skema BUSINESS SCHOOL | AI R&D CENTER CANADA

Correct Inference and Learning in the Presence of Incomplete Data

Specification and Use of Missingness Mechanisms as Causal Networks

Leopoldo Bertossi

Skema Canada, Montreal

Skema, PRISM Seminar, Dec. 2022

Missingness Mechanisms

- The general problem is about doing meaningful statistical inference from datasets that suffer from missing values (MVs)
- A "model" of MVs: Missingness Mechanisms (MMs) (Donald Rubin; [12, 7])
- MMs allow us to characterize, classify, investigate and attack problems related to MVs in datasets [4, 2]
- Think of census-like data: set of records (units) with attribute values

  Attributes are "variables", possibly taking `na` under certain conditions, and possibly according to a (possibly unknown) probability law

Dependencies w.r.t. MVs at intra-record level, as usual

Having a MV for Z may stochastically depend on X

	X	Y		Z
r_1:	x_{11}	y_{12}	$\cdots$	$z_{1,M}$
r_2:	x_{21}	y_{22}	$\cdots$	$z_{2,M} \longleftarrow$ `na`
			$\cdots$	
r_N:	x_{N1}	y_{12}	$\cdots$	$z_{N,M}$

$\mathbb{I}^Z$ denotes Bernouilli variable taking value 1 iff Z takes value `na`: $\mathbb{I}^Z$ may depend on X

- An MM (partially) describes *why* and *how* there are MVs, and under what conditions

  "Employees making more that 200K are likely to hide (not provide) their salaries"

  "People not providing marital status are prone not to provide number of children" (and maybe viceversa)

- Missingness Mechanisms:

1. Missing Completely at Random: (MCAR)

   For each record and each attribute there is a fixed and same probability that the value is `na`

   E.g. each respondent throws the same dice and returns `na` if 6 is obtained

2. Missing at Random: (MAR)

   Missingness of an attribute value depends randomly on other non-missing attribute values

   "MV for Salary *randomly depends on non-missing values for* Age, Education, *and* Race*"*

   "CTO's are less likely to reveal their salaries"

   (observed variable "position" influences observation of quantitative "salary")

3. Missing Not Completely at Random: (MNCAR)

   None of the above

   Some relevant special cases appear ...

3.1. Missing Depending on Variable Itself

- The likelihood of a MV for the variable depends only on the variable itself

  "People with high salary are less prone to reveal their salaries"

- A.k.a. "censorship" when *all* people with high salaries are likely not to report their salaries
- A usually difficult scenario
- Mitigate the problem by adding predictors for this variable that explain it

  Say variables that explain the high salaries, such as education level, race, etc.; falling back to case 2.

  May have to extrapolate beyond the spectrum of observed variables

3.2. Missing Depending on Unobserved Predictors

- Missingness for the outcome variable depends on variables that are not recorded at all or contain MVs

  "Disposition to provide the salary depends on paranoia, which is not considered in the study"

  "Disposition to provide the salary depends on education level, for which there are non-responses"

- MMs not about fixing MVs, but can can be used as a basis for it
- MMs can be detected, learned, modeled, and used
- MMs can be used for MV imputation, for statistical inferences, model building, e.g. ML models, under MVs, etc.
- Imputation: Filling in for MVs

  There are different techniques

  Not without problems ...
- Much of what follows has to do with modeling and using MMs for correct statistical inference

General Goal

- Develop, implement and apply ML models/algorithm that are based on incomplete training data or incomplete input data
  - Values for variables (features) are unobserved
  - Avoid discarding records with MVs or doing imputation
  - Incomplete data participate as such in training process
- Avoid if possible expensive methods, such as EM, sampling, etc.
- Use light-weight methods that appeal to an extra source of information
- Additional knowledge comes in the form of MMs

  They have to be properly modeled, specified and used

Specifying MMs via Missingness Graphs

- MMs can be represented as "missingness graphs" (MGs) (Mohan and Pearl [8])
- MGs are graphical models, like Bayesian Networks, representing directional stochastic dependencies, in particular, in relation to MVs
- Notation:
  - Always observed variable X denoted with X^o
  - Variable X that may have `na`-values denoted with X^m

