Current state of Structured Missingness

Robin Mitra

Alan Turing Institute/Cardiff University

November 10, 2021

Outline

Introduction

- Missing data are a common unavoidable problem in many fields.
- The presence of missing values can complicate typical analyses that analysts wish to perform on the data.
- Many standard statistical procedures will not be possible to be applied if missing values are present.
- Complete case analysis is the simplest method to handle missing data (throw away any unit/row in the data that has missing values).
- However this can result in biased analyses as well as substantial reduction in efficiency.

Missing data pattern

- How the missing data are spread out throughout a dataset are referred to as a missing data pattern.
- There are two broad types of missing data patterns, monotone and non-monotone missing patterns.
- Typically monotone patterns are easier to handle in that they allow a cleaner decomposition of the missing and observed parts of the data.

Missing data pattern - non monotone illustration

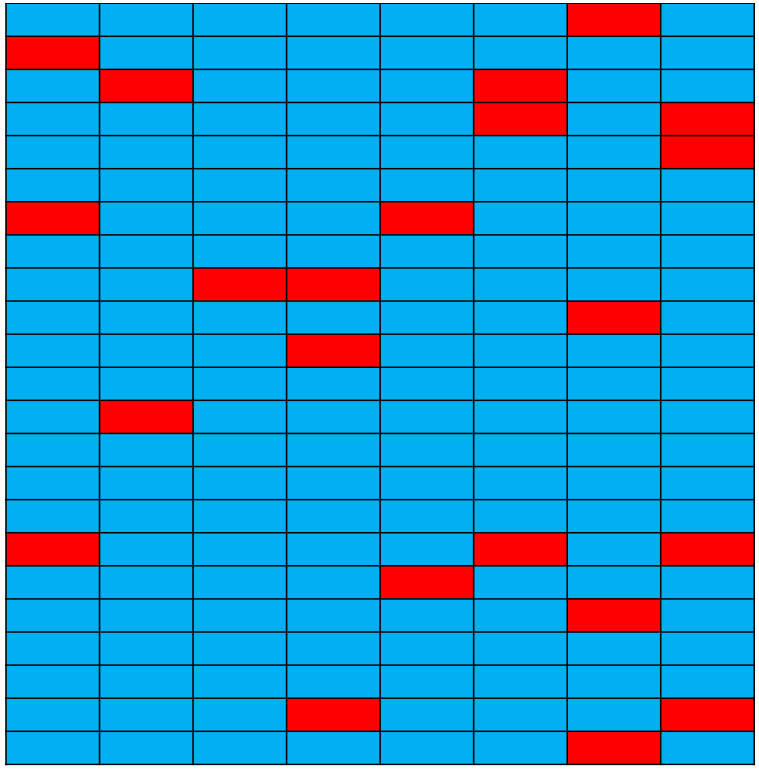

Figure: Illustration of a non-monotone missing pattern

Missing data pattern - monotone illustration

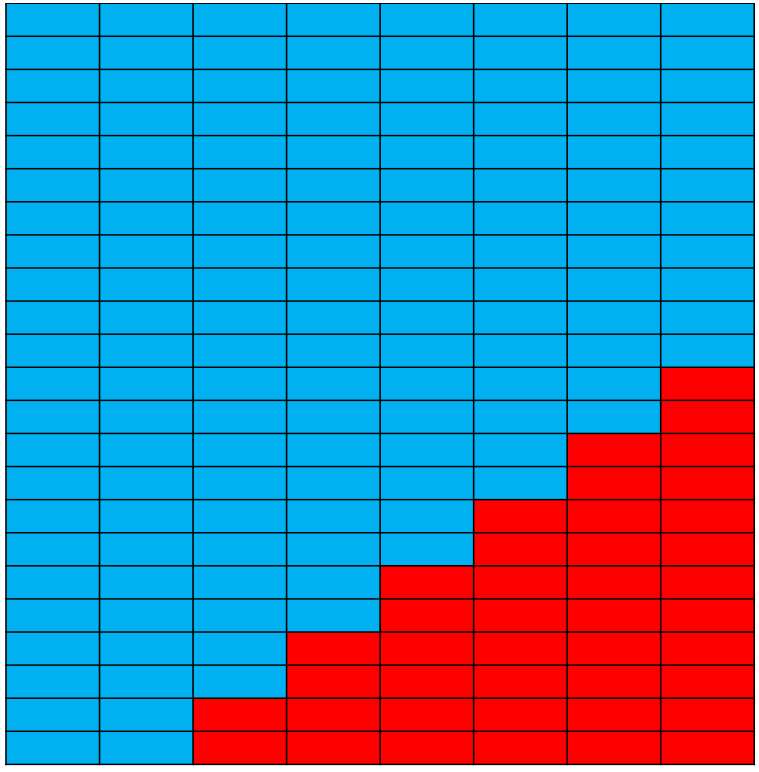

Figure: Illustration of a monotone missing pattern

Missing data mechanism - mathematical definition

- Not all missing data are the same.
- We need to consider the process or mechanism that causes the missing data.
- Suppose Y is the variable in the data that has missing values, with X some other variable(s).
- We define a missing data indicator M_i such that $M_i = 1$ when Y_i is missing, and 0 otherwise, for unit i in the data.
- We model the process that generates the missing data, $p(M|X, Y, \phi)$, and have various scenarios.
  - $p(M|X, Y, \phi) = p(M|\phi)$ - MCAR.
  - $p(M|X, Y, \phi) = p(M|X, \phi)$ - MAR.
  - $p(M|X, Y, \phi) = p(M|X, Y, \phi)$ - NMAR.

Structured missingness - definitions and examples

