Missingness-aware and Sample-based Query Processing

PhD Student: Moulay Idris DAOUADI

LIMOS, CNRS, Clermont Auvergne University

April 11, 2024

Table of Contents

Introduction

- Not reported values for some variables in a relational DB.
- Usually noted by 'na' (not available).
- Usually appears as a result of:
  - **Data entry errors:** Human factors (deletion)
  - **System issues:** Glitches (technical errors)
  - **Non-response in surveys:** Silence (respondent choice)
  - **Data transformation and integration issues:** Mismatch (variable disparities)
  - **Natural causes:** Environment (sensor disruptions)

Introduction

Example of DB with missing values

employee	Name	Age	Salary	Phone
t_1	John	25	20K	555-1234
t_2	Jane	48	na	555-5678
t_3	Mike	39	55K	na
t_4	Emily	41	na	555-4321
t_5	Chris	60	60K	555-8765

There are different types of nulls

- non-available values (exists but unknown), e.g: Jane's salary
- non-applicable values, e.g: Mike doesn't have a phone

⇒ in our work we consider the first case of null only.

- we can overcome the second case by decomposing the relation into multiple relations.

tid	Name	Age	Salary
t_1	John	25	20K
t_2	Jane	48	na
t_3	Mike	39	55K
t_4	Emily	41	na
t_5	Chris	60	60K

tid	Phone
t_1	555-1234
t_2	555-5678
t_4	555-4321
t_5	555-8765

employee	Name	Age	Salary
t_1	John	na	20K
t_2	Jane	na	na
t_3	Mike	41	55K
t_4	Emily	51	na
t_5	Chris	na	60K

- SQL uses *null* instead of *na*.
- Consider the following two queries:
  1. Q1: *SELECT SUM(Salary) FROM employee WHERE Age* $\leq$ *41*
  2. Q2: *SELECT SUM(Salary) FROM employee*
- The QA:
  1. QA1 $=$ 55K
  2. QA2 $=$ 20K $+$ 55K $+$ 60K $=$ 135K

$\Rightarrow$ SQL might exclude tuples with NULL values in filtering (three valued logic [1]).

$\Rightarrow$ SQL always ignores the null while dealing with aggregate queries.

p	q	$p \lor q$	$p \land q$	$p = q$
True	True	True	True	True
True	False	True	False	False
True	Unknown	True	Unknown	Unknown
False	True	True	False	False
False	False	False	False	True
False	Unknown	Unknown	False	Unknown
Unknown	True	True	Unknown	Unknown
Unknown	False	Unknown	False	Unknown
Unknown	Unknown	Unknown	Unknown	Unknown

p	$\neg p$
True	False
False	True
Unknown	Unknown

1. Deletion:
    - Exclude from the analysis the rows with missing values corresponding to the variables of interests [2].
    - Deletion is performed with the assumption that the missing data occurs randomly and does not adhere to a specific pattern.
    - Shows biased statistical results when:
        - high rate of missing data.
        - A non-random pattern of missingness.
    - Two known methods: pairwise deletion, listwise deletion.
    - **Listwise deletion:**

employee	**Name**	**Age**	**Salary**
t_1	John	na	20K
t_2	Jane	na	na
t_3	Mike	41	55K
t_4	Emily	51	na
t_5	Chris	na	60K

- SELECT SUM(Salary) FROM employee

1. Deletion:
   - Exclude from the analysis the rows with missing values corresponding to the variables of interests [2].
   - Deletion is performed with the assumption that the missing data occurs randomly and does not adhere to a specific pattern.
   - Shows biased statistical results when:
     - high rate of missing data.
     - A non-random pattern of missingness.
   - Two known methods: pairwise deletion, listwise deletion.
   - **Listwise deletion:**

employee	**Name**	**Age**	**Salary**
t_1	~~John~~	~~na~~	~~20K~~
t_2	~~Jane~~	~~na~~	~~na~~
t_3	Mike	41	55K
t_4	~~Emily~~	~~51~~	~~na~~
t_5	~~Chris~~	~~na~~	~~60K~~