    The variable does have values, but they may not be observed; as such it does not exhibit MVs
  - $X^{m\star}$ denotes its *proxy* showing observed and non-observed values (as `na`)
  - Variables X^m have associated *missingness indicator functions* $\mathbb{I}^X$ returning 1 or 0 depending on X's value being non-observed or observed

Example: X is a person's salary that may not be observed when below 100K:

X^m : $x_1^m = 40,\ x_2^m = 120,\ x_3^m = 80$ (has values, but not all observed)

$X^{m\star}$: $x_1^\star = \mathtt{na},\ x_2^\star = 120,\ x_3^\star = \mathtt{na}$ (coincides with X^m when observed)

$\mathbb{I}^X$: $\mathbb{I}^X(x_1) = 1,\ \mathbb{I}^X(x_2) = 0,\ \mathbb{I}^X(x_3) = 1$

- In general, $\mathbb{I}^X$ represents causal MM for X^m (better, its modelling represents ...)

  Fully observed variables: Of the form $X^o,\ Y^{m\star},\ \mathbb{I}^Z$

  Partially observed variables: Of the form V^m

- For example, for variables X, Y:

Y^o	X^m	$X^{m\star}$	$\mathbb{I}^X$
y_1	x_1	x_1	0
y_2	?	na	1
y_2	x_2	x_2	0
y_1	?	na	1

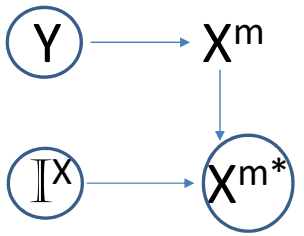
Observed distribution (sample) is for $Y^o, X^{m\star}, \mathbb{I}^X$

- In general, we have observations for joint variables $\mathbf{Y}^o, \mathbf{X}^{m\star}, \mathbb{I}^{\mathbf{X}}$

  There can be dependencies among $\mathbf{Y}^o, \mathbf{X}^m, \mathbf{X}^{m\star}, \mathbb{I}^{\mathbf{X}}$

Example: (cont.)

Y^o	X^m	$X^{m\star}$	$\mathbb{I}^X$
y_1	x_1	x_1	0
y_2	?	na	1
y_2	x_2	x_2	0
y_1	?	na	1



X^m depends on Y, and observed proxy $X^{m\star}$ depends upon X^m and $\mathbb{I}^X$
(fully observed variables encircled)

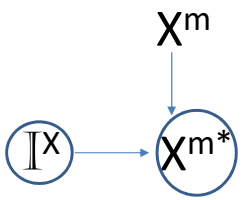
MVs in X depend only on X itself, and randomly, in that $\mathbb{I}^X$ does not depend on anything

- Analysis of MGs allows decision about correct recoverability/estimation of probabilities, and do the job in positive cases [8]

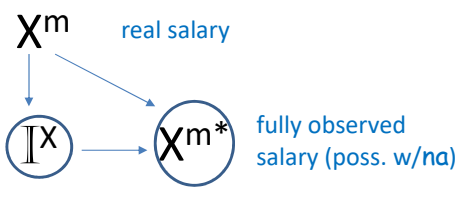
Analysis of a MG tells us *what kind of MMs* we are confronting

- Some common dependencies:

$$X^{m\star} = \begin{cases} \texttt{na} & \text{if } \mathbb{I}^X = 1 \\ X^m & \text{if } \mathbb{I}^X = 0 \end{cases}$$



as above



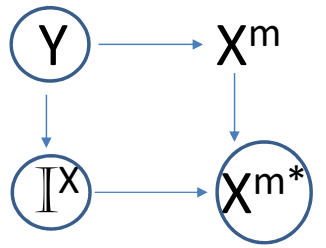
censorship

- MMs can be expressed and characterized via MGs:
  - MCAR: Missing Completely at Random (example: previous page, top)

