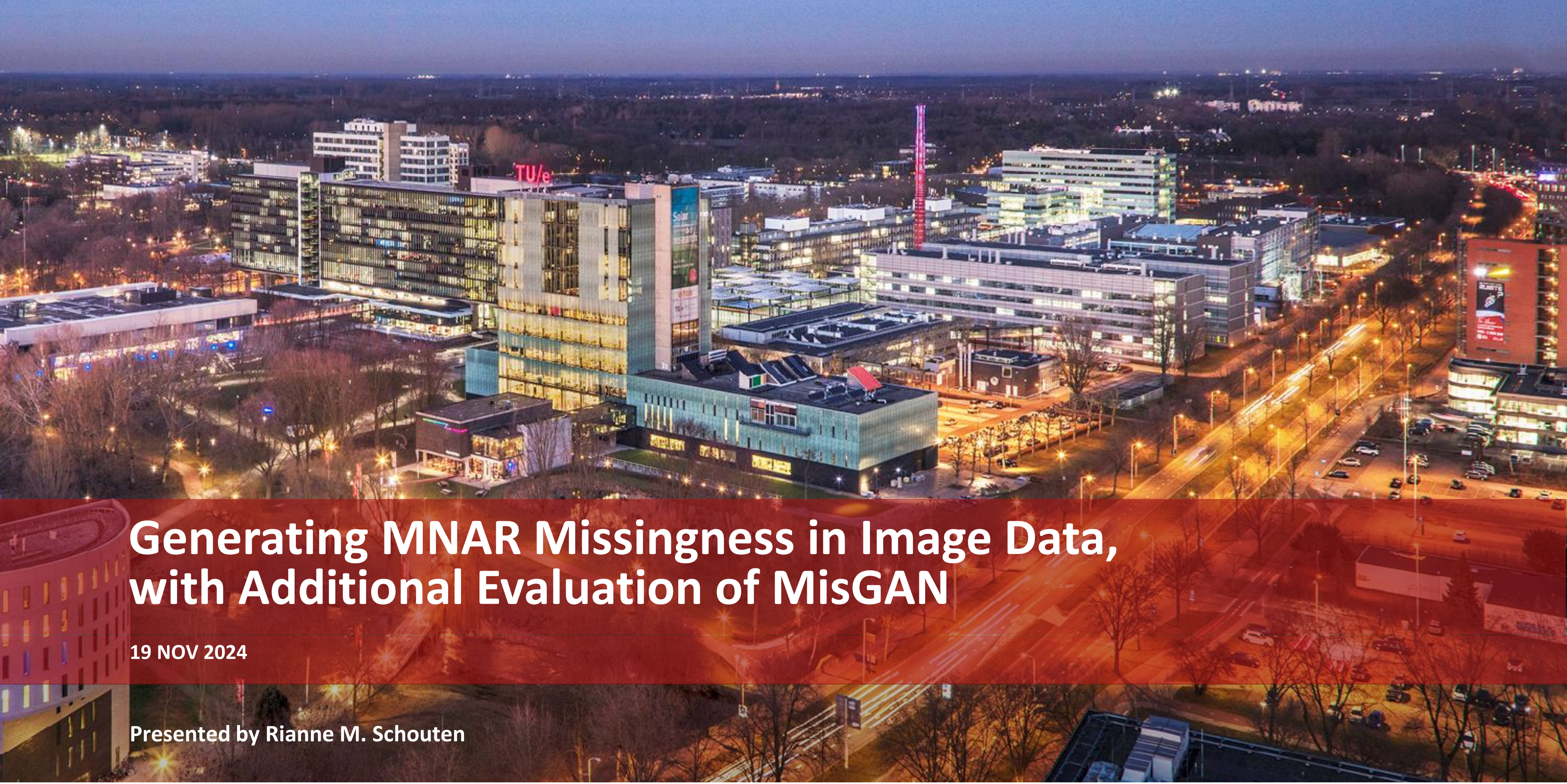


Generating MNAR Missingness in Image Data, with Additional Evaluation of MisGAN

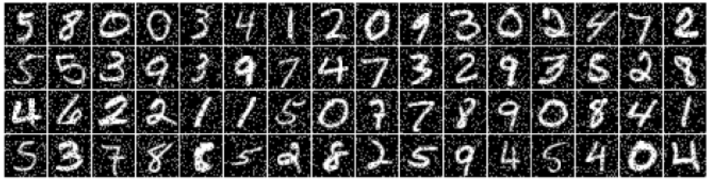
19 NOV 2024

Presented by Rianne M. Schouten

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

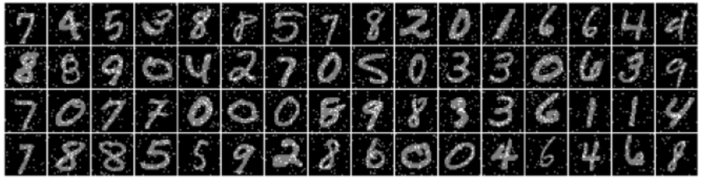
Our team from Eindhoven University of Technology

Natasha T.J. van den Berg, Bram O. Broekgaarden, Dionysia P.A. Mahieu, Jolijn G.M.J. Martens, Jonas M. Niederle, Rianne M. Schouten, Wouter Duivesteijn



(a) Masked images.