     - SELECT SUM(Salary) FROM employee

1. Deletion:
    - Exclude from the analysis the rows with missing values corresponding to the variables of interests [2].
    - Deletion is performed with the assumption that the missing data occurs randomly and does not adhere to a specific pattern.
    - Shows biased statistical results when:
        - high rate of missing data.
        - A non-random pattern of missingness.
    - Two known methods: pairwise deletion, listwise deletion.
    - **Pairwise deletion:**

employee	**Name**	**Age**	**Salary**
t_1	John	na	20K
t_2	~~Jane~~	~~na~~	~~na~~
t_3	Mike	41	55K
t_4	~~Emily~~	~~51~~	~~na~~
t_5	Chris	na	60K

- SELECT SUM(Salary) FROM employee

2. Imputation:
   - Replacing missing values with substituted values.
   - The quality of the analysis depends on the chosen imputation method.
   - There are several different approaches to imputing missing values:
     - Median, mean, K-nearest neighbors (kNN), regression imputation, multiple imputation...

employee	Name	Age	Salary
t_1	John	46	20K
t_2	Jane	46	45K
t_3	Mike	41	55K
t_4	Emily	51	45K
t_5	Chris	46	60K

Table: e.g: imputation using the mean

3. Missingness mechanism:
    - We can model missing values (MVs) using missingness mechanism (MM).
    - A MM describes why and how there are MVs, and under what conditions.
    - We identify 3 classes for MM (Missing Completely At Random, Missing At Random and Missing Not At Random) [5].

1. Missing Completely At Random (MCAR): Every record and attribute share a fixed uniform probability that the value is '*na*'
    - E.g: going down the column and throwing a dice if the dice value is 1 we record '*na*'.
2. Missing At Random (MAR): Missingness of an attribute value depends randomly on other non-missing attribute values
    - E.g: the presence of Missing Values (MVs) in salary is based upon the values assumed by the fully observed variable Age. To illustrate, consider the scenario where we roll a die, but now only when Age is between 50 and 60. If the die lands on 1, we record '*na*' for Salary.

3. Missing Not At Random (MNAR): None of the above
    - Missing Depending on Variable Itself: the probability of a variable having a missing value is solely determined by the variable itself.
        - E.g: Those individuals with the highest salaries are more inclined to keep it private.
    - Missing Depending on partially observed variable: the presence of missing values in the outcome variable is influenced by another variable that includes missing values
        - E.g: Elderly individuals are more prone to keep their income private (with age being a factor contributing to the missingness of income information). Nevertheless, the variable age itself also exhibits some missing values (as certain individuals did not report their ages).

- [3] used the missingness graph (another way to represent MM) to recover joint/conditional distribution from an incomplete database.

By having

- An incomplete DB
- the MMs.
- A query Q

⇒ how can we provide a better QA using the MMs?

Definitions (preliminaries)

Definitions (preliminaries)

- Missingness graph
- Probabilistic database PDB
- Block independent probabilistic database
- Semantics of query answering on PDB

Definitions (preliminaries)

Missingness graph

- We can describe MMs through m-graph (MG) [4].
- Lets consider the observed database D^*,

$employee^+$	*Age*	*Salary*
t_1	25	20K
t_2	48	na
t_3	39	55K
t_4	41	na
t_5	60	60K

Table: D^*

Definitions (preliminaries)

Missingness graph

- We can describe MMs through m-graph (MG) [4].
- Lets consider the observed database D^*,

$employee^+$	Age^o	$Salary^*$
t_1	25	20K
t_2	48	na
t_3	39	55K
t_4	41	na
t_5	60	60K

Table: D^*

Definitions (preliminaries)

Missingness graph

- We can describe MMs through m-graph (MG) [4].
- Lets consider the observed database D^*, and its complete version D^{om}