    MVs independent from variables with MVs and fully observed: $(\mathbf{Y}^o, \mathbf{X}^m) \perp\!\!\!\perp \mathbb{I}^{\mathbf{X}}$
  - MAR: Missing at Random

    MVs may depend on observed variables, but not on those with MVs

    MVs are independent from partially observed variables given the fully observed variables: $(\mathbf{X}^m \perp\!\!\!\perp \mathbb{I}^{\mathbf{X}})|\mathbf{Y}^o$



$(X^m \perp\!\!\!\perp \mathbb{I}^X)|Y$

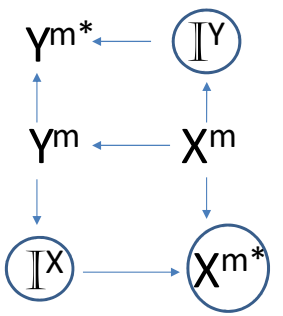
  - MNCAR: Missing Not Completely at Random

    None of the above

    Example on previous page, for salary, bottom-right

    $\mathbb{I}^X$ depends on not fully observed X^m

Another example:



- MMs can be identified from the structure of a MG
- Notion of *d-separation* plays a role here

  Syntactic, graphical analysis of BNs to detect/identify (in)dependencies [10]

  Exploit independence assumptions represented by MGs
- Language of MGs is more expressive, formal and precise than that of MMs
- MGs allow to decide what "queries" ($P(X,Y), P(Y), P(Y|X)$) can be correctly estimated (unbiased, convergent to real value) from a dataset and how [8, 9]
- Problem: How to estimate distributions that involve the $\mathbf{X}^m$?

  Most prominently: Overall join distribution $P(\mathbf{Y}^o, \mathbf{X}^m)$?
- Particulary appropriate for MCAR and MAR

  Most prominently, this can also be applied to MNCAR cases

- Sufficient conditions for recovering $P(\mathbf{Y}^o, \mathbf{X}^m)$ from joint distribution $P(\mathbf{Y}^o, \mathbf{X}^{m\star}, \mathbb{I}^{\mathbf{X}})$ or estimating it from sample for fully-observed $P(\mathbf{Y}^o, \mathbf{X}^{m\star}, \mathbb{I}^{\mathbf{X}})$? [8, 9]

D

ID	Age^o	Gender^o	Obesity^{m*}	$\mathbb{I}^{\text{Ob}}$
1	16	F	yes	0
2	15	F	na	1
3	15	M	na	1
4	14	F	no	0
5	13	M	no	0
6	15	M	yes	0
7	14	F	yes	0

Empirical data distribution

with $|D| = N$

$$P_D(a, g, ob^{m\star}, i^{Ob}) := \frac{\text{number of records of form } \langle a,g,ob^{m\star},i^{Ob}\rangle}{N} =: \frac{\#_D(\langle id;a,g,ob^{m\star},i^{Ob}\rangle)}{N}$$

- Reformulation: How to use $P_D(\mathbf{Y}^o, \mathbf{X}^{m\star}, \mathbb{I}^{\mathbf{X}})$ to estimate $P(\underbrace{\mathbf{Y}^o, \mathbf{X}^m}_{\mathbf{Z}})$?
- Can be done (or not) for cases identified on page 12 expressed via MGs

MCAR: $P(\mathbf{Z}) = P(\mathbf{Y}^o, \mathbf{X}^m | \mathbb{I}^{\mathbf{X}} = 0) = P(\mathbf{Y}^o, \mathbf{X}^{m\star} | \mathbb{I}^{\mathbf{X}} = 0) \approx P_D(\mathbf{Y}^o, \mathbf{X}^{m\star} | \mathbb{I}^{\mathbf{X}} = 0)$

(recall: $\mathbb{I}^X = 0$ iff X observed iff $X^m, X^{m\star}$ coincide)

