



Imputation Scores

Meta-Lina Spohn, Jeffrey Näf

ETH Zurich

16.05.2022

Joint work with



(a) *Loris Michel*



(b) *Nicolai Meinshausen*

Outline

- 1 Problem Motivation and Toy Example
- 2 A Solution: Imputation Scores (I-Scores)
- 3 A Specific I-Score: DR I-Score
- 4 Empirical Results


- 1 Problem Motivation and Toy Example
- 2 A Solution: Imputation Scores (I-Scores)
- 3 A Specific I-Score: DR I-Score
- 4 Empirical Results

Problem Motivation

- How to choose the ‘best’ imputation method in a given application?

Problem Motivation

- As noted in Muzellec et al.¹: “A desirable property of imputation methods is that they should preserve the joint and marginal distributions.”

¹Boris Muzellec et al. “Missing data imputation using optimal transport”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7130–7140. 

Problem Motivation

- As noted in Muzellec et al.¹: “A desirable property of imputation methods is that they should preserve the joint and marginal distributions.”
- *Goal*: If x^* is a complete observation and x an imputed observation, want them to be realizations of the same distribution.

¹Muzellec et al., “Missing data imputation using optimal transport”.

Problem Motivation

- As noted in Muzellec et al.¹: “A desirable property of imputation methods is that they should preserve the joint and marginal distributions.”
- *Goal*: If x^* is a complete observation and x an imputed observation, want them to be realizations of the same distribution.

⇒ Imputation Scores (I-Scores).

¹Muzellec et al., “Missing data imputation using optimal transport”.

Toy Example

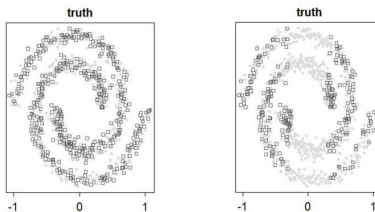


Figure 2: *Left: MCAR, Right: MAR*

- Toy motivating example: noisy version of the spiral in 2D.

Toy Example

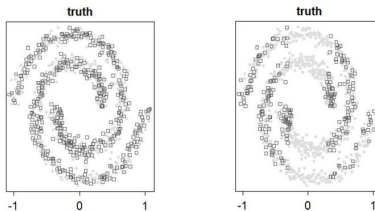


Figure 2: *Left: MCAR, Right: MAR*

- Toy motivating example: noisy version of the spiral in 2D.
- Generated 1000 observations, then applied MAR and MCAR mechanisms to set observations to NA.

Toy Example

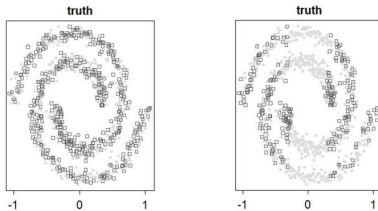


Figure 2: *Left: MCAR, Right: MAR*

- Toy motivating example: noisy version of the spiral in 2D.
- Generated 1000 observations, then applied MAR and MCAR mechanisms to set observations to NA.
- MCAR: every value is set to NA with $p_{miss} = 0.3$.

Toy Example

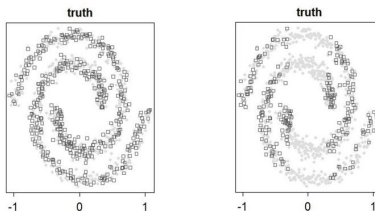


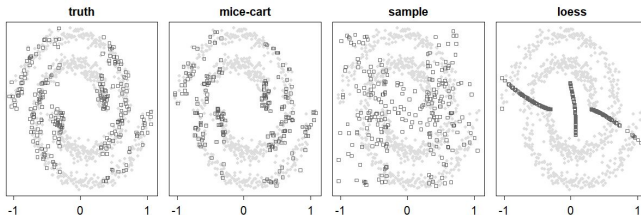
Figure 2: *Left: MCAR, Right: MAR*

- Toy motivating example: noisy version of the spiral in 2D.
- Generated 1000 observations, then applied MAR and MCAR mechanisms to set observations to NA.
- MCAR: every value is set to NA with $p_{miss} = 0.3$.
- MAR: the variable X_2 is set to NA with $p_{miss} = 0.3$ if the corresponding $|X_1| > 0.3$ and observed otherwise. The variable X_1 is set to NA with $p_{miss} = 0.3$ if $|X_2| < 0.3$.

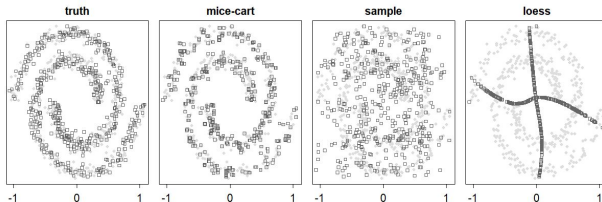
Toy Example

- Apply three imputation methods:
 - ① “loess”: estimating conditional expectations $\mathbb{E}[X_1|X_2]$ and $\mathbb{E}[X_2|X_1]$ on the complete cases with local regression, and imputing by predicting from X_1 (if X_2 is missing) or from X_2 (if X_1 is missing),
 - ② “sample”: random sampling an observed value for each missing entry,
 - ③ “mice-cart”: mice combined with a single tree in each iteration.

Toy Example



(a) MAR



(b) MCAR

Toy Example

- Try to quantify this visually obtained ranking: compute our *DR I-Score* (defined later) for all methods.

