Near-optimal rate of consistency for linear models with missing values

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, Erwan Scornet

Near-optimal rate of consistency for linear models with missing values

Alexis Ayme [1] **Claire Boyer** [1 2] **Aymeric Dieuleveut** [3] **Erwan Scornet** [3]

Abstract

Missing values arise in most real-world data sets due to the aggregation of multiple sources and intrinsically missing information (sensor failure, unanswered questions in surveys...). In fact, the very nature of missing values usually prevents us from running standard learning algorithms. In this paper, we focus on the extensively-studied linear models, but in presence of missing values, which turns out to be quite a challenging task. Indeed, the Bayes predictor can be decomposed as a sum of predictors corresponding to each missing pattern. This eventually requires to solve a number of learning tasks, exponential in the number of input features, which makes predictions impossible for current real-world datasets. First, we propose a rigorous setting to analyze a least-square type estimator and establish a bound on the excess risk which increases exponentially in the dimension. Consequently, we leverage the missing data distribution to propose a new algorithm, and derive associated adaptive risk bounds that turn out to be minimax optimal. Numerical experiments highlight the benefits of our method compared to state-of-the-art algorithms used for predictions with missing values.

1. Introduction

Missing values are more and more present as the size of datasets increases. These missing values can occur for a variety of reasons, such as sensor failures, refusals to answer poll questions, or aggregations of data coming from different sources (with different methods of data collection). There may be different processes of missing value generation on the same dataset, which makes the task of data cleaning difficult or impossible without creating large biases. In his leading work, Rubin (1976) distinguishes three missing values scenarios: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR), depending on the links between the observed variables, the missing ones, and the missing pattern.

In the linear regression framework, most of the literature focuses on *parameter estimation* (Little, 1992; Jones, 1996; Robins et al., 1994), using sometimes a sparse prior leading to the Lasso estimator (Loh & Wainwright, 2012) or the Dantzig selector (Rosenbaum & Tsybakov, 2010). Note that the robust estimation literature (Dalalyan & Thompson, 2019; Chen & Caramanis, 2013) could be also used to handle missing values, as the latter can be reinterpreted as a multiplicative noise in linear models. Besides, Sportisse et al. (2020) adapt and theoretically study the famous stochastic gradient algorithm for model estimation in online linear regression.

On the other hand, *prediction* with missing values in a parametric framework -even under a linear model- is in fact not an easy task. Indeed, the prediction task is distinct from model estimation: estimated model parameters cannot be directly used to predict on a test sample containing missing values as well. As a matter of fact, the occurrence of missing data turns the linear regression problem into a semi-discrete one of very high complexity. Finally, establishing risk bounds -even without missing values- for random designs is already a challenge as studied in papers (Györfi et al., 2006; Audibert & Catoni, 2011; Dieuleveut et al., 2017) and more recently in (Mourtada, 2019).

Related work. There is actually little work on prediction with missing values. Pelckmans et al. (2005) adapt the SVM classifier to the case of missing values. Josse et al. (2019) study the consistency of imputation strategies prior to non-parametric learning methods. Prediction under linear models has been studied in (Le Morvan et al., 2020a;b), by exploiting the peculiar pattern-by-pattern structure of the Bayes predictor (i.e. decomposable into predictors specific to each missing pattern), and estimating it when the input variables are assumed to be Gaussian. Le Morvan et al. (2020b) obtain risk bounds, that suffer from the curse of dimensionality, and are actually not compatible with their Gaussian assumption. Agarwal et al. (2019) investigates

[1]Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation (LPSM), F-75005 Paris, France [2]MOKAPLAN, INRIA Paris [3]CMAP, UMR7641, Ecole Polytechnique, IP Paris, 91128 Palaiseau, France. Correspondence to: Alexis Ayme <alexis.ayme@sorbonne-universite.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

the PCR strategy to deal with missing values in a high-dimensional setting.

Contributions. In this paper, we study pattern-by-pattern predictors for regression with missing input variables. First, we provide a synthetic overview of all assumptions that allow to obtain a pattern-wise linear Bayes predictor, and we propose a detailed study on how these assumptions are related (Section 2). Second, we provide a distribution-free excess risk bound for a least-square estimator handling unbounded features (Section 3), but suffering from the curse of dimensionality. We therefore introduce a novel thresholded estimator for which we establish an excess risk bound adaptive to the missing pattern distribution (Section 4). The latter actually applies to all types of missing data (MCAR, MAR, MNAR) and is shown to be minimax optimal. We exhibit three settings in which our bound is precisely evaluated, improving upon state-of-the-art results. Finally, we experimentally illustrate our method on three different simulation settings, outperforming existing competitors designed to handle missing values, both in terms of predictive performance and computational time (Section 5). All the proofs of theoretical results can be found in the supplementary materials.

Notations. For $n \in \mathbb{N}$, we denote $[n] = \{1, \dots, n\}$. We use $\lesssim$ to denote inequality up to a universal constant. We denote $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. For any $x \in \mathbb{R}^d$ and for any set $J \subset [d]$ of indices, we let x_J be the subvector of x composed of the components indexed by J. $|J|$ denotes the cardinal of a discrete set J.

2. Typology of missing value and its consequence on the Bayes predictor

2.1. Setting

In a context of regression, we observe $n \in \mathbb{N}$ input/output observations $(X_i, Y_i)_{i\in[n]}$, i.i.d. copies of a generic pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, assuming that the underlying model linking Y to X is linear.

Assumption 1 (Linear Model)**.** $Y = \beta_0 + \beta^\top X + \epsilon$, with a Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent of X.

The (unknown) model parameters are therefore $(\beta_0, \beta) \in \mathbb{R}^{d+1}$. Although standard linear regression is a well-understood problem in statistics, we consider here that only a fraction of the components of X is available: to the data $X \in \mathbb{R}^d$ one associates the missing values pattern $M \in \{0, 1\}^d$, such that $M_j = 1$ if and only if X_j is missing. Let $\mathcal{M} = \{0, 1\}^d$ be the set of missing values patterns. For $m \in \mathcal{M}$, we denote by $\mathrm{obs(m)}$ (resp. $\mathrm{mis(m)}$) the set of indexes of the observed variables (resp. the missing variables) and $X_{\mathrm{obs(m)}}$ (resp. $X_{\mathrm{mis(m)}}$) the vector of observed components (resp. unobserved components) of X. Thus, under a linear model with missing covariates, our goal is to predict Y given $\left(X_{\mathrm{obs(m)}}, M\right)$, denoted Z in the sequel.

2.2. Bayes predictor

The Bayes predictor for the quadratic loss can be decomposed according to the possible missing data patterns, as

$$f^\star(Z) = \mathbb{E}\left[Y|Z\right] = \mathbb{E}\left[Y|X_{\mathrm{obs(m)}}, M\right] = \sum_{m\in\mathcal{M}} f^\star_m(X_{\mathrm{obs(m)}})\mathbb{1}_{M=m},$$

where $f^\star_m(X_{\mathrm{obs(m)}}) := \mathbb{E}\left[Y|X_{\mathrm{obs(m)}}, M = m\right]$ can be seen as the Bayes predictor conditionally on the event "$M = m$". Under Assumption 1, $f^\star_m$ can be written as

$$f^\star_m(X_{\mathrm{obs(m)}}) = \beta_0 + \beta^\top_{\mathrm{obs(m)}} X_{\mathrm{obs(m)}} + \beta^\top_{\mathrm{mis}(m)} \mathbb{E}\left[X_{\mathrm{mis(m)}}|X_{\mathrm{obs(m)}}, M = m\right].$$

Thus, $f^\star_m$ remains linear in the observed variables X_{obs}, provided that $x \mapsto \mathbb{E}\left[X_{\mathrm{mis(m)}}|X_{\mathrm{obs(m)}} = x, M = m\right]$ is a linear function. This is not always true as shown in the following example.

Example 2.1. Let $Y = X_1 + X_2 + X_3 + \epsilon$, where $X_3 = X_2 e^{X_1}$. Then

$$f^\star_{(0,0,1)}(X_1, X_2) = X_1 + X_2 + X_2 e^{X_1},$$

where $m = (0, 0, 1)$ is the missing value pattern where only X_1 and X_2 are observed. Despite Assumption 1, the predictor $f^\star_{(0,0,1)}$ is not linear in the observed covariates, due to the non-linear link between the observed variables X_1, X_2 and the missing one X_3. Therefore, linear regression with missing data is hard to analyze without any additional assumptions on the joint distribution (X, M).

2.3. Data scenarios

There exist two main approaches for modelling the joint distribution of X and M: selection models (Heckman, 2012) and pattern-mixture ones (Little, 1993).

Selection models. They rely on the following factorization of the joint distribution $\mathbb{P}(X, M) = \mathbb{P}(X)\mathbb{P}(M|X)$. Therefore, in selection models, one specifies the distributions of X (the most common ones being Assumptions 2 and 3 below) and $M|X$ (Assumptions 4, 5 or 6 below).

Assumption 2 (Independent covariates)**.** The covariates $\{X_j\}_{j\in[d]}$ are mutually independent.

Assumption 3 (Gaussian covariates)**.** There exist $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ such that $X \sim \mathcal{N}(\mu, \Sigma)$.

Note that this latter assumption excludes the pathological Example 2.1. Regarding the distribution of $M|X$, Rubin (1976) introduces the three following missingness mechanisms.

Assumption 4 (Missing Completely At Random - MCAR)**.** For all $m \in \mathcal{M}$, $\mathbb{P}(M = m|X) = \mathbb{P}(M = m)$.

Assumption 5 (Missing At Random - MAR)**.** For all $m \in \mathcal{M}$, $\mathbb{P}(M = m|X) = \mathbb{P}\left(M = m|X_{\text{obs(m)}}\right)$.

Assumption 6 (Missing Non At Random - MNAR)**.** The missing pattern M depends on the full vector X (thus, on the observed and missing entries).

To illustrate these scenarios, consider the simple situation of a survey with two variables, *Income* and *Age*, with missing values only on the *Income* variable. The MCAR setting (Assumption 4) holds when the missing values are independent of any value (e.g. respondents have forgotten to fill the form). The MAR situation (Assumption 5) is verified when missing values on *Income* depend on the values of *Age* (e.g. older respondents would be less inclined to reveal their income). The MNAR scenario (Assumption 6) allows the occurrence of the missing values on *Income* to depend on the values of the income itself (e.g. poor and rich respondents would be less inclined to reveal their income). A particular case of the last example consists in considering that the missingness mechanism for a given variable is only dictated by its underlying value.

Assumption 7 (Gaussian Self-Masking)**.** For all $m \in \mathcal{M}$, $\mathbb{P}(M = m|X) = \prod_{j=1}^{d} \mathbb{P}(M_j = m_j|X_j)$ and for $j \in [d]$,

$$\mathbb{P}(M_j = 1|X_j) \propto \exp\left(-\frac{1}{2}\frac{(X_j - \widetilde{\mu}_j)^2}{\widetilde{\sigma}_j^2}\right).$$

Pattern-mixture models. Such models rely on the following factorization of the joint distribution $\mathbb{P}(X, M) = \mathbb{P}(M)\mathbb{P}(X|M)$. Therefore, in pattern-mixture models, one specifies the distributions of M and $X|M$: one can therefore appeal to the Gaussian pattern mixture model (GPMM).

Assumption 8 (Gaussian Pattern Mixture Model-GPMM)**.** For all $m \in \mathcal{M}$, $X|(M = m) \sim \mathcal{N}(\mu^{(m)}, \Sigma^{(m)})$.

2.4. Linearity of the Bayes predictor

In this subsection, we give an overview of the properties that ensure the linearity of $f_m^\star$ based on the assumptions defined in Section 2.3.

Definition 2.2. Consider the vector space of linear predictors in the observed variables i.e. $f \in \mathcal{F}_b$ if $f(., m)$ is linear for all $m \in \mathcal{M}$. The dimension of $\mathcal{F}_b$ is $p := 2^{d-1}(d + 2)$.

Proposition 2.3. *[Le Morvan et al. 2020b;a, resp. Prop. 4.1 and Prop. 2.1] Under Assumption 1 and one of the following hypotheses*

1. *Gaussian covariates with M(C)AR mechanisms (Assumption 3 and (4 or 5)),*
2. *Gaussian covariates with Gausian Self-Masking mechanisms (Assumption 3 and 7),*
3. *Gaussian Pattern Mixture Model (Assumption 8),*
4. *Independent covariates (Assumption 2).*

Then $f^\star \in \mathcal{F}_b$ i.e. for all $m \in \mathcal{M}$ there exist $\delta_0^{(m)} \in \mathbb{R}$ and $\delta^{(m)} \in \mathbb{R}^{|obs(m)|}$ such that

$$f_m^\star(X_{\text{obs(m)}}) = \delta_0^{(m)} + \left(\delta^{(m)}\right)^\top X_{\text{obs(m)}}.$$

Proposition 2.3 is summarized in Figure 1: there is indeed a wide variety of possible assumptions such that $f^\star \in \mathcal{F}_b$. Note that the covariates independence (Assumption 2) allows to get the Bayes predictor linearity beyond Gaussian models. Furthermore, Assumptions 2, 7, and 8 may not only include MAR but MNAR scenarios as well, the latter known to be challenging in an inference setting. However, this comes at the cost of an exponential growth of the dimension p of $\mathcal{F}_b$ with the ambient dimension d ($p = 2^{d-1}(d + 2)$).

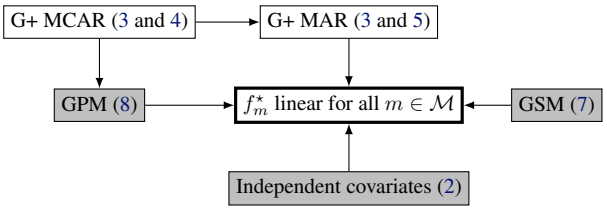


Figure 1. Links between the different assumptions to get the linearity of the Bayes predictor as in Proposition 2.3. Scenarios that may contain MNAR cases are depicted in gray.

2.5. Links between Gaussian PMM & selection models.

In this subsection, we investigate the links between the assumptions of Proposition 2.3, summarized in Figure 2.

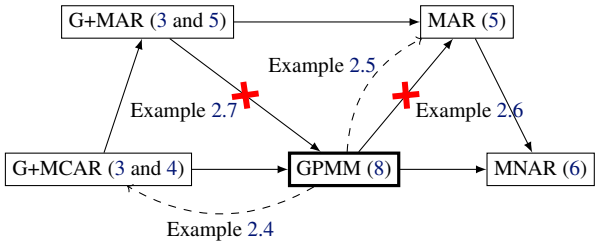


Figure 2. Links between Gaussian pattern mixture models (GPMM) and Gaussian selection models. Solid arrows correspond to inclusions, dotted (resp. crossed) arrows illustrate a partial inclusion (resp. non-inclusion).

First, we remark that GPMM may implicitly encode for M(C)AR and MNAR scenarios.

Example 2.4 (From GPMM to MCAR). Consider a GPMM such that there exist μ and Σ such that $\mu^{(m)} = \mu$ and $\Sigma^{(m)} = \Sigma$ for all $m \in \mathcal{M}$. One can show that the latter is necessary and sufficient to get a MCAR dataset (Assumption 4) with Gaussian covariates X (Assumption 3).

Example 2.5 (From GPMM to MAR). Consider a subset of always observed variables indexed by $J \subset [d]$ (i.e. $\mathbb{P}(M_j = 0) = 1$ for $j \in J$) and a GPMM such that

$$\begin{cases} \mu_J^{(m)} & = \mu_m \ \sim \mathcal{U}([-1,1]^{|J|}) \\ \mu_{J^c}^{(m)} & = \mu \in \mathbb{R}^{|J^c|} \text{ (fixed)} \end{cases}$$

and

$$\begin{cases} \Sigma_{J,J}^{(m)} & = \Sigma_m \in \mathbb{R}^{|J|\times|J|} \\ \Sigma_{J^c,J^c}^{(m)} & = \Sigma \in \mathbb{R}^{|J^c|\times|J^c|} \text{ (fixed)} \\ \Sigma_{J^c,J}^{(m)} & = 0, \end{cases}$$

where $\Sigma_m \in \mathbb{R}^{J\times J}$ can depend on m. In such a case, for all $m \in \mathcal{M}$, $\mathbb{P}(M = m|X) = \mathbb{P}(M = m|X_J)$ with X_J always observed, thus the missing mechanism can be qualified of MAR. Furthermore, note that as soon as there exist $m, m' \in \mathcal{M}$, such that $\mu_m \neq \mu_{m'}$ or $\Sigma_m \neq \Sigma'_m$ then the dataset is ensured not to be MCAR.

Example 2.6 (From GPMM to MNAR). Consider a GPMM such that for all $m \in \mathcal{M}$, $\Sigma^{(m)} = I_d$ and $\mu^{(m)}$ is uniformly drawn at random in $[-1,1]^d$. In such a case, the missing mechanism is MNAR and almost surely not MAR.

Note that, Gaussian linear models with MAR missing values (Assumptions 3 and 5) are not necessarily included in Gaussian pattern mixture models (Assumption 8). This is in particular highlighted by the following example.