Example: Assume for variables in table on page 14: $\mathbb{I}^{Ob} \perp\!\!\!\perp (\text{Age}^o, \text{Gender}^o, \text{Obesity}^{m\star})$:

$$P(\text{Age}^o, \text{Gender}^o, \text{Obesity}^m) \approx P_D(\text{Age}^o, \text{Gender}^o, \text{Obesity}^m, \mathbb{I}^{\text{Ob}} = 0) / P_D(\mathbb{I}^{\text{Ob}} = 0)$$

$$P(16, \text{F}, \text{yes}) \approx \frac{\frac{1}{7} \times \#_D(\langle 16, \text{F}, \text{yes}, 0\rangle)}{\frac{1}{7} \times \#_D(\langle 0\rangle)} = \frac{1}{5}$$

$$P(14, \text{F}, \text{no}) \approx \frac{\#_D(\langle 14, \text{F}, \text{no}, 0\rangle)}{\#_D(\langle 0\rangle)} = \frac{1}{5} \qquad P(13, \text{M}, \text{no}) \approx \frac{\#_D(\langle 13, \text{M}, \text{no}, 0\rangle)}{\#_D(\langle 0\rangle)} = \frac{1}{5}$$

$$P(15, \text{M}, \text{yes}) \approx \frac{\#_D(\langle 15, \text{M}, \text{yes}, 0\rangle)}{\#_D(\langle 0\rangle)} = \frac{1}{5} \qquad P(14, \text{F}, \text{yes}) \approx \frac{\#_D(\langle 14, \text{F}, \text{yes}, 0\rangle)}{\#_D(\langle 0\rangle)} = \frac{1}{5}$$

Here we obtain the same estimates as with *listwise deletion*, i.e. of tuples containing MVs, and then estimating with complete data

If we had two variables with MVs, the result could be different, corresponding to *available-case analysis*, i.e. deleting records only when both variables have MVs

MAR: $\mathbf{X}^o$ is all observed; $\mathbf{Y} = \mathbf{Y}^o \cup \mathbf{Y}^m$; $\mathbf{X}^{o\prime} := \mathbf{X}^o \setminus \mathbf{Y}^o$; equivalently $\mathbf{X}^o = \mathbf{Y}^o \cup \mathbf{X}^{o\prime}$

Marginal of $\mathbf{Y}$?

$$P(\mathbf{Y}) = \sum_{\mathbf{X}^{o\prime}} P(\mathbf{Y}^o, \mathbf{Y}^m, \mathbf{X}^{o\prime}) = \sum_{\mathbf{X}^{o\prime}} \Big(\underbrace{P(\mathbf{Y}^m | \mathbf{Y}^o, \mathbf{X}^{o\prime})}_{?} \times \underbrace{P(\mathbf{Y}^o, \mathbf{X}^{o\prime})}_{\text{easy, fully observed, with data}} \Big)$$

Under MAR, partially observed $\mathbf{Y}^m$ are conditioned to fully observed variables

Computing $P(\mathbf{Y}^m | \mathbf{Y}^o, \mathbf{X}^{o\prime})$ depends on the MGs (or MMs)

MAR $\Rightarrow$ each subset of data that fixes a value for $\mathbf{X}^o$ is locally MCAR

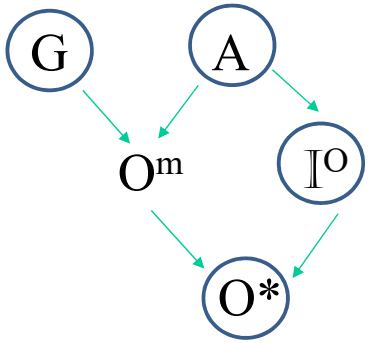
Continue locally as above with MCAR: $P(\mathbf{Y}^m | \mathbf{Y}^o, \mathbf{X}^{o\prime}) = P(\mathbf{Y}^{m\star} | \mathbf{Y}^o, \mathbf{X}^{o\prime}, \mathbb{I}^{\mathbf{Y}^m} = 0)$