Toy Example

- Try to quantify this visually obtained ranking: compute our *DR I-Score* (defined later) for all methods.
- Also compute the negative of RMSE (“-RMSE”).

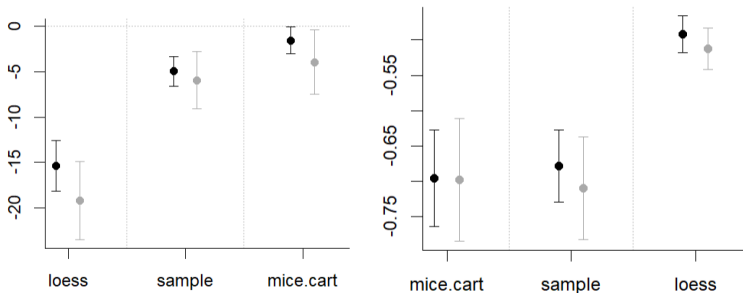
Toy Example

- Try to quantify this visually obtained ranking: compute our *DR I-Score* (defined later) for all methods.
- Also compute the negative of RMSE (“-RMSE”).
- *Note*: -RMSE is computed using the unobserved full data set, while our score does not.

Toy Example

- Try to quantify this visually obtained ranking: compute our *DR I-Score* (defined later) for all methods.
- Also compute the negative of RMSE (“-RMSE”).
- *Note*: -RMSE is computed using the unobserved full data set, while our score does not.
- Add approximated two-sided 95%-confidence intervals (CI) by Jackknife 1/2-subsampling.

Toy Example



(a) Estimated DR I-Score with Cls.

(b) Negative RMSE with Cls.

Figure 4: Black is MCAR and grey MAR case. In (a) we subtracted the score of the true data from the scores of the methods, thus the line at 0 represents the true data score.

- 1 Problem Motivation and Toy Example
- 2 A Solution: Imputation Scores (I-Scores)**
- 3 A Specific I-Score: DR I-Score
- 4 Empirical Results

Imputation Scores

- P refers to the distribution of X with missing values
- $P^* \in \mathcal{P}$ refers to the distribution of X^* without missing values.
- H refers to an imputation distribution.

Definition

Imputation Score (I-Score)

For a function $S_{NA}(\cdot, P) : \mathbb{R}^d \mapsto \mathbb{R}$ we define

$$S_{NA}(H, P) := \mathbb{E}_{X \sim H}[S_{NA}(X, P)],$$

the expectation over $X \sim H$. Such a function $S_{NA}(H, P)$ is a *proper I-Score* iff

$$S_{NA}(H, P) \leq S_{NA}(P^*, P),$$

for any imputation distribution H .

- 1 Problem Motivation and Toy Example
- 2 A Solution: Imputation Scores (I-Scores)
- 3 A Specific I-Score: DR I-Score**
- 4 Empirical Results

Density Ratio (DR) I-Score

Imputed Distribution

X1	X2	X3
56	41	-29
-26	77	-30
39	-2	-41
-21	-55	25
15	46	-89
-43	30	44
51	-38	-124
-55	-41	-22
-33	0	38
26	-81	13

Projection onto X2, X3

Imputed values
on projection

X2	X3
77	-30
-2	-41
-55	25
46	-89
-38	-124

Observed Distribution

X1	X2	X3
56	41	-29
-26	77	N/A
39	N/A	N/A
N/A	-55	N/A
N/A	N/A	-89
-43	30	44
N/A	-38	N/A
N/A	-41	-22
-33	0	38
N/A	-81	13

Fully observed
values on projection

Label 0

RF
classifier

$p/(1-p)$
(Density Ratio Estimate)

Estimate of
KL-Divergence

Average over
different projections

I-Score

X2	X3
41	-29
30	44
-41	-22
0	38
-81	13

Label 1

Figure 5: Illustration of the DR I-Score Algorithm

DR I-Score

Overall Summary: Average distributional differences between imputed and complete samples over different projections.

DR I-Score

Overall Summary: Average distributional differences between imputed and complete samples over different projections.

What to compare?

DR I-Score

Overall Summary: Average distributional differences between imputed and complete samples over different projections.

What to compare?

- *Problem:* Observe only incomplete data (access to P , not to P^*).

DR I-Score

Overall Summary: Average distributional differences between imputed and complete samples over different projections.

What to compare?

- *Problem:* Observe only incomplete data (access to P , not to P^*).
- ⇒ Idea: Compare imputed observations to fully observed ones.

DR I-Score

Overall Summary: Average distributional differences between imputed and complete samples over different projections.

What to compare?

- *Problem:* Observe only incomplete data (access to P , not to P^*).
- ⇒ Idea: Compare imputed observations to fully observed ones.
- That is we compare

fully observed samples from P , i.e. $P \mid M = \mathbf{0}$

against

samples with pattern $M = m$ from H , i.e. $H \mid M = m$.

DR I-Score

Example with 3 patterns: $M = (0, 0, 0)$, $M = (1, 0, 0)$, $M = (1, 1, 0)$:

13	10	3
NA	13	7
5	16	9
NA	NA	4
NA	5	8
-1	15	11
NA	16	18
NA	NA	10
9	14	16
NA	14	5

13	10	3
6	13	7
5	16	9
10	8	4
7	5	8
-1	15	11
14	16	18
17	12	10
9	14	16
12	14	5

Two comparisons:

13	10	3
5	16	9
-1	15	11
9	14	16

to

6	13	7
7	5	8
14	16	18
12	14	5

to

13	10	3
5	16	9
-1	15	11
9	14	16

10	8	4
17	12	10

DR I-Score

- *Problem 1*: There might be only very few fully observed samples.