Example 2.7 (G+MAR $\not\subseteq$ GPMM). Let $(X_1, X_2) \sim \mathcal{N}(0, I_2)$ such that X_1 is always observed ($M_1 = 0$) and X_2 is observed if and only if $X_1 \leq 0$ ($M_2 = \mathbb{1}_{X_1>0}$). This corresponds to a linear Gaussian model (Assumption 3), with a MAR missing variable X_2 (Assumption 5) since missing values on X_2 only depend on X_1 which is always observed. However, this cannot be a GPMM (Assumption 8) as the distribution of $X|(M = (0,1))$ is supported on a half space (preventing $X|M$ from being Gaussian).

3. A distribution-free bound on excess risk

In the framework of missing values, let the excess risk be

$$\mathcal{E}\left(\widehat{f}\right) := \mathbb{E}\left[\left(\widehat{f}(Z) - f^\star(Z)\right)^2 \middle| \mathcal{D}_n\right], \tag{1}$$

and its integrated version $\mathbb{E}[\mathcal{E}(\widehat{f})]$. This quantity measures the quality of performance made by a prediction function $\widehat{f}$ compared to the optimal predictor $f^\star$. Le Morvan et al. (2020b) propose a missing-pattern-distribution-free control on the integrated risk scaling in $d2^d/n$ for the least-square estimator, requiring (i) the Bayes predictor to be linear in the observed variables (Definition 2.2) and (ii) the covariates boundedness. Unfortunately, Proposition 2.3 highlights the convenience of Gaussian covariates to ensure the linearity of the Bayes predictor, and yet they are incompatible with (ii). We intend to fill this gap by providing a unified and more general analysis to include the case of unbounded covariates (Assumption 9 below) and where the Bayes predictor is assumed to be regular without being explicitly linear (Assumption 10).

Assumption 9 (Sub-Gaussian covariate)**.** There is a positive constant γ such that for all $j \in [d]$, $\mathbb{E}[X_j]^2 \leq \gamma$ and $X_j - \mathbb{E}[X_j]$ is γ-sub-Gaussian, that is

$$\forall t > 0, \qquad \mathbb{P}(|X_j - \mathbb{E}[X_j]| > t) \leq e^{-\frac{t^2}{2\gamma}}. \tag{2}$$

Assumption 10 (Lipschitz)**.** There exists $B > 0$ such that for all $m \in \mathcal{M}$, $f_m^\star$ is B-Lipschitz for the ℓ^∞-norm, and $|f_m^\star(0)| \leq B$.

According to Assumption 10, one can control the ℓ^∞-norm of Bayes predictors on an ℓ^∞-ball. For instance, this assumption is easily verified when Bayes predictors are linear functions. Since covariates are assumed to be unbounded (Assumption 9), one should consider the set

$$\mathcal{K}_D := \left\{(x_{\text{obs(m)}}, m), \quad \left\|x_{\text{obs(m)}}\right\|_\infty \leq D\right\}, \tag{3}$$

for some $D > \sqrt{\gamma}$, which consists in taking the covariates with all observed components in an ℓ^∞-ball of radius D.

Under Assumption 9, an observation Z falls into the bounded set $\mathcal{K}_D$ with high probability (see Lemma A.5). One can then adapt the results in (Györfi et al., 2006; Audibert & Catoni, 2011) when X is on a bounded set to the sub-Gaussian case. To do so, consider the modified least-squares (D-LS) estimator taking into account only the observations falling into $\mathcal{K}_D$:

$$\hat{f}^{(D\text{-LS})} \in \arg\min_{f\in\mathcal{F}_b} \sum_{Z_i\in\mathcal{K}_D} (f(Z_i) - Y_i)^2 \tag{4}$$

if $\mathcal{K}_D \neq \emptyset$, and $\hat{f}^{(D\text{-LS})} = 0$ otherwise. Computing $\hat{f}^{(D\text{-LS})}$ amounts to perform one ordinary least-square procedure per missing pattern (as $\mathcal{F}_b$ is composed of functions that are linear on each missing pattern). Finally, for technical purposes, to ensure that the prediction is bounded, we consider the *clipped* estimator at level L, $T_L\hat{f} := (-L) \vee \hat{f} \wedge L$.

Theorem 3.1. *Under Assumptions 9 and 10, choosing* $D = \sqrt{\gamma}(1 + \sqrt{\gamma\log(n)})$*, and* $L = (D+1)(B+1)$ *leads to*

$$\mathbb{E}\left[\mathcal{E}\left(T_L\hat{f}^{(D\text{-LS})}\right)\right] \lesssim (\log(n)+1)(\sigma_{\text{na}}^2 \vee L^2)2^d\frac{d}{n} + A_{\mathcal{F}_b}, \tag{5}$$

where $\hat{f}^{(D\text{-LS})}$ is the estimator defined in (4), and

$$\begin{cases} \sigma_{na}^2 := \sup_{z \in \mathrm{Supp}(Z)} \mathbb{V}\left[Y|Z=z\right] \\ A_{\mathcal{F}_b} := \inf_{f \in \mathcal{F}_b} \mathbb{E}\left[\left(f(Z) - f^\star(Z)\right)^2\right]. \end{cases} \tag{6}$$

Theorem 3.1 is the first theoretical result that provides a control on the excess risk of a least-square-type predictor under very general assumptions on the input variables distribution and without any assumption on the missing pattern distribution. This result only relies on concentration and regularity arguments. Note that leaving the approximation error aside, the obtained upper bound is the multiplication of three terms. The first factor $(\log(n)+1)$ is due to (Györfi et al., 2006, Theorem 11.3) on which our result is built upon. The second factor $\sigma_{\text{na}}^2 \vee L^2$ should be seen as a tight bound for $\mathbb{E}[Y^2]$, which corresponds to the risk of the trivial predictor (predicting 0 for any value of Z). Note that the coefficient L logarithmically depends on n: the truncation of the predictor should be less stringent with an increasing number of observations. The rate of convergence is eventually dictated by the factor $2^d \frac{d}{n}$, which remains problematic as it grows exponentially with the dimension. It reflects the fact that a different regression model is required for each missing value pattern. Overall, the bound ensures that when $n > d2^d$, the least-square predictor is better than the zero one. This curse of dimensionality is unfortunately unavoidable as it naturally arises for some specific distributions (worst-case scenario of the GPMM framework).

In the framework of Proposition 2.3 (cases 1-3), Assumption 9 and 10 trivially hold and the Bayes predictor is ensured to be linear. This wipes out the approximation error in Theorem 3.1 as underlined in the following result.

Corollary 3.2. *Under Assumptions [3 and (4 or 5)] or 8, with the same choice of D and L as in Theorem 3.1 with $B = \max_{m \in \mathcal{M}} \max[|\delta_0^{(m)}|, \|\delta^{(m)}\|_1]$, we have*

$$\mathbb{E}\left[\mathcal{E}\left(T_L \hat{f}^{(D\text{-LS})}\right)\right] \lesssim (\log(n)+1)\left(\sigma_{na}^2 \vee L^2\right) 2^d \frac{d}{n},$$

where the sub-Gaussian parameter γ in L, D is

$$\gamma = \begin{cases} \max\limits_{j \in [d]} \mathbb{E}\left[X_j^2\right] & \text{(Assumption 3)}, \\ \max\limits_{\substack{m \in \mathcal{M} \\ j \in [d]}} \mathbb{E}\left[X_j^2 | M = m\right] & \text{(Assumption 8)}. \end{cases}$$

To ease the readability, we define when possible

$$a_n := (\log(n)+1)\left(\sigma_{\text{na}}^2 \vee L^2\right), \tag{7}$$

which logarithmically grows with n and depends on the distribution of (X, Y).

4. Main result: an excess risk bound adaptive to the missing pattern distribution

The error bound obtained in Theorem 3.1 holds for any missing pattern distribution. For instance, when all the 2^d missing patterns are equiprobable, the bound of Theorem 3.1 appears sharp -as one should actually perform 2^d "independent" regressions- and then suffers from the curse of dimensionality. However, this bound is pessimistic when some missing patterns are not observed or, more generally, when the missing pattern distribution is non-uniform, i.e. of low entropy. In this section, we leverage the distribution of the missing patterns in order to derive better theoretical bounds compared to Theorem 3.1. To this end, we propose a refined version of the predictor introduced in Equation (4).

4.1. Regression only on high frequency missing patterns

For any missing pattern $m \in \mathcal{M}$, we denote $E_m = \{i \in [n], M_i = m\}$ and $\mathcal{D}_n^{(m)} = ((X_{i,\text{obs(m)}}, Y_i))_{i \in E_m}$ respectively the observation indices and the sub-sample with missing pattern m. For any $m \in \mathcal{M}$, we build an estimator $\widetilde{f}_m$ of $f_m^\star$ as

$$\widetilde{f}_m \in \arg\min_{f \in \mathcal{F}_m} \sum_{i \in \overline{E}_m} (f(X_i) - Y_i)^2 \tag{8}$$

if $\overline{E}_m := \left\{i \in E_m, \left\|X_{\text{obs(m)}}\right\|_\infty \leq D\right\}$ is non-empty, and $\widetilde{f}_m = 0$ otherwise. The global predictor is then obtained by combining the previous pattern-by-pattern predictors for all patterns $m \in \mathcal{M}$ that appear with a frequency $\widehat{p}_m := \frac{|E_m|}{n}$ larger than a threshold $\tau \in [0,1]$,

$$\widehat{f}^{(\tau)}(Z) = \sum_{m \in \mathcal{M}} \widetilde{f}_m\left(X_{\text{obs(m)}}\right) \mathbb{1}_{\widehat{p}_m > \tau} \mathbb{1}_{M=m}. \tag{9}$$

Contrary to the naive estimator $\hat{f}^{(D\text{-LS})}$ defined in (4), computing $\widehat{f}^{(\tau)}$ may not require to perform up to 2^d linear regressions. Indeed, linear regressions are only computed for patterns with a frequency larger than the threshold τ. This new predictor (9) enjoys the following risk bounds.

Theorem 4.1. *Under the same assumptions as in Theorem 3.1, for any $\tau \geq 1/n$, the generalization bound for the predictor $\widehat{f}^{(\tau)}$ defined in (9), reads as*

$$\mathbb{E}\left[\mathcal{E}\left(T_L \widehat{f}^{(\tau)}\right)\right] \lesssim a_n \left(1 \vee \frac{d}{n\tau}\right) \mathfrak{C}_p(\tau) + A_{\mathcal{F}_b}. \tag{10}$$

where a_n is defined in (7), and with the missing patterns distribution complexity $\mathfrak{C}_p(\tau)$ *defined by*

$$\mathfrak{C}_p(\tau) := \sum_{m \in \mathcal{M}} p_m \wedge \tau. \tag{11}$$

The upper bound in Inequality (10) is minimal for the choice $\tau = d/n$ which leads to

$$\mathbb{E}\left[\mathcal{E}\left(T_L \widehat{f}^{(d/n)}\right)\right] \lesssim a_n \mathfrak{C}_p\left(\frac{d}{n}\right) + A_{\mathcal{F}_b}. \qquad (12)$$

Theorem 4.1 is the first result controlling the excess risk of a pattern-by-pattern least-square-type predictor with a bound depending on the missing pattern distribution through the complexity $\mathfrak{C}_p$, and holds for any type of missing patterns. Theorem 4.1 improves over Theorem 3.1, as the pattern distribution complexity $\mathfrak{C}_p$ is a lower bound of $2^d d/n$. Note that choosing $\tau = d/n$ is relevant only in the case where $d < n$ (otherwise, the proposed predictor is the zero one). The adaptivity of $\mathfrak{C}_p$ to the missing pattern distribution is illustrated in the following examples.

4.2. Examples

In this subsection, we compute the quantity $\mathfrak{C}_p\left(\frac{d}{n}\right)$, driving the bound obtained in Theorem 4.1, for different missing data settings. We focus on the case $d \leq n \leq d2^d$, i.e. when we have enough observations for statistical guarantees in standard linear regression (w/out missing values) but not enough when missing values occur (setting of Theorem 3.1.)

4.2.1. Example 1: Few frequent missing patterns

One can actually write another characterization of the complexity $\mathfrak{C}_p$, as precised in the following lemma.

Lemma 4.2. *For any distribution p on the missing patterns*

$$\mathfrak{C}_p\left(\frac{d}{n}\right) = \inf_{\mathcal{B}\subset\mathcal{M}}\left\{\mathrm{Card}(\mathcal{B})\frac{d}{n} + \mathbb{P}\left(M \in \mathcal{B}^c\right)\right\},$$

where $\mathbb{P}\left(M \in \mathcal{B}^c\right) = \sum_{m\in\mathcal{B}^c} p_m$.

The proof can be found in Appendix C.4. To illustrate this lemma, consider a subset $\mathcal{B} \subset \mathcal{M}$ of small cardinality $|\mathcal{B}|$, so that only missing patterns in $\mathcal{B}$ are very frequent and that the other missing patterns occur with a residual probability $\delta = \mathbb{P}\left(M \in \mathcal{B}^c\right)$. Lemma 4.2 entails that

$$\mathfrak{C}_p\left(\frac{d}{n}\right) \leq |\mathcal{B}|\frac{d}{n} + \delta. \qquad (13)$$

and thus by Theorem 4.1,

$$\mathbb{E}\left[\mathcal{E}\left(T_L \hat{f}^{(d/n)}\right)\right] \lesssim a_n|\mathcal{B}|\frac{d}{n} + a_n\delta + A_{\mathcal{F}_b}. \qquad (14)$$

This bound clearly improves upon Theorem 3.1, as the complexity is now controlled by $|\mathcal{B}|\frac{d}{n}$ instead of $2^d\frac{d}{n}$. This bound reflects the good learning ability of the regressor $\hat{f}^{(d/n)}$ when there are few frequent missing patterns.

Note that Lemma 4.2 applies to any missing data mechanisms. In particular, MCAR, MAR and MNAR scenarios can be exemplified through the setting developed in this section, so that the upper bound (14) is very generic. The next two examples make use of this bound in two more specific scenarios, resulting in even more informative bounds.

4.2.2. Example 2: The Bernoulli model

Assume that the distribution p of missing value patterns is $p = \mathcal{B}(\epsilon_1) \otimes \cdots \otimes \mathcal{B}(\epsilon_d)$ for $\epsilon_j \in [0,1]$ with $j \in [d]$, so that components $(M_j)_j$ are independent and of distribution $M_j \sim \mathcal{B}(\epsilon_j)$. The model is said homogeneous when $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_d = \epsilon \in [0,1]$, and heterogeneous otherwise. Note that in such a setting, the missing mechanisms can be still of MCAR, MAR or MNAR nature.

Consider a homogeneous Bernoulli model with $\epsilon < 1/2$. Consequently, the most frequent patterns are those with the least missing values. For a given $s \in [d]$, define $\mathcal{B}_s$ the set of missing patterns with less than s missing values. Therefore, Equation (13) reads as

$$\mathfrak{C}_p\left(\frac{d}{n}\right) \leq |\mathcal{B}_s|\frac{d}{n} + \delta_s, \qquad (15)$$

where $\delta_s = \mathbb{P}\left(M \in \mathcal{B}_s^c\right)$ is the probability of having a pattern with more than s missing values. Controlling each of these terms gives the following lemma.

Lemma 4.3. *Under a homogeneous Bernoulli model with proportion ϵ of missing data, one has*

$$\mathfrak{C}_p\left(\frac{d}{n}\right) \leq \inf_{s\in[d]}\left(\frac{d}{n} + \epsilon^s\right)\left(\frac{ed}{s}\right)^s.$$

One can then obtain a version of Theorem 4.1 in the case of a Bernoulli model, by optimizing s in Lemma 4.3.

Proposition 4.4. *Under the assumptions of Theorem 3.1,*

$$\mathbb{E}\left[\mathcal{E}\left(T_L \hat{f}^{(d/n)}\right)\right] \lesssim a_n\left(\frac{ed}{s_\epsilon\left(d/n\right)}\right)^{s_\epsilon(d/n)}\frac{d}{n} + A_{\mathcal{F}_b},$$

with $s_\epsilon\left(d/n\right) := 1 \vee \left\lfloor \frac{\log\left(\frac{n}{d}\right)}{\log(\epsilon^{-1})} \right\rfloor \wedge d$.

Here, $s_\epsilon\left(d/n\right)$, being in $[d]$, can be interpreted as a *hidden dimension* (relative to the missing pattern distribution). Indeed, the initial complexity scaling as 2^d in Theorem 3.1 is replaced by $(\frac{ed}{s_\epsilon(d/n)})^{s_\epsilon(d/n)}$ in Proposition 4.4 for this Bernoulli model.

Observe that the bound improves as ϵ decreases, for example for $\epsilon \leq \frac{d}{n}$, the excess risk bound scales as $\frac{d^2}{n}$. This again highlights the benefit of adaptivity in Theorem 4.1, which allows us to obtain a bound that improves when the fraction of missing data decreases below a certain level.

We extend the result above to the *heterogeneous* case in Appendix C.2.3, and provide a discussion on the comparison

between the complexities for homogeneous and heterogeneous Bernoulli models that share *the same* overall fraction of missing data ϵ in Appendix C.2.1.

4.2.3. Example 3: Database Merge Model

Consider a context of multi-sources data, where for instance a medical register results from merging d-dimensional data coming from h different hospitals:

1. each hospital $k \in \{1, \dots, h\}$ has its own measurement protocol, resulting in the missing pattern P_k ($P_{k,j} = 1$ if measure $j \in \{1, \dots, d\}$, is not performed in hospital k). Note that this missing pattern is shared by all the patients in care in hospital k.
2. in addition, for each measure $j \in \{1, \dots, d\}$, the measuring device may make a protocol-independent error, that produces a missing value with probability η.