Fig. 4. Masked (top) and imputed (bottom) MNIST images under 15% missingness for MCAR. Masking is shown in gray.

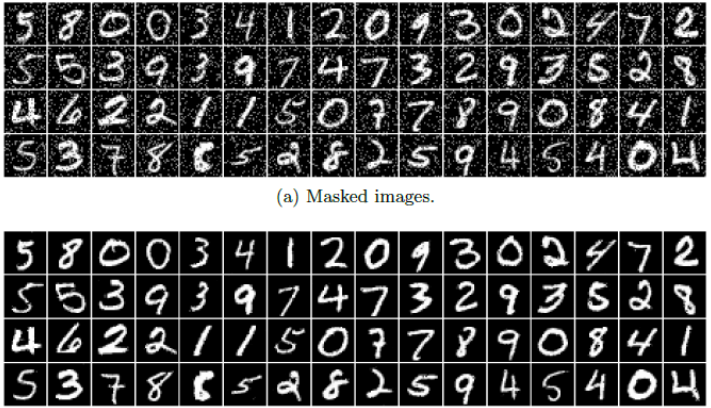


(a) Masked images.

Fig. 6. Masked (top) and imputed (bottom) MNIST images under 15% missingness for MNAR with $\rho = 0.2$. Masking is shown in gray.

Our team from Eindhoven University of Technology

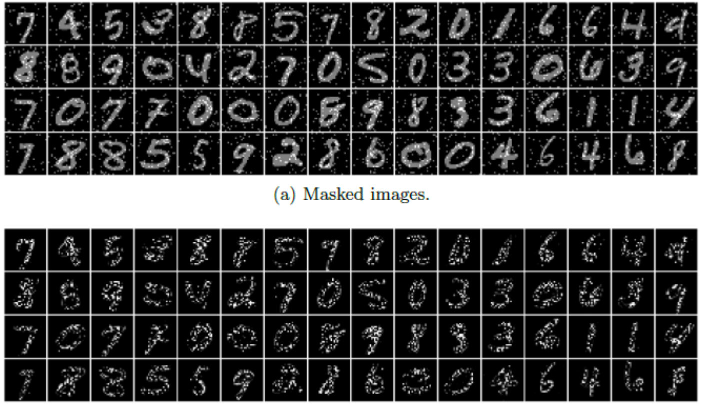
Natasha T.J. van den Berg, Bram O. Broekgaarden, Dionysia P.A. Mahieu, Jolijn G.M.J. Martens, Jonas M. Niederle, Rianne M. Schouten, Wouter Duivesteijn



(a) Masked images.

(b) Imputed images.

Fig. 4. Masked (top) and imputed (bottom) MNIST images under 15% missingness for MCAR. Masking is shown in gray.

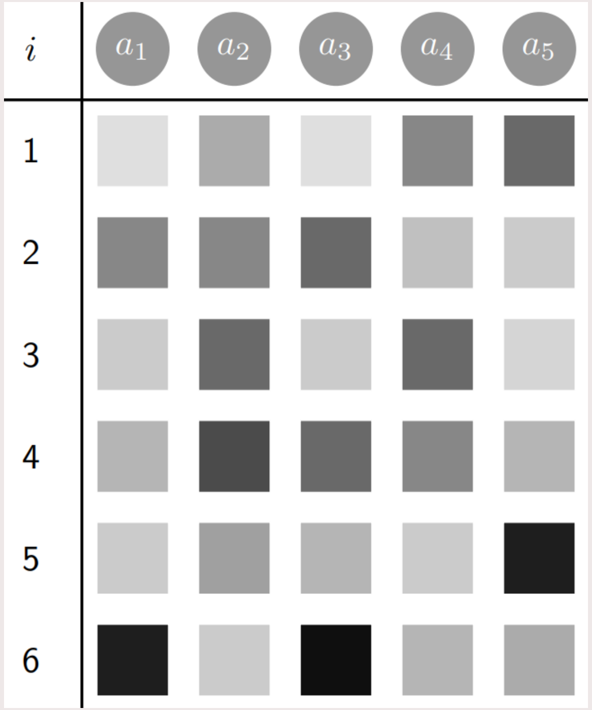


(a) Masked images.