- Structured missingness refers to relationships or structure characterising the missing data process
- There have been various references in the literature to structured missingness with different terms used to define the missing data depending on the contextual example
- Skip patterns in questionnaires
  - Suppose there is a two part question in a survey, e.g. Q1 Do you have a job? If yes got to Q2, if No skip to Q3.
  - Those who answer No to Q1 will be structurally missing in Q2
- Missing by design
  - Consider a longitudinal study on survival of transplant patients.
  - Part way through the study BMI begins to be collected.
  - For all patients entering prior to the date they will be missing a BMI measurement due to a change in data collection practices.

Structured missingness - clinico-genomic data

- In clinico-genomic data there are a variety of ways structured missingness can occur.
- Combining data sets results in missing data due to different numbers of people and measurements recorded in each data set.
- E.g. the Foundation Medicine database is only 20% the size of flatiron database.
- Patients may only receive a battery of tests depending on whether the clinician deems it necessary, so patients would either have a set measurements from all these tests or be missing on all the measurements (missing by test type).
- The above example also gives rise to the name "Block Missing data" to describe this type of missing pattern.

Structured Missing data pattern

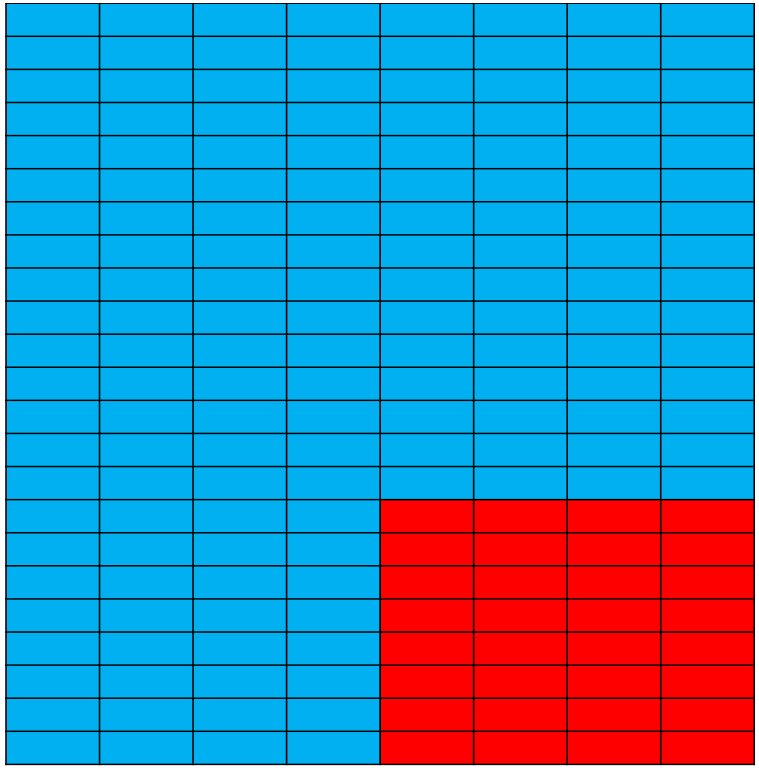

Figure: Illustration of a Structured "block" missing pattern

Structured Missingness and missing patterns

- From the above Structured Missingness illustration we see it results in a special case of a monotone missing data pattern.
- While this can be convenient, in terms of a clean decomposition, other challenges can arise.
- The units all missing the same variables can lead to a great amount of information loss than if the same amount of missing data was spread more evenly throughout the data.