$employee^+$	Age^o	$Salary^m$
t_1	25	20K
t_2	48	60K
t_3	39	55K
t_4	41	70K
t_5	60	60K

Table: D^{om}

$employee^+$	Age^o	$Salary^*$
t_1	25	20K
t_2	48	na
t_3	39	55K
t_4	41	na
t_5	60	60K

Table: D^*

Definitions (preliminaries)

Missingness graph

- We can describe MMs through m-graph (MG) [4].
- Lets consider the observed database D^*, and its complete version D^{om}

$employee^+$	Age^o	$Salary^m$
t_1	25	20K
t_2	48	60K
t_3	39	55K
t_4	41	70K
t_5	60	60K

Table: D^{om}

$employee^+$	Age^o	$Salary^*$	R_{salary}
t_1	25	20K	0
t_2	48	na	1
t_3	39	55K	0
t_4	41	na	1
t_5	60	60K	0

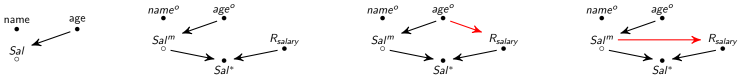
Table: The expanded format of D^*

- In the MG, every partially observed attribute will have three associated variables:
  - X^*: the available incomplete version of the partially observed variables.
  - X^m: the unavailable complete version of X^*
  - R_X: an indicator that takes 0 and 1 as values, where $X^m = X^*$ when $R_X = 0$

Definitions (preliminaries)

Missingness graph

- MGs are employed to depict the stochastic dependencies among the variables, particularly concerning MVs.
- It represents dependencies between the variables.
- It models the messingness mechanism.



(a) (b) (c) (d)

Definitions (preliminaries)

Missingness graph (m-graph)

Let G(V,E) be the causal DAG where V $= V^o \cup V^m \cup V^* \cup R$

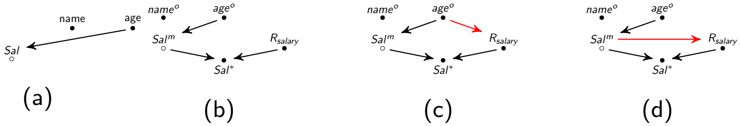
- V^o: the set of variables that are observed in all records.
- The variables that indicates missing values in the database is denoted X^* (proxy variable).
- X^m represents the unobserved complete version of X^*.
- R^X are the indicator variables, taking values 0 or 1, where each x^m is associated with an indicator variable.

$$v_i^* = f(r_{v_i}, v_i) = \begin{cases} v_i & \text{if } r_{v_i} = 0 \\ na & \text{if } r_{v_i} = 1 \end{cases}$$

Definitions (preliminaries)

Missingess graph and missingness mechanism

- We can represent the missingness mechanism by utilizing variable dependencies within the m-graph
- Missing Completely At Random (MCAR): if $(V^m \cup V^o \perp\!\!\!\perp R)$.
- Missing At Random (MAR): if $(V^m \perp\!\!\!\perp R | V^o)$.
- Missing Not At Random (MNAR): Data that are not MAR or MCAR fall under the MNAR category.



Definitions (preliminaries)

Missingness graph

The m-graph is viewed as a Bayesian network, capturing absolute and conditional probabilities of variables.

I^B

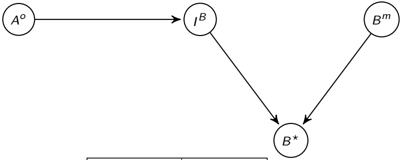
$\mathbf{A^o}$	$\mathbf{I^B}$	**P**
a_1	0	$p^{I^B}_{0a_1}$
a_1	1	$p^{I^B}_{1a_1}$
...	...	...
a_{l_a}	0	$p^{I^B}_{0a_{l_a}}$
a_{l_a}	1	$p^{I^B}_{1a_{l_a}}$

