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

A General remarks and proof of Proposition 3.1

A.1 Notation: letter cases

One letter refers to one quantity, with different cases: U is a random variable, while u is a constant. $\mathbf{U}_n$ is a (random) sample, and $\mathbf{u}_n$ is a realisation of that sample. u_j is the j-th coordinate of u, and if $\mathcal{J}$ is a set, $u_{\mathcal{J}}$ denotes the subvector with indices in $\mathcal{J}$.

A.2 Gaussian vectors

In assumption 4.1, conditionnally to M, X is Gaussian. It is useful to remind that in that case, for two subsets of indices $\mathcal{I}$ and $\mathcal{J}$, conditional distributions can be written as

$$X_{\mathcal{I}}|(X_{\mathcal{J}}, M) \sim \mathcal{N}(\mu^M_{\mathcal{I}|\mathcal{J}}, \Sigma^M_{\mathcal{I}|\mathcal{J}}) \tag{6}$$

with

$$\begin{cases} \mu^M_{\mathcal{I}|\mathcal{J}} & = \mu^M_{\mathcal{I}} + \Sigma^M_{\mathcal{I}\mathcal{J}}(\Sigma^M_{\mathcal{J}\mathcal{J}})^{-1}(X_{\mathcal{J}} - \mu^M_{\mathcal{J}}) \\ \Sigma^M_{\mathcal{I}|\mathcal{J}} & = \Sigma^M_{\mathcal{I}\mathcal{I}} - \Sigma^M_{\mathcal{I}\mathcal{J}}(\Sigma^M_{\mathcal{J}\mathcal{J}})^{-1}(\Sigma^M_{\mathcal{I}\mathcal{J}})^{\mathsf{T}}. \end{cases}$$

In particular, for all pattern m, for all $k \in mis(m)$,

$$\mathbb{E}\left[X_k \mid M = m, X_{obs(m)}\right] = \mu^m_k + \Sigma^m_{k,obs(m)}\left(\Sigma^m_{obs(m)}\right)^{-1}\left(X_{obs(m)} - \mu^m_{obs(m)}\right).$$

A.3 Proof of Proposition 3.1

Solving a linear regression problem with optimal imputation constants $c^\star = (c^\star_j)_{j\in[\![1,d]\!]}$ can be written as

$$\begin{aligned} &(\beta^\star, c^\star) \in \operatorname{argmin}_{\beta,c\in\mathbb{R}^d}\mathbb{E}\left[\left(Y - \left(\beta_0 + \sum_{j=1}^d \beta_j\left(X_j \mathbb{1}_{M_j=0} + c_j \mathbb{1}_{M_j=1}\right)\right)\right)^2\right] \\ \iff &(\beta^\star, c^\star) \in \operatorname{argmin}_{\beta,c\in\mathbb{R}^d}\mathbb{E}\left[\left(Y - \left(\beta_0 + \sum_{j=1}^d \beta_j X_j \mathbb{1}_{M_j=0} + \sum_{j=1}^d \beta_j c_j \mathbb{1}_{M_j=1}\right)\right)^2\right], \end{aligned}$$

where the terms $X_j\mathbb{1}_{M_j=0}$ is equal to the variable X_j, imputed by zero if X_j is missing and $\beta_j c_j$ is the linear coefficient associated to the variable $\mathbb{1}_{M_j=1}$. Therefore, the linear regression coefficient $\beta^\star = (\beta^\star_j)_{j\in[\![1,d]\!]}$ and the optimal imputation constants $c^\star = (c^\star_j)_{j\in[\![1,d]\!]}$ can be solved via the linear regression problem with inputs $(X_j)_{j\in[\![1,d]\!]}, (\mathbb{1}_{M_j=1})_{j\in[\![1,d]\!]}$ where the first set of d coefficients are the $(\beta^\star_j)_{j\in[\![1,d]\!]}$ and the second set of coefficients are equal to $(\beta^\star_j c^\star_j)_{j\in[\![1,d]\!]}$.

B Bayes estimate and Bayes risk

Proof of Proposition 4.1.

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

where, by Equation 6,

$$\mathbb{E}[X_{mis(M)} \mid M, X_{obs(M)}] = \mu^M_{mis(M)} + \Sigma^M_{mis(M),obs(M)} \left(\Sigma^M_{obs(M)}\right)^{-1} \left(X_{obs(M)} - \mu^M_{obs(M)}\right).$$

Hence,

$$\begin{aligned}\mathbb{E}[Y|Z] &= \beta_0 + \beta_{mis(M)}^\mathsf{T} \left(\mu^M_{mis(M)} - \Sigma^M_{mis(M),obs(M)} \left(\Sigma^M_{obs(M)}\right)^{-1} \mu^M_{obs(M)}\right) \\ &\quad + \left(\beta_{obs(M)}^\mathsf{T} + \beta_{mis(M)}^\mathsf{T} \, \Sigma^M_{mis(M),obs(M)} \left(\Sigma^M_{obs(M)}\right)^{-1}\right) X_{obs(M)} \\ &= \delta^M_{obs(M),0} + \left(\delta^M_{obs(M)}\right)^\mathsf{T} X_{obs(M)},\end{aligned}$$

by setting

$$\begin{aligned}\delta^M_{obs(M),0} &= \beta_0 + \beta_{mis(M)}^\mathsf{T} \left(\mu^M_{mis(M)} - \Sigma^M_{mis(M),obs(M)} \left(\Sigma^M_{obs(M)}\right)^{-1} \mu^M_{obs(M)}\right) \\ \delta^M_{obs(M)} &= \beta_{obs(M)} + \beta_{mis(M)}^\mathsf{T} \, \Sigma^M_{mis(M),obs(M)} \left(\Sigma^M_{obs(M)}\right)^{-1}.\end{aligned}$$

Therefore, $E[Y|Z]$ takes the form,

$$\begin{aligned}\mathbb{E}[Y|Z] &= \sum_{m \in \{0,1\}^d} \left[\delta^m_{obs(m),0} + \left(\delta^m_{obs(m)}\right)^\mathsf{T} X_{obs(m)}\right] \mathbb{1}_{M=m} \\ &= \langle W, \delta \rangle.\end{aligned}$$