For an entry of the merged medical register, call P (taking values in $P_1, \dots, P_h$) the missing pattern coding for the protocol effective in the hospital where this information has been collected, and $N \in \{0,1\}^d$ the missing pattern coding for the measurement failure. Therefore, the eventual missing value pattern can be decomposed as,

$$(1 - M) = (1 - P_H) \odot (1 - N), \tag{16}$$

where $\odot$ is the Hadamard product.

This model is compatible with MNAR missing data mechanisms. Indeed, the missing pattern may be informative about the missing data values, as it encloses information about the hospital where the data is collected, and thereby may depend on a certain type of population distribution (geographical location, level of wealth...) frequenting the above hospital. Theorem 4.1 can be adapted in such a setting as follows.

Proposition 4.5. *Under Assumptions of Theorem 3.1,*

$$\mathbb{E}\left[\mathcal{E}\left(T_L \hat{f}^{(d/n)}\right)\right] \lesssim a_n \left(\frac{ed}{s_\eta\left(d/n\right)}\right)^{s_\eta(d/n)} h\frac{d}{n} + A_{\mathcal{F}_b},$$

where s_η is defined in Proposition 4.4.

See Appendix C.3 for the proof. The excess risk bound in Proposition 4.4 encompasses a term similar to that of the Bernoulli case involving only the measurement failure probability η here, whereas the number of protocols h linearly intervenes. To understand why this could be an advantage, consider two hospitals ($h = 2$) in which only 50% of the variables are systematically measured, and assume that the probability of measurement failure η equals 0.01. The overall proportion ϵ of missing values in the merged dataset is therefore high, i.e. $\epsilon = 1 - 0.99/2 \simeq 0.5$. Altogether, the bound in Proposition 4.5 (controlled via s_η) improves upon the one of Proposition 4.4 (controlled via s_ϵ) by a factor $\epsilon/(h\eta) = 25$. This means that the bound in Proposition 4.4 does not suffer from the resulting proportion ϵ of missing values, and mostly depends on the probability η of measurement failure. This outlines the great plasticity of the complexity $\mathfrak{C}_p$ even in regimes with a large proportion of missing values, by leveraging the missing value structure.

4.3. Minimax aspects

In this section we discuss the optimality of the risk bound obtained for $\hat{f}^{(\tau)}$. To this end, we consider the class below.

Definition 4.6. The class of problems $\mathcal{P}_p(\sigma, R)$ is assumed to satisfy the following conditions: for all $\mathbb{P} \in \mathcal{P}_p(\sigma, R)$

1. $\forall m \in \mathcal{M}, \mathbb{P}\left(M = m\right) = p_m$,
2. $Y = \langle \beta, X \rangle + \epsilon$ where $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$,
3. Assumptions 9 and 10 hold with $B^2(\gamma + 1) \leq 3R^2$,
4. $A_{\mathcal{F}_b} = 0$.

Note that this class of problems includes the Gaussian case (Assumption 3) with M(C)AR, or GPMM (Assumption 8). For this large class of problems, the excess risk can be upper bounded by Theorem 4.1 at the rate $a_n \mathfrak{C}_p(d/n)$, with $a_n \leq (\log(n) + 1)^2(R^2 \vee \sigma_{\text{na}}^2)$. The following result provides a *lower bound* on the excess risk of the best possible predictor in the worst-case scenario of class $\mathcal{P}_p(\sigma, R)$: it exhibits the same dependency on the complexity $\mathfrak{C}_p$.

Theorem 4.7. *Consider a distribution p on $\mathcal{M}$, then $R, \sigma > 0$ and c be such that $16e^{-\frac{1}{4}\left(\frac{R}{d\sigma}\right)^2} \leq c$. Therefore,*

$$(1 - c)\sigma^2 \mathfrak{C}_p\left(\frac{1}{n}\right) \lesssim \min_{\hat{f}} \max_{\mathbb{P} \in \mathcal{P}_p(\sigma, R)} \mathbb{E}_{\mathbb{P}}\left[\mathcal{E}\left(\hat{f}\right)\right].$$

where the minimum is over all predictor $\hat{f}$.

This result highlights the relevancy of the complexity $\mathfrak{C}_p$ in the control of the excess risk. Since $\mathfrak{C}_p\left(\frac{1}{n}\right) \geq d^{-1}\mathfrak{C}_p\left(\frac{d}{n}\right)$, the lower bound in Theorem 4.7 is sharp up to a factor d. Note that if the distribution of missing patterns is uniform, one gets $\mathfrak{C}_p\left(\frac{1}{n}\right) = \frac{2^d}{n}$, meaning that the upper-bound of Theorem 4.1 cannot be improved in full generality. Restricting the considered class to the MAR ones does not impact the lower bound, as outlined in what follows.

Corollary 4.8. *Assume that one component of X is always observed. Then $\mathcal{P}_p(\sigma, R) \cap \mathcal{P}_{MAR}$ is non empty and*

$$(1 - c)\sigma^2 \mathfrak{C}_p\left(\frac{1}{n}\right) \lesssim \min_{\hat{f}} \max_{\mathbb{P} \in \mathcal{P}_p(\sigma, R) \cap \mathcal{P}_{\text{MAR}}} \mathbb{E}_{\mathbb{P}}\left[\mathcal{E}\left(\hat{f}\right)\right].$$

This lower bound is of the same order as that of the lower bound in Theorem 4.7, meaning that the worst-case scenario of $\mathcal{P}_p(\sigma, R)$ is as hard as the one of $\mathcal{P}_p(\sigma, R)$ restricted to MAR settings. While the MAR hypothesis facilitates the inference framework (the former actually originates from the latter, see Rubin 1976), Corollary 4.8 emphasizes that MAR scenarios do not help prediction purposes.

5. Numerical experiments

In this section, we numerically evaluate the performance of several regressors on varying missing data scenarios.

Regressors. More specifically, we compare the following five regression methods. First, we consider two baselines consisting in imputation followed by standard linear regression (on the completed data): for **Cst-imp+LR** we learn optimal imputation constants for each variable (note that this is equivalent to performing a LR of Y on $\left(X_{\text{obs(m)}}, M\right)$, see (Le Morvan et al., 2020b, Proposition 3.1)); for **MICE+LR**, the imputation is performed by the scikit-learn `IterativeImputer` which relies on MICE (Van Buuren & Groothuis-Oudshoorn, 2011). Moreover, we add two pattern-by-pattern methods, that learn one regression model per pattern as defined in Equation (9): for all patterns having at least one observation in **P-by-P imp** (i.e., $\tau = n^{-1}$ which matches the regressor in (4)), and with $\tau = d/n$ for **Thresholded P-by-P imp**. For both, the technical ℓ^∞-ball condition is not considered in numerical experiments. Finally **NeuMiss** (Le Morvan et al., 2020a) is a neural network which architecture is specifically designed to handle missing data in linear regression.

Data generation settings. We consider three different settings in dimension $d = 8$ with increasing difficulty: (a) **MCAR Bernoulli** in which X and M are independent, M is generated according to the homogeneous Bernoulli Model of Section 4.2.2 with missing value proportion $\epsilon = 10\%$ and $X \sim \mathcal{N}(\mu, \Sigma)$ where $\mu \neq 0$ and $\Sigma \neq I$; (b) **MAR** in which X is separated into two blocks of components $X^{(1)}, X^{(2)}$ each of size 4, $X^{(1)}$ is a Gaussian isotropic vector that is always observed and the missing pattern associated to $X^{(2)}$ is $M^{(2)} = \mathbb{1}_{X^{(1)}>0}$, and $X^{(2)}|X^{(1)} \sim \mathcal{N}(M^{(2)}, \Sigma)$ where $\Sigma \neq I$. (c) **MNAR-GPMM** in which (X, M) is distributed according to Assumption 8 with 7 non-null probability missing patterns. See Appendix E.1 for details.

Results. The results are presented in Figure 3. First, the P-by-P methods (with and without threshold) and NeuMiss are the only ones that are Bayes consistent regardless of the scenario (the excess risk tends to 0 on Figures 3(a,b,c)). This was indeed expected for the P-and-P methods as pointed out in Theorem 3.1. Neumiss provides similar performances at least in the MCAR and MAR settings, but its computational complexity, even in dimension $d = 8$, prevents from reaching large sample sizes (see Appendix E.2). All the previous methods clearly outperform the MICE+LR strategy as soon as the data are not MCAR anymore, by exploiting the information contained in the missing pattern. Note that the Cst-imp method poorly performs whatever the data setting is: this could be explained by the fact that the model includes $2d$ parameters, which is not sufficient to learn the correlations between the variables (which would require d^2 parameters at least). Secondly, we remark the benefit of thresholding in P-by-P methods (see Theorem 4.1): **Thresholded P-by-P** outperforms the unthresholded version in particular for a small number of samples. Thresholding thus acts as a regularizer, by avoiding overfitting on the least frequent missing patterns.

The behavior of P-by-P methods is discussed in higher dimensional regimes and on a real dataset in Appendix E.3.

6. Conclusion

In this paper, we propose a wide panel of data settings to study linear models with missing data. Contrary to most previous works, we focus on the prediction problem by evaluating the quadratic risk of linear models. We propose a new thresholded predictor coming with strong theoretical guarantees: the upper bound on its excess risk holds under very mild assumptions on the data, while integrating the complexity $\mathfrak{C}_p$ of the missing pattern distribution. This quantity is interesting on its own as it describes the influence of the missing data distribution on the predictive performances. We establish a lower bound on the excess risk of the best possible predictor on a class of data distributions previously considered for the upper bound analysis: the lower bound involves the same complexity $\mathfrak{C}_p$. This, supported with several examples, highlights the sharpness of our results. Numerical experiments emphasizes the improvement of our pattern-by-pattern estimator compared to state-of-the-art algorithms.

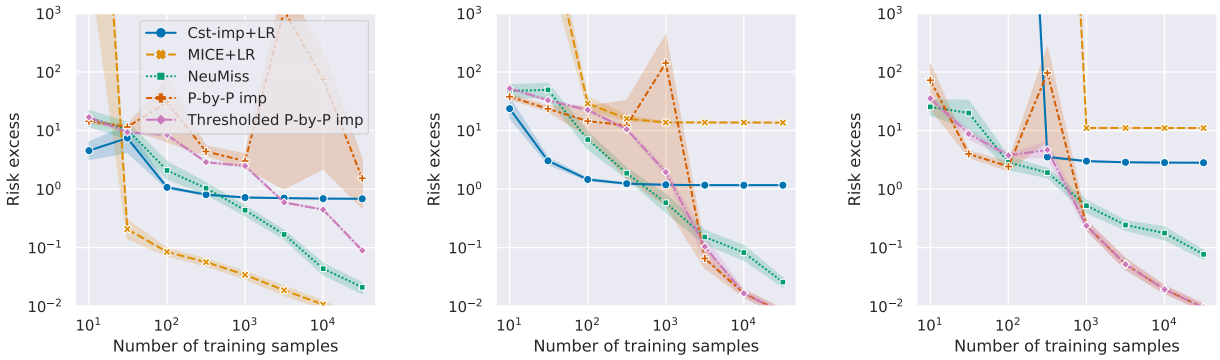
Training thresholded pattern-by-pattern predictors is a way to regularize the learning process highly complex when missing data occur. Other types of regularization should be investigated to break the induced curse of dimensionality. However, the lower bound on the minimax predictor suggest that current assumptions are not strong enough to obtain better guarantees. In the formalism of prediction with missing values, finding suitable assumptions on the missing patterns still remains an open question.

References

Agarwal, A., Shah, D., Shen, D., and Song, D. On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Audibert, J.-Y. and Catoni, O. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.



(a) MCAR (b) MAR (c) MNAR-GPMM

Figure 3. Excess risk w.r.t. the number of training samples. The curve represents the averaged excess risk over 100 repetitions within a 95% confidence interval.

Chen, Y. and Caramanis, C. Noisy and missing data regression: Distribution-oblivious support recovery. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 383–391, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/chen13d.html.

Dalalyan, A. S. and Thompson, P. Outlier-robust estimation of a sparse linear model using ell-1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems 32*, pp. 13188–13198, 2019. URL http://arxiv.org/pdf/1904.06288.

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Dieuleveut, A., Flammarion, N., and Bach, F. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Heckman, J. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *NBER Book Chapters*, 5, 02 2012.

Jones, M. P. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91:222–230, 1996.

Josse, J., Prost, N., Scornet, E., and Varoquaux, G. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

Le Morvan, M., Josse, J., Moreau, T., Scornet, E., and Varoquaux, G. NeuMiss networks: differentiable programming for supervised learning with missing values. In *NeurIPS 2020 - 34th Conference on Neural Information Processing Systems*, Vancouver / Virtual, Canada, December 2020a. URL https://hal.archives-ouvertes.fr/hal-02888867.

Le Morvan, M., Prost, N., Josse, J., Scornet, E., and Varoquaux, G. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pp. 3165–3174. PMLR, 2020b.

Little, R. J. Regression with missing x's: a review. *Journal of the American statistical association*, 87(420):1227–1237, 1992.

Little, R. J. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

Loh, P.-L. and Wainwright, M. J. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3): 1637 – 1664, 2012. doi: 10.1214/12-AOS1018. URL https://doi.org/10.1214/12-AOS1018.

Massart, P. *Concentration inequalities and model selection*. Springer, 2007.

Mourtada, J. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv preprint arXiv:1912.10754*, 2019.

Pelckmans, K., De Brabanter, J., Suykens, J. A., and De Moor, B. Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5-6):684–692, 2005.

Petersen, K. B., Pedersen, M. S., et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Rényi, A. et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Rosenbaum, M. and Tsybakov, A. B. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620 – 2651, 2010. doi: 10.1214/10-AOS793. URL https://doi.org/10.1214/10-AOS793.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Sportisse, A., Boyer, C., Dieuleveut, A., and Josses, J. Debiasing averaged stochastic gradient descent to handle missing values. *Advances in Neural Information Processing Systems*, 33, 2020.

Van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

A. Proofs for Section 2

A.1. Key lemma on Binomial law

Let's begin by a useful lemma on binomial law, which can be found in (Devroye et al., 2013, Lemma A2 p 587).

Lemma A.1. *Let $B \sim \mathcal{B}(p, n)$, we have*

$$\frac{1}{1+np} \leq \mathbb{E}\left[\frac{1}{1+B}\right] \leq \frac{1}{p(n+1)}, \tag{17}$$

and

$$\mathbb{E}\left[\frac{\mathbb{1}\{B>0\}}{B}\right] \leq \frac{2}{p(n+1)}. \tag{18}$$

Proof.

- Lower bound of (17): we use Jensen inequality

$$\frac{1}{1+np} = \frac{1}{1+\mathbb{E}B} \leq \mathbb{E}\left[\frac{1}{1+B}\right].$$

- Upper bound of (17),

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1+B}\right] &= \sum_{i=0}^{n} \binom{n}{i} \frac{1}{1+i} p^i (1-p)^{n-i} \\
&= \sum_{i=0}^{n} \frac{n!}{i!(n-i)!(1+i)} p^i (1-p)^{n-i} \\
&= \frac{1}{(n+1)\,p} \sum_{i=0}^{n} \frac{(n+1)!}{(i+1)!(n+1-i-1)!} p^{i+1} (1-p)^{n-i} \\
&= \frac{1}{(n+1)\,p} \sum_{i=0}^{n} \binom{n+1}{i+1} p^{i+1} (1-p)^{n+1-i-1} \\
&\leq \frac{1}{(n+1)\,p},
\end{aligned}$$

  using binomial formula.

- For (18), we use that $1/x \leq 2/(x+1)$ on $x \geq 1$ and previous result.

□

A.2. A key intermediate result on regression

First, let us mention a very useful theorem for analyzing the quadratic risk in the regression framework.

Theorem A.2 (Theorem 11.3 in (Györfi et al., 2006))**.** *Let two random variables X and Y be such that*

$$Y = f^\star(X) + \epsilon$$

and

$$\begin{cases} \|f^\star\|_\infty = \sup_{x \in \mathrm{Supp(X)}} |f^\star(x)| \leq L \\ \sigma^2 = \sup_{x \in \mathrm{Supp(X)}} \mathbb{V}\left[Y|X=x\right] < \infty \end{cases}$$

for some $L > 0$. Let $\mathcal{F}$ be a linear vector space of function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$. Define the estimate by $T_L f_n(x) = (-L) \vee f_n(x) \wedge L$ where

$$f_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2. \tag{19}$$

Then

$$\mathbb{E}\left[\left(T_L f_n(X) - f^\star(X)\right)^2\right] \leq c \max\left\{\sigma^2, L^2\right\} \frac{dim(\mathcal{F})(1+\log(n))}{n} + 8 \inf_{f\in\mathcal{F}} \mathbb{E}\left[\left(f(X) - f^\star(X)\right)^2\right], \tag{20}$$

for some universal constant c.

The main drawback of this theorem is that it is only useful if the support of X is bounded. Assumption 8 requires the covariates to be unbounded as they are assumed to be Gaussian. However, the covariates are on a bounded set with a high probability. The following corollary is adapted to this case.

