Supervised Learning with Missing Values

Gaël Varoquaux

Inria

with Julie Josse, Erwan Scornet, Marine Le Morvan, Nicolas Prost, & Thomas Moreau

Missing values

Partially observed exemplars

- Non-response in questionnaires
- Missing correspondences across tables
- Measurements not performed (*eg* due to patient urgency)

Ubiquitous in health and social sciences

Missing values

Partially observed exemplars

- Non-response in questionnaires
- Missing correspondences across tables
- Measurements not performed (*eg* due to patient urgency)

Ubiquitous in health and social sciences

How to build predictive models on such data?

Outline

1. Settings
   - Supervised learning theory
   - Classical missing-values framework
2. Adapting learning procedures
3. Linear mechanism, non-linear predictor
4. Differentiable programming: a neural architecture

1 Settings

- Supervised learning theory
- Classical missing-values framework

Based on [Josse... 2019] "On the consistency of supervised learning with missing values"

1 Settings

Supervised learning theory

Classical missing-values framework

Supervised learning settings

- Given n pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ drawn *i.i.d.*
  find a function $f : \mathcal{X} \to \mathcal{Y}$ such that $f(x) \approx y$

  Notation: $\hat{y} \stackrel{\text{def}}{=} f(x)$

Risk minimization

- Loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- Bayes predictor: $$f^\star \in \underset{f:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}} \mathbb{E}\big[l\big(f(x), y\big)\big]$$
- For quadratic loss, $f^\star(x) = \mathbb{E}[y|x]$

Supervised learning procedures

A *learning* procedure gives $\hat{f}_n$ from $\mathcal{D}_{n,\mathrm{train}} = \{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$

Bayes consistency

- A Bayes-consistent procedure asymptotically gives a Bayes predictor

$$\mathbb{E}[\ell(\hat{f}_n(\mathbf{X}), Y)] \xrightarrow[n\to\infty]{} \mathbb{E}[\ell(f^\star(\mathbf{X}), Y)]$$

Empirical risk minimization

- Estimation of f:

$$\hat{f}_n \in \underset{f\in\mathcal{F}}{\mathrm{argmin}} \sum_{i=1}^{n} l\big(f(x_i), y_i\big)$$

1 Settings

Supervised learning theory

Classical missing-values framework

Notations

Full data	$\mathbf{X} \in \mathbb{R}^d$
Missingness indicator	$\mathbf{M} \in \{0,1\}^d$, $M_j = 1$ iff X_j is not observed
Incomplete data	$\widetilde{\mathbf{X}} \in \bigotimes_{j=1}^{d} (\mathcal{X}_j \cup \{\mathrm{NA}\})$,
	$\widetilde{\mathbf{X}} = \mathbf{X} \odot (\mathbf{1} - \mathbf{M}) + \mathrm{NA} \odot \mathbf{M}$

Example realization	$\mathbf{x} = (1.1, 2.3, -3.1, 8, 5.27)$
	$\mathbf{m} = (0, 1, 0, 0, 1)$
	$\widetilde{\mathbf{x}} = (1.1,\ \mathrm{NA},\ -3.1,\ 8,\ \mathrm{NA})$
Observed fraction	$\mathbf{x}_o = (1.1,\ \cdot\ ,\ -3.1,\ 8,\ \cdot\ )$
Unobserved fraction	$\mathbf{x}_m = (\ \cdot\ ,\ 2.3,\ \cdot\ ,\ \cdot\ ,\ 5.27)$

Missing values and parametric likelihoods [Rubin 1976]

Model **a)** a distribution f_θ for the complete data $\mathbf{x}$
b) a random process g_ϕ generating a mask $\mathbf{m}$

Statistical inference: estimate θ

(full likelihood) $$\mathcal{L}_1(\theta, \phi) = \prod_{i=1}^{n} \int f_\theta(\mathbf{x}_{i,o}, \mathbf{x}_{i,m})\, g_\phi(\mathbf{m}_i | \mathbf{x}_{i,o}, \mathbf{x}_{i,m})\, \mathrm{d}\mathbf{x}_{i,m}$$

Expectation over
missing-values mechanism

Missing values and parametric likelihoods [Rubin 1976]