□

Proof of Proposition 4.2. The polynomial expression is given by

$$\begin{aligned}
\mathbb{E}[Y|Z] &= \sum_{m\in\{0,1\}^d} \mathbb{1}_{M=m} \times \left(\delta_0^m + \sum_{j=1}^d \mathbb{1}_{j\in obs(m)}\delta_j^m X_j\right) \\
&= \sum_{m\in\{0,1\}^d} \prod_{k=1}^d \left(1-(M_k-m_k)^2\right) \times \left(\delta_0^m + \sum_{j=1}^d (1-M_j)\delta_j^m X_j\right) \\
&= \sum_{m\in\{0,1\}^d} \prod_{k=1}^d (1-M_k-m_k+2M_km_k) \times \left(\delta_0^m + \sum_{j=1}^d (1-M_j)\delta_j^m X_j\right) \\
&= \sum_{m\in\{0,1\}^d} \sum_{\substack{\mathcal{S}_1\sqcup\mathcal{S}_2\sqcup\mathcal{S}_3\\ \sqcup\mathcal{S}_4=[\![1,d]\!]}} (-1)^{|\mathcal{S}_2|+|\mathcal{S}_3|}2^{|\mathcal{S}_4|} \prod_{\substack{k_3\in\mathcal{S}_3,\\ k_4\in\mathcal{S}_4}} m_{k_3}m_{k_4} \prod_{\substack{k_2\in\mathcal{S}_2,\\ k_4\in\mathcal{S}_4}} M_{k_2}M_{k_4} \times \left(\delta_0^m + \sum_{j=1}^d (1-M_j)\delta_j^m X_j\right) \\
&\quad \text{(where } \mathcal{S}_1\sqcup\mathcal{S}_2\sqcup\mathcal{S}_3\sqcup\mathcal{S}_4 \text{ is a partition of } [\![1,d]\!]\text{)} \\
&= \sum_{\substack{\mathcal{S}_1\sqcup\mathcal{S}_2\sqcup\mathcal{S}_3\\ \sqcup\mathcal{S}_4=[\![1,d]\!]}} (-1)^{|\mathcal{S}_2|+|\mathcal{S}_3|}2^{|\mathcal{S}_4|} \sum_{\substack{m\in\{0,1\}^d\\ obs(m)\subset\mathcal{S}_3^c\cap\mathcal{S}_4^c}} 1 \times \prod_{k_2\in\mathcal{S}_2,k_4\in\mathcal{S}_4} M_{k_2}M_{k_4} \times \left(\delta_0^m + \sum_{j=1}^d (1-M_j)\delta_j^m X_j\right) \\
&= \sum_{\substack{\mathcal{S}_1\sqcup\mathcal{S}_2\sqcup\mathcal{S}_3\\ \sqcup\mathcal{S}_4=[\![1,d]\!]}} \prod_{k_2\in\mathcal{S}_2,k_4\in\mathcal{S}_4} M_{k_2}M_{k_4} \times \left((-1)^{|\mathcal{S}_2|+|\mathcal{S}_3|}2^{|\mathcal{S}_4|} \sum_{\substack{m\in\{0,1\}^d\\ obs(m)\subset\mathcal{S}_3^c\cap\mathcal{S}_4^c}} \left(\delta_0^m + \sum_{j=1}^d (1-M_j)\delta_j^m X_j\right)\right) \\
&= \sum_{\substack{\mathcal{S}_1\sqcup\mathcal{S}_2\sqcup\mathcal{S}_3\\ \sqcup\mathcal{S}_4=[\![1,d]\!]}} \prod_{k_2\in\mathcal{S}_2,k_4\in\mathcal{S}_4} M_{k_2}M_{k_4} \times \left(\zeta_0^{\mathcal{S}_2,\mathcal{S}_3,\mathcal{S}_4} + \sum_{j=1}^d (1-M_j)\zeta_j^{\mathcal{S}_2,\mathcal{S}_3,\mathcal{S}_4} X_j\right) \\
&= \sum_{\mathcal{S}_2\sqcup\mathcal{S}_4\subset[\![1,d]\!]} \prod_{k_2\in\mathcal{S}_2,k_4\in\mathcal{S}_4} M_{k_2}M_{k_4} \times \sum_{\mathcal{S}_1\sqcup\mathcal{S}_3=(\mathcal{S}_2\sqcup\mathcal{S}_4)^c} \left(\zeta_0^{\mathcal{S}_2,\mathcal{S}_3,\mathcal{S}_4} + \sum_{j=1}^d (1-M_j)\zeta_j^{\mathcal{S}_2,\mathcal{S}_3,\mathcal{S}_4} X_j\right) \\
&\quad \text{(reindexing } \mathcal{S}=\mathcal{S}_2\sqcup\mathcal{S}_4\text{)} \\
&= \sum_{\mathcal{S}\subset[\![1,d]\!]} \prod_{k\in\mathcal{S}} M_k \times \left(\zeta_0^{\mathcal{S}} + \sum_{j=1}^d (1-M_j)\zeta_j^{\mathcal{S}} X_j\right).
\end{aligned}$$

Finally, the expression of noise(Z) results from

$$X_{mis(M)}|X_{obs(M)}, M=m \sim \mathcal{N}(\mu_M, T_M)$$

where the conditional expectation μ_M has been given above and

$$T_M = \Sigma_{mis(M)} - \Sigma_{mis(M),obs(M)} \left(\Sigma_{obs(M)}\right)^{-1} \Sigma_{obs(M),mis(M)}.$$