Corollary A.3 (Unbounded case)**.** *Let two random variable X, Y and a subset $\mathcal{K} \subset \mathbb{R}^d$ be such that*

$$\begin{cases} \|f^\star\|_{\infty,\mathcal{K}} = \sup_{x\in\mathcal{K}} |f^\star(x)| \leq L_{\mathcal{K}} \\ \sigma^2 = \sup_{x\in \mathrm{Supp(X)}} \mathbb{V}\left[Y|X=x\right] < \infty \end{cases}$$

for some $L_{\mathcal{K}} > 0$. Let $\mathcal{F}$ be a linear vector space of function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$. Define

$$f_{\mathcal{K}}(x) := T_L f_{n,\mathcal{K}}(x) \mathbb{1}_{x\in\mathcal{K}}, \tag{21}$$

where

$$\begin{cases} f_{n,\mathcal{K}} \in \arg\min_{f\in\mathcal{F}} \sum_{X_i\in\mathcal{K}} (f(X_i) - Y_i)^2 & \text{if } \exists i \in [n], X_i \in \mathcal{K}, \\ f_{n,\mathcal{K}} = 0 & \text{else.} \end{cases} \tag{22}$$

Then

$$\begin{aligned} \mathbb{E}\left[\left(f_{\mathcal{K}}(X) - f^\star(X)\right)^2\right] \leq c \max\left\{\sigma^2, L_{\mathcal{K}}^2\right\} \frac{dim(\mathcal{F})(1+\log(n))}{n} \\ +8 \inf_{f\in\mathcal{F}} \mathbb{E}\left[\mathbb{1}_{X\in\mathcal{K}} \left(f(X) - f^\star(X)\right)^2\right] + R_{\mathcal{K}}, \end{aligned} \tag{23}$$

where $p_{\mathcal{K}} = \mathbb{P}\left(X \in \mathcal{K}\right)$ and $R_{\mathcal{K}} = (1-p_{\mathcal{K}})^n \mathbb{E}\left[\mathbb{1}_{X\in\mathcal{K}} f^\star(X)^2\right] + \mathbb{E}\left[\mathbb{1}_{X\notin\mathcal{K}} f^\star(X)^2\right]$.

Compared to Theorem A.2, the bound obtained in the previous corollary includes an additional term $R_{\mathcal{K}}$. The extension of Theorem A.2 to the unbounded covariates case as done in Corollary A.3 will be therefore informative only if this new term $R_{\mathcal{K}}$ remains of small order compared to the other ones. In the next corollary, we will apply it for sub Gaussian covariates.

In particular, under assumption $\mathbb{E}\left[f^*\left(Z\right)^4\right] < +\infty$, we can use Cauchy-Schwarz inequality to obtain

$$R_{\mathcal{K}} \leq 2(1-p_K)^{1/2} \mathbb{E}\left[f^*\left(Z\right)^4\right]^{1/2}. \tag{24}$$

Proof of Corollary A.3 The main idea of the proof is to consider the subsample of observations that are in $\mathcal{K}$:

$$E_{\mathcal{K}} = \left\{i \in [n], X_i \in \mathcal{K}\right\}. \tag{25}$$

Step 1: Law on subsample

Let's start with a useful lemma to describe the elements of subsample induced by $E_{\mathcal{K}}$:

Lemma A.4. *Let $S = (X_i)_{i\in\mathbb{N}}$ be a sequence of independent variables with same distribution and $i_S = \inf\{i, X_i \in \mathcal{K}\}$. We suppose that $p_{\mathcal{K}} > 0$, then $i_S < \infty$ almost surely and X_{i_S} has the same distribution as $X|(X \in \mathcal{K})$.*

Proof. $\mathbb{P}\left(i_S > k\right) = p_K^k$ thus $\sum \mathbb{P}\left(i_S > k\right)$ is convergent. Borel-Cantelli lemma shows that $i_S < \infty$ almost surely.

Consider a bounded function ϕ:

$$\begin{aligned}
\mathbb{E}\left[\phi(X_{i_S})\right] &= \sum_{k=1}^{\infty} \mathbb{P}\left(i_S = k\right) \mathbb{E}\left[\phi(X_i) | i_S = k\right] \\
&= \sum_{k=1}^{\infty} \mathbb{P}\left(i_S = k\right) \mathbb{E}\left[\phi(X_k) | X_k \in \mathcal{K}; X_1, ..., X_{k-1} \notin \mathcal{K}\right] \\
&= \sum_{k=1}^{\infty} \mathbb{P}\left(i_S = k\right) \mathbb{E}\left[\phi(X_k) | X_k \in \mathcal{K}\right] && \text{because } X_k \text{ does not depend on } X_j, j < k \\
&= \mathbb{E}\left[\phi(X_1) | X_1 \in \mathcal{K}\right].
\end{aligned}$$

This concludes the lemma. □

Thanks to Lemma A.4, we can show $(X_i, Y_i) \sim \text{Law}\left((X, Y) | (X \in \mathcal{K})\right)$ for all $i \in E_{\mathcal{K}}$.
Let $(\widetilde{X}, \widetilde{Y}) \sim \text{Law}\left((X, Y) | X \in \mathcal{K}\right)$. Using the same notations of lemma, we can write the Bayes predictor for the regression problem involving the "conditional" data $(\widetilde{Y}, \widetilde{X})$: For all $x \in \mathcal{K}$,

$$\begin{aligned}
\widetilde{f}^{\star}(x) &= \mathbb{E}\left[\widetilde{Y} | \widetilde{X} = x\right] \\
&= \mathbb{E}\left[Y_{i_S} | X_{i_S} = x\right] \\
&= \sum_{k=1}^{\infty} \mathbb{P}\left(i_S = k\right) \mathbb{E}\left[Y_k | X_k = x \in \mathcal{K}; X_1, ..., X_{i-1} \notin \mathcal{K}\right] \\
&= \sum_{k=1}^{\infty} \mathbb{P}\left(i_S = k\right) \mathbb{E}\left[Y_k | X_k = x \in \mathcal{K}\right] && \text{because } X_k \text{ does not depend on } X_j, j < k \\
&= \mathbb{E}\left[Y | X = x\right] \\
&= f^{\star}(x).
\end{aligned}$$

Thus,

$$\forall x \in \text{Supp}\left(\text{X}\right), \quad \widetilde{\text{f}}^{\star}(\text{x}) = \mathbb{1}_{\text{x} \in \mathcal{K}} \text{f}^{\star}(\text{x}). \tag{26}$$

Step 2: Decomposition of excess risk
We can decompose:

$$\mathbb{E}\left[\left(f_{\mathcal{K}}(X) - f^{\star}(X)\right)^2\right] \leq \mathbb{E}\left[\mathbb{1}_{X \in \mathcal{K}} \left(f_{\mathcal{K}}(X) - f^{\star}(X)\right)^2\right] + \mathbb{E}\left[\mathbb{1}_{X \notin \mathcal{K}} \left(f_{\mathcal{K}}(X) - f^{\star}(X)\right)^2\right] \tag{27}$$

$$\leq p_{\mathcal{K}} \mathbb{E}\left[\left(f_{\mathcal{K}}(X) - f^{\star}(X)\right)^2 | X \in \mathcal{K}\right] + \mathbb{E}\left[\mathbb{1}_{X \notin \mathcal{K}} f^{\star}(X)^2\right], \tag{28}$$

using definition for the second term. We will bound the first term using Theorem A.2 by conditioning according to $E_{\mathcal{K}}$.

$$\mathbb{E}\left[\left(f_{\mathcal{K}}(X) - f^{\star}(X)\right)^2 | X \in \mathcal{K}\right] = \mathbb{E}\left[\left(f_{\mathcal{K}}(\widetilde{X}) - \widetilde{f}^{\star}(\widetilde{X})\right)^2\right] \qquad \text{by definition} \tag{29}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(f_{\mathcal{K}}(\widetilde{X}) - \widetilde{f}^{\star}(\widetilde{X})\right)^2 | E_{\mathcal{K}}\right]\right]. \tag{30}$$

Let E be a subset of $[n]$:

- If E is empty,

$$\mathbb{E}\left[\left(f_{\mathcal{K}}(\widetilde{X}) - \widetilde{f}^{\star}(\widetilde{X})\right)^2 | E_{\mathcal{K}} = E\right] = \mathbb{E}\left[\widetilde{f}^{\star}(\widetilde{X})^2\right] = \mathbb{E}\left[f^{\star}(\widetilde{X})^2\right], \tag{31}$$

  using that $f_{\mathcal{K}} = 0$ if $E_{\mathcal{K}}$ is empty.

- If E is non-empty, $f_{\mathcal{K}}$ is the clipped OLS estimator for the problem $(\widetilde{Y}, \widetilde{X})$ and the dataset $(Y_i, X_i)_{i\in E}$, thus

$$\mathbb{E}\left[\left(f_{\mathcal{K}}(\widetilde{X}) - f^\star(\widetilde{X})\right)^2 |E_{\mathcal{K}} = E\right] \leq c\max\left\{\sigma^2, L_{\mathcal{K}}^2\right\} \frac{\dim(\mathcal{F})(1+\log(|E|))}{|E|} + 8 \inf_{f\in\mathcal{F}} \mathbb{E}\left[\left(f(\widetilde{X}) - f^\star(\widetilde{X})\right)^2\right], \tag{32}$$

  using Theorem A.2.

Therefore,

$$\mathbb{E}\left[\left(f_{\mathcal{K}}(\widetilde{X}) - f^\star(\widetilde{X})\right)^2 |E_{\mathcal{K}}\right] \leq c\max\left\{\sigma^2, L_{\mathcal{K}}^2\right\} \frac{\dim(\mathcal{F})(1+\log(n))}{|E_{\mathcal{K}}|} \mathbb{1}_{|E_{\mathcal{K}}|>0} + \mathbb{1}_{|E_{\mathcal{K}}|=0}\mathbb{E}\left[f^\star(\widetilde{X})^2\right] + 8 \inf_{f\in\mathcal{F}} \mathbb{E}\left[\left(f(\widetilde{X}) - f^\star(\widetilde{X})\right)^2\right], \tag{33}$$

using $\log(|E_{\mathcal{K}}|) \leq \log(n)$. Moreover $|E_{\mathcal{K}}|$ is a binomial distribution with parameters $p_{\mathcal{K}}$ and n, thus

$$\mathbb{E}\left[\left(f_{\mathcal{K}}(\widetilde{X}) - f^\star(\widetilde{X})\right)^2\right] \leq c\max\left\{\sigma^2, L_{\mathcal{K}}^2\right\} \frac{2\dim(\mathcal{F})(1+\log(n))}{(n+1)p_{\mathcal{K}}} + (1-p_{\mathcal{K}})^n\mathbb{E}\left[f^\star(\widetilde{X})^2\right] + 8 \inf_{f\in\mathcal{F}} \mathbb{E}\left[\left(f(\widetilde{X}) - f^\star(\widetilde{X})\right)^2\right], \tag{34}$$

using the expectation of the inverse of a binomial distribution. By combining (34) and (28), and using

$$\begin{aligned}\mathbb{E}\left[\left(f(\widetilde{X}) - f^\star(\widetilde{X})\right)^2\right] &= \mathbb{E}\left[\left(f(X) - f^\star(X)\right)^2 |X \in \mathcal{K}\right] \\ &= p_{\mathcal{K}}^{-1}\mathbb{E}\left[\mathbb{1}_{X\in\mathcal{K}}\left(f(X) - f^\star(X)\right)^2\right],\end{aligned}$$

we find

$$\mathbb{E}\left[\left(f_{\mathcal{K}}(X) - f^\star(X)\right)^2\right] \leq c\max\left\{\sigma^2, L_{\mathcal{K}}^2\right\} \frac{\dim(\mathcal{F})(1+\log(n))}{n} + 8 \inf_{f\in\mathcal{F}} \mathbb{E}\left[\mathbb{1}_{X\in\mathcal{K}}\left(f(X) - f^\star(X)\right)^2\right] + R_{\mathcal{K}}, \tag{35}$$

where $p_{\mathcal{K}} = \mathbb{P}\left(X \in \mathcal{K}\right)$ and $R_{\mathcal{K}} = (1-p_{\mathcal{K}})^n\mathbb{E}\left[\mathbb{1}_{X\in\mathcal{K}} f^\star(X)^2\right] + \mathbb{E}\left[\mathbb{1}_{X\notin\mathcal{K}} f^\star(X)^2\right]$.

Lemma A.5. *Under Assumption 9, we have for all $D > \sqrt{\gamma}$,*

$$\forall i \in [p], \qquad \forall D > \sqrt{\gamma}, \qquad \mathbb{P}\left[|X_i| > D\right] \leq 2\exp\left[-\frac{\left(D-\sqrt{\gamma}\right)^2}{2\gamma}\right], \tag{36}$$

and

$$\mathbb{P}\left(Z \notin \mathcal{K}_D\right) \leq 2d\exp\left(-\frac{\left(D-\sqrt{\gamma}\right)^2}{2\gamma}\right). \tag{37}$$

Proof. Let's fix $m \in \mathcal{M}$ and $i \in [p]$, by Assumption 9: By bounding the tail of a Sub-Gaussian distribution,

$$\mathbb{P}\left(|X_i - \mathbb{E}X_i| > D\right) \leq 2\exp\left(-\frac{D^2}{2\gamma}\right). \tag{38}$$

Since $|X_i| + \sqrt{\gamma} \geq |X_i| + |\mathbb{E}X_i| \geq |X_i - \mathbb{E}X_i|$, we have

$$\forall i \in [p], \forall D > \sqrt{\gamma}, \qquad \mathbb{P}\left(|X_i| > D\right) \leq 2\exp\left(-\frac{\left(D-\sqrt{\gamma}\right)^2}{2\gamma}\right).$$

For the second point, we make a union bound and we use (36):

$$\begin{aligned}\mathbb{P}\left(Z \notin \mathcal{K}_D\right) =& \mathbb{P}\left(\left\|X_{\text{obs(m)}}\right\|_\infty > D\right) \\ &\leq \mathbb{P}\left(\|X\|_\infty > D\right) \\ &\leq \sum_{i=1}^{d} \mathbb{P}\left(|X_i| > D\right) \\ &\leq 2d \exp\left(-\frac{\left(D-\sqrt{\gamma}\right)^2}{2\gamma}\right).\end{aligned}$$

□

A.3. Proof of Theorem 3.1

The proof consists in verifying the assumptions of the corollary A.3 and to bound the additional term $R_{\mathcal{K}_D}$.

- $\sup_{Z\in\mathcal{K}_D} |f^\star(Z)|$ upper bound:

$$\begin{aligned}\sup_{Z\in\mathcal{K}_D} |f^\star(Z)| &= \sup_{m\in\mathcal{M}} \sup_{\left\|x_{\text{obs(m)}}\right\|_\infty \leq D} \left|f_m^\star(x_{\text{obs(m)}})\right| \\ &= \sup_{m\in\mathcal{M}} \sup_{\left\|x_{\text{obs(m)}}\right\|_\infty \leq D} \left|f_m^\star(x_{\text{obs(m)}}) - f_m^\star(0)\right| + |f_m^\star(0)| \\ &= \sup_{m\in\mathcal{M}} \sup_{\left\|x_{\text{obs(m)}}\right\|_\infty \leq D} B\left\|x_{\text{obs(m)}}\right\|_\infty + B,\end{aligned}$$

  using Assumption 10, thus

$$\sup_{Z\in\mathcal{K}_D} |f^\star(Z)| \leq (D+1)B \leq L. \tag{39}$$

- $R_{\mathcal{K}_D}$ upper bound: From (24)

$$\begin{aligned}R_{\mathcal{K}} &\leq 2(1-p_{\mathcal{K}_D})^{1/2}\mathbb{E}\left[f^*\left(Z\right)^4\right]^{1/2} \\ &\leq 2\sqrt{\mathbb{P}\left(Z \notin \mathcal{K}_D\right)\mathbb{E}\left[f^*\left(Z\right)^4\right]} \\ &\leq 2\sqrt{2d}\exp\left(-\frac{\left(D-\sqrt{\gamma}\right)^2}{4\gamma}\right)\sqrt{\mathbb{E}\left[f^*\left(Z\right)^4\right]}.\end{aligned} \tag{40}$$

  It remains to bound

$$\begin{aligned}\mathbb{E}\left[f^*\left(Z\right)^4\right] &= \sum_{m\in\mathcal{M}} \mathbb{P}\left(M=m\right)\mathbb{E}\left[f_m^*\left(X_{\text{obs(m)}}\right)^4 | M=m\right] \\ &= \sum_{m\in\mathcal{M}} \mathbb{P}\left(M=m\right)\mathbb{E}\left[\left|\left|f_m^*\left(X_{\text{obs(m)}}\right) - f_m^*\left(0\right)\right| + \left|f_m^*\left(0\right)\right|\right|^4 | M=m\right] \\ &\leq 8\sum_{m\in\mathcal{M}} \mathbb{P}\left(M=m\right)\mathbb{E}\left[B^4\left\|X_{\text{obs(m)}}\right\|_2^4 + B^4 | M=m\right] \qquad \text{by Assumption 10} \\ &\leq 8B^4\sum_{m\in\mathcal{M}} \mathbb{P}\left(M=m\right)\mathbb{E}\left[\|X\|_\infty^4 | M=m\right] + B^4. \\ &\leq 8B^4\left(\mathbb{E}\left[\|X\|_\infty^4\right] + 1\right).\end{aligned} \tag{41}$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\|X\|_\infty^4\right] &\leq \mathbb{E}\left[\left(\|X-\mathbb{E}X\|_\infty + \|\mathbb{E}X\|_\infty\right)^4\right] && \text{by triangular inequality}\\
&\leq 8\mathbb{E}\left[\|X-\mathbb{E}X\|_\infty^4 + \|\mathbb{E}X\|_\infty^4\right]\\
&\leq 8\mathbb{E}\left[\|X-\mathbb{E}X\|_\infty^4\right] + 8\gamma^2 && \text{by Assumption 9}\\
&\leq 8\mathbb{E}\left[\max_{j\in[d]}|X_j-\mathbb{E}X_j|^4\right] + 8\gamma^2\\
&\leq 8\sum_{j=1}^{d}\mathbb{E}\left[|X_j-\mathbb{E}X_j|^4\right] + 8\gamma^2\\
&\leq 8d\left(2!\,(4\gamma)^2\right) + 8\gamma^2 && (42)\\
&\leq 257 d\gamma^2. && (43)
\end{aligned}$$

We have used moment's characterization (Boucheron et al., 2013, Theorem 2.2) in (42). Using (40), (41) and (43), we have

$$R_{\mathcal{K}_D} \leq 64B^2 d^{1/2}\left(\gamma+1\right)\exp\left(-\frac{\left(D-\sqrt{\gamma}\right)^2}{4\gamma}\right).$$

The choice $D = \sqrt{\gamma} + \sqrt{4\gamma\log(n)}$ leads to:

$$R_{\mathcal{K}_D} \leq 64B^2\left(\gamma+1\right)\frac{d}{n}. \tag{44}$$

Since Assumptions of Corollary A.3 are satisfied, we have

$$\begin{aligned}
\mathbb{E}\left[\left(T_L\hat{f}^{(D\text{-LS})}(Z) - f^*(Z)\right)^2\right] &\leq c\left[\max\left[\sigma_{\text{na}}^2, B^2(D+1)^2\right]\frac{2^d d(\log(n)+1)}{n} + 64B^2\left(\gamma+1\right)\frac{d}{n}\right] + 8A_{\mathcal{F}_b}\\
&\leq c_2\max\left(\sigma_{\text{na}}^2, L^2\right)\frac{2^d d\left(\log(n)+1\right)}{n} + 8A_{\mathcal{F}_b}.
\end{aligned}$$

A.4. Proof of Corollary 3.2

We just have to check the assumptions of the Theorem 3.1.