DR I-Score

- *Problem 1*: There might be only very few fully observed samples.
- ⇒ Idea: Use random projections in the variable space.

DR I-Score

- *Problem 1*: There might be only very few fully observed samples.
- ⇒ Idea: Use random projections in the variable space.
- *Problem 2*: Need to measure the distance between two distributions in a way that yields enough detection power to be meaningful.

DR I-Score

- *Problem 1*: There might be only very few fully observed samples.
- ⇒ Idea: Use random projections in the variable space.
- *Problem 2*: Need to measure the distance between two distributions in a way that yields enough detection power to be meaningful.
- ⇒ Idea: Use the Kullback-Leibler Divergence (KL-Divergence), estimated from samples using a classifier.

DR I-Score

The Density Ratio (DR) I-Score contains 3 main steps:

- *Step 1*: Random projections to solve Problem 1
- *Step 2*: KL-Divergence estimation to solve Problem 2
- *Step 3*: Forming an overall score

DR I-Score

Step 1: Random projections to solve Problem 1

- In most data sets of reasonable size, it is hard to find complete cases.

DR I-Score

Step 1: Random projections to solve Problem 1

- In most data sets of reasonable size, it is hard to find complete cases.
- Considering observations that are projected into a lower-dimensional space allows us to recover more complete cases.

DR I-Score

Step 1: Random projections to solve Problem 1

- In most data sets of reasonable size, it is hard to find complete cases.
- Considering observations that are projected into a lower-dimensional space allows us to recover more complete cases.
- Example: $x = (\text{NA}, 1, \text{NA}, 2)$ is not complete, but if we project it to dimensions $A = \{2, 4\}$, $x_A = (1, 2)$ is complete.

DR I-Score

Step 2: KL-Divergence estimation to solve Problem 2

- For two densities p , h the negative KL-Divergence is

$$-D_{KL}(p, h) = \mathbb{E}_{X \sim H} \left[\log \left(\frac{p(X)}{h(X)} \right) \right].$$

DR I-Score

Step 2: KL-Divergence estimation to solve Problem 2

- For two densities p , h the negative KL-Divergence is

$$-D_{KL}(p, h) = \mathbb{E}_{X \sim H} \left[\log \left(\frac{p(X)}{h(X)} \right) \right].$$

- Need to estimate a density ratio.

DR I-Score

Step 2: KL-Divergence estimation to solve Problem 2

- For two densities p , h the negative KL-Divergence is

$$-D_{KL}(p, h) = \mathbb{E}_{X \sim H} \left[\log \left(\frac{p(X)}{h(X)} \right) \right].$$

- Need to estimate a density ratio.
- Instead of estimating p and h , we estimate $p(x)/h(x)$ directly using *classification*.

DR I-Score

Step 3: Forming an overall score

- The final score of an imputation is built out of estimated negative KL-Divergences, averaged over multiple random projections.

DR I-Score

Step 3: Forming an overall score

- The final score of an imputation is built out of estimated negative KL-Divergences, averaged over multiple random projections.
- Can apply the I-Scores algorithm to multiple imputation methods and select the one with the highest score.

DR I-Score

Step 3: Forming an overall score

- The final score of an imputation is built out of estimated negative KL-Divergences, averaged over multiple random projections.
 - Can apply the I-Scores algorithm to multiple imputation methods and select the one with the highest score.
- ⇒ I-Score can guide the selection of the best imputation for a data set at hand.

DR I-Score

Step 3: Forming an overall score

- The final score of an imputation is built out of estimated negative KL-Divergences, averaged over multiple random projections.
 - Can apply the I-Scores algorithm to multiple imputation methods and select the one with the highest score.
- ⇒ I-Score can guide the selection of the best imputation for a data set at hand.
- ⇒ R-package `Iscores` available on CRAN.

DR I-Score

Imputed Distribution

X1	X2	X3
56	41	-29
-26	77	-30
39	-2	-41
-21	-55	25
15	46	-89
-43	30	44
51	-38	-124
-55	-41	-22
-33	0	38
26	-81	13

Projection onto X2, X3

Imputed values on projection

X2	X3
77	-30
-2	-41
-55	25
46	-89
-38	-124

Observed Distribution

X1	X2	X3
56	41	-29
-26	77	N/A
39	N/A	N/A
N/A	-55	N/A
N/A	N/A	-89
-43	30	44
N/A	-38	N/A
N/A	-41	-22
-33	0	38
N/A	-81	13

Fully observed values on projection

Label 0

X2	X3
41	-29
30	44
-41	-22
0	38
-81	13

Label 1

RF classifier

$p/(1-p)$
(Density Ratio Estimate)

Estimate of KL-Divergence

Average over different projections

I-Score

Figure 6: Illustration of the DR I-Score Algorithm

DR I-Score

Given a projection A and a pattern M_A on A , we define for $X_A \sim H_A | M_A = m_A$,

$$S_{NA}^*(X_A, P_A; M_A) = \log \left(\frac{p_A(X_A | M_A = \mathbf{0})}{h_A(X_A | M_A = m_A)} \right).$$

Definition

Density Ratio I-Score

We define the DR I-Score of the imputation distribution H by

$$S_{NA}^*(H, P) = \mathbb{E}_{A \sim \mathcal{X}, M_A \sim P_A^M, X_A \sim H_{M_A}} [S_{NA}^*(X_A, P_A; M_A)].$$

Theoretical Consideration: Propriety

- The population version of the DR I-Score is proper, that is it holds

$$S_{NA}^*(H, P) \leq S_{NA}^*(P^*, P).$$

Theoretical Consideration: Propriety

- The population version of the DR I-Score is proper, that is it holds

$$S_{NA}^*(H, P) \leq S_{NA}^*(P^*, P).$$

- This is true under *missing at random* (MAR) on every projection in the set \mathcal{A} of possible projections.