So, $P(\mathbf{Y})$ can be recovered from distributions for fully observed variables, and estimated therethrough: (closed form)

$$P(\mathbf{Y}) \approx \sum_{\mathbf{X}^{o\prime}} P_D(\mathbf{Y}^{m\star} | \mathbf{Y}^o, \mathbf{X}^{o\prime}, \mathbb{I}^{\mathbf{Y}^m} = 0) \times P_D(\mathbf{Y}^o, \mathbf{X}^{o\prime}) \qquad (1)$$

Example:

G	A	$O^{m\star}$	$\mathbb{I}^O$	$P(G, A, O^{m\star}, \mathbb{I}^O)$
M	10 - 13	Y	0	p_1
M	13 - 15	Y	0	p_2
M	15 - 18	Y	0	p_3
M	10 - 13	N	0	p_4
M	13 - 15	N	0	p_5
M	15 - 18	N	0	p_6
F	10 - 13	Y	0	p_7
F	13 - 15	Y	0	p_8
F	15 - 18	Y	0	p_9
F	10 - 13	N	0	p_{10}
F	13 - 15	N	0	p_{11}
F	15 - 18	N	0	p_{12}
M	10 - 13	na	1	p_{13}
M	13 - 15	na	1	p_{14}
M	15 - 18	na	1	p_{15}
F	10 - 13	na	1	p_{16}
F	13 - 15	na	1	p_{17}
F	15 - 18	na	1	p_{18}



A MAR case

p_i's can be real (asymptotic frequencies) or empirical frequencies of the form $\frac{\#\text{record in } D}{N}$ with sample size $N >> 18$

Want to recover/estimate $P(G^o, A^o, O^m)$ from $P(G^o, A^o, O^{m\star}, \mathbb{I}^O)$

Going *ad hoc*: [8]

$$P(G^o, A^o, O^m) = P(G^o, O^m|A^o) \times \underbrace{P(A^o)}_{\text{observed, easy marginal}} \quad \text{(factorization suggested by MG)}$$

$$\underbrace{=}_{\text{MAR, check MG}} P(G^o, O^m|A^o, \mathbb{I}^O = 0)P(A^o) = \underbrace{P(G^o, O^{m\star}|A^o, \mathbb{I}^O = 0)}_{\text{all observed}} P(A^o)$$

$$\approx P_D(G^o, O^{m\star}|A^o, \mathbb{I}^O = 0)P_D(A^o) \tag{2}$$

E.g. $P(\text{M,10-13,Y}) \approx P_D(\text{M,Y}|\text{10-13}, \mathbb{I}^O = 0)P_D(\text{10-13}) = \frac{P_D(\text{M,Y,10-13},\mathbb{I}^O=0)}{P_D(\text{10-13},\mathbb{I}^O=0)} P_D(\text{10-13})$

$$= \frac{p_1}{p_1+p_4+p_7+p_{10}}(p_1 + p_4 + p_7 + p_{10} + p_{13} + p_{16}) \quad (\neq p_1)$$

Can be seen as deleting certain records, but re-weighting the remaining ones

Called "direct-deletion" approaches

(2) can be obtained from general (1): $\mathbf{Y} = \{G^o, A^o, O^m\}$, $\mathbf{X}^{o\prime} = \mathbf{X}^o \setminus \mathbf{Y}^o = \emptyset$

$$P(\mathbf{Y}) \approx \sum_{\mathbf{X}^{o\prime}} P_D(\mathbf{Y}^{m\star}|\mathbf{Y}^o, \mathbf{X}^{o\prime}, \mathbb{I}^{\mathbf{Y}^m} = 0) \times P_D(\mathbf{Y}^o, \mathbf{X}^{o\prime})$$

In fact: $P(G^o, A^o, O^m) \approx P_D(O^{m\star}|G^o, A^o, \mathbb{I}^{O^m} = 0) P_D(G^o, A^o)$ (from (1))

$$= \frac{P_D(O^{m\star}, G^o, A^o, \mathbb{I}^{O^m}=0)}{P_D(G^o, A^o, \mathbb{I}^{O^m}=0)} P_D(G^o, A^o)$$

$$\underbrace{=}_{MG} \frac{P_D(O^{m\star}, G^o, A^o, \mathbb{I}^{O^m}=0)}{P_D(G^o) P_D(A^o, \mathbb{I}^{O^m}=0)} P_D(G^o) P_D(A^o) = \frac{P_D(O^{m\star}, G^o, A^o, \mathbb{I}^{O^m}=0)}{P_D(A^o, \mathbb{I}^{O^m}=0)} P_D(A^o)$$