- Under Assumptions of Proposition 2.3 $f_m^\star$ is linear:

$$\begin{aligned}
|f_m(x) - f_m(y)| &\leq \left|\left\langle x-y, \delta^{(m)}\right\rangle\right|\\
&= \|x-y\|_\infty\left|\left\langle\frac{x-y}{\|x-y\|_\infty}, \delta^{(m)}\right\rangle\right|\\
&\leq \|x-y\|_\infty\sup_{\|u\|_\infty\leq 1}\left|\left\langle u, \delta^{(m)}\right\rangle\right|\\
&= \|x-y\|_\infty\left\|\delta^{(m)}\right\|_1,
\end{aligned}$$

  thus 10 is verified with $B = \max_{m\in\mathcal{M}}\max\left[|\delta_0^{(m)}|, \left\|\delta^{(m)}\right\|_1\right]$.

- Let's check Assumption 9:

– Under Assumption 3 and (5 or 4), for $j \in [d]$, X_j is Gaussian, thus $X_j - \mathbb{E}X_j$ is $\mathbb{V}[X_j]$-Sub-Gaussian and

$$\begin{cases} \mathbb{E}\left[X_j\right]^2 \leq \gamma \\ \mathbb{V}\left[X_j\right] \leq \gamma, \end{cases}$$

where $\gamma = \max_{j\in[d]} \mathbb{E}\left[X_j^2\right]$.

– Under Assumption 8: Let γ such that for all $m \in \mathcal{M}$ and $j \in [d]$,

$$\mathbb{E}\left[X_j^2 | M = m\right] \leq \gamma \tag{45}$$

We will use moment characterisation (Boucheron et al., 2013, Theorem 2.1) to proove that the tail of X_j is Sub-Gausian.

$$\begin{aligned} \mathbb{E}\left[\left(X_j - \mathbb{E}X_j\right)^{2q}\right] &\leq \mathbb{E}\left[X_j^{2q}\right] \\ &= \sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[X_j^{2q} | M = m\right] \end{aligned}$$

By assumption, $X_j | M = m$ is Gaussian, we denote by $e_{j,m}$ is expectancy, we have $e_{j,m}^2 \leq \gamma$, from this it follow that

$$\begin{aligned} \mathbb{E}\left[X_j^{2q} | M = m\right] &= \sum_{k=0}^{2q} \binom{2q}{k} \mathbb{E}\left[\left(X_j - e_{j,m}\right)^k | M = m\right] e_{j,m}^{2q-k} \\ &= \sum_{k=0}^{q} \binom{2q}{2k} \mathbb{E}\left[\left(X_j - e_{j,m}\right)^{2k} | M = m\right] e_{j,m}^{2q-2k} \\ &\leq \sum_{k=0}^{q} \binom{2q}{2k} \mathbb{E}\left[\left(X_j - e_{j,m}\right)^{2k} | M = m\right] \gamma^{2q-2k} \\ &\leq \sum_{k=0}^{q} \binom{2q}{2k} \frac{(2k)!}{2^k k!} \gamma^{2k} \gamma^{2q-2k} \\ &\leq \gamma^{2q} (2q)! \sum_{k=0}^{q} \frac{1}{2^k} \\ &\leq 2\gamma^{2q} (2q)!. \end{aligned}$$

Thus,

$$\mathbb{E}\left[\left(X_j - \mathbb{E}X_j\right)^{2q}\right] \leq (2\gamma)^{2q} (2q)!.$$

This conclude that $X_j - \mathbb{E}X_j$ is 8γ-Sub-Gaussian.

B. Proofs of the main result from Section 4

B.1. Intermediate results

The following lemma allows to control a key quantity, that appears when we try to separate the frequent patterns among those that are less frequent.

Lemma B.1. *Let's define*

$$R_{\tau,p}(n) := \mathbb{E}\left[\sum_{m\in\mathcal{M}} p_m \left(\frac{\mathbb{1}_{|E_m|>\tau}}{|E_m|} d + \mathbb{1}_{|E_m|=0}\right)\right],$$

where $E_m = \{i \in [n], M_i = m\}$. *We have for* $\tau \geq 1/n$*:*

$$R_{\tau,p}(n) \leq 5 \max\left(1, \frac{d}{n\tau}\right) \mathfrak{C}_p(\tau). \tag{46}$$

Proof. • Case 1: if $\tau > 1$ then $R_{\tau,p}(\tau) = 1$ and $\mathfrak{C}_p(\tau) = 1$ thus

$$R_{\tau,p}(n) \leq 5\mathfrak{C}_p(\tau).$$

• Case 2: We suppose that $\tau \leq 1$, we have $\frac{d}{n\tau} = 1$. We denote by K_τ the cardinal number of $\{m \in \mathcal{M}, p_m > \tau\}$,

$$\begin{aligned} R_{\tau,p}(n) &\leq \mathbb{E}\left[\sum_{m\in\mathcal{M}} p_m \left(\frac{\mathbb{1}_{|E_m|>n\tau}}{|E_m|}d + \mathbb{1}_{|E_m|<n\tau}\right)\right] \\ &= \mathbb{E}\left[\sum_{m:p_m>\tau} p_m \left(\frac{\mathbb{1}_{|E_m|>n\tau}}{|E_m|}d + \mathbb{1}_{|E_m|<n\tau}\right)\right] + \mathbb{E}\left[\sum_{m:p_m\leq\tau} p_m \left(\frac{\mathbb{1}_{|E_m|>n\tau}}{|E_m|}d + \mathbb{1}_{|E_m|<n\tau}\right)\right]. \end{aligned}$$

Using that $\frac{d}{n\tau} = 1$, we have $\frac{\mathbb{1}_{|E_m|>n\tau}}{|E_m|}d + \mathbb{1}_{|E_m|<n\tau} \leq \max\left(1, \frac{d}{n\tau}\right)$. Thus,

$$\begin{aligned} R_{\tau,p}(n) &\leq \mathbb{E}\left[\sum_{m:p_m>\tau} p_m \left(\frac{\mathbb{1}_{|E_m|>n\tau}}{|E_m|}d + \mathbb{1}_{|E_m|\leq n\tau}\right)\right] + \max\left(1, \frac{d}{n\tau}\right) \sum_{m:p_m\leq\tau} p_m \\ &\leq \mathbb{E}\left[\sum_{m:p_m>\tau} p_m \frac{\mathbb{1}_{|E_m|>n\tau}}{|E_m|}d\right] + \sum_{m:p_m>\tau} p_m \mathbb{P}\left(|E_m| \leq n\tau\right) + \max\left(1, \frac{d}{n\tau}\right) \sum_{m:p_m\leq\tau} p_m. \end{aligned} \tag{47}$$

– First term: We use Lemma A.1, $\mathbb{E}\left[\frac{\mathbb{1}_{|E_m|>1}}{|E_m|}\right] \leq \frac{2}{p_m(n+1)}$ because $|E_m| \sim \mathcal{B}(n, p_m)$.

$$\mathbb{E}\left[\sum_{m:p_m>\tau} p_m \frac{\mathbb{1}_{|E_m|>n\tau}}{|E_m|}d\right] \leq 2K_\tau \frac{d}{n} \leq 2K_\tau\tau. \tag{48}$$

– Second term:

$$\begin{aligned} \sum_{m:p_m>\tau} p_m \mathbb{P}\left(|E_m| \leq n\tau\right) &\leq \sum_{m:p_m>\tau} p_m \left(\mathbb{P}\left(|E_m| = 0\right) + \mathbb{P}\left(\frac{\mathbb{1}_{|E_m|>0}}{|E_m|} \geq 1/\tau n\right)\right) \\ &\leq \sum_{m:p_m>\tau} p_m \left((1-p_m)^n + \tau n \mathbb{E}\left[\frac{\mathbb{1}_{|E_m|>0}}{|E_m|}\right]\right) && \text{by Markov inequality} \\ &\leq \sum_{m:p_m>\tau} p_m (1-p_m)^n + \sum_{m:p_m>\tau} p_m \tau n \frac{2}{p_m(n+1)} \\ &\leq \frac{K_\tau}{n} + 2\tau K_\tau. && \text{by optimization} \\ &\leq 3\tau K_\tau && \text{because } 1/n \leq \tau. \end{aligned} \tag{49}$$

Combining (47),(48) and (49), we find

$$R_{\tau,p}(n) \leq 5\tau K_\tau + \max\left(1, \frac{d}{n\tau}\right) \sum_{m:p_m\leq\tau} p_m \leq 5 \max\left(1, \frac{d}{n\tau}\right) \mathfrak{C}_p(\tau).$$

□

The next lemma is particularly useful to show that the optimal threshold is $\tau = d/n$.

Lemma B.2. *For all $\tau > 0$,*

$$\mathfrak{C}_p\left(\frac{d}{n}\right) \leq \max\left(1, \frac{d}{n\tau}\right) \mathfrak{C}_p(\tau).$$

Proof. Remark that for any τ,

$$\begin{aligned}\mathfrak{C}_p\left(\frac{d}{n}\right) &= \sum_{m\in\mathcal{M}} p_m \wedge \frac{d}{n} = \sum_{m\in\mathcal{M}} p_m \wedge \left(\frac{d}{n\tau}\tau\right) \\ &\leq \sum_{m\in\mathcal{M}} p_m \wedge \left(\max\left(1, \frac{d}{n\tau}\right)\tau\right) \\ &\leq \max\left(1, \frac{d}{n\tau}\right)\mathfrak{C}_p\left(\tau\right).\end{aligned}$$

Using that $\max\left(1, \frac{d}{n\tau}\right) \geq 1$. □

B.2. Proof of Theorem 4.1

Proof. In this proof we use same notations and some results of the proof of Theorem 3.1. We consider on each $m \in \mathcal{M}$, $\mathcal{K}_{m,D} := \left\{x_{\text{obs(m)}}, \left\|x_{\text{obs(m)}}\right\|_\infty \leq D\right\}$. From (39), we have for $L = (D+1)B$

$$\forall x \in \mathcal{K}_{m,D}, \qquad |f_m^\star(x)| \leq L. \tag{50}$$

Let's begin by a decomposition of excess risk.

$$\begin{aligned}\mathbb{E}\left[\left(T_L\widehat{f}^{(\tau)}(Z) - f^\star(Z)\right)^2\right] &= \sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[\left(T_L\widehat{f}^{(\tau)}(Z) - f^\star(Z)\right)^2 | M = m\right] \\ &= \sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[\left(T_L\widehat{f}_m^{(\tau)}(X_{\text{obs(m)}}) - f_m^\star(X_{\text{obs(m)}})\right)^2 | M = m\right],\end{aligned}$$

where $f_m^\star(X_{\text{obs(m)}}) := \widetilde{f}_m\left(X_{\text{obs(m)}}\right)\mathbb{1}_{\widehat{p}_m > \tau}$. Using Corollary A.3 on each m with

$$\mathcal{K} = \mathcal{K}_{m,D} = \left\{x_{\text{obs(m)}}, \left\|x_{\text{obs(m)}}\right\|_\infty \leq D\right\},$$

we have

$$\begin{aligned}\mathbb{E}\left[\left(T_L\widehat{f}_m^{(\tau)}(X_{\text{obs(m)}}) - f_m^\star(X_{\text{obs(m)}})\right)^2 | M = m, E_m\right] \leq\ & \mathbb{1}_{|E_m|\leq\tau n}\mathbb{E}\left[f_m^\star(X_{\text{obs(m)}})^2 | M = m\right] \\ & + c\sigma_{\text{na}}^2 \vee L^2\frac{d}{|E_m|}\mathbb{1}_{|E_m|>\tau n} \\ & + 8\text{Approx}\left(f_m^\star, \mathcal{F}_m\right) + R_{\mathcal{K}_{m,D}}.\end{aligned}$$

We split the first term:

$$\begin{aligned}\mathbb{1}_{|E_m|\leq\tau n}\mathbb{E}\left[f_m^\star(X_{\text{obs(m)}})^2 | M = m\right] =\ & \mathbb{1}_{|E_m|\leq\tau n}\mathbb{E}\left[\mathbb{1}_{\left\|x_{\text{obs(m)}}\right\|_\infty \leq D} f_m^\star(X_{\text{obs(m)}})^2 | M = m\right] \\ & + \mathbb{1}_{|E_m|\leq\tau n}\mathbb{E}\left[\mathbb{1}_{\left\|x_{\text{obs(m)}}\right\|_\infty > D} f_m^\star(X_{\text{obs(m)}})^2 | M = m\right].\end{aligned} \tag{51}$$

If $X_{\text{obs(m)}} \in \mathcal{K}_{m,D}$ then $f_m^\star(X_{\text{obs(m)}})^2 \leq L^2$ and the second term is smaller than $R_{\mathcal{K}_{m,D}}$. Thus,

$$\mathbb{1}_{|E_m|\leq\tau n}\mathbb{E}\left[f_m^\star(X_{\text{obs(m)}})^2 | M = m\right] \leq \mathbb{1}_{|E_m|\leq\tau n}L^2 + R_{\mathcal{K}_{m,D}}. \tag{52}$$

By combining,

$$\begin{aligned}\mathbb{E}\left[\left(T_L\widehat{f}_m^{(\tau)}(X_{\text{obs(m)}}) - f_m^\star(X_{\text{obs(m)}})\right)^2 | M = m, E_m\right] \leq\ & c\sigma_{\text{na}}^2 \vee L^2\left(\frac{d}{|E_m|}\mathbb{1}_{|E_m|>\tau n} + \mathbb{1}_{|E_m|\leq\tau n}\right) \\ & + 8\text{Approx}\left(f_m^\star, \mathcal{F}_m\right) + 2R_{\mathcal{K}_{m,D}}.\end{aligned}$$

By summing and taking expectation, we obtain

$$\mathbb{E}\left[\left(T_L\widehat{f}^{(\tau)}(Z)-f^\star(Z)\right)^2\right] = c\sigma_{\text{na}}^2 \vee L^2 \sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[\left(\frac{d}{|E_m|}\mathbb{1}_{|E_m|>\tau n}+\mathbb{1}_{|E_m|\leq\tau n}\right)\right] + 8A_{\mathcal{F}_b} + 2R_{\mathcal{K}_D}. \tag{53}$$

We have used that $\sum_{m\in\mathcal{M}} p_m \text{Approx}\left(f_m^\star, \mathcal{F}_m\right) = A_{\mathcal{F}_b}$ and $\sum_{m\in\mathcal{M}} p_m R_{\mathcal{K}_{m,D}} = R_{\mathcal{K}_D}$. From Lemma B.1 we have

$$\sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[\left(\frac{d}{|E_m|}\mathbb{1}_{|E_m|>\tau n}+\mathbb{1}_{|E_m|\leq\tau n}\right)\right] \leq 5\max\left(1,\frac{d}{n\tau}\right)\mathfrak{C}_p(\tau).$$

We recall that we have from (44) and (60):

$$R_{\mathcal{K}_D} \leq 64L^2\frac{d}{n} \leq 64L^2\mathfrak{C}_p(d/n) \leq 64L^2\max\left(1,\frac{d}{n\tau}\right)\mathfrak{C}_p(\tau), \tag{54}$$

using Lemma B.2. This concludes on (10).