Theoretical Consideration: Propriety

- The population version of the DR I-Score is proper, that is it holds

$$S_{NA}^*(H, P) \leq S_{NA}^*(P^*, P).$$

- This is true under *missing at random (MAR)* on every projection in the set \mathcal{A} of possible projections.
- In particular, it is valid if
 - (i) the missingness mechanism is MCAR,
 - (ii) the missingness mechanism is MAR and $\mathcal{A} = \{1, \dots, p\}$,
 - (iii) it is known that blocks of data are jointly MAR, and the set of projections \mathcal{A} is chosen such that the blocks are contained as a whole in the projections.

Theoretical Consideration: Propriety

- The population version of the DR I-Score is proper, that is it holds

$$S_{NA}^*(H, P) \leq S_{NA}^*(P^*, P).$$

- This is true under *missing at random (MAR)* on every projection in the set \mathcal{A} of possible projections.
- In particular, it is valid if
 - (i) the missingness mechanism is MCAR,
 - (ii) the missingness mechanism is MAR and $\mathcal{A} = \{1, \dots, p\}$,
 - (iii) it is known that blocks of data are jointly MAR, and the set of projections \mathcal{A} is chosen such that the blocks are contained as a whole in the projections.
- This condition is both surprising and revealing.

Theoretical Consideration: Propriety

For a missingness pattern $m \in \mathcal{M}$, $o(x, m) = (x_j)_{j \in \{1, \dots, d\}: m_j=0}$ subsets the observed elements of x according to m , while $o^c(x, m) = (x_j)_{j \in \{1, \dots, d\}: m_j=1}$, subsets the missing elements.

Theoretical Consideration: Propriety

For a missingness pattern $m \in \mathcal{M}$, $o(x, m) = (x_j)_{j \in \{1, \dots, d\}: m_j=0}$ subsets the observed elements of x according to m , while $o^c(x, m) = (x_j)_{j \in \{1, \dots, d\}: m_j=1}$, subsets the missing elements.

$$x = (x_1, x_2, x_3, x_4, x_5), \quad m = (1, 1, 0, 1, 0)$$

$$\implies o(x, m) = (x_3, x_5)$$

$$\implies o^c(x, m) = (x_1, x_2, x_4)$$

Theoretical Consideration: Back to the Example

Example with 3 patterns: $M = (0, 0, 0)$, $M = (1, 0, 0)$, $M = (1, 1, 0)$:

13	10	3
NA	13	7
5	16	9
NA	NA	4
NA	5	8
-1	15	11
NA	16	18
NA	NA	10
9	14	16
NA	14	5

13	10	3
6	13	7
5	16	9
10	8	4
7	5	8
-1	15	11
14	16	18
17	12	10
9	14	16
12	14	5

Two comparisons:

13	10	3
5	16	9
-1	15	11
9	14	16

to

6	13	7
7	5	8
14	16	18
12	14	5

13	10	3
5	16	9
-1	15	11
9	14	16

to

10	8	4
17	12	10

Theoretical Consideration: Propriety

Thus on the right, we compare

13	10	3
5	16	9
-1	15	11
9	14	16

to

6	13	7
7	5	8
14	16	18
12	14	5

- The left observations have distribution $P | M = (0, 0, 0)$, while the right have $H | M = (1, 0, 0)$. These will be different in general.
- With the KL-Divergence it is enough to assume

$$p^*(o^c(x, m) | o(x, m), M = \mathbf{0}) = p^*(o^c(x, m) | o(x, m), M = m),$$

to obtain propriety (in this case $m = (1, 0, 0)$).

Theoretical Consideration: Propriety

With the KL-Divergence it is enough to assume

$$p^*(o^c(x, m) | o(x, m), M = \mathbf{0}) = p^*(o^c(x, m) | o(x, m), M = m),$$

to obtain propriety (in this case $m = (1, 0, 0)$).

This is because

$$\text{KL}(\text{pattern } m) = \text{KL}(\text{observed in pattern } m) + \mathbb{E}_{\text{observed}}[\text{KL}(\text{missing given observed in pattern } m)]$$

Theoretical consideration: Propriety

- It is not enough to have MAR overall (unless the set of projections just corresponds to the full projection).

Theoretical consideration: Propriety

- It is not enough to have MAR overall (unless the set of projections just corresponds to the full projection).
- Once we start removing variables, a MAR data set can become MNAR.

Theoretical consideration: Propriety

- It is not enough to have MAR overall (unless the set of projections just corresponds to the full projection).
- Once we start removing variables, a MAR data set can become MNAR.
- The price to pay for using projections: The more projections, the more power, but the less likely MAR holds on each projection.