(b) Imputed images.

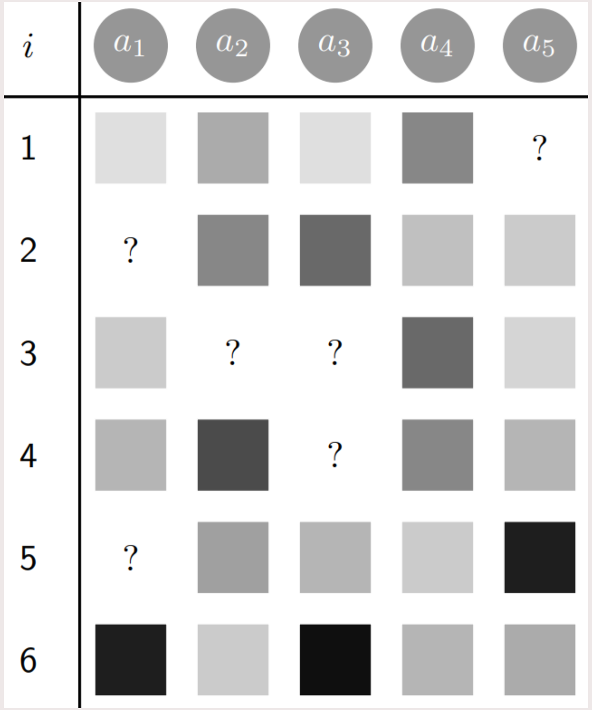
Fig. 6. Masked (top) and imputed (bottom) MNIST images under 15% missingness for MNAR with $\rho = 0.2$. Masking is shown in gray.

Missing data mechanisms



Dataset $D \in \mathbb{R}^{n \times d}$

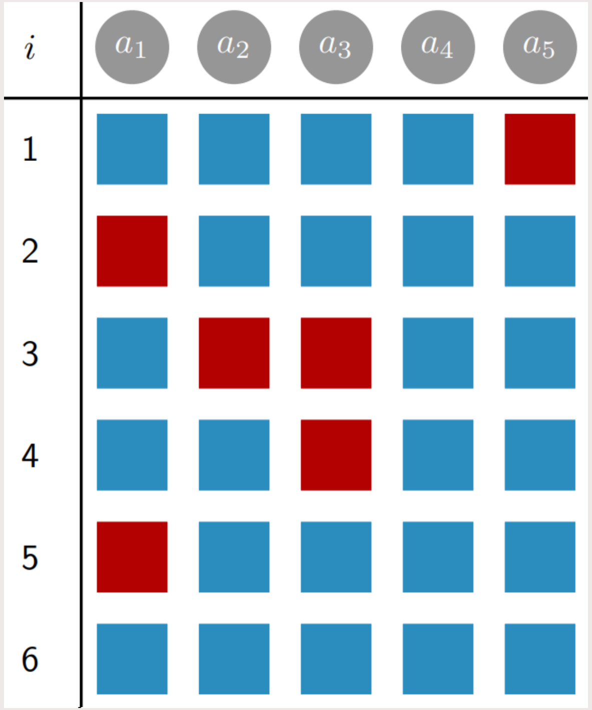
Missing data mechanisms



Dataset $D \in \mathbb{R}^{n \times p}$

Missing data indicator $\mathrm{R} \in \{0,1\}^{n \times d}, R_{ij} = 1$ if $D_{ij} = missing$

Missing data mechanisms



Dataset $D \in \mathbb{R}^{n \times p}$

Missing data indicator $\mathrm{R} \in \{0,1\}^{n \times d}, R_{ij} = 1$ if $D_{ij} = missing$

$D_{ij} \in \boldsymbol{D^{mis}}$ if $R_{ij} = 1$, $D_{ij} \in \boldsymbol{D^{obs}}$ if $R_{ij} = 0$

$D = \{\boldsymbol{D^{mis}}, \boldsymbol{D^{obs}}\}$

Missing data mechanism: $\Pr(\mathrm{R} \mid \boldsymbol{D^{mis}}, \boldsymbol{D^{obs}}, \psi)$