Model a) a distribution f_θ for the complete data $\mathbf{x}$
b) a random process g_ϕ generating a mask $\mathbf{m}$

Statistical inference: estimate θ

(full likelihood) $$\mathcal{L}_1(\theta, \phi) = \prod_{i=1}^{n} \int f_\theta(\mathbf{x}_{i,o}, \mathbf{x}_{i,m})\, g_\phi(\mathbf{m}_i | \mathbf{x}_{i,o}, \mathbf{x}_{i,m})\, \mathrm{d}\mathbf{x}_{i,m}$$

Expectation over
missing-values mechanism

(ignoring missing mechanism) $$\mathcal{L}_2(\theta) = \prod_{i=1}^{n} \int f_\theta(\mathbf{x}_{i,o}, \mathbf{x}_{i,m})\, \mathrm{d}\mathbf{x}_{i,m}$$

Ignorable missingness [Rubin 1976]

Definition: **Missing at random** situation (MAR)

for non-observed values, the probability of missingness does not depend on this non-observed value.

[Rubin 1976], modern formulation in [Josse... 2019]

$$\text{observed}(\mathbf{x}', \mathbf{m}_i) = \text{observed}(\mathbf{x}_i, \mathbf{m}_i) \quad \Rightarrow \quad g_\phi(\mathbf{m}_i|\mathbf{x}') = g_\phi(\mathbf{m}_i|\mathbf{x}_i)$$

Theorem [Rubin 1976], in MAR, maximizing likelihood that ignores the missing mechanism gives the same maximum-likelihood estimates θ for of model a) as the full likelihood.

Ignorable missingness [Rubin 1976]

Definition: **Missing at random** situation (MAR)

for non-observed values, the probability of missingness does not depend on this non-observed value.

[Rubin 1976], modern formulation in [Josse... 2019]

$$\text{observed}(\mathbf{x}', \mathbf{m}_i) = \text{observed}(\mathbf{x}_i, \mathbf{m}_i) \quad \Rightarrow \quad g_\phi(\mathbf{m}_i|\mathbf{x}') = g_\phi(\mathbf{m}_i|\mathbf{x}_i)$$

Special case: **Missing completely at random** (MCAR)

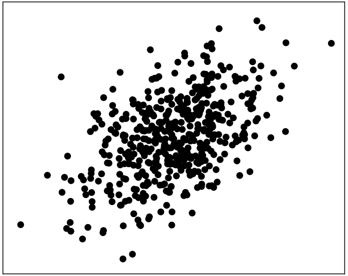
$\mathbf{M}$ is independent of $\mathbf{X}$

Missing Not at Random situation (MNAR)

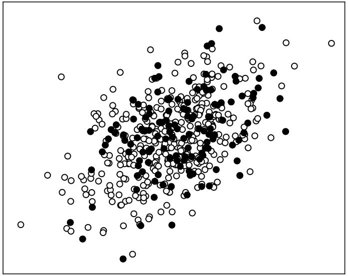
Missingness **not ignorable** $\Rightarrow$ Hard

must explicitly model the mechanism

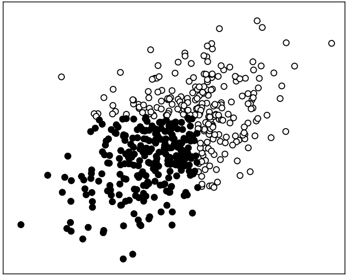
Missing-values settings



Complete



MCAR



MNAR (censored)

Expectation maximization

Optimize likelihood $\mathcal{L}_2(\theta)$ (ignoring missing mechanism) by alternating:

- Expectation in Likelihood over unobserved values, using parameters $\theta^{(t)}$
- Maximization of the resulting expression over θ to give $\theta^{(t+1)}$

Estimation procedures that build upon ignorability

Expectation maximization

Optimize likelihood $\mathcal{L}_2(\theta)$ (ignoring missing mechanism) by alternating:

- Expectation in Likelihood over unobserved values, using parameters $\theta^{(t)}$
- Maximization of the resulting expression over θ to give $\theta^{(t+1)}$