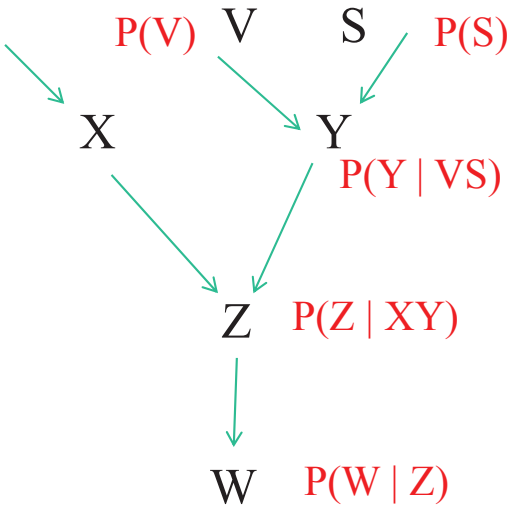
$$= P_D(O^{m\star}, G^o|A^o, \mathbb{I}^{O^m} = 0) P_D(A^o), \quad \text{which coincides with (2)}$$

- Recoverability/estimation methods based on MGs can be applied to MNCAR cases that are out of reach from other approaches [8, 14]

Bayesian Networks and Incomplete Data

- Bayesian Networks (BNs) have probability distribution associated to the nodes

  Fixed the structure, they become the parameters to be estimated from data, possibly incomplete



- MGs representing MMs can be applied as above, and integrated with the BNs being learned [14]

  Estimation can be efficiently done on the expanded neighborhoods (with local MGs) of BN and sample

  Avoiding inference or complex local iterative processes

- Excellent performance and accuracy is shown in [14]

  Based on MG-based estimation and network factorization

Some Remarks

- MG-based methods for MCAR and MAR are quite general

  Method can handle vast classes of MNCAR cases (still work to do)

  Some cases for BNs in [8] (not in [14])

- When MG analysis tells us unbiased recoverability/estimation impossible, no statistical method will do

  Put up with it or enrich the model by bringing in auxiliary variables

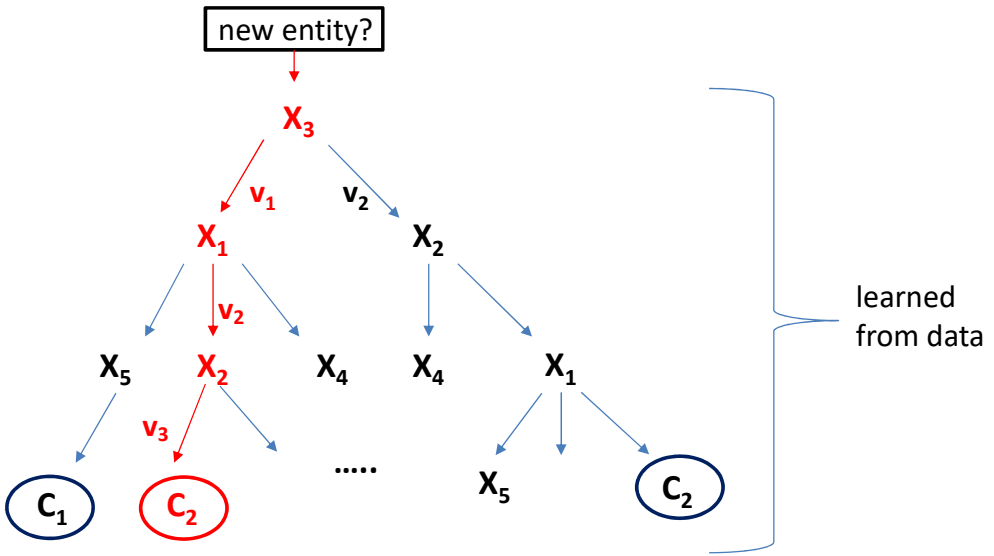
  (not quite clear how to proceed about this)