□

C Bayes Risk

Proposition C.1. *The Bayes risk associated to the Bayes estimator of proposition 4.1 is given by*

$$\mathbb{E}\left[\left(Y - f^\star(Z)\right)^2\right] = \sigma^2 + \sum_{m\in\{0,1\}^d} \mathbb{P}(M=m)\Lambda_m,$$

with

$$\begin{aligned}\Lambda_m = &\left(\gamma^m_{obs(m)}\right)^\mathsf{T} \Sigma^m_{obs(m)} \gamma^m_{obs(m)} + \beta^\mathsf{T}_{mis(m)} \Sigma^m_{mis(m)} \beta_{mis(m)} - 2\left(\gamma^m_{obs(m)}\right)^\mathsf{T} \Sigma^m_{obs(m),mis(m)} \beta_{mis(m)} \\ &+ \left(\gamma^m_{obs(m),0}\right)^2 + \left(\left(\gamma^m_{obs(m)}\right)^\mathsf{T} \mu^m_{obs(m)}\right)^2 + \left(\beta^\mathsf{T}_{mis(m)} \mu^m_{mis(m)}\right)^2 + 2\gamma^m_{obs(m),0}\left(\gamma^m_{obs(m)}\right)^\mathsf{T} \mu^m_{obs(m)} \\ &- 2\gamma^m_{obs(m),0} \beta^\mathsf{T}_{mis(m)} \mu^m_{mis(m)} - 2\left(\gamma^m_{obs(m)}\right)^\mathsf{T} \mu^m_{obs(m)} \beta^\mathsf{T}_{mis(m)} \mu^m_{mis(m)},\end{aligned}$$

where $\gamma^m_{obs(m)}$ is a function of the regression coefficients on the missing variables and the means and covariances given M.

Proof of Proposition C.1.

$$\begin{aligned}&\mathbb{E}\left[\left(\mathbb{E}[Y|Z] - Y\right)^2\right] \\ &= \mathbb{E}\left[\left(\delta^M_{obs(M),0} + \left(\delta^M_{obs(M)}\right)^\mathsf{T} X_{obs(M)} - \beta_0 - \beta^\mathsf{T} X - \varepsilon\right)^2\right] \\ &= \mathbb{E}\left[\left(\delta^M_{obs(M),0} - \beta_0 + \left(\delta^M_{obs(M)} - \beta_{obs(M)}\right)^\mathsf{T} X_{obs(M)} - \beta^\mathsf{T}_{mis(M)} X_{mis(M)} - \varepsilon\right)^2\right].\end{aligned}$$

By posing

$$\begin{cases} \gamma^M_{obs(M),0} &= \delta^M_{obs(M),0} - \beta_0 \quad &= \beta^\mathsf{T}_{mis(M)}\left(\mu^M_{mis(M)} - \Sigma^M_{mis(M),obs(M)}\left(\Sigma^M_{obs(M)}\right)^{-1}\mu^M_{obs(M)}\right) \\ \gamma^M_{obs(M)} &= \delta^M_{obs(M)} - \beta_{obs(M)} \quad &= \beta^\mathsf{T}_{mis(M)}\,\Sigma^M_{mis(M),obs(M)}\left(\Sigma^M_{obs(M)}\right)^{-1}, \end{cases}$$

one has

$$\begin{aligned}&\mathbb{E}\left[\left(\mathbb{E}[Y|Z] - Y\right)^2\right] \\ &= \mathbb{E}\left[\left(\gamma^M_{obs(M),0} + \left(\gamma^M_{obs(M)}\right)^\mathsf{T} X_{obs(M)} - \beta^\mathsf{T}_{mis(M)} X_{mis(M)} - \varepsilon\right)^2\right] \\ &= \sum_{m\in:\{0,1\}^d} \mathbb{P}(M=m)\cdot\mathbb{E}\left[\left(\gamma^m_{obs(m),0} + \left(\gamma^m_{obs(m)}\right)^\mathsf{T} X_{obs(m)} - \beta^\mathsf{T}_{mis(m)} X_{mis(m)} - \varepsilon\right)^2 \,\middle|\, M=m\right] \\ &= \sum_{m\in:\{0,1\}^d} \mathbb{P}(M=m)\cdot\left[\sigma^2 + \mathrm{Var}\left(\left(\gamma^m_{obs(m)}\right)^\mathsf{T} X_{obs(m)} - \beta^\mathsf{T}_{mis(m)} X_{mis(m)} \,\middle|\, M=m\right)\right. \\ &\qquad \left. + \left(\gamma^m_{obs(m),0} + \left(\gamma^m_{obs(m)}\right)^\mathsf{T}\mathbb{E}\left[X_{obs(m)}|M=m\right] - \beta^\mathsf{T}_{mis(m)}\mathbb{E}\left[X_{mis(m)}|M=m\right]\right)^2\right] \\ &= \sigma^2 + \sum_{m\in:\{0,1\}^d} \mathbb{P}(M=m)\ \Lambda_m\end{aligned}$$

□

D Proof of Theorem 5.1 and Theorem 5.2

Theorem 11.3 in Györfi et al. (2006) allows us to bound the risk of the linear estimator, even in the misspecified case. We recall it here for the sake of completeness.

Theorem D.1 (Theorem 11.3 in Györfi et al. (2006)). *Assume that*

$$Y = f^\star(X) + \varepsilon,$$

where $\|f^\star\|_\infty < L$ *and* $\mathbb{V}[\varepsilon|X] < \sigma^2$ *almost surely. Let* $\mathcal{F}$ *be the space of linear function* $f : [-1,1]^d \to \mathbb{R}$*. Then, letting* $T_L f_n$ *be the linear regression* f_n *estimated via OLS, clipped at* $\pm L$*, we have*

$$\mathbb{E}[(T_L f_n(X) - f^\star(X))^2] \leq c \max\{\sigma^2, L^2\} \frac{d(1+\log n)}{n} + 8 \inf_{f \in \mathcal{F}} \mathbb{E}[(f(X) - f^\star(X))^2],$$

for some universal constant c*.*

Proof of Theorem 5.1. Since Assumptions of Theorem 11.3 in Györfi et al. (2006) are satisfied, we have

$$\mathbb{E}[(T_L f_{\hat{\beta}_{\text{expanded}}}(Z) - f^\star(Z))^2] \leq c \max\{\sigma^2, L^2\} \frac{p(1+\log n)}{n} + 8 \inf_{f \in \mathcal{F}} \mathbb{E}[(f(Z) - f^\star(Z))^2],$$

Since the model is well-specified, the second term is null. Besides,

$$\begin{aligned}\mathbb{E}[(Y - T_L f_{\hat{\beta}_{\text{expanded}}}(Z))^2] &\leq \mathbb{E}[(Y - f^\star(Z))^2] + \mathbb{E}[(T_L f_{\hat{\beta}_{\text{expanded}}}(Z) - f^\star(Z))^2] \\ &\leq \sigma^2 + c \max\{\sigma^2, L^2\} \frac{p(1+\log n)}{n},\end{aligned}$$

which concludes the proof since the full linear model has $p = 2^{d-1}(d+2)$ parameters.