Theoretical consideration: Propriety

- It is not enough to have MAR overall (unless the set of projections just corresponds to the full projection).
- Once we start removing variables, a MAR data set can become MNAR.
- The price to pay for using projections: The more projections, the more power, but the less likely MAR holds on each projection.
- Practically, we observed propriety in most cases.

- 1 Problem Motivation and Toy Example
- 2 A Solution: Imputation Scores (I-Scores)
- 3 A Specific I-Score: DR I-Score
- 4 Empirical Results**

Empirical Results: Propriety

Test empirical propriety of the DR I-Score: the non-inferiority of the true data score.

Empirical Results: Propriety

Test empirical propriety of the DR I-Score: the non-inferiority of the true data score.

- Used 9 imputation methods H that are easily usable in R and 15 real world data sets, with varying n and d .

Empirical Results: Propriety

Test empirical propriety of the DR I-Score: the non-inferiority of the true data score.

- Used 9 imputation methods H that are easily usable in R and 15 real world data sets, with varying n and d .
- We score the true data by $\widehat{S}_{NA}^*(P^*, P)$ and the imputed data by $\widehat{S}_{NA}^*(H, P)$ for each imputation distribution H and consider the difference $D_H := \widehat{S}_{NA}^*(H, P) - \widehat{S}_{NA}^*(P^*, P)$.

Empirical Results: Propriety

Test empirical propriety of the DR I-Score: the non-inferiority of the true data score.

- Used 9 imputation methods H that are easily usable in R and 15 real world data sets, with varying n and d .
- We score the true data by $\widehat{S}_{NA}^*(P^*, P)$ and the imputed data by $\widehat{S}_{NA}^*(H, P)$ for each imputation distribution H and consider the difference $D_H := \widehat{S}_{NA}^*(H, P) - \widehat{S}_{NA}^*(P^*, P)$.
- We want to test for all H :

$$H_0 : D_H = 0 \text{ vs } H_A : D_H > 0.$$

Empirical Results: Propriety

Test empirical propriety of the DR I-Score: the non-inferiority of the true data score.

- Used 9 imputation methods H that are easily usable in R and 15 real world data sets, with varying n and d .
- We score the true data by $\widehat{S}_{NA}^*(P^*, P)$ and the imputed data by $\widehat{S}_{NA}^*(H, P)$ for each imputation distribution H and consider the difference $D_H := \widehat{S}_{NA}^*(H, P) - \widehat{S}_{NA}^*(P^*, P)$.
- We want to test for all H :

$$H_0 : D_H = 0 \text{ vs } H_A : D_H > 0.$$

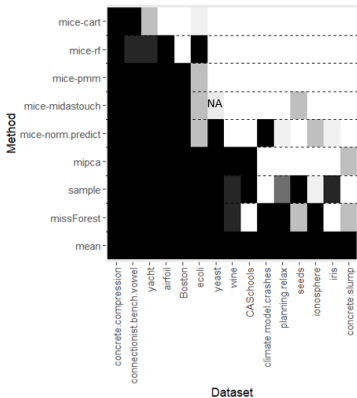
- We assume that approximately

$$D_H \stackrel{H_0}{\approx} \mathcal{N}(0, \sigma^2(D_H)),$$

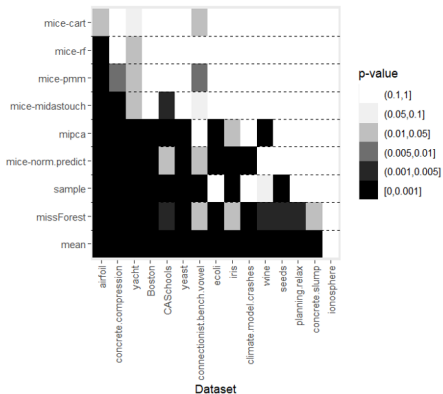
where we estimate $\sigma(D_H)$ with the Jackknife variance estimator using 30 times 1/2-subsampling.

Empirical Results: Assessment of Methods

Assessment of Methods with DR I-Score: Reverse the alternative hypothesis $H_0 : D_H = 0$ vs $H_A : D_H < 0$.



(a) MAR



(b) MCAR

Figure 7: We used $p_{miss} = 0.2$ and $m = 5$.

Conclusion

- The DR I-Score can evaluate imputation methods when
 - ① the target is the true data distribution,
 - ② there is no access to the true data, underlying the missing values,
 - ③ we do not want to artificially mask observations for the evaluation,
 - ④ the data is continuous and/or discrete,
 - ⑤ there are no complete observations (under certain conditions).

Conclusion

- The DR I-Score can evaluate imputation methods when
 - ① the target is the true data distribution,
 - ② there is no access to the true data, underlying the missing values,
 - ③ we do not want to artificially mask observations for the evaluation,
 - ④ the data is continuous and/or discrete,
 - ⑤ there are no complete observations (under certain conditions).
- Mice-Cart seems to broadly perform well

Conclusion

- The DR I-Score can evaluate imputation methods when
 - ① the target is the true data distribution,
 - ② there is no access to the true data, underlying the missing values,
 - ③ we do not want to artificially mask observations for the evaluation,
 - ④ the data is continuous and/or discrete,
 - ⑤ there are no complete observations (under certain conditions).
- Mice-Cart seems to broadly perform well
- Shortcomings: No finite sample guarantees, MAR on each projection

Conclusion

- The DR I-Score can evaluate imputation methods when
 - ① the target is the true data distribution,
 - ② there is no access to the true data, underlying the missing values,
 - ③ we do not want to artificially mask observations for the evaluation,
 - ④ the data is continuous and/or discrete,
 - ⑤ there are no complete observations (under certain conditions).
- Mice-Cart seems to broadly perform well
- Shortcomings: No finite sample guarantees, MAR on each projection

Questions and Comments?