Imputation & plug-in estimation

1. Use a routine to compute $\mathcal{P}(x_{i,m}|x_{i,o})$
2. Create a complete data (emulating the expectation in $\mathcal{L}_2$)
3. Apply standard routine to maximize likelihood of complete data

   Bonus: monte-carlo approximation by multiple-imputations

Estimation procedures that build upon ignorability

Expectation maximization

Imputation & plug-in estimation

In prediction settings,
procedures must be adapted to work out-of-sample
[Josse... 2019]

The predictive model is applied on partially-observed test data

Supervised learning with missing values

Focus on risks not likelihood

- Missing values at test time

$\Rightarrow f$ must predict on missing values

$$f : \mathcal{X} \rightarrow \mathcal{Y} \qquad \hat{f}_n \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} l\big(f(x_i), y_i\big)$$

- Semi discrete space $\mathcal{X} = \bigotimes_{j=1}^{d} (\mathcal{X}_j \cup \{\text{NA}\})$

2 Adapting learning procedures

[Josse... 2019] “On the consistency of supervised learning with missing values”

Test-time imputation

Theorem [Josse... 2019], given $f^\star$, Bayes predictor on *fully-observed* data,

$$f^\star_{\mathrm{MI}}(\widetilde{\mathbf{x}}) = \mathbb{E}_{\mathbf{X}_m|\mathbf{X}_o=\mathbf{x}_o}\left[f^\star(\mathbf{X}_m, \mathbf{x}_o)\right],$$

is a Bayes-optimal predictor in MAR settings.

The expectation can be computed by sampling multiple imputations.

Note: single imputation is not, in general, consistent

Train-time constant imputation

(constant imputation) $$X'_1 = X_1 \mathbb{1}_{M_1=0} + \alpha \mathbb{1}_{M_1=1}.$$

Assumption (Regression model) $Y = f^\star(\mathbf{X}) + \varepsilon$, with $\mathbf{X}$ has a continuous density $g > 0$ on $[0, 1]^d$, $\|f^\star\|_\infty < \infty$, and $\varepsilon \perp\!\!\!\perp (\mathbf{X}, M_1)$

Assumption (Missingness pattern - MAR) $X_2, \ldots, X_d$ fully observed and missingness M_1 on X_1 satisfies $M_1 \perp\!\!\!\perp X_1 | X_2, \ldots, X_d$ and is such that the function $(x_2, \ldots, x_d) \mapsto \mathbb{P}[M_1 = 1 | X_2 = x_2, \ldots, X_d = x_d]$ is continuous.

Train-time constant imputation

(constant imputation) $$x_1' = x_1 \mathbb{1}_{M_1=0} + \alpha \mathbb{1}_{M_1=1}.$$

Theorem [Josse... 2019], The Bayes predictor after constant imputation, $$f^{\star}_{\mathrm{SI}}(\mathbf{x}') = \mathbb{E}[Y|X' = x'],$$ is equal to the Bayes predictor on the original data almost everywhere.

Corollary constant imputation followed by universally-consistent learner is a procedure consistent almost everywhere.[1]

[1]Almost everywhere because input data landing exactly on imputation constant α will be mistaken for an NA.

Adapting supervised learning procedures

- Different trade offs than statistical inference
- Good imputation is not necessary

Also in [Josse... 2019]

- Risk of tree-based models which can optimize naturally for inputs in semi-discrete spaces.

3 Linear mechanism, non-linear predictor

The seemingly-simple case of data generated from a linear mechanism.

Linear predictor on linearly-generated data with missing values: non consistency and solutions [Le Morvan... 2020b]

Linear mechanism and missing data

Settings $y = X\,w$, Z is observed: X masked by M

The best predictor may not be linear

Example

Let $Y = X_1 + X_2 + \varepsilon$, where $X_2 = \exp(X_1) + \varepsilon_1$.