To address the second statement of Theorem 5.1, recall that in our setting, the dimension d is fixed and does not grow to infinity with n. Let $\mathcal{M} = \{m \in \{0,1\}^d, \mathbb{P}[M = m] > 0\}$ and, for all $m \in \mathcal{M}$, $N_m = |\{i : M_i = m\}|$. Note that, the estimator in Theorem 5.1 is nothing but $|\mathcal{M}|$ linear estimators, each one being fitted on data corresponding to a specific missing pattern $m \in \mathcal{M}$. Thus, according to Tsybakov (2003), we know that, there exists constants $c_1, c_2 > 0$, such that, for each missing pattern $m \in \mathcal{M}$, we have the lower bound,

$$\mathbb{E}[(Y - T_L f_{\hat{\beta}_{\text{expanded}}}(Z))^2 | M = m, N_m] \geq \sigma^2 + c_1 \frac{d + 1 - \|m\|_0}{N_m} \mathbb{1}_{N_m \geq 1} + c_2 \mathbb{1}_{N_m = 0}.$$

Taking the expectation with respect to $N_m \sim B(n, \mathbb{P}[M = m])$ and according to Lemma 4.1 in Györfi et al. (2006), we have, for all $m \in \mathcal{M}$,

$$\mathbb{E}[(Y - T_L f_{\hat{\beta}_{\text{expanded}}}(Z))^2 | M = m] \geq \sigma^2 + c_1 \frac{2(d + 1 - \|m\|_0)}{(n+1)\mathbb{P}[M = m]} + c_2 (1 - \mathbb{P}[M = m])^n.$$

Consequently,

$$\begin{aligned}R(T_L f_{\hat{\beta}_{\text{expanded}}}) &= \sum_{m \in \mathcal{M}} \mathbb{E}[(Y - T_L f_{\hat{\beta}_{\text{expanded}}}(Z))^2 | M = m] \mathbb{P}[M = m] \\ &\geq \sigma^2 + \frac{2c_1}{n+1} \sum_{m \in \mathcal{M}} (d + 1 - \|m\|_0) + c_2 \sum_{m \in \mathcal{M}} (1 - \mathbb{P}[M = m])^n \mathbb{P}[M = m] \\ &\geq \sigma^2 + \frac{2c_1 |\mathcal{M}|}{n+1} + c_2 (1 - \min_{m \in \mathcal{M}} \mathbb{P}[M = m])^n.\end{aligned}$$

By assumption, there exists a constant c, such that, for all n large enough, we have

$$R(T_L f_{\hat{\beta}_{\text{expanded}}}) \geq \sigma^2 + \frac{2^d c}{n+1}.$$

□

Proof of Theorem 5.2. As above,

$$\begin{aligned}R(T_L f_{\hat{\beta}_{\text{approx}}}) &\leq \sigma^2 + \mathbb{E}[(f^\star(Z) - T_L f_{\hat{\beta}_{\text{approx}}}(Z))^2] \\ &\leq \sigma^2 + c \max\{\sigma^2, L^2\} \frac{2d(1+\log n)}{n} + 8\mathbb{E}[(f^\star(Z) - f_{\beta^\star_{\text{approx}}}(Z))^2].\end{aligned}$$

To upper bound the last term, note that, for any β_{approx} we have

$$\begin{aligned}
&\mathbb{E}[(f^\star(Z) - f_{\beta_{\text{approx}}}(Z))^2] \\
&= \mathbb{E}\Bigg[\beta_{0,0,\text{approx}} + \sum_{j=1}^{d} \beta_{0,j,\text{approx}} \mathbb{1}_{M_j=1} - \sum_{m\in\{0,1\}^d} \beta^\star_{0,m,\text{expanded}} \mathbb{1}_{M=m} \\
&\quad + \Big(\beta_{1,\text{approx}} - \sum_{m\in\{0,1\}^d} \beta^\star_{1,m,\text{expanded}} \mathbb{1}_{M=m}\Big) X_1 \\
&\quad + \ldots + \Big(\beta_{d,\text{approx}} - \sum_{m\in\{0,1\}^d} \beta^\star_{d,m,\text{expanded}} \mathbb{1}_{M=m}\Big) X_d \Bigg]^2 .
\end{aligned}$$

Denoting by X_{approx} the design matrix X where each element `na` has been replaced by zero, and using a triangle inequality, we have

$$\begin{aligned}
&\mathbb{E}[(W\beta^\star_{\text{full}} - X_{\text{approx}}\beta_{\text{approx}})^2] \\
&\leq (d+1)\mathbb{E}\Bigg[\beta_{0,0,\text{approx}} + \sum_{j=1}^{d} \beta_{0,j,\text{approx}} \mathbb{1}_{M_j=1} - \sum_{m\in\{0,1\}^d} \beta^\star_{0,m,\text{expanded}} \mathbb{1}_{M=m}\Bigg]^2 \\
&\quad + (d+1)\sum_{j=1}^{d} \mathbb{E}\Bigg[\Big(\beta_{j,\text{approx}} - \sum_{m\in\{0,1\}^d} \beta^\star_{j,m,\text{expanded}} \mathbb{1}_{M=m}\Big) X_j\Bigg]^2
\end{aligned}$$

Now, set for all j, $\beta_{0,j,\text{approx}} = 0$ and for all $j = 1, \ldots, d$,

$$\beta_{j,\text{approx}} = \mathbb{E}\left[\sum_{m\in\{0,1\}^d} \beta^\star_{j,m,\text{expanded}} \mathbb{1}_{M=m}\right]$$

and also

$$\beta_{0,0,\text{approx}} = \mathbb{E}\left[\sum_{m\in\{0,1\}^d} \beta^\star_{0,m,\text{expanded}} \mathbb{1}_{M=m}\right].$$

