.1 & 3.2 \\ \texttt{NA} & 0.1 \\ 4.6 & \texttt{NA} \\ 4.0 & 0.9 \\ \texttt{NA} & 2.2 \end{pmatrix} \end{array}, \quad \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{pmatrix} 1.1 & 3.2 \\ C_1 & 0.1 \\ 4.6 & C_2 \\ 4.0 & 0.9 \\ C_1 & 2.2 \end{pmatrix} \end{array} \Leftrightarrow \begin{array}{c} \begin{array}{cccc} X_1 & M_1 & X_2 & M_2 \end{array} \\ \begin{pmatrix} 1.1 & 0 & 3.2 & 0 \\ 0 & 1 & 0.1 & 0 \\ 4.6 & 0 & 0 & 1 \\ 4.0 & 0 & 0.9 & 0 \\ 0 & 1 & 2.2 & 0 \end{pmatrix} \end{array}.$$

Indeed,

$$\beta_j \left\{X_j(1-M_j) + c_j M_j\right\} = \beta_j X_j(1-M_j) + \left\{\beta_j c_j\right\} M_j.$$

Finite sample bounds for linear predictors

The Bayes predictor and its linear approximation offer different bias-variance tradeoffs.

Assumption

- $Y = f_{\text{Bayes}}(Z) + \text{noise}(Z)$ *where* $\text{noise}(Z)$ *is a centred noise conditional on* Z *and such that there exists* $\sigma^2 > 0$ *satisfying* $\mathbb{V}[Y|Z] \leq \sigma^2$ *almost surely,*
- $\|f_{\text{Bayes}}\|_\infty < L$*,*
- $\text{Supp}(X) \subset [-1, 1]^d$*.*

This assumption is required for the next two results.

Finite sample bounds for linear predictors

Under these assumptions:

Theorem

- *The risk of the OLS estimate clipped at L for the* ***expanded model*** *satisfies*

$$\frac{2^d c_1}{n+1} \leq R(T_L f_{\hat{\beta}_{expanded}}) - \sigma^2 \leq c \max\{\sigma^2, L^2\} \frac{2^{d-1}(d+2)(1+\log n)}{n}$$

- *The risk of the OLS estimate clipped at L for the* ***linear approximation model*** *satisfies*

$$R(T_L f_{\hat{\beta}_{approx}}) - \sigma^2 \leq c \max\{\sigma^2, L^2\} \frac{2d(1+\log n)}{n} + 64(d+1)^2 L^2$$

It follows that the risk of the expanded model is lower than that of the linear approximation model if:

$$n \geq \frac{2^d}{d}$$

Outline

Why a Multilayer perceptron?

A Multilayer Perceptron with:

- Rectified Linear Units activation functions for hidden units ($ReLU(x) = \max(0, x)$),
- Identity activation for the output unit,

produces a prediction function that is piecewise affine.

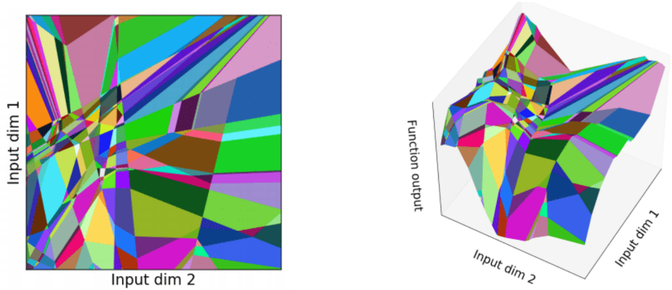


Figure from Hanin et al. 2019

Bayes consistency of the MLP

Theorem (MLP)

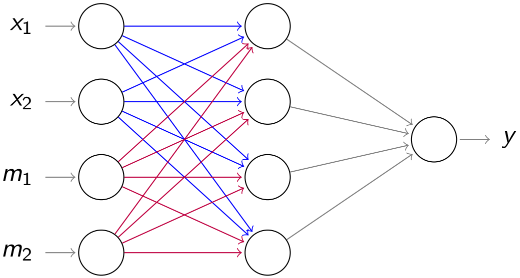
Assume that the Bayes predictor takes the form described earlier (expanded Bayes Predictor). A MLP:

- *with one hidden layer containing 2^d hidden units*
- *ReLU activation functions*
- *which is fed with the concatenated vector (X, M) where X is imputed by zero*

is Bayes consistent.

Proof: We show that there exists a configuration of the parameters of the MLP so that the resulting predictor is the Bayes predictor.