Thank you!

Buuren, S. van. *Flexible Imputation of Missing Data. Second Edition.*

Boca Raton, FL: Chapman & Hall/CRC Press, 2018.

Doove, L.L., S. Van Buuren, and E. Dusseldorp. “Recursive partitioning for missing data imputation in the presence of interaction effects”. In: *Computational Statistics & Data Analysis* 72.C (2014), pp. 92–104.

Muzellec, Boris et al. “Missing data imputation using optimal transport”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7130–7140.

Backup Slides

Further Empirical Results

Relevancy of DR I-Score:

- One would hope that the methods chosen by our score perform well on a wide range of targets, even though it was not designed to select for any of these targets specifically.
- We focus on *average coverage and average width* of marginal confidence intervals for each NA value, obtained by the m multiple imputations.

Empirical Results: Relevancy of DR I-Score

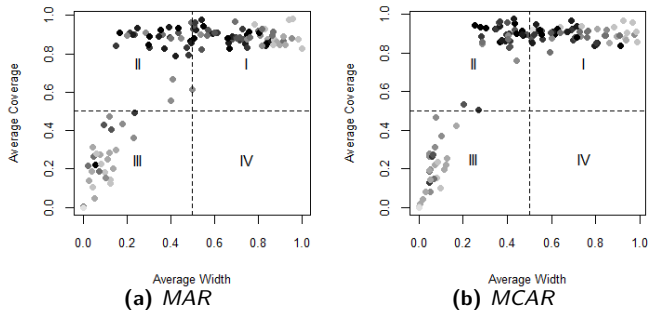


Figure 8: Average coverage plotted against average width for the 9 methods applied to the 15 data sets (total = $9 \times 15 = 135$ points). The darkness indicates the rank induced by the DR I-Score (the darker, the higher the rank). We used the missingness mechanism MAR in (a) and MCAR in (b) with $p_{miss} = 0.2$, $m = 20$.

Empirical Results: Relevancy of DR I-Score

method/quadrant	I	II	III
cart	0.47	0.53	0
pmm	0.53	0.47	0
midastouch	0.67	0.33	0
rf	0.73	0.27	0
mipca	0.87	0.13	0
sample	0.87	0.13	0
norm.predict	0	0.13	0.87
mean	0	0	1
missForest	0	0	1

(a) MAR

method/quadrant	I	II	III
pmm	0.40	0.60	0
cart	0.47	0.53	0
midastouch	0.53	0.47	0
rf	0.60	0.40	0
mipca	0.87	0.13	0
norm.predict	0	0.13	0.87
sample	0.93	0.07	0
mean	0	0	1
missForest	0	0	1

(b) MCAR

Table 1: The fraction of times each method appeared in the quadrants I, II and III in the MAR case (a) and the MCAR case (b).

- Approach 1: Fix a target quantity
- ⇒ **Problem**: Often the target is not clear, or there are several targets²
- Approach 2: Add more NAs
- ⇒ **Problem**: If predictive accuracy is measured with quadratic loss, imputations based on the conditional mean will rank highest.

²S. van Buuren. *Flexible Imputation of Missing Data. Second Edition*. Boca Raton, FL: Chapman & Hall/CRC Press, 2018, Chapter 2.3.4.▶

Theory

How to estimate the KL-Divergence

- The classifier with smallest overall error is the *Bayes* classifier:

$$\phi(x) = \begin{cases} 1, & \text{if } p(x) \geq 1/2 \\ 0, & \text{if } p(x) < 1/2, \end{cases}$$

where

$$p(x) = \frac{\pi f(x)}{\pi f(x) + (1 - \pi)h(x)},$$

and π is the (prior) probability of observing samples from density f .

- So: Using Bayes-Formula, if we ensure balanced samples (i.e. $\pi = 1/2$) and we can obtain an estimate \hat{p} of p ,

$$\frac{\hat{p}}{1 - \hat{p}},$$

is an estimate of the density ratio $f(x)/h(x)$.

Classification

- assign label 1 to the fully observed observations, and label 0 to the imputed observations,
- estimate $p(\hat{x})$, the probability of being labeled 1 given x ,
- Get a density ratio estimation $f(x)/h(x)$ through $p(\hat{x})/(1 - p(\hat{x}))$,
- Evaluating at samples from H and taking averages yields a KL-Divergence estimate.

Detailed Notation

- Let \mathcal{A} be a subset of the power set $2^{\{1, \dots, d\}}$, that denotes the set of all possible projections, such that each $A \in \mathcal{A}$ describes a set of variables we project onto.
- The projections are chosen randomly according to \mathcal{K} with support \mathcal{A}
- For each A we define P_A^M the distribution over the missingness patterns in P_A with support \mathcal{M}_A .
- $m_A \in \mathcal{M}_A$ is a given missingness pattern on the projection with associated probability $\mathbb{P}(M_A = m_A) = P_A^M(m_A)$.
- For any distribution $H \in \mathcal{H}_P$ we can then consider the conditional distribution $H_A | M_A = m_A$, i.e. the distribution of an imputation H , given the missingness pattern m_A on the projection A . Abbreviated with H_{m_A} , so that the density of H_{m_A} is given as $h_{m_A}(x_A) := h_A(x_A | M = m_A)$.
- Denoting with $\mathbf{0}$ the vector of zeros, we similarly write $p_A(x_A | M_A = \mathbf{0})$ to mean the density of the fully observed part of P , $P | M = \mathbf{0}$, projected to A .