Therefore, for this choice of β_{approx},

$$\begin{aligned}
&\mathbb{E}[(W\beta^\star_{\text{full}} - X_{\text{approx}}\beta_{\text{approx}})^2] \\
&\leq (d+1)\mathbb{V}\left[\sum_{m\in\{0,1\}^d} \beta^\star_{0,m,\text{expanded}} \mathbb{1}_{M=m}\right] + (d+1)\|X\|_\infty^2 \sum_{j=1}^{d} \mathbb{V}\Big[\sum_{m\in\{0,1\}^d} \beta^\star_{j,m,\text{expanded}} \mathbb{1}_{M=m}\Big] \\
&\leq 8(d+1)^2\|f^\star\|_\infty^2.
\end{aligned}$$

Finally, by definition of $\beta^\star_{\text{approx}}$, we have

$$\begin{aligned}
\mathbb{E}[(f^\star(Z) - f_{\beta^\star_{\text{approx}}}(Z))^2] &\leq \mathbb{E}[(f^\star(Z) - f_{\beta_{\text{approx}}}(Z))^2] \\
&\leq 8(d+1)^2\|f^\star\|_\infty^2.
\end{aligned}$$

Finally,

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

since $\|f^\star\|_\infty \leq L$, according to Assumption 5.1. □

E Proof of Theorem 6.1

Let $W^{(1)} \in \mathbb{R}^{2^d \times 2d}$ be the weight matrix connecting the input layer to the hidden layer, and $W^{(2)} \in \mathbb{R}^{2^d}$ the matrix connecting the hidden layer to the output unit. Let $b^{(1)} \in \mathbb{R}^{2^d}$ be the bias for the hidden layer and $b^{(2)} \in \mathbb{R}$ the bias for the output unit. With these notations, the activations of the hidden layer read:

$$\forall k \in [\![1, 2^d]\!],\ a_k = W^{(1)}_{k,\cdot}(X, M) + b^{(1)}_k$$

Splitting $W^{(1)}$ into two parts $W^{(X)}, W^{(M)} \in \mathbb{R}^{2^d \times d}$, the activations can be rewritten as:

$$\forall k \in [\![1, 2^d]\!],\ a_k = W^{(X)}_{k,\cdot} X + W^{(M)}_{k,\cdot} M + b^{(1)}_k$$

Case 1: Suppose that $\forall k \in [\![1, 2^d]\!]$, $\forall j \in [\![1, d]\!]$, $W^{(X)}_{k,j} \neq 0$.
With this assumption, the activations can be reparametrized by posing $G_{k,j} = W^{(M)}_{k,j} / W^{(X)}_{k,j}$, which gives:

$$\begin{aligned}\forall k \in [\![1, 2^d]\!],\ a_k &= W^{(X)}_{k,\cdot} X + W^{(X)}_{k,\cdot} \odot G_{k,\cdot} M + b^{(1)}_k \\ &= W^{(X)}_{k,obs(M)} X_{obs(M)} + W^{(X)}_{k,mis(M)} G_{k,mis(M)} + b^{(1)}_k\end{aligned}$$

and the predictor for an input $(x, m) \in \mathbb{R}^d \times \{0, 1\}^d$ is given by:

$$\begin{aligned}y(x, m) &= \sum_{k=1}^{2^d} W^{(2)}_k ReLU(a^{(1)}_k) + b^{(2)} \\ &= \sum_{k=1}^{2^d} W^{(2)}_k ReLU(W^{(X)}_{k,obs(m)} x_{obs(m)} + W^{(X)}_{k,mis(m)} G_{k,mis(m)} + b^{(1)}_k) + b^{(2)}\end{aligned}$$

We will now show that there exists a configuration of the weights $W^{(X)}$, G, $W^{(2)}$, $b^{(1)}$ and $b^{(2)}$ such that the predictor y is exactly the Bayes predictor. To do this, we will first show that we can choose G and $b^{(1)}$ such that the points with a given missing-values pattern all activate one single hidden unit, and conversely, a hidden unit can only be activated by a single missing-values pattern. This setting amounts to having one linear regression per missing-values pattern. Then, we will show that $W^{(X)}$ and $W^{(2)}$ can be chosen so that for each missing-values pattern, the slope and bias match those of the Bayes predictor.

One to one correspondence between missing-values pattern and hidden unit In this part, $W^{(X)}$, $W^{(2)}$ and $b^{(2)}$ are considered to be fixed to arbitrary values. We denote by m_k, $k \in [\![1, 2^d]\!]$, the possible values taken by the mask vector M. There is a one-to-one correspondence between missing-values pattern and hidden unit if G and $b^{(1)}$ satisfy the following system of 2^{2d} inequations:

$$\forall x \in \text{supp}(X),\ \forall k \in [\![1, 2^d]\!],\ \begin{cases} W^{(X)}_{k,obs(m_k)} x_{obs(m_k)} + W^{(X)}_{k,mis(m_k)} G_{k,mis(m_k)} + b^{(1)}_k \geq 0 & (7) \\ W^{(X)}_{k,obs(m')} x_{obs(m')} + W^{(X)}_{k,mis(m')} G_{k,mis(m')} + b^{(1)}_k \leq 0 \qquad \forall m' \neq m_k & (8) \end{cases}$$

i.e., missing-values pattern m_k activates the k^{th} hidden unit but no other missing-values pattern activates it.

Hereafter, we suppose that the support of the data is finite, so that there exist $M \in \mathbb{R}^+$ such that for any $j \in [\![1, d]\!]$, $|x_j| < M$. As a result, we have:

$$\begin{aligned}\left| W^{(X)}_{k,obs(m_k)} x_{obs(m_k)} \right| &\leq M \sum_{j \in obs(m_k)} \left| W^{(X)}_{k,j} \right| \\ &\leq M \,|\, obs(m_k) \,|\, \max_{j \in obs(m_k)} \left| W^{(X)}_{k,j} \right| \\ &= K_k \,|\, obs(m_k) \,|\end{aligned}$$

where we define $K_k = M \max_{j \in obs(m_k)} \left| W_{k,j}^{(X)} \right|$. We also define $I_k^{(1)} \in \mathbb{R}$ such that:

$$\forall j \in mis(m_k),\ W_{k,j}^{(X)} G_{k,j} = I_k^{(1)} \tag{9}$$

Then satisfying inequation (7) implies satisfying the following inequation:

$$\forall k \in [\![1, 2^d]\!],\ -|\, obs(m_k)\,|\, K_k + |\, mis(m_k)\,|\, I_k^{(1)} + b_k^{(1)} \geq 0 \tag{10}$$