B^m

$\mathbf{B^m}$	**P**
b_1	p^B_1
...	...
b_{l_b}	$p^B_{l_b}$

A^o

$\mathbf{A^o}$	**P**
a_1	p^A_1
...	...
a_{l_a}	$p^A_{l_a}$



$B^\star$

$\mathbf{B^m}$	$\mathbf{I^B}$	$B^\star$	**P**
b_1	0	b_1	1
b_1	1	na	1
...	...	...	...
b_{l_b}	0	b_{l_b}	1
b_{l_b}	1	na	1

- Data: Traditional relational data coupled with probabilities quantifying the level of uncertainty.
- Queries: standard SQL queries, whose answers are annotated with output probabilities

Definition

A *probabilistic database* is a probability space $D = (W, P)$ where W is the set of possible worlds (the set of the possible database instances) and P is a probability over W.

⇒ Motivation: Given the impracticality of enumerating all potential worlds, there is a necessity for an optimized representation.

employee	*Age*^{o}	*Salary*^{m}	p
t_1	25	20K	1
t_2	48	20K	p_2^1
		...	...
		90K	p_2^k
t_3	39	55K	1
t_4	41	20	p_4^1
		...	...
		90K	p_4^k
t_5	60	60K	1

Table: Example of BIPDB

- Tuples from the same block are disjoint.
- Tuples from different blocks are independent.

Definition

A *block* is a probability space (B, P) where B the *domain* of the block is a set of tuples sharing the same identifier. This value is called the identifier of the block.

A *Block-Independent Probabilistic Database* (BIPDB) is a set of blocks with different identifiers. We can see a BIPDB $\{(B_i, P_i) \mid 1 \leq i \leq n\}$ as a probabilistic database space (W, P), where:

- $W = \left\{ \bigcup_{j=1}^{n} \{t_j\} \mid (t_1, \ldots t_n) \in \prod_{i=1}^{n} B_i \right\}$
- $\forall w = \{t_1, \ldots t_n\} \in W,\ P(w) = \prod_{i=1}^{n} P_i(t_i)$

Problem statement

Problem statement

Problem formulation

Given:

- an incomplete database D^*
- an qualitative MG
- a scalar aggregate query Q

$\Rightarrow$ How can we build a BID?

$\Rightarrow$ How can we use the BID for imputation?

Problem statement

The process of building BIPDB

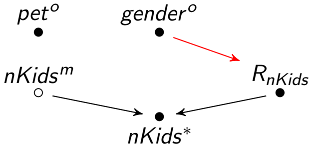


Figure: The m-graph

$Person*$	pet^o	$gender^o$	$nKids^*$
t_1	y	M	0
t_2	y	M	0
t_3	y	F	na
t_4	y	F	0
t_5	y	F	na
t_6	y	F	1
t_7	y	F	na
t_8	y	F	2

Figure: The observed database D^*

$Person^+$	pet^o	$gender^o$	$nKids^*$	R^{nKids}
t_1	y	M	0	0
t_2	y	M	0	0
t_3	y	F	na	1
t_4	y	F	0	0
t_5	y	F	na	1
t_6	y	F	1	0
t_7	y	F	na	1
t_8	y	F	2	0

Figure: The expanded schema instance D^+

- We have the probabilities: $P(pet^o)$, $P(gender^o)$, $P(nKids^m)$, $P(Rnkids \mid gender^o)$, $P(nKids^* \mid R_{nKids}, nkids^m)$

Problem statement

The process of building BIPDB

- Given the domains of $pet^o, gender^o, nKids^m$, such that:

  $$\text{dom}(pet^o) = \{y, n\}, \text{dom}(gender^o) = \{M, F\}, \text{dom}(nKids^m) = \{0, 1, 2\}$$

- For each tuple $t \in D^*$ we build a block B_t
  - $B_t = \mathsf{X}_{C \in Sort(D^*)} V_C^t$, where $V_C^t = \begin{cases} dom(C), & \text{if } t[C] = na \\ \{t[C]\}, & \text{otherwise} \end{cases}$
  - E.g: tuple $t_1 = (y, M, 0)$; $B_{t_1} = \{y\}x\{M\}x\{0\} = \{(y, M, 0)\}$
  - E.g: tuple $t_3 = (y, F, na)$;
    $B_{t_3} = \{y\}x\{F\}x\{0, 1, 2\} = \{(y, F, 0), (y, F, 1), (y, F, 2)\}$