The optimal choice of τ to minimize the upper bound (10) is $\tau = d/n$, by a direct application of Lemma B.2. □

C. Properties of $\mathfrak{C}_p$ and examples

C.1. Insight on $\mathfrak{C}_p$

In this section, we will enumerate a number of results on $\mathfrak{C}_p$. In particular, thanks to the link with the notion of entropy, and the properties linking structure and complexity of distribution p, we can deal with examples such as the homogeneous and heterogeneous Bernoulli Model.

C.1.1. Link with entropies

Computing $\mathfrak{C}_p\left(\frac{d}{n}\right)$ explicitly can be tricky and requires the knowledge of the distribution p of the missing data patterns. The purpose of the following development is to control this complexity with generic bounds.

Definition C.1. Let $b > 0$, let $\mathcal{P}_b(\mathcal{M})$ be the set of $p \in \mathcal{P}(\mathcal{M})$ such that for all $m \in \mathcal{M}$, $p_m \leq b$. We define $\mathcal{G}_b$ the set of function $g : (1/b, +\infty) \longrightarrow \mathbb{R}_+^\star$ such that

$$\begin{aligned} (G_1): &\quad x \longmapsto g(x) && \text{is non decreasing} \\ (G_2): &\quad x \longmapsto xg(1/x) && \text{is non decreasing.} \end{aligned}$$

And, for all $p \in \mathcal{P}_b(\mathcal{M})$, set

$$H_g(p) := \sum_{m\in\mathcal{M}} p_m g(1/p_m).$$

Depending on the choice of g, the quantities $H_g(p)$ can convey some characteristics of the distribution p. For example, if $g = \text{id}$, $H_g(p)$ falls down to the cardinal of the support. If now $g = \log$, this leads to the standard Shannon entropy. Note that if we rewrite (11) as

$$\forall\tau\in(0,1), \quad \mathfrak{C}_p(\tau) := \sum_{m\in\mathcal{M}} p_m \min\left(1,\frac{\tau}{p_m}\right), \tag{55}$$

then, $\mathfrak{C}_p(\tau) = H_g(p)$ for $g(x) = \min(1, \tau x)$. This gives the intuition of the following result.

Theorem C.2. *Let $b > 0$, for all $p \in \mathcal{P}_b(\mathcal{M})$ and $\tau \in (0, b)$,*

$$\mathfrak{C}_p(\tau) = \inf_{g\in\mathcal{G}_b} \frac{H_g(p)}{g(1/\tau)}. \tag{56}$$

The reformulation of $\mathfrak{C}_p$ provided by Theorem C.2 gives us a great diversity of possible upper bounds on $\mathfrak{C}_p$. The following table presents different upper bounds obtained for different choices of functions.

Name	g	$\mathfrak{C}_p(\tau)$ upper bound	Related entropy
Cardinal (or Hartley)	$g(x) = x$	$\text{card}(\mathcal{M})\,\tau$	$\text{Ent}_1(p) = \log(\text{card}(\mathcal{M}))$
Shannon	$g(x) = \log x$	$\frac{\text{Ent}_0(p)}{\log(1/\tau)}$	$\text{Ent}_0(p) = \sum p_m \log(1/p_m)$
$\alpha-$Renyi	$g(x) = x^{1-\alpha}$	$\left(\tau e^{\text{Ent}_\alpha(p)}\right)^{1-\alpha}$	$\text{Ent}_\alpha(p) = \frac{1}{1-\alpha}\log\sum p_m^\alpha$
$\alpha-$Bertrand	$g(x) = x^{1-\alpha}\log^\alpha(x)$	$\frac{\tau^{1-\alpha}}{\log^\alpha(1/\tau)}\sum_{m\in\mathcal{M}}(p_m\log(1/p_m))^\alpha$	Na

Table 1. Upper bounds on $\mathfrak{C}_p(\tau)$. Note that the parameter α is in $(0,1)$ and that Shannon and Bertrand's upper bounds are verified only for $p \in \mathcal{P}_{1/e}$.

The bound based on the cardinality of $\mathcal{M}$ is a classical one and suffers from the curse of dimensionality when the cardinality is too large. The Shannon bound is cardinal-free and adapts with the entropy of p. Therefore, even if the cardinal scales exponentially in the dimension, when the entropy is low, the corresponding bound is more relevant than the classical bound, all the more so as when τ is large (this dependence being only logarithmic). The Renyi bound, obtained with $g(x) = x^{1-\alpha}$, is a good compromise between the two previous ones: it is smaller than the cardinal for large τ and decreases rapidly as τ decreases.

Remark C.3 (Rényi entropy). Depending on the considered τ, upper-bounds provided in Table 1 may be more or less relevant. First, note that the first three upper bounds (Hartley-Shannon-Renyi) of Table 1 are informative, i.e. strictly less than 1, if and only if

$$\tau < e^{-\text{Ent}_\alpha(p)}, \tag{57}$$

for $\alpha \in [0,1]$, where Rényi's entropy is defined by

$$\text{Ent}_\alpha(p) := \frac{1}{1-\alpha}\log\left(\sum_{m\in\mathcal{M}} p_m^\alpha\right). \tag{58}$$

Note that Shannon and Hartley's entropies can be reformulated as limiting cases of Renyi's entropy (Rényi et al., 1961) when $\alpha = 1$ and $\alpha = 0$. Note also that all of these entropies are one when the distribution p of the missing patterns is uniform. As soon as the latter is non-uniform, different regimes for these entropies can be identified. Indeed, for very small τ (less than $\min_m p_m$), Hartley's bound (i.e. the cardinal-type bound) is the lowest one. For larger τ, Rényi's bound is bounded from above by Hartley's one (i.e. the cardinal-type bound) and from below by Shannon's one. Furthermore, remark that Renyi's Entropy is non-increasing in α (see (Rényi et al., 1961)), so given (57), as τ decreases, the Shannon's bound is the first one to be informative (less than 1), followed by Rényi's one, in turn, followed by Hartley's one. The advantage of an entropic form is that you can use the additivity property which is very useful for dealing with examples.

Remark C.4. A number of properties other than Theorem C.2 are very useful for dealing with certain distributions that have a particular structure (for example defined as a tensor product).

The proof of Theorem C.2 is based on the following lemma.

Lemma C.5. *Let $b > 0$, $\tau, p \in (0,b)$ and $g \in \mathcal{G}_b$, one has*

$$\min(p,\tau) \le \frac{pg(1/p)}{g(1/\tau)}. \tag{59}$$

Proof.

- If $p < \tau$,

$$\begin{aligned}\min(\tau,p) &= p \\ &= \frac{pg(1/p)}{g(1/p)} \\ &\le \frac{pg(1/p)}{g(1/\tau)} \qquad \text{using that } 1/\tau < 1/p \text{ and condition } (G_1).\end{aligned}$$

- If $\tau \leq p$,

$$\begin{aligned}\min(\tau, p) &= \tau \\ &= \frac{\tau g(1/\tau)}{g(1/\tau)} \\ &\leq \frac{p g(1/p)}{g(1/\tau)} && \text{using that } \tau \leq p \text{ and condition } (G_2).\end{aligned}$$

□

Theorem C.2 is just an application of this lemma for each term of $\mathfrak{C}_p(\tau)$ in (11).

C.1.2. Some properties of $\mathfrak{C}_p(\tau)$

Proposition C.6.

1. *$\mathfrak{C}_p$ is non-decreasing, concave, and for all $\lambda > 1$ and $\tau \in (0,1)$:*
$$\mathfrak{C}_p(\lambda\tau) \leq \lambda \mathfrak{C}_p(\tau). \tag{60}$$
2. *For $\tau \in (0,1)$:*
$$\tau \leq \mathfrak{C}_p(\tau). \tag{61}$$
3. *Let p, q be two distributions with countable supports, for all $\tau \in (0,1)$,*
$$\mathfrak{C}_{p\otimes q}(\tau) \leq \mathfrak{C}_p(\mathfrak{C}_q(\tau)). \tag{62}$$
4. *"Data processing inequality": Let $f : \mathcal{M} \longrightarrow \mathcal{M}'$, we denote by p^f the distribution of $f(M)$ when $M \sim p$. We have for all $\tau \in (0,1]$*
$$\mathfrak{C}_{p^f}(\tau) \leq \mathfrak{C}_p(\tau). \tag{63}$$

The first inequality reads backward, it is less expensive to increase the argument than to increase the factor before $\mathfrak{C}_p$. This inequality is illustrated in Theorem 4.1 with the optimal choice of threshold. The second inequality (61) gives us a lower bound. Inequalities (62) and (63) will help us to deal with examples that involve several combined processes of missing data generation such as the database merge model of Section 4.2.3.

Proof.

- Proof of (60): For $\lambda \geq 1$, $\min(p_m, \lambda\tau) \leq \lambda \min(p_m, \tau)$, this conclude that $\mathfrak{C}_p(\lambda\tau) \leq \lambda\mathfrak{C}_p(\tau)$.
- Proof of (61): We use (60) with $\lambda = 1/\tau > 1$.
- Proof of (62):

$$\begin{aligned}\mathfrak{C}_{p\otimes q}(\tau) &= \sum_{m,m'} \min(p_m q_{m'}, \tau) \\ &= \sum_m \sum_{m'} q_{m'} \min\left(p_m, \frac{\tau}{q_{m'}}\right) \\ &= \sum_m \sum_{m'} q_{m'} \min\left(p_m, \min\left(\frac{\tau}{q_{m'}}, 1\right)\right) && \text{because } p_m \leq 1 \\ &\leq \sum_m \min\left(p_m, \sum_{m'} q_{m'} \min\left(\frac{\tau}{q_{m'}}, 1\right)\right) && \text{using Jensen inequality} \\ &\leq \mathfrak{C}_p(\mathfrak{C}_q(\tau)) && \text{using definition.}\end{aligned}$$

- Proof of (63): We will use $\min(a+b,c) \leq \min(a,c) + \min(b,c)$ for $a,b,c \geq 0$.

$$\begin{aligned}
\mathfrak{C}_{p^f}(\tau) &= \sum_{k \in \text{Supp}(p^f)} \min\left(p^f_k, \tau\right) \\
&\leq \sum_{k \in \text{Supp}(p^f)} \min\left(\sum_{m:f(m)=k} p_m, \tau\right) \\
&\leq \sum_{k \in \text{Supp}(p^f)} \sum_{m:f(m)=k} \min\left(p_m, \tau\right) \\
&\leq \sum_{m \in \text{Supp}(p)} \min\left(p_m, \tau\right) \\
&\leq \mathfrak{C}_p(\tau).
\end{aligned}$$

□

C.2. Bernoulli Model

It is assumed that the components of M are independent, and for $j \in [d]$, $M_j \sim \mathcal{B}(\epsilon_j)$ where $\epsilon_j \in [0,1]$. The distribution p of missing value pattern is $p = \mathcal{B}(\epsilon_1) \otimes \cdots \otimes \mathcal{B}(\epsilon_d)$. Let's define $\bar{\epsilon} := \frac{1}{d}\sum_{j=1}^{d} \epsilon_j$, the average proportion of missing values. When $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_d = \bar{\epsilon}$, the model is homogenous, otherwise it is heterogenous.

C.2.1. Numerical experiments.

The quantity $\mathfrak{C}_p\left(\frac{d}{n}\right)$ can be compared graphically for different missing pattern distributions of the Bernoulli model. In particular, we have chosen $d = 4$ and

- p_A: Homogeneous Bernoulli with $\bar{\epsilon} = 0.5$,
- p_B: Homogeneous Bernoulli with $\bar{\epsilon} = 0.15$,
- p_C: Heterogeneous Bernoulli with $\bar{\epsilon} = 0.15$ ($\epsilon_1 = 0.3, \epsilon_2 = 0.1, \epsilon_3 = 0.05, \epsilon_4 = 0.05$),
- p_D: Homogeneous Bernoulli with $\bar{\epsilon} = 0.10$.

Note that p_A matches with the uniform distribution over all missing patterns. Figure 4 highlights three key points:

1. The distribution p_A, which corresponds to the uniform distribution on $\mathcal{M}$, is the worst in terms of complexity $\mathfrak{C}_p$.
2. The complexity seems to increase with the proportion of missing data $\bar{\epsilon}$ for homogeneous Bernoulli.
3. The comparison between homogeneous and heterogeneous does not seem relevant because p_B and p_C have the same proportion of missing values and each has a regime with a better $\mathfrak{C}_p$ than the other.

C.2.2. Proof of Proposition 4.4 on the Homogeneous case

Proof. From (15), we must bound $|\mathcal{B}_s|$ and δ_s. For $|\mathcal{B}_s|$, from (Massart, 2007, Proposition 2.5)

$$|\mathcal{B}_s| = \sum_{k=0}^{s} \binom{n}{k} \leq \left(\frac{ed}{s}\right)^s. \tag{64}$$

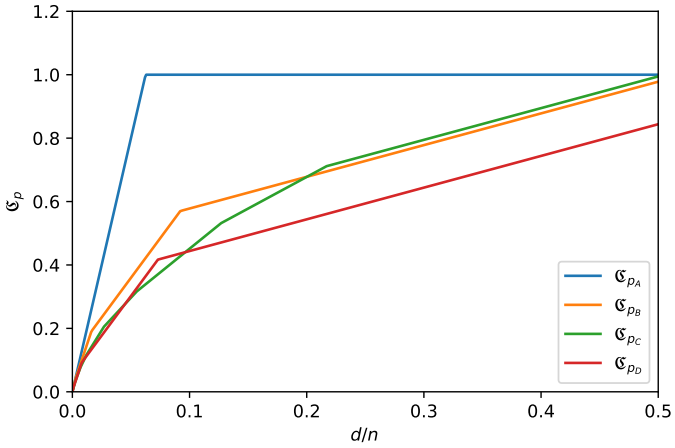


Figure 4. $\mathfrak{C}_p$ as a function of $\frac{d}{n}$.

Let $B \sim \mathcal{B}(\epsilon, d)$, we have $\delta_s = \mathbb{P}(B > s)$. Let $t > 0$, from Markov inequality

$$\begin{aligned}
\mathbb{P}(B > s) = \mathbb{P}\left(t^B > t^s\right) &\leq \frac{\mathbb{E}\left[t^B\right]}{t^s} \\
&= \frac{(\epsilon t + (1-\epsilon))^d}{t^s} \\
&\leq \frac{(\epsilon t + 1)^d}{t^s} \\
&= \exp\left(d \log(1 + \epsilon t) - s \log t\right) \\
&\leq \exp\left(d\epsilon t - s \log t\right).
\end{aligned}$$

The optimal choice $t = \frac{s}{\epsilon d}$, leads to

$$\mathbb{P}(B > s) \leq \epsilon^s \left(\frac{d}{s}\right)^s. \tag{65}$$

Combining (65) and (64), we have,

$$\mathfrak{C}_p\left(\frac{d}{n}\right) \leq \left(\epsilon^s + \frac{d}{n}\right)\left(\frac{ed}{s}\right)^s.$$

□

C.2.3. Heterogeneous Case

Under a certain constraint, the same result as in the homogeneous case can be formulated for the heterogeneous case.