Similarly, we define a quantity $I_k^{(2)} \in \mathbb{R}$ which satisfies:

$$\forall j \in obs(m_k),\ W_{k,j}^{(X)} G_{k,j} = I_k^{(2)} \tag{11}$$

A missing-values pattern $m' \neq m_k$ differs from m_k by a set of entries $\mathcal{J} \subseteq mis(m_k)$ which are missing in m_k but observed in m', and a set of entries $\mathcal{L} \subseteq obs(m_k)$ which are observed in m_k but missing in m'. We will call a pair $\mathcal{J} \subseteq mis(m_k)$, $\mathcal{L} \subseteq obs(m_k)$ such that $|\, \mathcal{J} \cup \mathcal{L}\,| \neq 0$ a *feasible* pair. With these quantities, satisfying inequation 8 implies satisfying the following inequation:

$$\forall k \in [\![1, 2^d]\!],\ \forall (\mathcal{J}, \mathcal{L}) \text{ feasible},\ (|\, obs(m_k)\,| + |\, \mathcal{J}\,| - |\, \mathcal{L}\,|)\, K_k + (|\, mis(m_k)\,| - |\, \mathcal{J}\,|)\, I_k^{(1)} + |\, \mathcal{L}\,|\, I_k^{(2)} + b_k^{(1)} \leq 0 \tag{12}$$

Thus, by (10) and (12), a one to one correspondence between missing-values pattern and hidden unit is possible if there exists $I_k^{(1)}$, $I_k^{(2)}$, $b_k^{(1)}$ such that:

$$\forall k \in [\![1, 2^d]\!], \begin{cases} |\, mis(m_k)\,|\, I_k^{(1)} + b_k^{(1)} \geq |\, obs(m_k)\,|\, K_k \\ |\, mis(m_k)\,|\, I_k^{(1)} + b_k^{(1)} \leq -|\, obs(m_k)\,|\, K_k - (|\, \mathcal{J}\,| - |\, \mathcal{L}\,|) K_k + |\, \mathcal{J}\,|\, I_k^{(1)} - |\, \mathcal{L}\,|\, I_k^{(2)} \end{cases} \quad \forall (\mathcal{J}, \mathcal{L}) \text{ feasible} \tag{13}$$

Because $b_k^{(1)}$ can be any value, this set of inequations admits a solution if for any feasible $(\mathcal{J}, \mathcal{L})$:

$$\begin{aligned} & |\, obs(m_k)\,|\, K_k < -|\, obs(m_k)\,|\, K_k - (|\, \mathcal{J}\,| - |\, \mathcal{L}\,|) K_k + |\, \mathcal{J}\,|\, I_k^{(1)} - |\, \mathcal{L}\,|\, I_k^{(2)} \\ \iff & 2\,|\, obs(m_k)\,|\, K_k + (|\, \mathcal{J}\,| - |\, \mathcal{L}\,|) K_k < |\, \mathcal{J}\,|\, I_k^{(1)} - |\, \mathcal{L}\,|\, I_k^{(2)} \\ \iff & \begin{cases} \frac{-2|\, obs(m_k)\,| K_k}{|\, \mathcal{L}\,|} + K_k > I_k^{(2)} & \text{if } |\, \mathcal{J}\,| = 0 \\ \frac{2|\, obs(m_k)\,| K_k}{|\, \mathcal{J}\,|} + K_k < I_k^{(1)} & \text{if } |\, \mathcal{L}\,| = 0 \\ I_k^{(1)} > K_k + \frac{|\, obs(m_k)\,| K_k}{|\, \mathcal{J}\,|} \quad \text{and} \quad I_k^{(2)} < K_k - \frac{|\, obs(m_k)\,| K_k}{|\, \mathcal{L}\,|} & \text{otherwise} \end{cases} \end{aligned}$$

Satisfying these inequalities for any feasible $(\mathcal{J}, \mathcal{L})$ can be achieved by choosing:

$$I_k^{(1)} > (1 + 2\,|\, obs(m_k\,|)) K_k \tag{14}$$

$$I_k^{(2)} < (1 - 2\,|\, obs(m_k\,|)) K_k \tag{15}$$

To conclude, it is possible to achieve a one to one correspondence between missing-values pattern and hidden unit by choosing G and $b^{(1)}$ such that for the k^{th} hidden unit:

$$\begin{cases} I_k^{(1)} > (1 + 2\,|\, obs(m_k\,|)) K_k & \text{by 9 and 14} \\ I_k^{(2)} < (1 - 2\,|\, obs(m_k\,|)) K_k & \text{by 11 and 15} \\ b_k^{(1)} \text{ satisfies 13} \end{cases} \tag{16}$$

Equating slopes and biases with that of the Bayes predictor We just showed that it is possible to choose G and $b^{(1)}$ such that the points with a given missing-values pattern all activate one single hidden unit, and conversely, a hidden unit can only be activated by a single missing-values pattern. As a consequence, the predictor for an input $(x, m_k) \in \mathbb{R}^d \times \{0, 1\}^d$ is given by:

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

For each missing-values pattern, it is now easy to choose $W^{(X)}_{k,obs(m_k)}$ and $W^{(2)}$ so that the slopes and biases of this linear function match those of the Bayes predictor defined in proposition 4.1. Let $\beta_k \in \mathbb{R}^{|\, obs(m_k)\,|}$ and $\alpha_k \in \mathbb{R}$ be the slope and bias of the Bayes predictor for missing-values pattern m_k. Then setting

$$\begin{cases} W^{(2)}_k \left(W^{(X)}_{k,mis(m_k)} G_{k,mis(m_k)} + b^{(1)}_k \right) + b^{(2)} = \alpha_k \\ W^{(2)}_k W^{(X)}_{k,obs(m_k)} = \beta_k \end{cases} \tag{17}$$

equates the slope and bias of the MLP to those of the bias predictor.