When only X_1 is observed, the model can be rewritten as

$$Y = X_1 + \exp(X_1) + \varepsilon + \varepsilon_1,$$

[Le Morvan... 2020b]

Linear mechanism, missing data, and Gaussian variates

Assumption Gaussian pattern mixture model

X conditional on M is Gaussian: for all $m \in \{0, 1\}^d$, there exist μ_m and Σ_m such that

$$X \mid (M = m) \sim \mathcal{N}(\mu_m, \Sigma_m).$$

Proposition The optimal predictor is a polynomial of X and cross-products of M, with 2^d terms.

$$f^\star(Z) = \beta^\star_{0,0} + \sum_{j=1}^{d} \beta^\star_{j,0} M_j + \sum_{j=1}^{d} \beta^\star_{j,1} M_j X_j + \sum_{i=1}^{d} \sum_{j=1}^{d} \beta^\star_{i,j,2} M_i M_j X_j + \dots$$

Estimation and finite-sample bounds

Polynomial fitting is linear fitting on expended basis

Theorem Estimating the polynomial coefficients with ordinary least squares leads to a risk R of order $\mathcal{O}(2^d/n)$:

$$\sigma^2 + \frac{2^d\, c_1}{n+1} \;\leq\; R \;\leq\; c\,\sigma^2 \frac{2^{d-1}(d+2)(1+\log n)}{n} + \sigma^2.$$

[Le Morvan... 2020b]

Multi-layer perceptron

The Bayes predictor is piece-wise affine

Theorem: Feeding the concatenated vector $(X \odot (\mathbf{1} - M), M)$ to a Multi-Layer Perceptron with ReLU non-linearities and width 2^d is Bayes consistent.

Heuristic: Reducing the width of the network controls model complexity.

[Le Morvan... 2020b]