Proposition C.7. *Under the assumptions of Theorem 3.1, and the Heterogeneous Bernoulli Model, if*

$$s_{\bar{\epsilon}}(d/n) = 1 \vee \left\lfloor \frac{\log\left(\frac{n}{d}\right)}{\log(\bar{\epsilon}^{-1})} \right\rfloor \wedge d \geq \bar{\epsilon} d,$$

then

$$\mathbb{E}\left[\mathcal{E}\left(T_L \hat{f}^{(d/n)}\right)\right] \lesssim a_n \left(\frac{ed}{s_{\bar{\epsilon}}(d/n)}\right)^{s_{\bar{\epsilon}}(d/n)} \frac{d}{n} + A_{\mathcal{F}_b}.$$

Proof of C.7. Let's $\tau = d/n$, it is sufficient to show that

$$\mathfrak{C}_p(\tau) \leq \left(\frac{ed}{s_{\bar{\epsilon}}(\tau)}\right)^{s_{\bar{\epsilon}}(\tau)} \tau.$$

We remark that $\epsilon \longmapsto \text{Ent}_\alpha(\mathcal{B}(\epsilon)) = \frac{1}{1-\alpha}\log\left(\epsilon^\alpha + (1-\epsilon)^\alpha\right)$ is concave for $\alpha \in (0,1)$. Thus Renyi entropy of p takes the form

$$\begin{aligned}
\text{Ent}_\alpha(p) &= \sum_{i=1}^{n} \text{Ent}_\alpha(\mathcal{B}(\epsilon_i)) && \text{using additivity of Renyi entropy} \\
&\leq d\text{Ent}_\alpha(\mathcal{B}(\bar{\epsilon})) && \text{using Jensen inequality} \\
&= \frac{d}{1-\alpha}\log\left(\bar{\epsilon}^\alpha + (1-\bar{\epsilon})^\alpha\right),
\end{aligned}$$

using Jensen inequality. From Renyi's bound of Table 1, for all $\alpha \in (0,1)$ this leads to

$$\begin{aligned}
\mathfrak{C}_p(\tau) &\leq \left(\bar{\epsilon}^\alpha + (1-\bar{\epsilon})^\alpha\right)^d \tau^{1-\alpha} \\
&\leq \left(\bar{\epsilon}^\alpha + 1\right)^d \tau^{1-\alpha} \\
&\leq e^{\log(1+\bar{\epsilon}^\alpha)d} \tau^{1-\alpha} \\
&\leq e^{\bar{\epsilon}^\alpha d + (1-\alpha)\log(\tau)}.
\end{aligned}$$

We can minimize function $\psi(\alpha) = \bar{\epsilon}^\alpha d + (1-\alpha)\log(\tau)$ on $(0,1)$:

$$\psi'(\alpha) = \log(\bar{\epsilon})\,\bar{\epsilon}^\alpha d - \log\tau.$$

The first order condition gives us that

$$\log(\bar{\epsilon})\,\bar{\epsilon}^{\alpha^\star} d - \log\tau = 0.$$

And then,

$$\bar{\epsilon}^{\alpha^\star} d = \frac{\log\tau}{\log\bar{\epsilon}} = s_{\bar{\epsilon}}(\tau).$$

Thus,

$$\alpha^\star = \frac{\log\left(\frac{s_{\bar{\epsilon}}(\tau)}{d}\right)}{\log(\bar{\epsilon})}.$$

We have $\alpha^\star \in (0,1)$ if and only if $d\bar{\epsilon} < s_{\bar{\epsilon}}(\tau) < d$. Under this condition, we have

$$\begin{aligned}
\psi(\alpha) &= s_{\bar{\epsilon}}(\tau) + \log(\tau)\left(1 - \frac{\log\left(\frac{s_{\bar{\epsilon}}(\tau)}{d}\right)}{\log(\bar{\epsilon})}\right) \\
&= s_{\bar{\epsilon}}(\tau) + \frac{\log(\tau)}{\log(\bar{\epsilon})}\left(\log(\bar{\epsilon}) - \log\left(\frac{s_{\bar{\epsilon}}(\tau)}{d}\right)\right) \\
&= s_{\bar{\epsilon}}(\tau)\left(1 + \log\left(\frac{d\bar{\epsilon}}{s_{\bar{\epsilon}}(\tau)}\right)\right) \\
&= s_{\bar{\epsilon}}(\tau)\log\left(\frac{ed\bar{\epsilon}}{s_{\bar{\epsilon}}(\tau)}\right).
\end{aligned}$$

The upper bound is therefore

$$\mathfrak{C}_p(\tau) \leq \left(\frac{ed\bar{\epsilon}}{s_{\bar{\epsilon}}(\tau)}\right)^{s_{\bar{\epsilon}}(\tau)} = \left(\frac{ed}{s_{\bar{\epsilon}}(\tau)}\right)^{s_{\bar{\epsilon}}(\tau)} \bar{\epsilon}^{s_{\bar{\epsilon}}(\tau)} = \left(\frac{ed}{s_{\bar{\epsilon}}(\tau)}\right)^{s_{\bar{\epsilon}}(\tau)} \tau. \tag{66}$$

□

C.3. Proof of Proposition 4.5

Proof. We denote by p_P (resp. p_N) the distribution of P (resp. N). Using (63), we have

$$\mathfrak{C}_p \leq \mathfrak{C}_{p_H \otimes p_N}.$$

Furthermore, (62) leads to,

$$\begin{aligned}
\mathfrak{C}_p\left(\frac{d}{n}\right) &\leq \mathfrak{C}_{p_H}\left(\mathfrak{C}_{p_N}\left(\frac{d}{n}\right)\right) \\
&\leq h\mathfrak{C}_{p_N}\left(\frac{d}{n}\right) && \text{because } |\mathrm{Supp(H)}| \leq \mathrm{h} \\
&\leq \left(\frac{ed}{s_\eta\left(d/n\right)}\right)^{s_\eta(d/n)} h\frac{d}{n},
\end{aligned}$$

using Proposition 4.4 on $\mathfrak{C}_{p_N}$.

□

C.4. Proof of Lemma 4.2

Proof. Let $\mathcal{B}$ be a subset of $\mathcal{M}$, we have

$$\begin{aligned}
\mathfrak{C}_p\left(\frac{d}{n}\right) &= \sum_{m\in\mathcal{M}} p_m \wedge \left(\frac{d}{n}\right) \\
&= \sum_{m\in\mathcal{B}} p_m \wedge \left(\frac{d}{n}\right) + \sum_{m\in\mathcal{B}^c} p_m \wedge \left(\frac{d}{n}\right) \\
&\leq \sum_{m\in\mathcal{B}} \frac{d}{n} + \sum_{m\in\mathcal{B}^c} p_m \\
&\leq |\mathcal{B}|\,\frac{d}{n} + \mathbb{P}\left(M \in \mathcal{B}\right).
\end{aligned}$$

We obtain equality with $\mathcal{B} = \{m \in \mathcal{M}, p_m > d/n\}$. Thus,

$$\mathfrak{C}_p\left(\frac{d}{n}\right) = \inf_{\mathcal{B}\subset\mathcal{M}} \left\{\mathrm{Card}(\mathcal{B})\frac{d}{n} + \mathbb{P}\left(M \in \mathcal{B}^c\right)\right\}.$$

□

D. Proof of Section 4.3

The purpose of this part is to establish the lower bounds of Section 4.3.

D.1. Preliminary lemmas.

We consider a set of identifiable models:

$$\mathcal{P}_\mathcal{I} := \{\mathbb{P}_\mu, \mu \in \mathcal{I}\},$$

where $\mathbb{P}_\mu$ is identifiable and $\mathcal{I}$ is a set of parameters. Let $X_1, ..., X_n$ be i.i.d. observations of $\mathbb{P}_\mu$. We define the quadratic risk of an estimator $\widehat{\mu}$ as:

$$r(\mu, \widehat{\mu}) := \mathbb{E}_{\mathcal{P}_\mu}\left[\left(\widehat{\mu}_n - \mu\right)^2\right]. \tag{67}$$

The first step is to lower bound the integrated quadratic risk according to a distribution Π on the set of parameters.

Lemma D.1. *We consider the class of models*

$$\mathcal{P} := \left\{\mathbb{P}_\mu \sim \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}\right\},$$

with σ^2 known. Let $\lambda > 0$ and consider $\Pi \sim \mathcal{N}(0, \lambda^2)$ as a prior distribution for μ. Then

$$\inf_{\widehat{\mu}} \mathbb{E}_{\mu\sim\Pi}\left[r\left(\widehat{\mu}, \mu\right)\right] = \frac{\lambda^2\sigma^2}{\sigma^2 + \lambda^2 n}, \tag{68}$$

where the infinimum is over all $\sigma(X_1, ..., X_n)$-measurable estimator $\widehat{\mu}$.

Proof.

$$\begin{aligned}
\inf_{\widehat{\mu}} \mathbb{E}_{\mu\sim\Pi}\left[r\left(\widehat{\mu}, \mu\right)\right] &= \inf_{\widehat{\mu}} \mathbb{E}_{\mu\sim\Pi}\left[\mathbb{E}_\mu\left[\left(\widehat{\mu} - \mu\right)^2\right]\right] \\
&= \inf_{\widehat{\mu}} \mathbb{E}\left[\mathbb{E}\left[\left(\widehat{\mu} - \mu\right)^2 | X_1, ..., X_n\right]\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[\mu | X_1, ..., X_n\right] - \mu\right)^2\right] \\
&= \mathbb{V}\left[\mathbb{E}\left[\mu | X_1, ..., X_n\right]\right],
\end{aligned}$$

because Bayes estimator $\mathbb{E}\left[\mu | X_1, ..., X_n\right]$ is optimal for the integrated Risk and unbiased. According prior Π, $(\mu, X_1, ..., X_n)$ is a gaussian vector with the following covariance matrix,

$$\Gamma = \begin{pmatrix} \lambda^2 & \lambda^2 & \cdots & \lambda^2 \\ \lambda^2 & \lambda^2 + \sigma^2 & \cdots & \lambda^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda^2 & \lambda^2 & \cdots & \lambda^2 + \sigma^2 \end{pmatrix}.$$

Thus, the variance of $\mathbb{E}\left[\mu | X_1, ..., X_n\right]$ is

$$\begin{aligned}
\mathbb{V}\left[\mathbb{E}\left[\mu | X_1, ..., X_n\right]\right] &= \lambda^2 - \left(\lambda^2, \cdots, \lambda^2\right) \begin{pmatrix} \lambda^2 + \sigma^2 & \cdots & \lambda^2 \\ \vdots & \ddots & \vdots \\ \lambda^2 & \cdots & \lambda^2 + \sigma^2 \end{pmatrix}^{-1} \begin{pmatrix} \lambda^2 \\ \vdots \\ \lambda^2 \end{pmatrix} \\
&= \lambda^2 - \left(\lambda^2, \cdots, \lambda^2\right) \left(\sigma^2 I_n + \lambda u u^T\right)^{-1} \begin{pmatrix} \lambda^2 \\ \vdots \\ \lambda^2 \end{pmatrix},
\end{aligned}$$

where $u = (1, ..., 1)^T$. The Sherman-Morrison formula (see (Petersen et al., 2008) for example) gives

$$\begin{aligned}
\left(\sigma^2 I_n + \lambda u u^T\right)^{-1} &= \frac{1}{\sigma^2} I_n + \frac{\lambda^2 u u^T / \sigma^2}{1 + \frac{\lambda^2 u^T u}{\sigma^2}} \\
&= \frac{1}{\sigma^2} I_n + \frac{\lambda^2 u u^T}{\sigma^2 + \lambda^2 n}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{V}\left[\mathbb{E}\left[\mu | X_1, ..., X_n\right]\right] &= \lambda^2 - \left(\frac{n\lambda^4}{\sigma^2} + \frac{n^2\lambda^4}{\sigma^2 + \lambda^2 n}\right) \\
&= \frac{\lambda^2\sigma^2}{\sigma^2 + \lambda^2 n}.
\end{aligned}$$

Thus,

$$\inf_{\widehat{\mu}} \mathbb{E}_{\mu\sim\Pi}\left[\mathbb{E}_\mu\left[\left(\widehat{\mu} - \mu\right)^2\right]\right] = \frac{\lambda^2\sigma^2}{\sigma^2 + \lambda^2 n}.$$

□

Remark D.2. Using the comparison between minimax and Bayes risks, this result can be used to prove that

$$\inf_{\widehat{\mu}} \sup_{\mu \in \mathbb{R}} r\left(\widehat{\mu}, \mu\right) \geq \frac{\lambda^2 \sigma^2}{\sigma^2 + \lambda^2 n}.$$

We obtain the classical result of the minimax estimation of a Gaussian mean where $\lambda \to \infty$:

$$\inf_{\widehat{\mu}} \sup_{\mu \in \mathbb{R}} r\left(\widehat{\mu}, \mu\right) = \frac{\sigma^2}{n}.$$

Note that this lower bound is only valid when there are no constraints on the parameter space. However, we are interested in guarantees when μ is bounded, this is the purpose of the following result.

Lemma D.3. *Let $\lambda > 0$ and $\Pi \sim \mathcal{N}(0, \lambda^2)$. Then*

$$\left|\mathbb{E}_{\mu \sim \Pi}[r\left(\mu, T_R \widehat{\mu}\right)] - \mathbb{E}_{\mu \sim \Pi}[r\left(T_R \mu, T_R \widehat{\mu}\right)]\right| \leq 8\lambda^2 e^{-\frac{1}{4}\left(\frac{R}{\lambda}\right)^2}. \tag{69}$$

Proof.

$$\begin{aligned}
\left|\mathbb{E}_{\mu \sim \Pi} r\left(\mu, T_R \widehat{\mu}\right) - \mathbb{E}_{\mu \sim \Pi} r\left(T_R \mu, T_R \widehat{\mu}\right)\right| &\leq \left|\left(T_R \widehat{\mu} + R\right)^2 \Pi\left[\mu < -R\right] + \left(T_R \widehat{\mu} - R\right)^2 \Pi\left[\mu > R\right]\right| \\
&\quad + \left|\int_{|\mu| > R} \left(T_R \widehat{\mu} + \mu\right)^2 d\Pi\right| \\
&\leq \int_{|\mu| > R} \left(6R^2 + 2\mu^2\right) d\Pi \\
&\leq 8\mathbb{E}_{\Pi}\left[\mathbb{1}_{|\mu| > R} \mu^2\right] \qquad (70) \\
&\leq 8\sqrt{\Pi\left(|\mu| > R\right) \mathbb{E}_{\Pi}\left[\mu^4\right]} \qquad (71) \\
&\leq 8\sqrt{e^{-\frac{R^2}{2\lambda^2}} \times 3\frac{\sigma^4}{n^2}} \qquad (72) \\
&\leq 8\lambda^2 e^{-\frac{1}{4}\left(\frac{R}{\lambda}\right)^2}.
\end{aligned}$$

We have used Cauchy Schwarz inequality in (71), moment and tail upper bound of Gaussian distribution in (72). □

D.2. Minimax estimation of a value per missing pattern

We consider the following Problem,

$$Y = f^\star(M) + \epsilon, \tag{73}$$

with $f^\star$ a deterministic function of the missing pattern M. We define $\widetilde{\mathcal{P}}_p(\sigma, R)$ as the set of $\mathbb{P}$ that satisfies:

1. $\mathbb{P}(M = m) = p_m$.
2. $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and ϵ is independent of M.
3. $\max_{m \in \mathcal{M}} |f^\star(m)| \leq R$

We denote by $\mathbb{P}_f$ the probability that satisfies the two first conditions with $f^\star = f$. We have the following minimax result on the estimation of $f^\star$.

Proposition D.4. *Let $R, \sigma, c > 0$ such that $c \leq 16e^{-\frac{1}{4}\left(\frac{R}{\sigma}\right)^2}$, then*

$$\inf_{\hat{f}} \sup_{\mathbb{P} \in \widetilde{\mathcal{P}}(\sigma, R)} \mathbb{E}\left[\left(\widehat{f}(M) - f^\star(M)\right)^2\right] \geq (1 - c)\sigma^2 \mathfrak{C}_p\left(\frac{1}{n}\right). \tag{74}$$

Proof. **Step 1**: Comparison with integrated risk and decomposition.

Let $\widehat{f}$ a estimator of $f^\star$. Without loss of generality, we can assume that $\widehat{f}$ belongs to $B_R := \{f | \forall m \in \mathcal{M}, |f(m)| \leq R\}$. Note Π a prior distribution for $f^\star$.

$$\begin{aligned}\sup_{\mathbb{P}\in\widetilde{\mathcal{P}}(\sigma,R)} \mathbb{E}_{\mathbb{P}}\left[\left(\widehat{f}(M) - f^\star(M)\right)^2\right] &= \sup_{f^\star\in B_R} \mathbb{E}_{\mathbb{P}_{f^\star}}\left[\left(\widehat{f}(M) - f^\star(M)\right)^2\right] \\ &\geq \mathbb{E}_{f^\star\sim\Pi}\mathbb{E}_{\mathbb{P}_{f^\star}}\left[\left(\widehat{f}(M) - T_R f^\star(M)\right)^2\right].\end{aligned}$$

We denote by $\mathbb{E}_\Pi = \mathbb{E}_{f^\star\sim\Pi}\mathbb{E}_{\mathbb{P}_{f^\star}}$.

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}(\sigma,R)} \mathbb{E}_{\mathbb{P}}\left[\left(\widehat{f}(M) - f^\star(M)\right)^2\right] &\geq \mathbb{E}_\Pi\left[\left(\widehat{f}(M) - T_R f^\star(M)\right)^2\right] \\ &\geq -\left|\mathbb{E}_\Pi\left[\left(\widehat{f}(M) - T_R f^\star(M)\right)^2\right] - \mathbb{E}_\Pi\left[\left(\widehat{f}(M) - f^\star(M)\right)^2\right]\right| \\ &\quad + \mathbb{E}_\Pi\left[\left(\widehat{f}(M) - f^\star(M)\right)^2\right].\end{aligned} \tag{75}$$

Step 2: Lower bound of the first term.

We choose $\Pi = \bigotimes_{m\in\mathcal{M}} \mathcal{N}(0, \lambda_m^2)$ where $0 \leq \lambda_m \leq \sigma$. Conditioning by M and using Fubini theorem, we obtain,

$$\begin{aligned}\mathbb{E}_\Pi\left[\left(\widehat{f}(M) - f^\star(M)\right)^2\right] &= \mathbb{E}_{f^\star\sim\Pi}\sum_{m\in\mathcal{M}} p_m \mathbb{E}_{\mathbb{P}_{f^\star}}\left[\left(\widehat{f}(m) - f^\star(m)\right)^2\right] \\ &= \sum_{m\in\mathcal{M}} p_m \mathbb{E}_\Pi\left[\left(\widehat{f}(m) - f^\star(m)\right)^2\right] \\ &= \sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[\mathbb{E}_\Pi\left[\left(\widehat{f}(m) - f^\star(m)\right)^2\right] | (M_i)_{i\in[n]}\right] \\ &\geq \sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[\mathbb{E}_\Pi\left[\left(\mathbb{E}\left[f^\star(m)|(Y_i)_{i\in[n]}\right] - f^\star(m)\right)^2\right] | (M_i)_{i\in[n]}\right] \qquad (76) \\ &= \sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[\mathbb{E}_\Pi\left[\left(\mathbb{E}\left[f^\star(m)|(Y_i)_{i\in E_m}\right] - f^\star(m)\right)^2\right] | (M_i)_{i\in[n]}\right] \qquad (77) \\ &= \sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[\mathbb{E}_\Pi\left[\mathbb{V}\left[f^\star(m)|(Y_i)_{i\in E_m}\right]\right] | (M_i)_{i\in[n]}\right].\end{aligned}$$

We have used variational definition of $\mathbb{E}\left[f(m)|(Y_i)_{i\in[n]}\right]$ in (76) and for a prior distribution Π, Y_i and Y_j are independent provided that $M_i \neq M_j$ in (77). Using Lemma D.1, we obtain,

$$\mathbb{E}_\Pi\left[\left(\widehat{f}(M) - f^\star(M)\right)^2\right] \geq \sum_{m\in\mathcal{M}} p_m \mathbb{E}\left[\frac{\lambda_m^2\sigma^2}{\sigma^2 + \lambda_m^2 |E_m|}\right].$$

Using Jensen inequality (and Lemma A.1 with $|E_m| \sim \mathcal{B}(n, p_m)$), we have

$$\mathbb{E}_\Pi\left[\left(\widehat{f}(M) - f^\star(M)\right)^2\right] \geq \sum_{m\in\mathcal{M}} p_m \frac{\lambda_m^2\sigma^2}{\sigma^2 + \lambda_m^2 n p_m}. \tag{78}$$

Step 3: Lower bound of the second term.