Compatible Imputations H

- P refers to the distribution of X with missing values
- $P^* \in \mathcal{P}$ refers to the distribution of X^* without missing values.
- For a missingness pattern $m \in \mathcal{M}$, $o(x, m) = (x_j)_{j \in \{1, \dots, d\}: m_j=0}$ subsets the observed elements of x according to m , while $o^c(x, m) = (x_j)_{j \in \{1, \dots, d\}: m_j=1}$, subsets the missing elements.
- We define $\mathcal{H}_P \subset \mathcal{P}$ to be the set of imputation distributions compatible with P :

$$\mathcal{H}_P := \{H \in \mathcal{P} : h(o(x, m)|M = m) = p(o(x, m)|M = m) \text{ for all } m \in \mathcal{M}\}.$$

- If $p_A(X_A | M_A = \mathbf{0}) = 0$ for a set of $X_A \sim H_{M_A}$ with nonzero probability, we take $S_{NA}^*(H, P) = -\infty$ as a convention.
- As an intuition, the DR I-Score given by (2) can be rewritten as

$$S_{NA}^*(H, P) = -\mathbb{E}_{A \sim \mathcal{X}, M_A \sim P_A^M} D_{KL}(h_{M_A} \parallel p_A(\cdot | M = \mathbf{0})).$$

The KL-Divergence between two distributions $P, Q \in \mathcal{P}$ on \mathbb{R}^d with densities p, q is defined by

$$D_{KL}(p \parallel q) := \int p(x) \log \left(\frac{p(x)}{q(x)} \right) d\mu(x).$$

Proposition DR I-Score Propriety

The DR I-Score is indeed a proper under the following assumption:

Proposition

Let $H \in \mathcal{H}_P$. If for all $A \in \mathcal{A}$,

$$p^*(o^c(x_A, m_A) | o(x_A, m_A), M = m'_A) = p^*(o^c(x_A, m_A) | o(x_A, m_A)),$$

for all $m'_A, m_A \in \mathcal{M}_A$,

then $S_{NA}^*(H, P)$ is a proper I-Score.

Proposition DR I-Score Propriety

- The condition in the proposition is simply the MAR condition on the projection A .
- The key insight is that for any imputation distribution $H \in \mathcal{H}_P$ and $m \in \mathcal{M}$ it holds that

$$h_m(x) = h(o^c(x, m) | o(x, m), M = m) p^*(o(x, m) | M = m),$$

by the definition of \mathcal{H}_P .

- This can be used to show that the DR I-Score factors into (i) an *irreducible part*, stemming from the difference in the observed parts $p^*(o(x, m) | M = m)$ and $p(o(x, m) | M = \mathbf{0})$, and (ii) into a score for the distance of the conditional distributions.

Empirical Details

Formula Empirical DR I-Score

- Let $\tilde{\mathcal{A}}$ be a set of random projections sampled from \mathcal{A} with \mathcal{K} .
- Let \mathcal{N}_{m_A} be the set of indices i such that $x_{i,A}$ has pattern m_A and assume to have an estimator $\hat{\pi}_{m_A}(x_A)$ of $\pi_{m_A}(x_A)$.
- Given an imputation method with $N \geq 1$ different imputed values x_i^1, \dots, x_i^N of the incomplete observations, the estimator of $S_{NA}^*(H, P)$ is given by

$$\hat{S}_{NA}^*(H, P) := \frac{1}{N} \sum_{j=1}^N \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{1}{|\tilde{\mathcal{A}}|} \sum_{A \in \tilde{\mathcal{A}}} \frac{1}{|\mathcal{N}_{m_A}|} \sum_{i \in \mathcal{N}_{m_A}} \left[\log \left(\frac{\hat{\pi}_{m_A}(x_{i,A}^j)}{1 - \hat{\pi}_{m_A}(x_{i,A}^j)} \right) \right]$$

yielding a score of the imputation performance of H , averaged over $N \geq 1$ imputations.

- For each projection A and pattern m_A , we split data into a training and test set.
- Make sure to have observations with pattern m_A in both sets. We then fit $\hat{\pi}_{m_A}$ on the training set and evaluate it on the test set.

Ensuring Class Balancing

- We follow a simple procedure to ensure the same number of observations in the training sets \mathcal{S}_A^P and $\mathcal{S}_{m_A}^H$.
- First if $\mathcal{S}_{m_A}^H$ has fewer elements than \mathcal{S}_A^P , but is “large enough” relative to \mathcal{S}_A^P , we simply upsample $\mathcal{S}_{m_A}^H$ with replacement until it contains the same number of elements as \mathcal{S}_A^P .
- The exact same procedure is applied if \mathcal{S}_A^P has fewer elements than $\mathcal{S}_{m_A}^H$.
- On the other hand, if the set $\mathcal{S}_{m_A}^H$ is much smaller than \mathcal{S}_A^P , say if $|\mathcal{S}_{m_A}^H| < \tau \cdot |\mathcal{S}_A^P|$, for some $\tau \in (0, 1)$, we sample with replacement observations from other patterns to $\mathcal{S}_{m_A}^H$.
- This is done to ensure that we do not upsample one or two observations.
- In practice it seems adding additional patterns in the training step of the classifier does not hurt propriety.