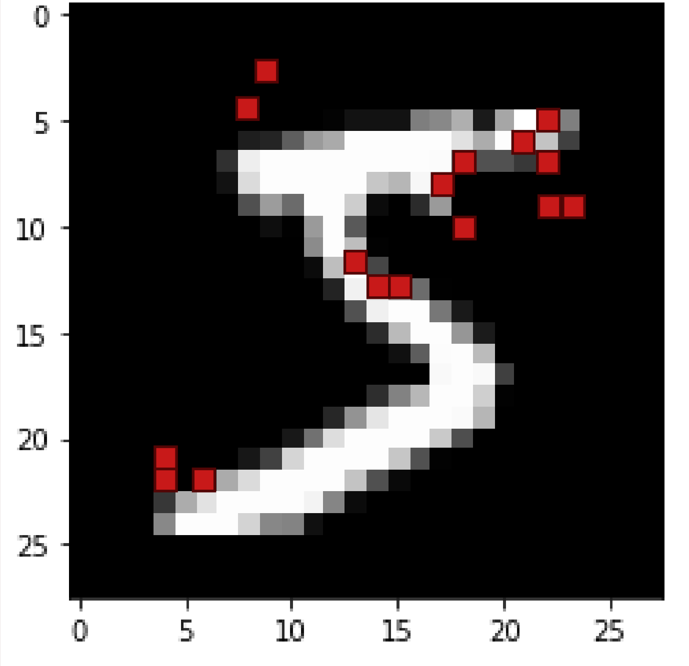
Missing Completely At Random : $\Pr(\mathrm{R} \mid \boldsymbol{D^{mis}}, \boldsymbol{D^{obs}}, \psi) = \Pr(\mathrm{R} \mid \psi)$

Missing At Random : $\Pr(\mathrm{R} \mid \boldsymbol{D^{mis}}, \boldsymbol{D^{obs}}, \psi) = \Pr(\mathrm{R} \mid \boldsymbol{D^{obs}}, \psi)$

Missing Not At Random : $\Pr(\mathrm{R} \mid \boldsymbol{D^{mis}}, \boldsymbol{D^{obs}}, \psi) = \Pr(\mathrm{R} \mid \boldsymbol{D^{mis}}, \boldsymbol{D^{obs}}, \psi)$

van Buuren, S.: Flexible imputation of missing data. CRC press (2018)

Missing data in images



Dataset $D = \{(\mathbf{x}_i, \mathbf{m}_i)\}_{i=1,2,..,n}$

An image $\mathbf{x}_i \in [0,1]^{28\times 28}$ (re-scaled from $\mathbb{R}$ to [0,1], black to white)

A mask $\mathbf{m}_i \in \{0,1\}^{28\times 28}$, $\mathbf{m}_{i,d} = 1$ if $\mathbf{x}_{i,d}$ is observed ($d = 1,2,\dots,28 \times 28$)

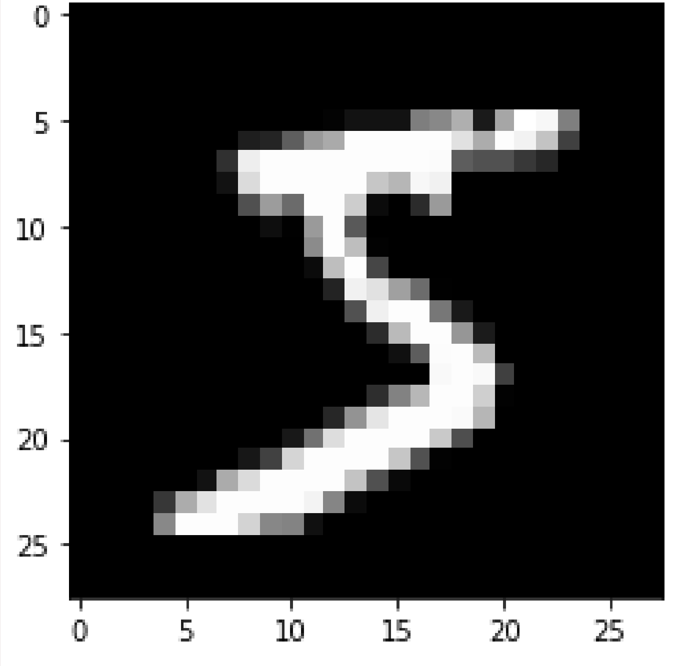
Remark that mask **m** is the complement of the missing data indicator R.

Amputation: the process of generating missing values in complete data

for tabular data

Schouten R.M., Lugtig, P. & Vink, G. (2018) Generating missing values for simulation purposes: A multivariate amputation procedure Journal of Statistical Computation and Simulation, 88(15): 1909-1930.