Construction of weights for which the MLP is the Bayes predictor We have shown that achieving a one to one correspondence between missing data pattern and hidden units involves satisfying a set of inequations on the weights (16), while equating the slopes and biases to those of the Bayes predictor involves another set of equations (17). To terminate the proof, it remains to be shown that the whole system of equations and inequations admits a solution.

We start by working on the one-to-one correspondence system of inequations (16). Let $\epsilon > 0$ be a parameter. Inequations (14) and (15) are satisfied by choosing:

$$I^{(1)}_k = (1 + 2\,|\, obs(m_k)\,|)K_k + \epsilon \tag{18}$$
$$I^{(2)}_k = (1 - 2\,|\, obs(m_k)\,|)K_k - \epsilon \tag{19}$$

According to the second inequation in (13), $b^{(1)}_k$ is upper bounded as:

$$b^{(1)}_k \leq -\,|\, obs(m_k)\,|\, K_k - |\, mis(m_k)\,|\, I^{(1)}_k - (|\, \mathcal{J}\,| - |\, \mathcal{L}\,|)K_k + |\, \mathcal{J}\,|\, I^{(1)}_k - |\, \mathcal{L}\,|\, I^{(2)}_k$$

This inequation can be simplified:

$$\begin{aligned} b^{(1)}_k &\leq -\,|\, obs(m_k)\,|\, K_k - |\, mis(m_k)\,|\, I^{(1)}_k + |\, \mathcal{J}\,|\, (I^{(1)}_k - K_k) + |\, \mathcal{L}\,|\, (K_k - I^{(2)}_k) \\ &= -\,|\, obs(m_k)\,|\, K_k - |\, mis(m_k)\,| + |\, \mathcal{J}\,|\, (2\,|\, obs(m_k)\,|\, K_k + \epsilon) + |\, \mathcal{L}\,|\, (2\,|\, obs(m_k)\,|\, K_k + \epsilon) \end{aligned}$$

The smallest upper bound is obtained for $|\, \mathcal{J} \cup \mathcal{L}\,| = 1$ which gives:

$$b^{(1)}_k \leq |\, obs(m_k)\,|\, K_k - |\, mis(m_k)\,| + \epsilon$$

According to the first inequation in (13), $b^{(1)}_k$ is also lower bounded as:

$$b^{(1)}_k \geq |\, obs(m_k)\,|\, K_k - |\, mis(m_k)\,|\, I^{(1)}_k$$

A valid choice for $b^{(1)}_k$ is the mean of its upper and lower bounds. We therefore choose to set:

$$b^{(1)}_k = |\, obs(m_k)\,|\, K_k - |\, mis(m_k)\,|\, I^{(1)}_k + \frac{\epsilon}{2} \tag{20}$$

To summarise, we can restate the conditions for one to one correspondence as:

$$\epsilon > 0 \tag{21}$$
$$I^{(1)}_k = (1 + 2\,|\, obs(m_k)\,|)K_k + \epsilon \tag{22}$$
$$I^{(2)}_k = (1 - 2\,|\, obs(m_k)\,|)K_k - \epsilon \tag{23}$$
$$b^{(1)}_k = |\, obs(m_k)\,|\, K_k - |\, mis(m_k)\,|\, I^{(1)}_k + \frac{\epsilon}{2} \tag{24}$$

We now turn to the slopes and biases equations (17). Replacing $b_k^{(1)}$ in the bias equation by its value in (24) we get:

$$\begin{aligned} & W_k^{(2)} \left(W_{k,mis(m_k)}^{(X)} G_{k,mis(m_k)} + b_k^{(1)} \right) + b^{(2)} = \alpha_k \\ \iff & W_k^{(2)} \left(|\, mis(m_k) \,|\, I_k^{(1)} + b_k^{(1)} \right) + b^{(2)} = \alpha_k \\ \iff & W_k^{(2)} \left(|\, obs(m_k) \,|\, K_k + \frac{\epsilon}{2} \right) + b^{(2)} = \alpha_k \end{aligned}$$

Putting together the one to one correspondence conditions, the slope and biases equations as well as the variable definitions, we get a set of 8 equations and 1 inequation:

$$\epsilon > 0 \tag{25}$$

$$I_k^{(1)} = (1 + 2\,|\, obs(m_k)\,|)K_k + \epsilon \tag{26}$$

$$I_k^{(2)} = (1 - 2\,|\, obs(m_k)\,|)K_k - \epsilon \tag{27}$$

$$b_k^{(1)} = |\, obs(m_k) \,|\, K_k - |\, mis(m_k) \,|\, I_k^{(1)} + \frac{\epsilon}{2} \tag{28}$$

$$W_k^{(2)} \left(|\, obs(m_k) \,|\, K_k + \frac{\epsilon}{2} \right) + b^{(2)} = \alpha_k \tag{29}$$

$$W_k^{(2)} W_{k,obs(m_k)}^{(X)} = \beta_k \tag{30}$$

$$K_k = M \max_{j \in obs(m_k)} \left| W_{k,j}^{(X)} \right| \tag{31}$$

$$\forall j \in mis(m_k),\ W_{k,j}^{(X)} G_{k,j} = I_k^{(1)} \tag{32}$$

$$\forall j \in obs(m_k),\ W_{k,j}^{(X)} G_{k,j} = I_k^{(2)} \tag{33}$$

One can verify that this system of inequations has a solution. Indeed, choose $W_{k,obs(m_k)}^{(X)}$ proportional to β_k so that equation (30) can be verified. This imposes a value for $W_k^{(2)}$ via (30) and a value for K_k via (31). In turn, it imposes a value for ϵ via (29): $\epsilon = 2 \left(\alpha_k - b^{(2)} - W_k^{(2)} \,|\, obs(m_k) \,|\, K_k \right)$. The value obtained for ϵ is positive if we choose $b^{(2)}$ sufficiently negative. Note that there is one single value of $b^{(2)}$ for all units so $b^{(2)}$ should be chosen by considering all units. Then K_k and ϵ impose $I_k^{(1)}$ and $I_k^{(2)}$ via (26) and (27). K_k, ϵ and $I_k^{(1)}$ impose $b_k^{(1)}$ via (28). Finally $W_{k,.}^{(X)}$, $I_k^{(1)}$ and $I_k^{(2)}$ impose G via (32) and (33).