Distribution over Projections

- We group samples wrt missingness pattern and for each of the groups we sample `num.proj` many projections from \mathcal{A} , adapted to the given pattern.
- $O_m^c \subseteq \{1, \dots, d\}$ is index set of variables with a missing value, such that $o^c(x_i, m) = x_{i, O_m^c}$ and similarly $O_m = \{1, \dots, d\} \setminus O_m^c$ the index set of variables without a missing value.
- Given m of a group of samples, we choose $\mathcal{A} = \mathcal{A}_m$ as the set of subsets A that satisfy $A \cap O_m^c \neq \emptyset$ and $A \setminus O_m^c \neq \emptyset$.
- In practice, we select at random a subset $\tilde{\mathcal{A}}_m$ of \mathcal{A} . We first sample a number r_1 in $\{1, \dots, |O_m^c|\}$ and a number r_2 in $\{1, \dots, d - r_1\}$. Then we obtain $A \in \tilde{\mathcal{A}}_m$ by taking the union of a random subset of size r_1 from O_m^c and a random subset of size r_2 from O_m .

Variance Estimation

- We divide \mathbf{X} randomly into two parts and compute the DR I-Score for a given imputation method for each part, obtaining $S^{(1)}$ and $S^{(2)}$.
- This is repeated B times to obtain scores $S_1^{(1)}, \dots, S_B^{(1)}$ and $S_1^{(2)}, \dots, S_B^{(2)}$.
- Let $\bar{S}_j = 1/2(S_j^{(1)} + S_j^{(2)})$ and let \hat{S} be the score of the original data set for a given imputation method.
- We estimate the variance as

$$\widehat{\text{Var}}(\hat{S}) = \frac{1}{B} \left(\sum_{j=1}^B \left(\bar{S}_j - \frac{1}{B} \sum_{j=1}^B \bar{S}_j \right)^2 \right).$$

- The approximate $(1 - \alpha)$ -Confidence Interval for our score is then given as

$$\hat{S} \pm q_{1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}(\hat{S})},$$

where $q_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of a standard normal distribution.

Data Sets

data set	n	d
airfoil	1503	6
Boston	506	14
CASchools	420	10
climate.model.crashes	540	19
concrete.compression	1030	9
concrete.slump	103	10
connectionist.bench.vowel	990	10
ecoli	336	5
ionosphere	351	32
iris	150	4
planning.relax	182	12
seeds	210	7
wine	178	13
yacht	308	7
yeast	1484	8

- 1) **missForest** is a multiple imputation method based on iterative use of RF, allowing for continuous and categorical data. After an initial mean-imputation, the variables are sorted according to their amount of missing values, starting with the lowest. For each variable as response, a RF is fitted based on the observed values. The missing values are then predicted with the RF. The imputation procedure is repeated until a stopping criterion is met.
- 2) **mipca** is a multiple imputation method with a PCA model. After an initialization step, an EM algorithm with parametric bootstrap is applied to iteratively update the PCA-parameter estimates and draw imputations from the predictive distribution. The algorithm is implemented in the function MIPCA of the R-package missMDA. We use the function `estim_ncpPCA` to estimate the number of dimensions for the principal component analysis by cross-validation.

Imputation Methods

- 3) **mean** is the simplest single imputation method considered. It imputes with the mean of the observed cases for numerical predictors and the mode of observed cases for categorical predictors. We use the implementation of the package `mice`.
- 4) **sample** is a multiple imputation method sampling at random a value of the observed observations in each variable to impute missing values. We use the implementation of the package `mice`.
- 5) **mice-cart** is a multiple imputation method cycling through the following steps multiple times (?): After an initial imputation through sampling of the observed values, a classification or regression tree is fitted. For each observation with missing values, the terminal node they end up according to the fitted tree is determined. A random member in this node is selected of which the observed value leads the imputation.
- 6) **mice-norm.predict** is a multiple imputation method cycling through the same steps as `mice-cart` with the adaptation that a linear regression is fitted and its predicted value is used as imputation.

- 7) **mice-pmm** Predictive Mean-Matching is a semi-parametric imputation approach. Based on the complete data, a linear regression model is estimated, followed by a parameter update step. Each missing value is filled with the observed value of a donor that is randomly selected among complete observations being close in predicted values to the predicted value of the case containing the missing value.
- 8) **mice-midastouch** is a multiple imputation method using an adaption of classical predictive mean-matching, where candidate donors have different probabilities to be drawn. The probability depends on the distance between the donor and the incomplete observation. A closeness parameter is adapted to the data.
- 9) **mice-rf** cycles through the same steps as mice-cart where one tree is fitted for every bootstrap sample. For each observation with NA, the terminal nodes in each tree are determined. A random member of the union of the terminal nodes is selected of which the observed value leads the imputation.

Results for testing empirical propriety of the DR I-Score:

- At level $\alpha = 0.05$ we found no single significant p-value not in the MCAR nor in the MAR case.
- At level $\alpha = 0.1$ we found in the MAR case two significant p-values for mice-rf and in the in the MCAR case one significant p-value for mice-cart.
- Theses results strongly indicate that the estimated DR I-Score is often proper.
- *Note:* In MAR case we did not verify the MAR assumption on the projections.