Amputation: the process of generating missing values in complete images



A complete image $X_i \in [0,1]^{28\times 28}$

A probability matrix $P_i \in [0,1]^{28\times 28}$ that contains missingness probabilities per pixel

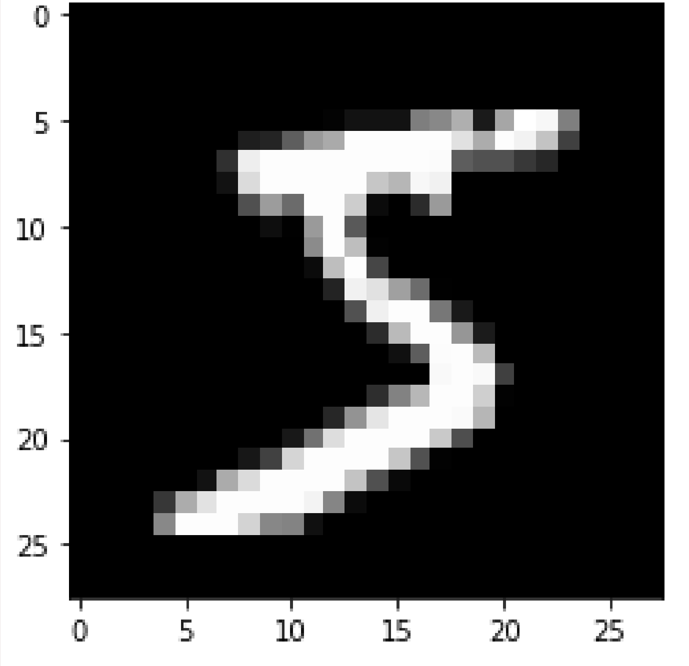
We sample from a Bernoulli distribution using the values in P,

$$\Pr\left(X_{i,d} \text{ is missing}\right) = P_{i,d} \; with \; d = 1,2,\ldots,28 \times 28$$

We then obtain an amputed (incomplete) image $\boldsymbol{X}_i \in [0,1]^{28\times 28}$ with corresponding mask M_i.

Amputation: Missing Completely At Random

$\Pr(\mathbf{M} \mid \mathbf{X}^{mis}, \mathbf{X}^{obs}, \psi) = \Pr(\mathbf{M} \mid \psi)$



A complete image $X_i \in [0,1]^{28\times 28}$

A probability matrix $P_i \in [0,1]^{28\times 28}$ that contains missingness probabilities per pixel

MCAR $P_{i,d} = \psi$ for all $d = 1,2,\ldots,28\times 28, \psi \in [0,1]$

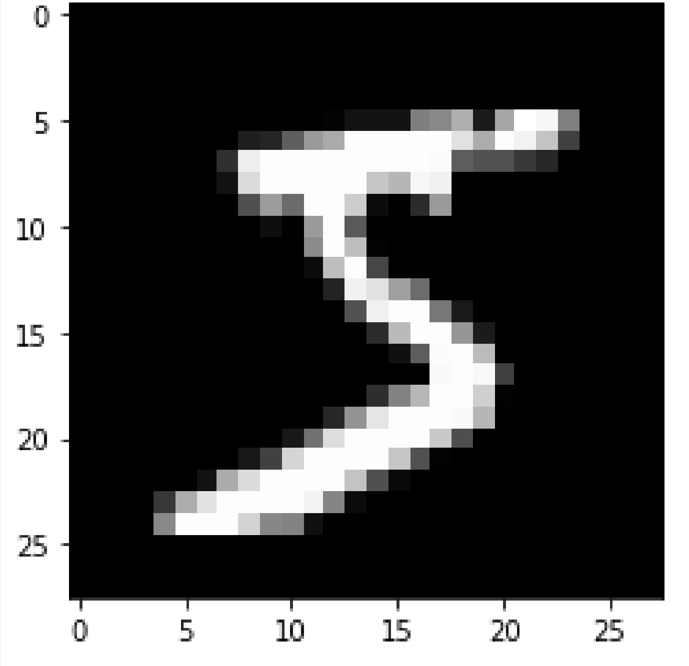
We sample from a Bernoulli distribution using the values in P,

$$\Pr\left(X_{i,d} \text{ is missing}\right) = P_{i,d} \; with \; d = 1,2,\ldots,28\times 28$$

TU/e

Amputation: Missing Not At Random

$$\Pr(\mathbf{M} \mid \mathbf{X}^{mis}, \mathbf{X}^{obs}, \psi) = \Pr(\mathbf{M} \mid \mathbf{X}^{mis}, \mathbf{X}^{obs}, \psi)$$



A complete image $X_i \in [0,1]^{28\times 28}$

A probability matrix $P_i \in [0,1]^{28\times 28}$ that contains missingness probabilities per pixel