Case 2: Suppose that $\exists k \in [\![1, 2^d]\!]$, $\exists j \in [\![1, d]\!] : W_{k,j}^{(X)} = 0$.

Recall that the proof which shows that we can achieve a one to one correspondence between missing-values pattern and hidden unit relies on the assumption that $\forall k \in [\![1, 2^d]\!]$, $\forall j \in [\![1, d]\!]$, $W_{k,j}^{(X)} \neq 0$. However, if there is a slope β_k of the Bayes predictor such that its j^{th} coefficient is 0, then we must choose $W_{k,j}^{(X)} = 0$ to achieve Bayes consistency. In such a case, we need to extend the one to one correspondence proof to the case where an entry of $W_{k,j}^{(X)}$ can be zero. It turns out to be easy.

In this case, we cannot pose $G_{k,j} = W_{k,j}^{(M)} / W_{k,j}^{(X)}$. Let $\mathcal{Z}_k \subseteq [\![1, d]\!]$ be the set of indices such that $\forall j \in \mathcal{Z}_k$, $W_{k,j}^{(X)} = 0$. The whole reasoning exposed in case 1 still holds if we replace $obs(m)$ by $obs(m) \setminus \mathcal{Z}_k$ and $mis(m)$ by $mis(m) \setminus \mathcal{Z}_k$.

F Complementary figures

F.1 Comparison at $n = 75\,000$

Figure 3 gives a box plot view of the behavior at $n = 75\,000$. It is complementary to the learning curves, though it carries the same information.

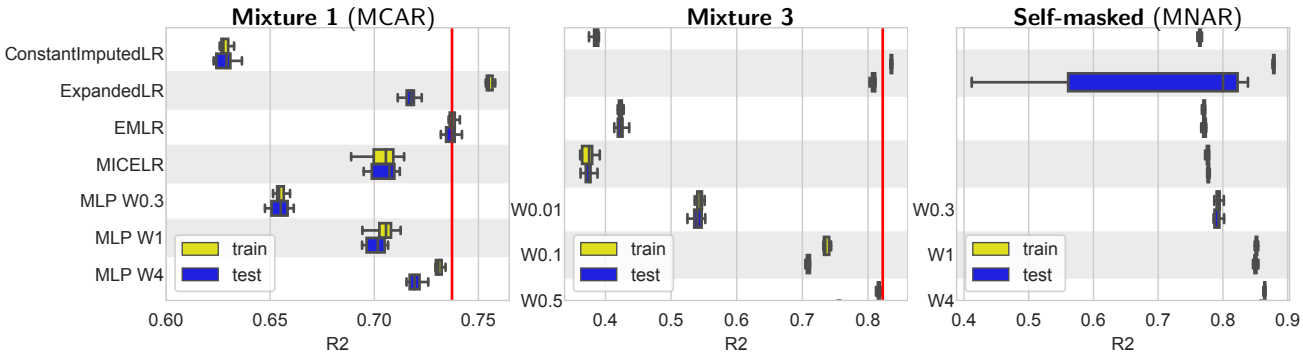

Figure 3: **Prediction accuracy** R2 score for the 3 data types with $n = 75,000$ training samples and in dimension $d = 10$. The quantities displayed are the mean and standard deviation over 5 repetitions.

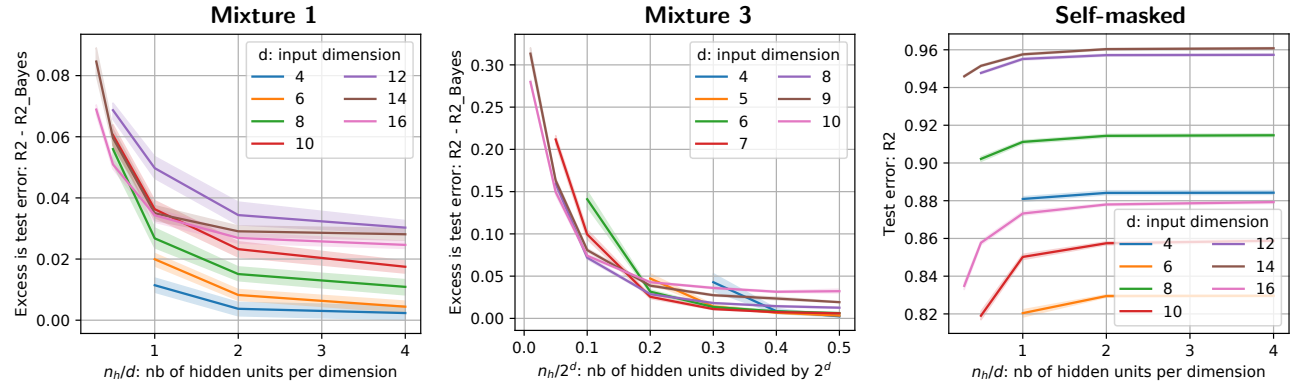

Figure 4: **Performance of the one hidden layer MLP as a function of its number of hidden units** For the mixtures of Gaussians, the performance is given as the difference between the R2 score of the MLP and that of the Bayes predictor. For each dimension d, multiple MLPs are trained, each with a different number of hidden units given by $q \times d$ for mixture 1 and self-masked, $q \times 2^d$ for mixture 3. 75,000 training samples were used for Mixture 1 and Self-masked and 375,000 for Mixture 3.

F.2 Experiments on growing MLP's width

Figure 4 shows the performance of the MLP in the various simulation scenarios as a function of the number of hidden units of the networks. In each scenario, the number of hidden units is taken proportional to a function of the input dimension d:

mixture 1 : $n_h \propto d$

mixture 3 : $n_h \propto 2^d$

selfmasked : $n_h \propto d$

These results show that the number of hidden units needed by the MLP to predict well are a function of the complexity of the underlying data-generating mechanism. Indeed, for the *mixture 1*, the MLP only needs $n_h \propto d$ while the missing values are MCAR, and therefore ignorable. For *selfmasked*, the challenge is to find the right set of thresholds, after which the prediction is relatively simple: the MLP also needs $n_h \propto d$. On the opposite, for *mixture 3*, the multiple Gaussians create couplings in the prediction function; as the consequence, the MLP needs $n_h \propto 2^d$.