Problem statement

The process of building the BIPDB

BID	pet^o	$gender^o$	$nKids^m$	P
B_1	y	M	0	p_1
B_2	y	M	0	p_2
B_3	y	F	0	p_3^1
	y	F	1	p_3^2
	y	F	2	p_3^3
B_4	y	F	0	p_4
B_5	y	F	0	p_5^1
	y	F	1	p_5^2
	y	F	2	p_5^3
B_6	y	F	1	p_6
t_7	y	F	0	p_7^1
	y	F	1	p_7^2
	y	F	2	p_7^3
B_8	y	F	2	p_8

Figure: Initial BIPDB

- If $|B_t| = 1$, then $p(t) = 1$.
- else: for all $\hat{t} \in B_t$
  - $P(\hat{t}) = P(nKids^m = \hat{t}[nKids^m] \mid gender^o = \hat{t}[gender^o], pet^o = \hat{t}[pet^o], nKids^\star = na)$

Problem statement

The process of building the BIPDB

bid	pet^o	$gender^o$	$nKids^m$	P
B_1	y	M	0	1
B_2	y	M	0	1
B_3	y	F	0	0.5
	y	F	1	0.25
	y	F	2	0.25
B_4	y	F	0	1
B_5	y	F	0	0.5
	y	F	1	0.25
	y	F	2	0.25
B_6	y	F	1	1
t_7	y	F	0	0.5
	y	F	1	0.25
	y	F	2	0.25
B_8	y	F	2	1

Figure: Final BIPDB

pet^o	$gender^o$	$nKids^m$
y	M	0
y	M	0
y	F	0
y	F	0
y	F	1
y	F	1
y	F	2
y	F	2

Table: A possible world I; P(I) = 0.5 x 0.25 x 0.25

Definition

In a probabilistic database (W, P), classes are defined as a partition of possible worlds using an equivalence relation, where two $w_i \sim w_j$ iff: $P_{w_i}(X^m \cup X^o) = P_{w_j}(X^m \cup X^o)$ where P_{w_i} is the empirical distribution. The probability of a class C is defined as $\sum_{w \in C} P(w)$.

- For query Q, worlds within the same class share a common QA.
- We want to evaluate our query on the class or classes with the highest probability in the BIPDB.
- Our intuition is that the most probable class will have the closest distribution to the one defined by the MG.

- Identical blocks are grouped into the same super-blocks.

Definition

A BIPDB D is *balanced* if for each S superblock in D, $(B, P) \in S$ and $t \in B$, $P(t) \times |S|$ is an integer.

Proof:

- The set of superblocks in D is $\{S_1, \dots, S_l\}$.
- Withing the superblock, the blocks share the same domain $B = \{t_1, \dots, t_m\}$.
- For each block (B, P_i) in S_i and tuple $t_j \in B$ $\exists$ integer $u_{i,j}$ such that $P_i(t_j) = \frac{u_{i,j}}{|S_i|}$
- A class C is identified by K_j (#occurence of t_j in C) where $\sum_{1 \leq j \leq m} K_j = | D |$

- $k_{i,j}$ is $\#t_j$ coming from superblock S_i where $k_j = \sum_{1 \leq i \leq l} k_{i,j}$
- for each $1 \leq i \leq l$, $\sum_{1 \leq j \leq m} k_{i,j} = |S_i|$