MCAR $P_{i,d} = \psi$ for all $d = 1,2,\ldots,28 \times 28, \psi \in [0,1]$

MNAR $P_{i,d} = f(X_{i,d}, \psi)$

$$P_i = (\mu_i \mathrm{J} - X_i)\rho + X_i + c\mathrm{J}$$

We sample from a Bernoulli distribution using the values in P,

$$\Pr\left(X_{i,d} \text{ is missing}\right) = P_{i,d} \; with \; d = 1,2,\ldots,28 \times 28$$

TU/e

Our proposed image amputation procedure

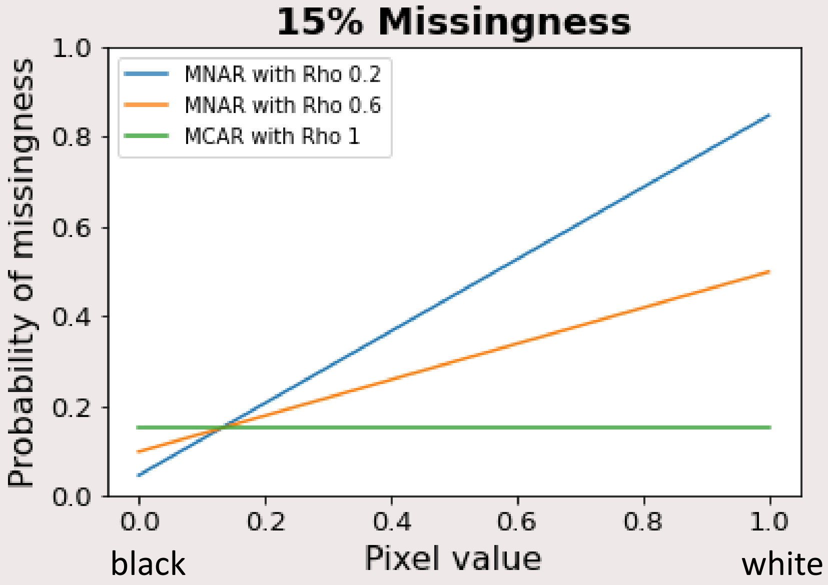
$$P_i = (\mu_i \mathrm{J} - X_i)\rho + X_i + c\mathrm{J}$$

$\mathrm{J} = \mathbb{I}^{28\times 28}$ is an all-ones matrix

μ_i is the mean pixel value of image X_i

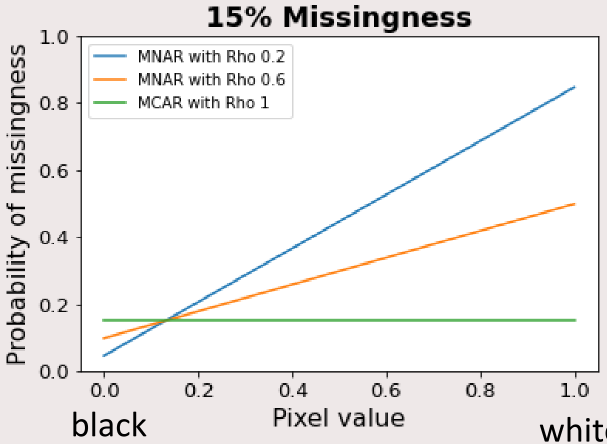
c is a constant that controls the overall missingness percentage over all images

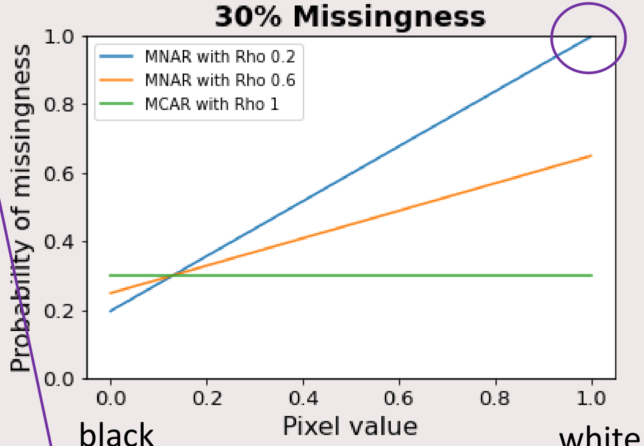
$\rho \in [0,1]$ determines the extent to which the missingness depends on the pixel value