Proof 1/3 - Learned imputations



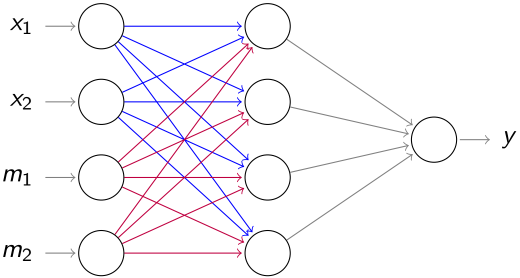
Parameters hidden layer:

$W^{(1)} = \left[W^{(X)}, W^{(M)}\right] \in \mathbb{R}^{4\times 4}$
$b^{(1)} \in \mathbb{R}^4$

Parameters output layer:

$W^{(2)} \in \mathbb{R}^4$
$b^{(2)} \in \mathbb{R}$

Proof 1/3 - Learned imputations



Parameters hidden layer:

$W^{(1)} = \left[W^{(X)}, W^{(M)}\right] \in \mathbb{R}^{4\times 4}$
$b^{(1)} \in \mathbb{R}^4$

Parameters output layer:

$W^{(2)} \in \mathbb{R}^4$
$b^{(2)} \in \mathbb{R}$

The activation of hidden unit k for input (x, m) is:

$$\begin{aligned} a_k &= W^{(X)}_{k,.}x + W^{(M)}_{k,.}m + b^{(1)}_k \\ &= W^{(X)}_{k,.}x + W^{(X)}_{k,.} \odot G_{k,.}m + b^{(1)}_k \\ &= W^{(X)}_{k,obs(m)}x_{obs(m)} + W^{(X)}_{k,mis(m)}G_{k,mis(m)} + b^{(1)}_k \end{aligned}$$

where G (reparametrization of $W^{(M)}$) can be seen as learned imputations.

Proof 2/3 - one-to-one mapping mdp/hidden unit

The proof shows that the parameters of the MLP can be chosen so that:

1. all points with missing data pattern m_k exclusively activate hidden unit k, and hidden unit k is exclusively activated by points with missing data pattern m_k.

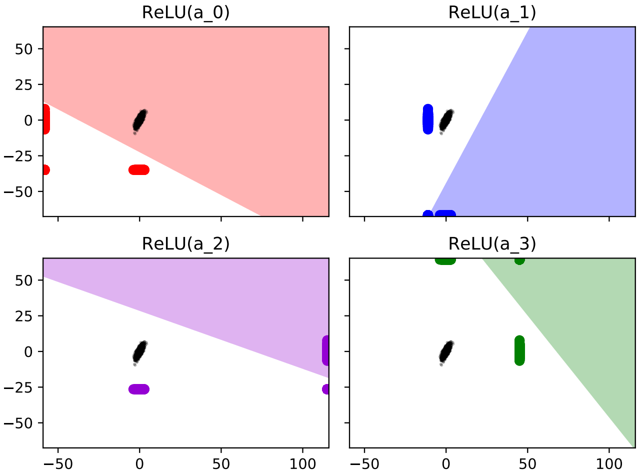
$$
\begin{aligned}
y(x, m_k) &= \sum_{h=1}^{2^d} W_h^{(2)} ReLU(a_h) + b^{(2)} \\
&= \sum_{h=1}^{2^d} W_h^{(2)} ReLU(W_{h,obs(m_k)}^{(X)} x_{obs(m_k)} + W_{h,mis(m_k)}^{(X)} G_{h,mis(m_k)} + b_h^{(1)}) + b^{(2)} \\
&= W_k^{(2)} \left(W_{k,obs(m_k)}^{(X)} x_{obs(m_k)} + W_{k,mis(m_k)}^{(X)} G_{k,mis(m_k)} + b_k^{(1)} \right) + b^{(2)}
\end{aligned}
$$

i.e, the MLP produces a predictor $y(x, m_k)$ that is linear per pattern.

2. The slopes and biases of $y(x, m_k)$ equal those of the Bayes predictor.

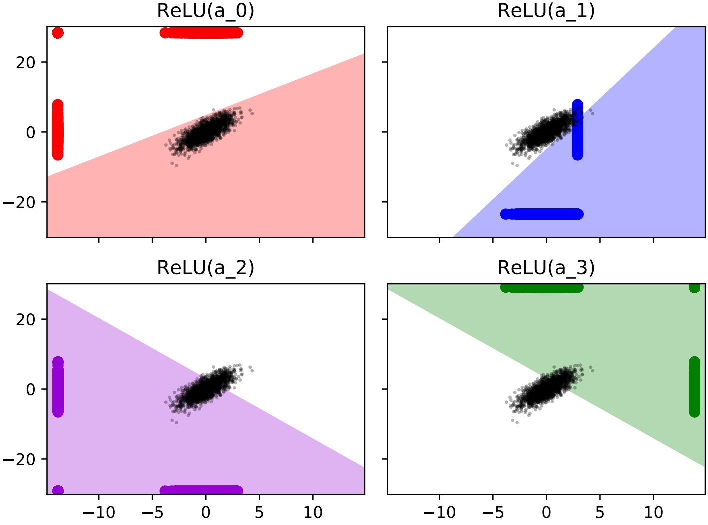
Proof 3/3 - vizualisation of a bayes consistent MLP

We simulated data (X, M) in 2 dimensions, and based on our proof, built a MLP (with 4 hidden units) that is Bayes consistent.



$$y(x, m) = W^{(2)}_{1,.} {\color{red}ReLU(a_0)} + W^{(2)}_{1,.} {\color{blue}ReLU(a_1)} + W^{(2)}_{2,.} {\color{purple}ReLU(a_2)} + W^{(2)}_{3,.} {\color{green}ReLU(a_3)} + b^{(2)}$$

Example of an optimized MLP in two dimensions.



$$y(x,m) = W^{(2)}_{1,.} \mathit{ReLU}(a_0) + W^{(2)}_{1,.} \mathit{ReLU}(a_1) + W^{(2)}_{2,.} \mathit{ReLU}(a_2) + W^{(2)}_{3,.} \mathit{ReLU}(a_3) + b^{(2)}$$

Trading off estimation and approximation error

Number of parameters of:

- a MLP with one hidden layer and 2^d units:

$$(d+1)2^{d+1}+1$$

- the expanded linear model:

$$(d+1)2^{d-1}$$

The MLP is slightly overparametrized, and the number of parameters is exponential in d.