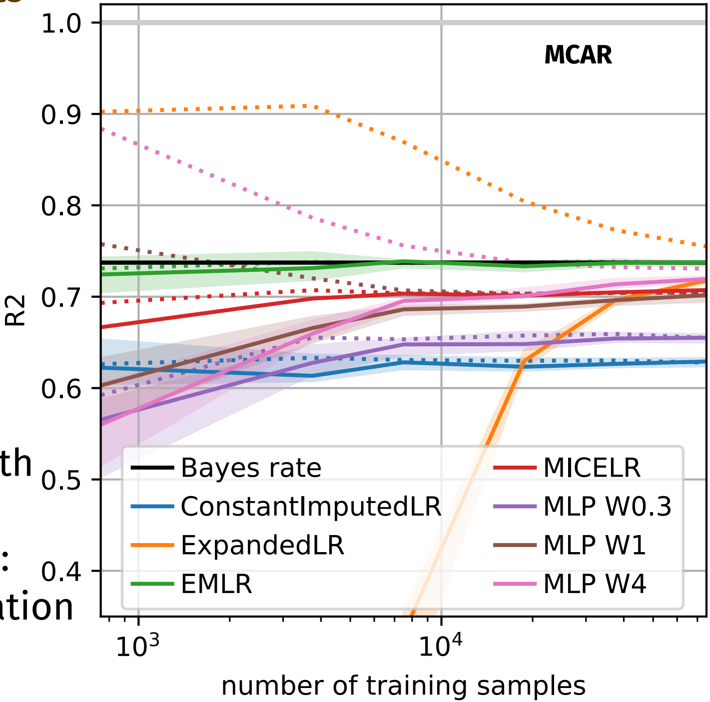
Experimental results

MICELR:
imputation

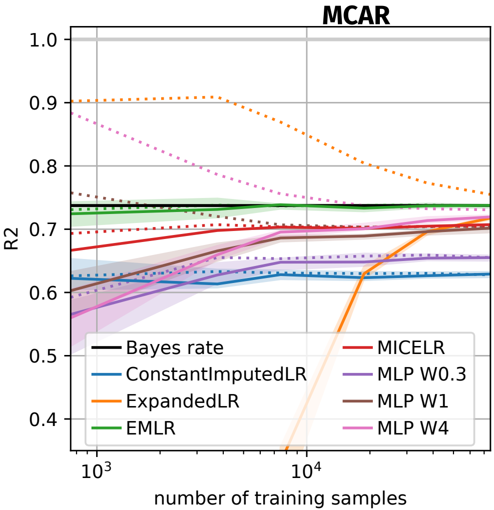
EMLR:
EM algorithm

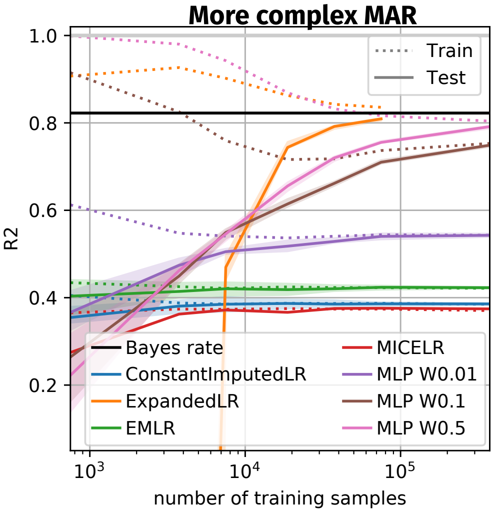
MLP W?:
MLP varying width

ConstantImputedLR:
Constant imputation

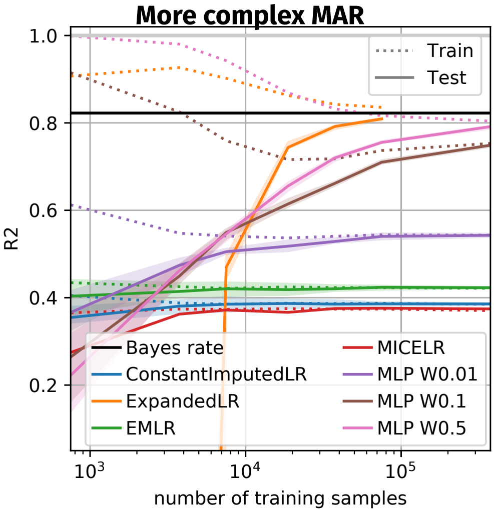


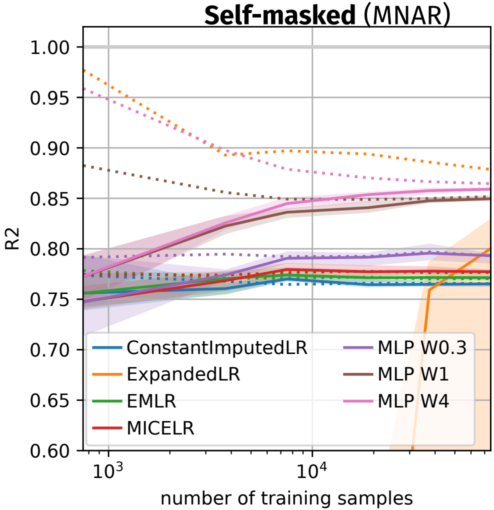
Experimental results





Experimental results





The Multi-layer perceptron is robust to violations of the model

Linear mechanism

- The linear predictor, even with constant imputation, is not consistent
- Basis expansion with polynomial of the mask is consistent, but $\mathcal{O}(2^d)$ sample complexity
- MLP is consistent, requiring 2^d width for high-entropy missing-values mechanism, but can adapt

4 Differentiable programming: a neural architecture

Craft a dedicated neural architecture to approximate the Bayes predictor

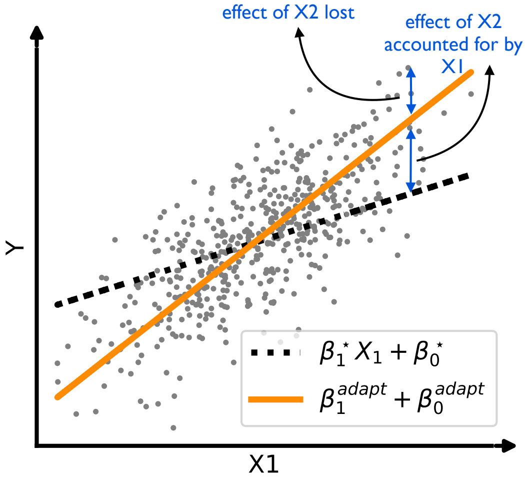
NeuMiss networks: differentiable programming for supervised learning with missing values [Le Morvan... 2020a]

Intuition: linear regression with missing values

$$Y = \beta_1^\star X_1 + \beta_2^\star X_2 + \beta_0^\star$$

$$\mathrm{cor}(X_1, X_2) = 0.5.$$

If X_2 is missing, the coefficient of X_1 should **compensate for the missingness of** X_2.