Our proposed image amputation procedure

$$P_i = (\mu_i \mathrm{J} - X_i)\rho + X_i + c\mathrm{J}$$





black white black white

μ_D = 0.131

c = 0.019

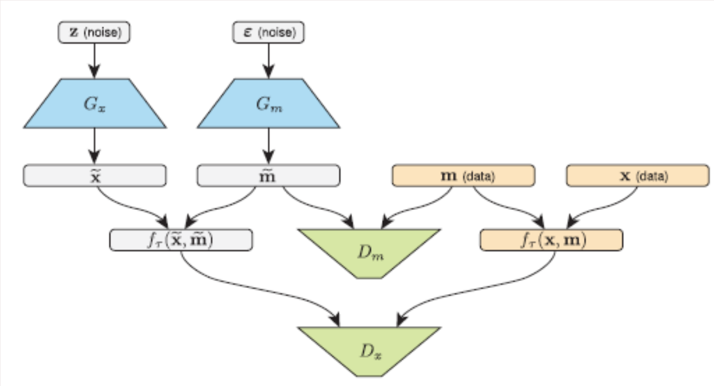
RIGHT type

c = 0.169

Schouten, R.M. and Vink, G. (2021) The dance of the mechanisms: How observed information influences the validity of missingness assumptions Sociological Methods & Research, 50(3): 1243-1258.

Evaluation

1. Take MNIST dataset, n = 60.000.
2. Apply amputation procedure for MCAR, MNAR 0.2 and MNAR 0.6.
3. Impute using MisGAN



Li, S.C., Jiang, B., Marlin, B.M.: MisGAN: Learning from incomplete data with generative adversarial networks. In: Proc. ICLR (2019)

f_τ imputes missing values with τ

Evaluation

1. Take MNIST dataset, n = 60.000.
2. Apply amputation procedure for MCAR, MNAR 0.2 and MNAR 0.6.
3. Impute using MisGAN
4. Calculate imputation accuracy of non-black pixel values

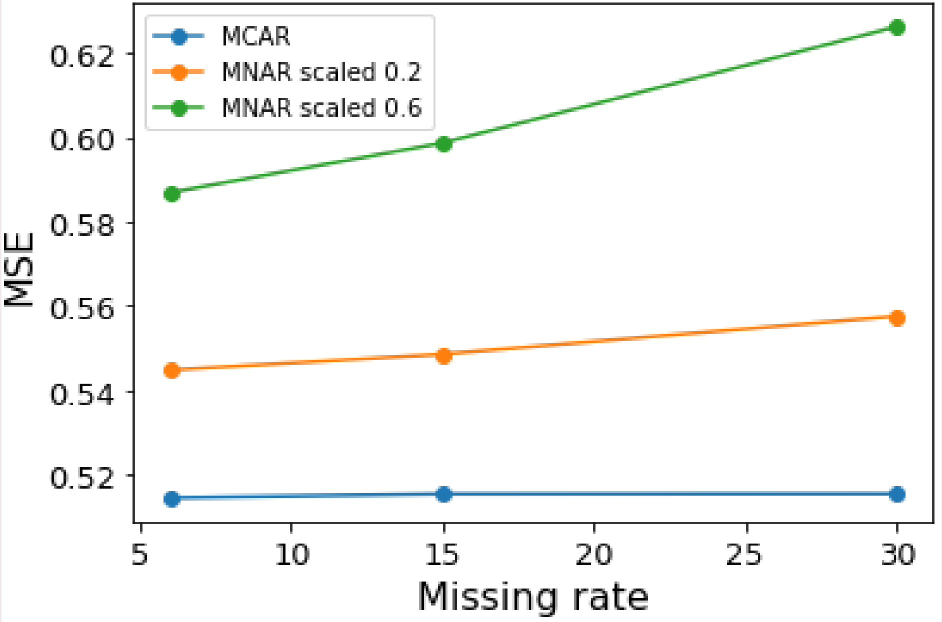
Alternatives are:
- prediction accuracy (can we still predict the class labels?)
- validity and efficiency of statistical estimates (but more applicable in the context of tabular data)-
- evaluating quality of generated images (FID)

Evaluation

1. Take MNIST dataset, n = 60.000.
2. Apply amputation procedure for MCAR, MNAR 0.2 and MNAR 0.6.
3. Impute using MisGAN
4. Calculate imputation accuracy of non-black pixel values

$$\mathrm{MSE}(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) = \frac{1}{k_i} \sum_{j=0}^{k_i - 1} \left((\boldsymbol{y}_i)_j - (\hat{\boldsymbol{y}}_i)_j \right)^2$$