However, contrary the the expanded linear model, the MLP provides a natural way to reduce the model capacity by reducing the number of hidden units.

Outline

Simulation models

The data (X, M) is generated according to 3 simulation models:

- **mixture 1**:
  - $P(X) = \mathcal{N}(\mu, \Sigma)$
  - $P(M) = \frac{1}{2^d}$
  - Gaussian pattern mixture model with 1 component
  - Corresponds to a Missing Completely At Random (MCAR) problem

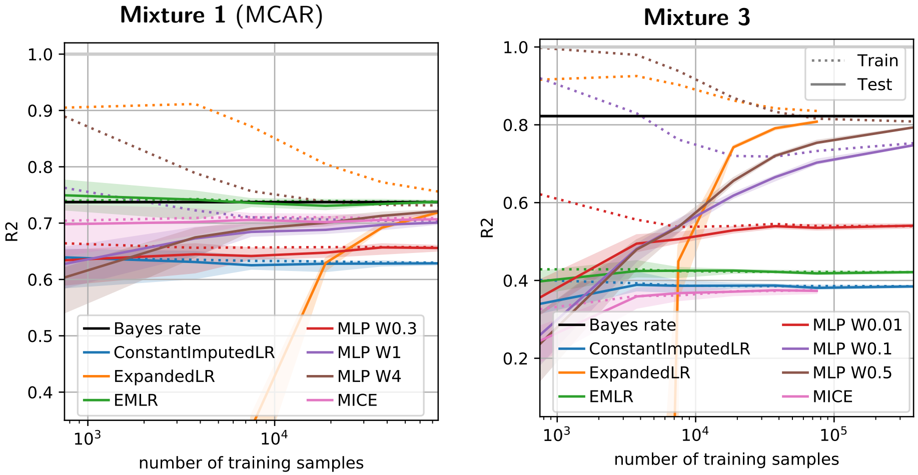
- **mixture 3**:
  - $P(X|M = m) = \mathcal{N}(\mu_m, \Sigma_m)$, with 3 distinct Gaussian components.
  - $P(M) = \frac{1}{2^d}$
  - Gaussian pattern mixture model (with 3 components)

- **selfmasking**:
  - $P(X) = \mathcal{N}(\mu, \Sigma)$
  - $P(M = 1|X_j) = \text{Probit}\left(\lambda_j \left(X_j - \mu_0\right)\right)$
  - Not an instance of pattern mixture model! (Theory does not hold)
  - Corresponds to a typical Missing Non At Random (MNAR) problem

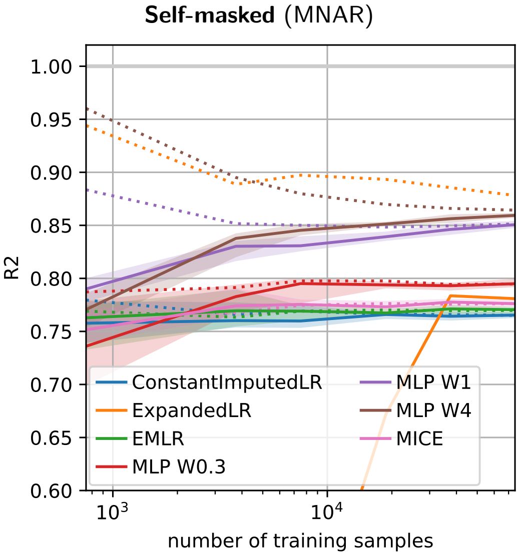
Estimation Approaches

- **EMLR**: EM is used to fit a multivariate normal distribution for the $(p+1)$-dimensional random variable $(X_1, ..., X_p, Y)$.
- **ConstantImputedLR**: Optimal imputation method.
- **MICE**: Conditional imputation with an iterative imputer (similar to the well known MICE) followed by linear regression.
- **ExpandedLR**: Expanded linear model.
- **MLP**: Multilayer perceptron with one hidden layer whose size is varied between and 1 and 2^d hidden units.

Learning curves: Gaussian mixtures



Learning curves: self-masking



Conclusion

Conclusion:

- The Bayes-optimal predictor is no longer a linear function of the data.
- It is explicit under Gaussian assumptions, but high-dimensional.
- Possible approximations include constant imputation and MLP, which can be consistent.
- The MLP adapts naturally to the complexity of the data.
- Our risk-minimisation strategy is robust to the missing-value mechanism.