The difficulty of supervised learning with missing values is to handle **up to** 2^d missing data patterns (i.e. 2^d possible inputs of varying length).

Expression of Bayes predictor

Assumptions: Linear model: $Y = \beta_0^\star + \sum_{j=1}^{d} \beta_j^\star X_j + \epsilon$

Gaussian data: $X \sim \mathcal{N}(\mu, \Sigma)$

MCAR settings

$$f^\star(X_{obs}, M) = \beta_0^\star + \left\langle \beta_{obs}^\star, X_{obs} \right\rangle + \left\langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs} (\Sigma_{obs})^{-1} (X_{obs} - \mu_{obs}) \right\rangle$$

Expression of Bayes predictor

Assumptions: Linear model: $Y = \beta_0^\star + \sum_{j=1}^{d} \beta_j^\star X_j + \epsilon$

Gaussian data: $X \sim \mathcal{N}(\mu, \Sigma)$

MCAR settings

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \left\langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs} (\Sigma_{obs})^{-1} (X_{obs} - \mu_{obs}) \right\rangle$$

Gaussian self-masking settings

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \Big\langle \beta_{mis}^\star, (Id + D_{mis} \Sigma_{mis|obs}^{-1})^{-1} \times \Big(\tilde{\mu}_{mis} + D_{mis} \Sigma_{mis|obs}^{-1} (\mu_{mis} + \Sigma_{mis,obs} (\Sigma_{obs})^{-1} (X_{obs} - \mu_{obs})) \Big) \Big\rangle$$

Expression of Bayes predictor

Main difficulty: approx. of Σ_{obs}^{-1}, for any missing data pattern!

NeuMann iterations: approximate Σ_{obs}^{-1} by unrolling the order-ℓ truncation of a NeuMann series:

$$S_{obs(m)}^{(\ell)} = (Id - \Sigma_{obs(m)})S_{obs(m)}^{(\ell-1)} + Id.$$

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \left\langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \right\rangle$$

Gaussian self-masking settings

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \Big\langle \beta_{mis}^\star, (Id + D_{mis}\Sigma_{mis|obs}^{-1})^{-1}$$
$$\times \left(\tilde{\mu}_{mis} + D_{mis}\Sigma_{mis|obs}^{-1}(\mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}))\right)\Big\rangle$$

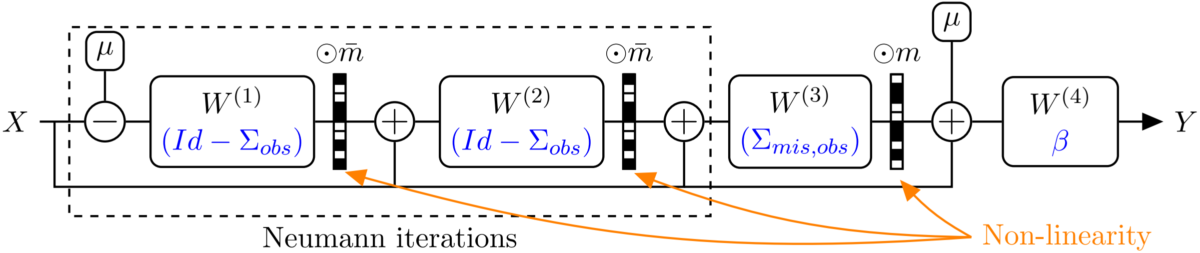
Proposition

Let ν be the smallest eigenvalue of Σ.
Assume that the spectral radius of Σ is < 1.

$$\mathbb{E}\left[\left(f^\star_\ell(X_{obs}, M) - f^\star(X_{obs}, M)\right)^2\right] \leq \frac{(1-\nu)^{2\ell}\|\beta^\star\|_2^2}{\nu}\,\mathbb{E}\left[\left\|Id - S^{(0)}_{obs(M)}\Sigma_{obs(M)}\right\|_2^2\right]$$