The probability of a class is obtained by the product of multinomial laws in each superblock:

$$\prod_{1 \leq i \leq l} \binom{|S_i|}{k_{i,1}} \left(\frac{u_{i,1}}{|S_i|}\right)^{k_{i,1}} \binom{|S_i|-k_{i,1}}{k_{i,2}} \left(\frac{u_{i,2}}{|S_i|}\right)^{k_{i,2}} \dots \binom{|S_i|-k_{i,1} \dots - k_{i,m-1}}{k_{i,m}} \left(\frac{u_{i,m}}{|S_i|}\right)^{k_{i,m}}$$

simplified as follow:

$$\frac{|S_1|! \dots |S_l|!}{|S_1|^{|S_1|} \dots |S_l|^{|S_l|}} \prod_{1 \leq i \leq l} \prod_{1 \leq j \leq m} \frac{u_{i,j}^{k_{i,j}}}{k_{i,j}!}$$

- To find the MPC we maximize for each fixed i, j the $k_{i,j}$ maximizing $\frac{u_{i,j}^{k_{i,j}}}{k_{i,j}!}$

Problem statement

Proof of the Most Probable Class (MPC) in a balanced BIPDB

$$\frac{u_{i,j}^{k_{i,j}}}{k_{i,j}!} = \frac{u_{i,j}}{1} \times \frac{u_{i,j}}{2} \times \cdots \times \frac{u_{i,j}}{k_{i,j}}$$

- The maximum is reached when $k_{i,j} = u_{i,j} - 1$ or $k_{i,j} = u_{i,j}$
- However; considering the constraints $1 \leq i \leq l$, $\sum_{1 \leq j \leq m} k_{i,j} = |S_i|$ will leave us with only one choice $k_{i,j} = u_{i,j}$
- With $k_{i,j}$ values, we can impute missing values.

Problem statement

Detailed example

bid	pet^o	$gender^o$	$nKids^m$	P
B_1	y	M	0	1
B_2	y	M	0	1
B_3	y	F	0	2/4
	y	F	1	1/4
	y	F	2	1/4
B_4	y	F	0	2/4
	y	F	1	1/4
	y	F	2	1/4
B_5	y	F	0	2/4
	y	F	1	1/4
	y	F	2	1/4
B_6	y	F	1	1
t_7	y	F	0	2/4
	y	F	1	1/4
	y	F	2	1/4
B_8	y	F	2	1

Figure: BIPDB

- $S_1 = \{B_3, B_4, B_5, B_7\}$
- $\mid S_1 \mid = 4$
- $k_{1,1} + k_{1,2} + k_{1,3} = 4$
- $u_{1,1} = \frac{2}{4}, u_{1,2} = \frac{1}{4}, u_{1,3} = \frac{1}{4}$
- The equivalent values of $k_{1,i}$ to find MPC:
  - $k_{1,1} = u_{1,1} = 2$
  - $k_{1,2} = u_{1,2} = 1$
  - $k_{1,3} = u_{1,3} = 1$

Problem statement

Detailed example

Person	pet^o	$gender^o$	$nKids^m$
t_1	y	M	0
t_2	y	M	0
t_3	y	F	0
t_4	y	F	0
t_5	y	F	1
t_6	y	F	1
t_7	y	F	2
t_8	y	F	2

Table: A world from the MPC

- Identical blocks are grouped into the same super-blocks.

Definition

A BIPDB D is *unbalanced* if $\exists$ superblock S_i in D, $(B, P_i) \in S_i$ and $t \in B$, $P_i(t) \times |S_i|$ is not an integer.

Proof:

- The set of superblocks in D is $\{S_1, \ldots, S_l\}$.
- Withing the superblock, the blocks share the same domain $B = \{t_1, \ldots, t_m\}$.
- For each block (B, P_i) in S_i and tuple $t_j \in B$ $\exists$ integer $u_{i,j}$ such that $P_i(t_j) = \frac{u_{i,j}}{\hat{s}_i}$
- A class C is identified by K_j (#occurence of t_j in C) where $\sum_{1 \leq j \leq m} K_j = \mid D \mid$

