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

(1) Problem setting
(2) The Bayes predictor
(3) Linear approximation
(4) Multilayer perceptron approximation
(5) Empirical study

## Outline

(1) Problem setting
(2) The Bayes predictor
(3) Linear approximation

4 Multilayer perceptron approximation
(5) Empirical study

## Notation

- $\mathbf{x}_{n} \in \mathbb{R}^{n \times d}$ : complete data (unavailable).
- $\mathbf{z}_{n} \in\{R \times \mathrm{na}\}^{\mathrm{n} \times \mathrm{d}}$ : 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.

$$
\mathbf{z}_{n}=\left(\begin{array}{cc}
9.1 & 8.5 \\
2.1 & \text { na } \\
\text { na } & 9.6 \\
\text { na } & \text { na }
\end{array}\right), \mathbf{x}_{n}=\left(\begin{array}{ll}
9.1 & 8.5 \\
2.1 & 3.5 \\
6.7 & 9.6 \\
4.2 & 5.5
\end{array}\right), \mathbf{m}_{n}=\left(\begin{array}{ll}
0 & 0 \\
0 & 1 \\
1 & 0 \\
1 & 1
\end{array}\right), \mathbf{y}_{n}=\left(\begin{array}{l}
4.6 \\
7.9 \\
8.3 \\
4.6
\end{array}\right)
$$

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:

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

## Problem setting

- Working hypothesis:

In this work, we assume that the response is linearly generated:
Assumption (Linear model)

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

- Problem formulation:

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

$$
\min _{f:\{\mathbb{R} \times \text { na }\}^{d} \rightarrow \mathbb{R}} \mathbb{E}\left[(Y-f(Z))^{2}\right],
$$

## Outline

## (1) Problem setting

(2) The Bayes predictor
(3) Linear approximation

4 Multilayer perceptron approximation
(5) Empirical study

## Characterizing optimal regressors: the Bayes predictor

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

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

- For the least squares loss, we know it is the conditional expectation of the response given the input:
$\checkmark$ In the complete case: $f^{\star}=\mathbb{E}[Y \mid X]=\langle\beta, X\rangle+\beta_{0}$.
$\checkmark$ In the incomplete case: $f^{\star}=\mathbb{E}[Y \mid Z]=\mathbb{E}\left[Y \mid M, X_{o b s(M)}\right]$
- In the incomplete case, the Bayes predictor need not be linear.


## Example

Let $Y=X_{1}+X_{2}+\varepsilon$, where $X_{2}=\exp \left(X_{1}\right)+\varepsilon_{1}$. Now, assume that only $X_{1}$ is observed. Then the Bayes predictor is:

$$
f\left(X_{1}\right)=X_{1}+\exp \left(X_{1}\right) .
$$

## The Bayes predictor for incomplete data

Assumption (Gaussian pattern mixture model)

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

## 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 $\beta,\left(\mu^{m}\right)_{m \in\{0,1\}^{d}}$ and $\left(\Sigma^{m}\right)_{m \in\{0,1\}^{d}}$, and the random variable $W \in \mathbb{R}^{p}$ is the concatenation of $j=1, \ldots, 2^{d}$ blocks, each one being

$$
\left(\mathbb{1}_{M=m_{j}}, X_{o b s\left(m_{j}\right)} \mathbb{1}_{M=m_{j}}\right) .
$$

where $W$ is an expansion of $Z$.

## The Bayes predictor for incomplete data

Assumption (Gaussian pattern mixture model)

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

## 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. $d=2$ )

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

## Outline of the proof

Under the linear assumption we have:

$$
\begin{aligned}
f^{\star}(Z) & =\mathbb{E}[Y \mid Z] \\
& =\mathbb{E}\left[\beta_{0}+\beta^{\top} X \mid Z\right] \\
& =\mathbb{E}\left[\beta_{0}+\beta^{\top} X \mid M, X_{o b s(M)}\right] \\
& =\beta_{0}+\beta_{o b s(M)}^{\top} X_{o b s(M)}^{\top}+\beta_{m i s(M)}^{\top} \mathbb{E}\left[X_{m i s(M)} \mid M, X_{o b s(M)}\right]
\end{aligned}
$$

Moreover under the Gaussian per pattern assumption,

$$
\mathbb{E}\left[X_{m i s(M)} \mid M, X_{o b s(M)}\right]=\theta+\Gamma^{\top} X_{o b s(M)}
$$

where $\theta$ and $\Gamma$ depend on $\mu^{M}$ and $\Sigma^{M}$.
Thus,

$$
f^{\star}(Z)=\beta_{0}+\beta_{m i s(M)}^{\top} \theta+\left(\beta_{o b s(M)}+\Gamma\right)^{\top} X_{o b s(M)}
$$

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

## The expanded linear model

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

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

Problem: the dimension of $W$ is

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

## Outline

## (1) Problem setting

(2) The Bayes predictor
(3) Linear approximation

## 4 Multilayer perceptron approximation

(5) Empirical study

## 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

$$
f_{\text {approx }}^{\star}(Z)=\beta_{0,0}^{\star}+\sum_{j=1}^{d} \beta_{j, 0}^{\star} M_{j}+\sum_{j=1}^{d} \beta_{j}^{\star} X_{j}\left(1-M_{j}\right) .
$$

## Estimation of the linear approximation model

- $f_{\text {approx }}^{\star}$ 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.