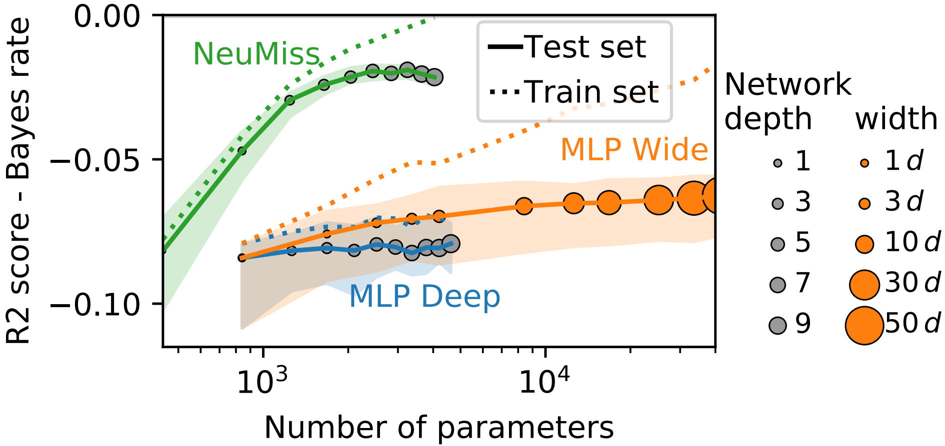
where $f_l^\star$ is the Bayes predictor, replacing the inverse by its order-l Neumann approximation.

NeuMiss: a dedicated architecture



A new type of non-linearity: the multiplication entrywise by the missingness indicator.

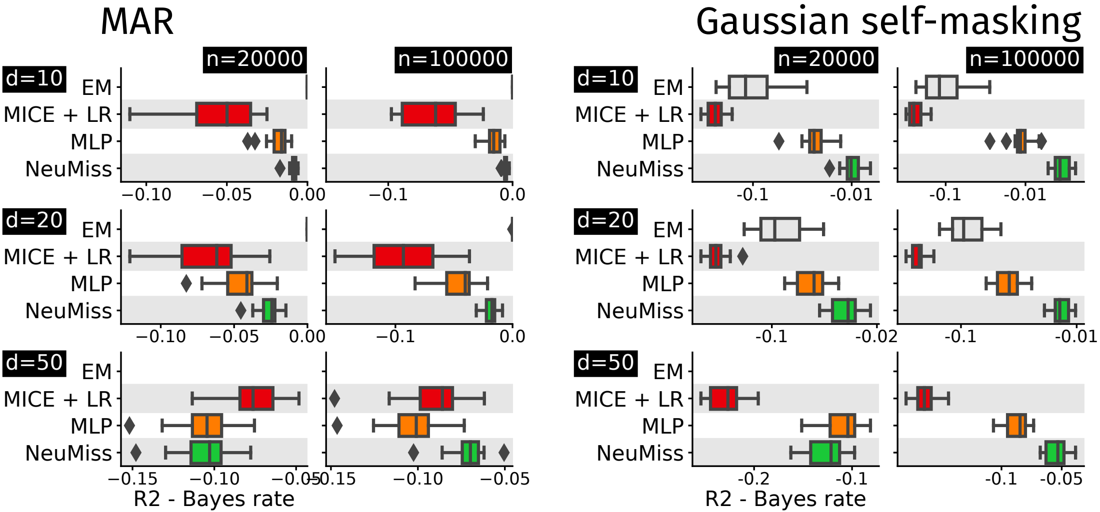
Empirical results: approximation efficiency



NeuMiss needs less samples to approximate well

(and predict well)

Empirical results: prediction performance



- NeuMiss prediction performance close to optimal
- NeuMiss is robust to the missing-data mechanism

Summary

Risk minimization good imputation is not necessary

Semi-discrete input $\Rightarrow$ optimization difficult

Formalisation [Josse... 2019]

Bayes predictors

2^d sub-models

$\Rightarrow$ complex model even for simple data-generating mechanisms

Tailored model:

functional form to capture dependencies between sub-models

Risk minimization can make it robust to missing-value mechanisms

References I

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

M. Le Morvan, J. Josse, T. Moreau, E. Scornet, and G. Varoquaux. Neumiss networks: differential programming for supervised learning with missing values. In *Advances in Neural Information Processing Systems 33*, 2020a.

M. Le Morvan, N. Prost, J. Josse, E. Scornet, and G. Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. *AISTATS*, 2020b.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.