Problem statement

Proof of the Most probable class (MPC) in an unbalanced BIPDB

- $k_{i,j}$ is $\#t_j$ coming from superblock S_i where $k_j = \sum_{1 \leq i \leq l} k_{i,j}$
- ~~for each $1 \leq i \leq l$, $\sum_{1 \leq j \leq m} k_{i,j} = |S_i|$~~

The probability of a class is obtained by the product of multinomial laws in each superblock:

$$\prod_{1 \leq i \leq l} \binom{|S_i|}{k_{i,1}} \left(\frac{u_{i,1}}{\hat{s}_i}\right)^{k_{i,1}} \binom{|S_i|-k_{i,1}}{k_{i,2}} \left(\frac{u_{i,2}}{\hat{s}_i}\right)^{k_{i,2}} \dots \binom{|S_i|-k_{i,1} \dots - k_{i,m-1}}{k_{i,m}} \left(\frac{u_{i,m}}{\hat{s}_i}\right)^{k_{i,m}}$$

simplified as follow:

$$\frac{|S_1|! \dots |S_l|!}{\hat{s}_1{}^{|S_1|} \dots \hat{s}_l{}^{|S_l|}} \prod_{1 \leq i \leq l} \prod_{1 \leq j \leq m} \frac{u_{i,j}^{k_{i,j}}}{k_{i,j}!}$$

- To find the MPC we maximize for each fixed i, j the $k_{i,j}$ maximizing $\frac{u_{i,j}^{k_{i,j}}}{k_{i,j}!}$

Problem statement

Proof of the Most Probable Class (MPC) in an unbalanced BIPDB

$$\frac{u_{i,j}^{k_{i,j}}}{k_{i,j}!} = \frac{u_{i,j}}{1} \times \frac{u_{i,j}}{2} \times \cdots \times \frac{u_{i,j}}{k_{i,j}}$$

- The maximum is reached when $k_{i,j} = u_{i,j} - 1$ or $k_{i,j} = u_{i,j}$
- We randomly select one of the most probable classes and designate an instance from it to represent the final imputation in our database.

Problem statement

Comparative study

- The k-Nearest Neighbors (KNN) imputation (sckit learn).
- Predictive Mean Matching (PMM) imputation (MICE package).
- Classification and Regression Trees (CART) imputation (MICE package).
- Mode imputation (sckit learn).

Starting from a complete database, introduce missing values for different missingness mechanism and under different missingness rate.

- Wasserstein Distance:

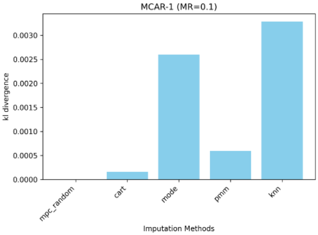
$$W_p(P, Q) = \left(\inf_{\pi \in \Gamma(P,Q)} \int_{R^d \times R^d} \|X - Y\|^p \, d\pi \right)^{1/p}$$

  - $\Gamma(P, Q)$ is the set of all joint probability measures on $R^d \times R^d$ whose marginals are P, Q

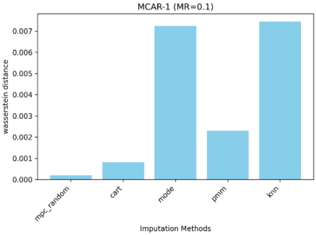
- The Kullback-Leibler Divergence (KL Divergence):

$$D_{KL}(P \parallel Q) = \sum_i P(i) \cdot \log \left(\frac{P(i)}{Q(i)} \right)$$