$$
\text { Given }\left(\begin{array}{cc}
X_{1} & X_{2} \\
1.1 & 3.2 \\
\text { NA } & 0.1 \\
4.6 & \text { NA } \\
4.0 & 0.9 \\
\text { NA } & 2.2
\end{array}\right), \quad\left(\begin{array}{cc}
X_{1} & X_{2} \\
1.1 & 3.2 \\
C_{1} & 0.1 \\
4.6 & C_{2} \\
4.0 & 0.9 \\
C_{1} & 2.2
\end{array}\right) \Leftrightarrow\left(\begin{array}{cccc}
X_{1} & M_{1} & X_{2} & M_{2} \\
1.1 & 0 & 3.2 & 0 \\
0 & 1 & 0.1 & 0 \\
4.6 & 0 & 0 & 1 \\
4.0 & 0 & 0.9 & 0 \\
0 & 1 & 2.2 & 0
\end{array}\right) .
$$

Indeed,

$$
\beta_{j}\left\{X_{j}\left(1-M_{j}\right)+c_{j} M_{j}\right\}=\beta_{j} X_{j}\left(1-M_{j}\right)+\left\{\beta_{j} c_{j}\right\} M_{j} .
$$

## Finite sample bounds for linear predictors

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

## Assumption

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

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

$$
\frac{2^{d} c_{1}}{n+1} \leq R\left(T_{L} f_{\hat{\beta}_{\text {expanded }}}\right)-\sigma^{2} \leq c \max \left\{\sigma^{2}, L^{2}\right\} \frac{2^{d-1}(d+2)(1+\log n)}{n}
$$

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

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

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

## (1) Problem setting

(2) The Bayes predictor
(3) Linear approximation

4 Multilayer perceptron approximation
(5) Empirical study

## Why a Multilayer perceptron?

A Multilayer Perceptron with:

- Rectified Linear Units activation functions for hidden units $(\operatorname{ReL} U(x)=\max (0, x))$,
- 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:

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

Parameters output layer:

$$
\begin{aligned}
& W^{(2)} \in \mathbb{R}^{4} \\
& b^{(2)} \in \mathbb{R}
\end{aligned}
$$

## Proof 1/3 - Learned imputations



Parameters hidden layer:

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

Parameters output layer:

$$
\begin{aligned}
& W^{(2)} \in \mathbb{R}^{4} \\
& b^{(2)} \in \mathbb{R}
\end{aligned}
$$

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

$$
\begin{aligned}
a_{k} & =W_{k, .}^{(X)} x+W_{k, .}^{(M)} m+b_{k}^{(1)} \\
& =W_{k, .}^{(X)} x+W_{k, .}^{(X)} \odot G_{k, .} m+b_{k}^{(1)} \\
& =W_{k, o b s(m)}^{(X)}{ }^{x_{o b s(m)}}+W_{k, \operatorname{mis}(m)}^{(X)} G_{k, m i s(m)}+b_{k}^{(1)}
\end{aligned}
$$

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 $m_{k}$.

$$
\begin{aligned}
y\left(x, m_{k}\right) & =\sum_{h=1}^{2^{d}} W_{h}^{(2)} \operatorname{ReLU}\left(a_{h}\right)+b^{(2)} \\
& =\sum_{h=1}^{2^{d}} W_{h}^{(2)} \operatorname{ReLU}\left(W_{h, o b s\left(m_{k}\right)^{(X)}}^{\left.x_{o b s\left(m_{k}\right)}+W_{h, \operatorname{mis}\left(m_{k}\right)}^{(X)} G_{h, m i s\left(m_{k}\right)}+b_{h}^{(1)}\right)+b^{(2)}}\right. \\
& =W_{k}^{(2)}\left(W_{k, o b s\left(m_{k}\right)}^{(X)} x_{o b s\left(m_{k}\right)}+W_{k, \operatorname{mis}\left(m_{k}\right)}^{(X)} G_{k, \operatorname{mis}\left(m_{k}\right)}+b_{k}^{(1)}\right)+b^{(2)}
\end{aligned}
$$

i.e, the MLP produces a predictor $y\left(x, m_{k}\right)$ that is linear per pattern.
(2) The slopes and biases of $y\left(x, m_{k}\right)$ 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.

$y(x, m)=W_{1, .}^{(2)} \operatorname{ReLU}\left(a_{0}\right)+W_{1, .}^{(2)} \operatorname{ReLU}\left(a_{1}\right)+W_{2, .}^{(2)} \operatorname{ReLU}\left(a_{2}\right)+W_{3, .}^{(2)} \operatorname{ReLU}\left(a_{3}\right)+b^{(2)}$

## Example of an optimized MLP in two dimensions.





$y(x, m)=W_{1, .}^{(2)} \operatorname{ReLU}\left(a_{0}\right)+W_{1, .}^{(2)} \operatorname{ReL} U\left(a_{1}\right)+W_{2, .}^{(2)} \operatorname{ReLU}\left(a_{2}\right)+W_{3, .}^{(2)} \operatorname{ReLU}\left(a_{3}\right)+b^{(2)}$

## 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

## (1) Problem setting

(2) The Bayes predictor
(3) Linear approximation

4 Multilayer perceptron approximation
(5) Empirical study

## Simulation models

The data $(\mathrm{X}, \mathrm{M})$ 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:
- $P(X \mid M=m)=\mathcal{N}\left(\mu_{m}, \Sigma_{m}\right)$, with 3 distinct Gaussian components.
- $P(M)=\frac{1}{2^{d}}$
- Gaussian pattern mixture model (with 3 components)
- selfmasking:
- $P(X)=\mathcal{N}(\mu, \Sigma)$
- $P\left(M=1 \mid X_{j}\right)=\operatorname{Probit}\left(\lambda_{j}\left(X_{j}-\mu_{0}\right)\right)$
- 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 $\left(X_{1}, \ldots, X_{p}, Y\right)$.
- 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

Mixture 1 (MCAR)


Mixture 3


## Learning curves: self-masking

Self-masked (MNAR)


## 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.