Using Lemma D.3, for $A = \left| \mathbb{E}_\Pi \left[\left(\widehat{f}(M) - T_R f^\star(M) \right)^2 \right] - \mathbb{E}_\Pi \left[\left(\widehat{f}(M) - f^\star(M) \right)^2 \right] \right|$, we have

$$\begin{aligned} A &\leq \sum_{m \in \mathcal{M}} p_m \left| \mathbb{E}_\Pi \left[\left(\widehat{f}(m) - T_R f^\star(m) \right)^2 \right] - \mathbb{E}_\Pi \left[\left(\widehat{f}(m) - f^\star(m) \right)^2 \right] \right| \\ &\leq \sum_{m \in \mathcal{M}} p_m 8 \lambda_m^2 e^{-\frac{1}{4}\left(\frac{R}{\lambda_m}\right)^2} \\ &\leq \frac{c}{2} \sum_{m \in \mathcal{M}} p_m \lambda_m^2, \end{aligned} \tag{79}$$

with $c \leq 16 e^{-\frac{1}{4}\left(\frac{R}{\sigma}\right)^2}$ and since $\lambda_m \leq \sigma$.

Step 4: Choice of λ_m and conclusion.

Combining (78) and (79) in (75), we obtain

$$\sup_{\mathbb{P} \in \widetilde{\mathcal{P}}(\sigma, R)} \mathbb{E}_\mathbb{P} \left[\left(\widehat{f}(M) - f^\star(M) \right)^2 \right] \geq \sum_{m \in \mathcal{M}} p_m \frac{\lambda_m^2 \sigma^2}{\sigma^2 + \lambda_m^2 n p_m} - \frac{c}{2} \sum_{m \in \mathcal{M}} p_m \lambda_m^2.$$

We choose

$$\begin{cases} \lambda_m^2 = \frac{\sigma^2}{p_m n} & \text{if } p_m > 1/n, \\ \lambda_m^2 = \sigma^2 & \text{if } p_m \leq 1/n. \end{cases}$$

The condition $\lambda_m \leq \sigma$ holds and $p_m \lambda_m^2 = \sigma^2 \min(p_m, 1/n)$, thus

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{P}(\sigma, R)} \mathbb{E}_\mathbb{P} \left[\left(\widehat{f}(M) - f^\star(M) \right)^2 \right] &\geq \sum_{m \in \mathcal{M}} p_m \frac{\sigma^2 \min(p_m, 1/n)}{1 + \min(p_m, 1)} - \frac{c}{2} \sigma^2 \sum_{m \in \mathcal{M}} \min(p_m, 1/n) \\ &\geq \sum_{m \in \mathcal{M}} p_m \frac{\sigma^2 \min(p_m, 1/n)}{2} - \frac{c}{2} \sigma^2 \sum_{m \in \mathcal{M}} \min(p_m, 1/n) \\ &\geq \frac{\sigma^2}{2} (1 - c) \, \mathfrak{C}_p(1/n). \end{aligned}$$

□

D.3. Proof of Section 4.3

Proof of Theorem 4.7. The idea is to reduce the prediction problem on class $\mathcal{P}_p(R, \sigma)$ to an estimation problem on class $\widetilde{\mathcal{P}}_p(R/d, \sigma)$ and then use Proposition D.4. We denote by m_0 the missing pattern without missing values.

Let $a \in [-1, 1]^\mathcal{M}$, we consider $\mathbb{P}_a \in \mathcal{P}_p(R, \sigma)$ which satisfies:

1. $\beta_0 = R a_{m_0}$
2. $\beta = R (1/d, ..., 1/d)^T$.
3. For all $m \neq m_0$, $X | M = m \sim \delta_{\mu^{(m)}}$ where $\mu^{(m)}_{mis(m)} = (a_m - a_{m_0})(1, 0, ..., 0)^T$ and $\mu^{(m)}_{\text{obs(m)}} = 0$ (δ denote the Dirac distribution).

These problems satisfy Assumption 9 with $\gamma = 2$.

Step 1: Recall that the Bayes predictor is given by

$$\begin{aligned} f^\star_m \left(X_{\text{obs(m)}} \right) &= \mathbb{E} \left[Y | X_{\text{obs(m)}}, M = m \right] \\ &= \mathbb{E} \left[\langle X, \beta \rangle | X, M = m \right] \\ &= \left\langle X_{\text{obs(m)}}, \beta_{\text{obs(m)}} \right\rangle + \mathbb{E} \left[\left\langle X_{mis(m)}, \beta_{mis(m)} \right\rangle | X, M = m \right] \end{aligned}$$

Using $X|M=m \sim \delta_{\mu^{(m)}}$, we have

$$f_m^\star\left(X_{\text{obs(m)}}\right) = Ra_m.$$

We have $f_m^\star$ $R-$lipschitz for ℓ_∞-norm (because f_m is constant) and $|f_m^\star(0)| \leq R$ then $\mathbb{P}_a$ satisfies Assumption 10 with $B = R$ and $B^2(\gamma+1) \leq 3R^2$.

Step 2: Problem reduction. For $\mathbb{P}_a$, $X_{\text{obs(M)}} = 0$ $\mathbb{P}_a$-a.s., then there are no information in $X_{\text{obs(M)}}$, all the information is contained in the missing patterns and $Z = \left(X_{\text{obs(M)}}, M\right) = (0, M)$ $\mathbb{P}_a$-a.s. i.e. we can ignore $X_{\text{obs(M)}}$. The Bayes predictor is

$$f^\star(M) = Ra_M,$$

and

$$Y = f^\star(M) + \epsilon.$$

This corresponds to Problem (73), and varying $a \in [-1,1]^{\mathcal{M}}$, we obtain the set $\widetilde{\mathcal{P}}_p(R/d, \sigma)$. Thus, using Proposition D.4

$$\begin{aligned}
\max_{\mathbb{P}\in\mathcal{P}_p(\sigma,R)} \mathbb{E}_{\mathbb{P}}\left[\left(f^\star(Z) - \hat{f}(Z)\right)^2\right] &= \max_{a\in[-1,1]^{\mathcal{M}}} \mathbb{E}_{\mathbb{P}_a}\left[\left(f^\star(M) - \hat{f}(M)\right)^2\right] \\
&= \max_{\mathbb{P}\in\widetilde{\mathcal{P}}_p(R/d,\sigma)} \mathbb{E}_{\mathbb{P}}\left[\left(f^\star(M) - \hat{f}(M)\right)^2\right] \\
&\geq (1-c)\frac{\sigma^2}{2}\mathfrak{C}_p\left(\frac{1}{n}\right).
\end{aligned}$$

□

Proof of Corollary 4.8. We will use the same method as in the previous proof. We need to find a subclass of problem MAR included in $\mathcal{P}_p(\sigma, R)$. Let $a \in [-1,1]^{\mathcal{M}}$, we denote by $\mathbb{P}_a$ the following problem.

1. $X_1 \sim \mathcal{N}(0,1)$.
2. $M = h(X_1)$ a.s. where h satisfies $\mathbb{P}\left(h(X_1) = m\right) = p_m$.
3. $X_{2:d}|X_1 \sim \delta_{\mu^{(h(X_1))}}$ where $\mu_{mis(m)}^{(h(X_1))} = a_{h(X_1)}(1,0,...,0)^T$ and $\mu_{(obs(m)}^{(h(X_1))} = 0$.
4. $\beta = R\left(0, 1/(d-1), ..., 1/(d-1)\right)^T$.

By construction, $\mathbb{P}_a$ is MAR, and Assumption 9 holds with $\gamma = 1$. With this new choice of $\mathbb{P}_a$, the rest of the proof is similar to the proof of Theorem 4.7.

□

E. Details on numerical experiments of Section 5

The codes of our numerical experiments are all available on a `github.com/AlexisAyme/minimax_linear_na`.

E.1. Details on data generation setting

In order for the simulations to be reproducible, here are the useful parameters to generate the dataset of Section 5.

Let $U \in \mathbb{R}^{8\times 8}$ be the diagonal matrix per block with each block equal to $(1)_{i,j\in[2]}$. For all scenarios $\beta = (1,1,1,1,1,1,1,1)$ and $\beta_0 = 0$.

(a) MCAR $\mu = (1,1,1,1,1,1,1,1)$, $\Sigma = U$, and $\sigma = 0.1$.

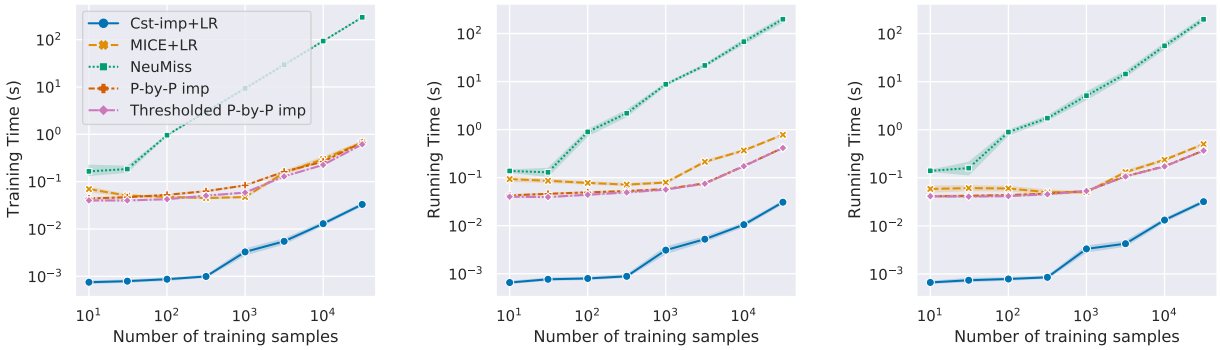
(b) MAR $\Sigma = U$ on corresponding block and $\sigma = 0.5$.

(c) GPMM. (X, M) is distributed according to Assumption 8 with $\sigma = 1$ and

- $p_{m_1} = 0.6$, $m_1 = (0, 1, 0, 1, 0, 0, 0, 0)$, $\mu_{m_1} = (0, 5, 4, -1, 0, 0, 0, 0)$, and $\Sigma_{m_1} = U$.
- $p_{m_2} = 0.3$, $m_2 = (1, 0, 1, 1, 0, 0, 0, 0)$, $\mu_{m_2} = (1, 3, 0, 2, 0, 0, 0, 0)$, and $\Sigma_{m_2} = (1)_{i,j\in[8]}$.
- $p_{m_3} = 0.02$, $m_3 = (0, 1, 1, 1, 0, 0, 0, 0)$, $\mu_{m_3} = (0, 5, 4, -1, 0, 0, 0, 0)$, and $\Sigma_{m_3} = I_8$.
- $p_{m_4} = 0.02$, $m_4 = (1, 1, 0, 1, 0, 0, 0, 0)$, $\mu_{m_4} = (0, 5, 0, -1, 0, 0, 0, 0)$, and $\Sigma_{m_4} = I_8$.
- $p_{m_5} = 0.02$, $m_5 = (1, 1, 0, 0, 0, 0, 0, 0)$, $\mu_{m_5} = (0, -10, 7, -1, 0, 0, 0, 0)$, and $\Sigma_{m_5} = I_8$.
- $p_{m_6} = 0.02$, $m_6 = (0, 1, 0, 0, 0, 0, 0, 0)$, $\mu_{m_6} = (0, 9, 0, -1, 0, 0, 0, 0)$, and $\Sigma_{m_6} = I_8$.
- $p_{m_7} = 0.02$, $m_7 = (0, 0, 1, 0, 0, 0, 0, 0)$, $\mu_{m_7} = (3, 0, 0, -1, 0, 0, 0, 0)$, and $\Sigma_{m_7} = I_8$.

E.2. Training Time

Figure 5 corresponds to the training time of the simulations in Section 5 and are associated with the curve in Figure 3. NeuMiss has a much more limiting training time than other methods. The most time-efficient method is also the most biased. Indeed, Cst-imp+LR does not adapt to any scenario (see Figure 3). The training times are similar for the other methods, but MICE+LR is only relevant for scenario (a).



(a) MCAR (b) MAR (c) MNAR-GPMM

Figure 5. Training time w.r.t. the number of training sample.

E.3. Additional experiments

E.3.1. High dimensional scenario

In this section, we assess the numerical performances of pattern-by-pattern regressors, more specifically when the ambient dimension of the input increases. To illustrate the theoretical generalization bounds, and to show that our methods can leverage from low-entropic missing pattern distributions, we consider a **GPMM** in which we make d vary, from 10 to 100, while keeping a same "complexity" of the missing pattern distribution. To this end, for each dimension d, we generate $50d$ samples always containing only 4 missing patterns: m_1, m_2, m_3, m_4 randomly generated with 50% missing values. According to the GPMM model, the distribution of X given $M = m_i$, for $i = 1, 2, 3, 4$, is $\mathcal{N}(\mu_{m_i}, \Sigma_{m_i})$, such that for all missing patterns $\mu_i = 0$, and

- $p_{m_1} = 0.4$, $\beta_1 = \beta$, and $\Sigma_{m_1} = I$;
- $p_{m_2} = 0.3$, $\beta_2 = 0$, and $\Sigma_{m_2} = T_d$;
- $p_{m_3} = 0.2$, $\beta_3 = -\beta$, and $\Sigma_{m_3} = T_d$;
- $p_{m_4} = 0.1$, $\beta_4 = 2\beta$, and $\Sigma_{m_4} = T_d$,

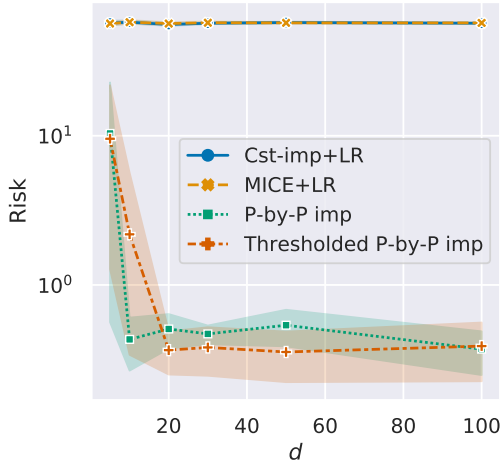


Figure 6. Risk w.r.t. the dimension d. The number of training sample is $n = 50d$.

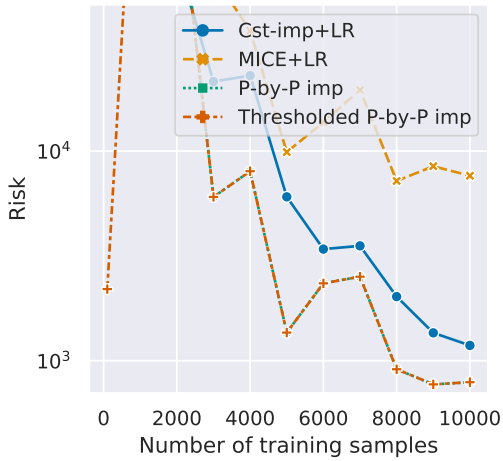


Figure 7. Risk w.r.t. the number of training samples.

where $\beta = ((1/2)^j)_{j\in[d]}$ and T_d is a symmetric Toeplitz matrix with coefficients $((1/2)^j)_{j\in[d]}$. The results are gathered in Figure 6. Despite the MNAR feature of the generated data, our pattern-by-pattern methods perform well in this setting, regardless of the increasing dimension: this highlights the compatibility of pattern-by-pattern regressors with high-dimensional scale provided that the "complexity" of the missing pattern distribution remains bounded.

E.3.2. Real dataset

We ran experiments on the real **superconductivity dataset** ($d = 81$) by simulating missing patterns corresponding to the aggregation of 2 datasets (complying with the example developed in Section 4.2.3 choosing $\eta = 0$) and constructed as follows. First a clustering on the outputs $(y_i)_i$ is performed resulting into two groups of data, the first (resp. second) one containing observations with high (resp. low) outputs. The first dataset associated to high outputs does not contain any missing entries in the input variables. The second one, associated to low outputs, presents a single missing pattern in which only half of the input variables are observed, in particular the ones that are the least correlated to the output y.

Figure 7 shows the results: our proposed predictors behave well in any sample size regimes, outperforming 2-step strategies. This exemplifies that even in cases easy in appearance where only two missing patterns are surveyed, pattern-by-pattern regressors remain relevant, being able to efficiently grasp complex dependency between the missing pattern distribution and the output.