miss rate $= 0.1$



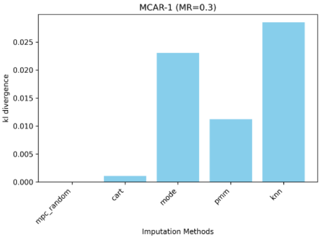
(a) KL divergence



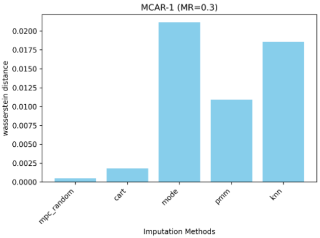
(b) wasserstein distance

Figure: comparing the distributionfor all competitors

miss rate $= 0.3$



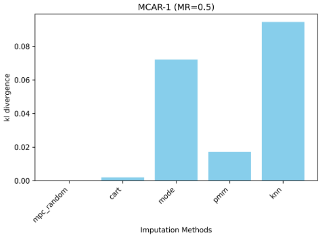
(a) KL divergence



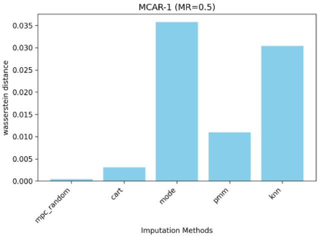
(b) wasserstein distance

Figure: comparing the distributionfor all competitors

miss rate = 0.5



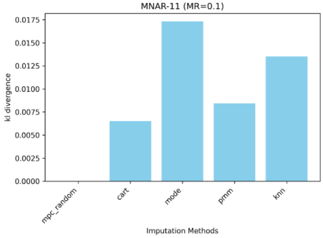
(a) KL divergence



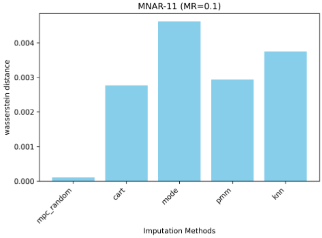
(b) wasserstein distance

Figure: comparing the distributionfor all competitors

miss rate $= 0.1$



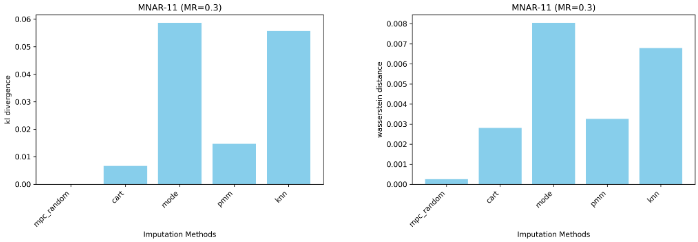
(a) KL divergence



(b) wasserstein distance

Figure: comparing the distributionfor all competitors

miss rate $= 0.3$



(a) KL divergence (b) wasserstein distance

Figure: comparing the distributionfor all competitors

miss rate = 0.5

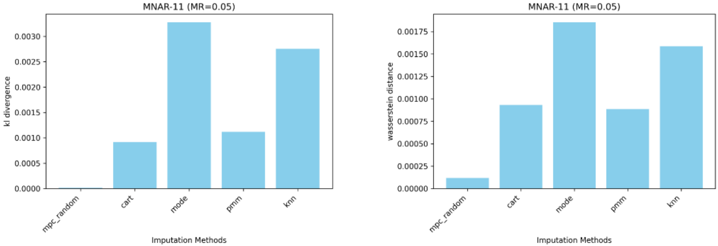


Figure: comparing the distributionfor all competitors

Perspective and future work

Perspective and future work

- Relaxation of the assumption:
  - What if we consider having only qualitative m-graph (no probabilities)?

  ⇒ [3] consistently recover the joint/conditional distribution for given m-graph.
  - Building a BID will depend on the recovered JD (ongoing work).
- Compre the QAs provided by the new imputed database with the most probable answer in the context of the probabilistic database.

References I

Leonid Libkin.
Sql's three-valued logic and certain answers.
ACM Trans. Database Syst., 41(1), mar 2016.

R. J Little.
Regression with missing x's: a review.
Journal of the American Statistical Association, 87, pages 1227–1237, 1992.

Karthika Mohan and Judea Pearl.
Graphical models for processing missing data.
Journal of the American Statistical Association, 116, 01 2018.

Karthika Mohan and Judea Pearl.
Graphical models for processing missing data.
Journal of the American Statistical Association, 116, 01 2018.

References II

Donald Rubin.
Inference and missing data.
pages 581–592, 1976.