Structured Missingness and Missing Mechanisms

- Structured missingness could lie within the exist taxonomy of mechanisms.
- MCAR - combining two clinico-genomic data sets that are random samples from the population. Units in both data sets will be fully observed, while the other units will be structurally missing in the data set they were not present in.
- MAR - a clinician may order different sets of tests depending on gender (observed), meaning males and females will have different structurally missing data.
- NMAR - a clinician may make a subjective assessment (unobserved) on which tests to order for a patient giving rise to informative (structurally) missing data.

Structured Missingness and Missing Mechanisms cont.

- Structured Missingness could also be defined as its own Missing Mechanism.
- Consider a data matrix X and corresponding missing indicator matrix M where an element of M equal to 1 or 0 implies the corresponding element in X is missing or observed respectively.
- Normally a mechanism here is defined by considering $p(M|X, \phi)$.
- E.g. $p(M|X, \phi) = p(M|X_{obs}, \phi)$.
- However, in Structured Missingness there is dependence between elements in M. E.g. columns j and k in M are completely dependent.
- E.g. a missing value in column j implies a missing value in column k and vice versa.
- Does this affect any of the theory upon which missing data methodology is built on, e.g. MAR and ignorability?

Dealing with Structured Missing data - beyond imputation

- Imputation is one method to consider here, e.g. using the MICE or missForest packages in R.
- Some further relevant imputation approaches here include, Audigier et al. (2018), Suresh et al. (2020), Dong et al. (2018).
- However, imputation methods may not always be optimal or even appropriate. E.g. suppose the structured missing data correspond to an impossible combination of attributes?
- Other approaches include, but are not limited to:
  - Likelihood/Bayesian models - e.g. formulate a model for the data and integrate over the missing data.
  - Linkages/calibration methods, e.g. probabilistically link records from different data sets to each other.
  - Possibly applying techniques from meta analysis?

Understanding Structured Missingness

- There may be complexities to appreciate around structural missingness itself. It might be the case columns in M are not deterministically related but some dependencies exist.

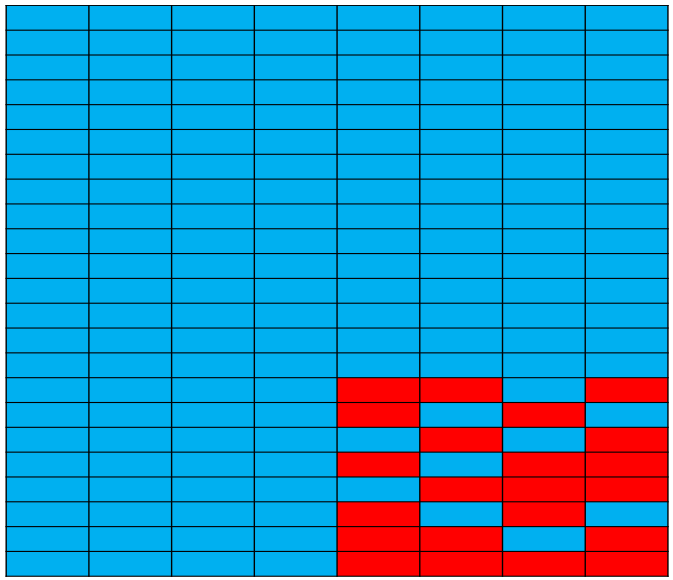

Figure: Illustration of an " almost block" missing pattern

- Tierney et al (2015) looks at using tree based methods to learn about the structure of missing data.

Open Challenges/some ideas

- Missing data are often not the key applied question of interest.
- Considering the analysis goal is an important factor in determining how to address structured missingness appropriately.
- E.g. if a multilevel, or hierarchical model may result in the missingness be deal with in a particular way.
- Is the research problem inferential or predictive, or both?!
- Can we assess/quantify what impact Structured Missingness has over non Structured Missing data, with respect to different analysis/uses of clinico-genomic data.
- The Clinico-genomic data set provides a rich example for possible use cases but the problem can extend beyond this.
- Ideally we can think creatively, e.g. can we consider the problem from a design (other?) points of view as well.

References

Audigier, Vincent, et al. "Multiple imputation for multilevel data with continuous and binary variables." Statistical Science 33.2 (2018): 160-183. APA

Dong, Xuesi, et al. "TOBMI: trans-omics block missing data imputation using a k-nearest neighbor weighted approach." Bioinformatics 35.8 (2019): 1278-1283

Suresh, Marcus, et al. "Sharpening the BLADE: missing data imputation using supervised machine learning." Australasian Joint Conference on Artificial Intelligence. Springer, Cham, 2019

Tierney NJ, Harden FA, Harden MJ, et al. "Using decision trees to understand structure in missing data." BMJ Open 2015;5