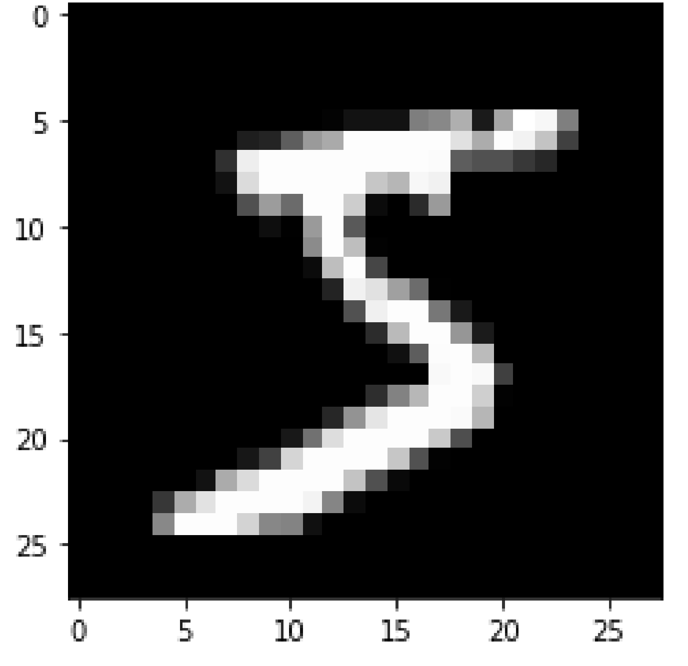
Results



Imputation MSE after MCAR amputation is not affected by missingness percentage.

Imputation MSE after MNAR amputation is affected by missingness percentage. The stronger the MNAR effect (0.2 vs 0.6), the more missingness percentage influences the imputation accuracy.

Future work: MAR



A complete image $X_i \in [0,1]^{28\times28}$

A probability matrix $P_i \in [0,1]^{28\times28}$ that contains missingness probabilities per pixel

MCAR $P_{i,d} = \psi$ for all $d = 1,2,\ldots,28 \times 28, \psi \in [0,1]$

MAR ←

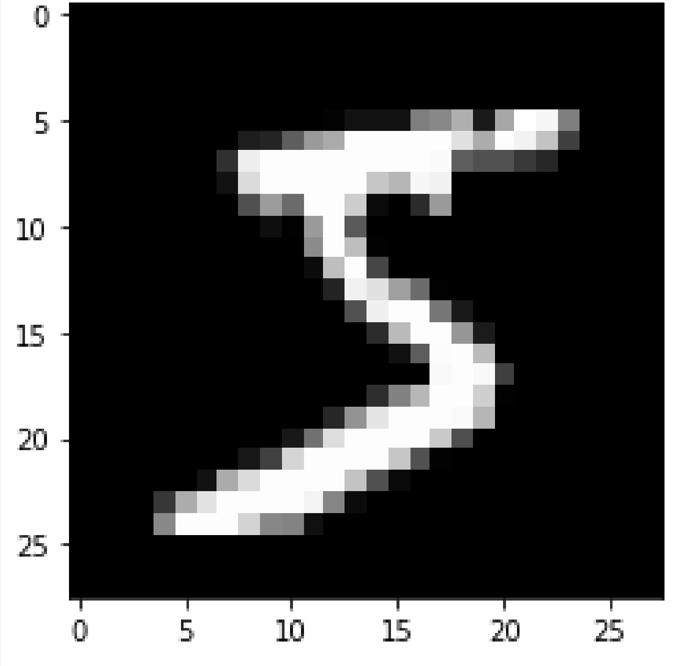
MNAR $P_{i,d} = f(X_{i,d}, \psi)$

$$P_i = (\mu_i J - X_i)\rho + X_i + cJ$$

We sample from a Bernoulli distribution using the values in P,

$$\Pr(X_{i,d} \text{ is missing}) = P_{i,d} \; with \; d = 1,2,\ldots,28 \times 28$$

Future work: MAR + consider mechanisms defined over entire image dataset



A complete image $X_i \in [0,1]^{28\times28}$

A probability matrix $P_i \in [0,1]^{28\times28}$ that contains missingness probabilities per pixel

MCAR $P_{i,d} = \psi$ for all $d = 1,2,\dots,28\times28, \psi \in [0,1]$

MAR $\longleftarrow$

MNAR $P_{i,d} = f(X_{i,d}, \psi)$

$$P_i = (\mu_i \mathrm{J} - X_i)\rho + X_i + c\mathrm{J}$$

We sample from a Bernoulli distribution using the values in P,

$$\Pr\left(X_{i,d} \text{ is missing}\right) = P_{i,d}\ with\ d = 1,2,\dots,28\times28$$

TU/e

Thank you!

r.m.schouten@tue.nl
https://rianneschouten.github.io

Van den Berg, N. T., Broekgaarden, B. O., Mahieu Dionysia, P., Martens, J. G., Niederle, J., **Schouten, R.M.**, & Duivesteijn, W. (2024) Generating MNAR missingness in image data, with additional evaluation of MisGAN.

Repository: https://github.com/RianneSchouten/misgan_mnar/