- Analysis of MG also tells us if its assumptions and consequences are "testable" or not [8]

Ongoing Research: Directions and Issues

- Learning Bayesian Network and Decision Tree Classifiers with Incomplete Data
- Data in the form of labeled examples for training may have MVs
- DTs inductively and iteratively built, determining tree-levels for placing variables

  E.g. on the basis of "information gain" (depends on data at hand)
- DT algorithms can be extended to obtain probabilities for class membership



- Develop algorithms for DT building with MV under different MMs (MGs)
- What about inputs with MVs?

- Develop a declarative language to specify MMs (and MGs)

  And reason with them

  And algorithms on those specifications to identify (in)dependencies and cases of MMs, e.g. in relation to d-separation

  In some cases, e.g. BNs, they could be integrated with initial (in)dependencies

  E.g. in the BN above: $V \perp\!\!\!\perp S$, $(Z \perp\!\!\!\perp V, S)|Y$
- Develop a "calculus of MMs" to reason with MMs and infer from them

  There are some approaches in relation to (in)dependence [11]
- Detecting/learning MMs?
- How much computation as above can be profitably done inside a relational DB?

  As *In-DB* ML [1]

- Impact of MV (and system building/use) on fairness requires much more research [5, 3]
- Aside ML and more generally:

  Apply all of above to deal with MVs in relational DBs, but not as usual SQL nulls with certain answers

  Take MVs in RDBs seriously and do statistical estimation (whenever possible)
- And in probabilistic DBs? [13]

  And Probabilistic DBs extended with probabilistic schemas? [6]

Valuable conversations with Long Nguyen, Foula Vagena and Guy Van den Broeck are much appreciated

References

[1] Abo Khamis, M., Ngo, H. Q. and Rudra, A. FAQ: Questions Asked Frequently. Proc. PODS, 2016.

[2] Bertossi, L. Missing Values in Data. Presentation at RelationalAI, Oct. 2018. slides

[3] Caton, S., Malisetty, S. and Haas, Ch. Impact of Imputation Strategies on Fairness in Machine Learning. J. Artificial Inteligence Research, 2022, 74:1011-1035.

[4] Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press, 2007.

[5] Goel, N., Amayuelas, A., Deshpande, A. and Sharma, A. The Importance of Modeling Data Missingness in Algorithmic Fairness: A Causal Perspective. arXiv:2012.11448, 2021.

[6] David Heckerman, D., Meek, Ch. and Koller, D. Probabilistic Entity-Relationship Models, PRMs, and Plate Models. Proc. 21st International Conference on Machine Learning, 2004.

[7] Little, R. J. A. and Rubin, D.B. *Statistical Analysis with Missing Data*. Wiley, 2002.

[8] Mohan, K. and Pearl, J. Graphical Models for Processing Missing Data. *Journal of the American Statistical Association*, 2021, 116(534):1023–1037. arXiv:1801.03583.

[9] Mohan, K., Pearl, J. and Tian, J. Graphical Models for Inference with Missing Data. Proc. NIPS, 2013.

[10] Pearl, J. and Paz, A. GRAPHOIDS: A Graph-Based Logic for Reasoning about Relevance Relations. Proc. ECAI, 1986.

[11] Pearl, J. and Verma, T. The Logic of Representing Dependencies by Directed Graphs. Proc. AAAI, 1987.

[12] Rubin, D. B. Inference and Missing Data. *Biometrika*, 1976, 63(3):581-592.

[13] Suciu, D., Olteanu, D., Re, C., Koch, C. Probabilistic Databases. Synthesis Lectures on Data Management, Morgan & Claypool Pub., 2011.

[14] Van den Broeck, G., Mohan, K. and Arthur Choi, A., Darwiche, A. and Pearl, P. Efficient Algorithms for Bayesian Network Parameter Learning from Incomplete Data. Proc. UAI, 2015.