Missing Data: Theory & Methods

Chapter 7
A Review

Lu Mao lmao@biostat.wisc.edu

Contents

Missingness Mechanisms and Identifiability

- Missing data are classified into three categories according to the mechanism by which missingness occurs.
- These are missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).
- In MCAR data, missing values occur in a completely haphazard way, that is, independent of the observed and unobserved data. For example, data records are accidentally lost, or the investigator randomly chooses a subset of study subjects to collect full data.
- MCAR is the simplest scenario of missing data. It is the easiest to deal with because it implies that the complete cases are a random sample of the population and thus a complete-case (CC) analysis would yield valid results.

Missingness Mechanisms and Identifiability

- However, that does not mean that CC analysis is always the wise thing to do with MCAR data. Because even though discarding incomplete cases incurs no bias, it throws away useful (partial) information contained therein. This leads to loss of statistical efficiency, particularly when the missing rate is high.
- Therefore it is sometimes more desirable to treat MCAR as a special case of MAR and try to extract information from both complete and incomplete cases.
- For MAR, missingness is allowed to depend on data through the observed values. It is a generalization of MCAR.
- But unlike MCAR, CC analysis need not be valid for MAR data, for the complete cases are generally a non-random sample of the population.

Missingness Mechanisms and Identifiability

- However, because of likelihood factorization, inference can be based on the observed-data likelihood without regard to the specific model under which missingness is generated.
- The observed-data likelihood is usually less tractable than the full-data likelihood because the former may involve averaging (integration) over missing values. Thus, specialized iterative procedures, e.g., EM, Gibbs sampler, etc., are generally needed to make likelihood-based inference.
- If missingness is allowed to depend on the missing values as well as the observed ones, the data are said to be NMAR, which is the most general case of missing data.
- Given its generality, NMAR would have been the natural assumption to adopt when analyzing missing data if not for the identifiability issue.

Missingness Mechanisms and Identifiability

- Technically, a parameter (or an assumption) is identifiable if the distribution of the observed data corresponds uniquely with the parameter (or determines whether the assumption is true).
- One can think of identifiability as the ability to pinpoint the parameter value (or the correctness of the assumption) with an infinite amount of data.
- Obviously, lack of identifiability removes the ground for statistical analysis. That is why estimation of an unidentifiable parameter (or testing an unidentifiable assumption) is an utterly ill posed problem.
- When data are MAR, the complete cases are representative of the sub-population within each level of the observed values, and thus, under certain regularity (positivity) conditions (See Proposition 1.1), identifiability is preserved from the full data to the observed data.

Missingness Mechanisms and Identifiability

- With NMAR, the missing values are allowed to differ from the observed one in (probabilistically) arbitrary ways. So one is unable to trace back the information contained in the full data, resulting in non-identifiability.
- The usual solution to the identifiability issue is to conduct analysis under the assumption of MAR, and then posit parameter selection models (such that the parameter is identifiable within the parametric model) to carry out sensitivity analyses.

Contents

The Likelihood-Based Analysis Methods

- We have introduced three likelihood-based methods for analysis of missing data: maximum likelihood estimation, Bayesian analysis, and multiple imputation.
- By "likelihood-based" we mean methods based on observed-data likelihood.
- If the observed-data likelihood is easy to work with, then one just works with the observed-data likelihood, under either the frequentist or Bayesian paradigms.
- More often than not, the observed-data likelihood is intractable and one wishes to convert the problem into a full-data one in order to utilize the existing techniques for full-data analysis.
- Thus, the logical first step would be to fill in, or "impute", missing values so that a full dataset can be formed for subsequent analysis.

The Likelihood-Based Analysis Methods

- This is the motivation behind various simple/improper imputation methods, that is, substituting missing data by some estimated values based on the observed data and perhaps some initial estimator of the parameter.
- This line of approach is not entirely satisfactory as the simple/improper imputation is usually not a "correct" way of filling in missing values. This is because the conditional distribution of the missing data given the observed data depends on the joint distribution of the full data, which in turn depends on the unknown parameter.
- In fact, there are two processes that are looped with each other: if we knew the missing values, we could deduce the correct parameter based on the full data; if we knew the true parameter, we could correctly fill in the missing values based on the true conditional distribution of the missing given the observed.

The Likelihood-Based Analysis Methods

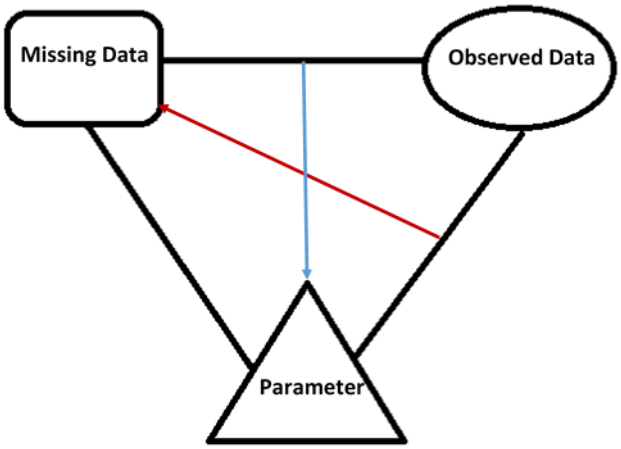

Figure 7.1: The triad of missing data, observed data, and parameter. Red arrow: imputing the missing data depends on the observed data and the parameter; blue arrow: estimating/drawing the parameter depends on the full data, i.e., observed and missing data.

The Likelihood-Based Analysis Methods

- A natural solution to the "three-body" problem is to iterate between the two-steps represented by the red and blue arrows.
- For computation of MLE using the EM algorithm, the red arrow is completed by replacing the unobserved parts of the full-data log-likelihood with their conditional expectations given the observed parts and the current iterate of the parameter, which is essentially the E step.
- The blue arrow is completed by finding the maximizer of the full-data log-likelihood "imputed" from the E step, and this, of course, constitutes the M step.

The Likelihood-Based Analysis Methods

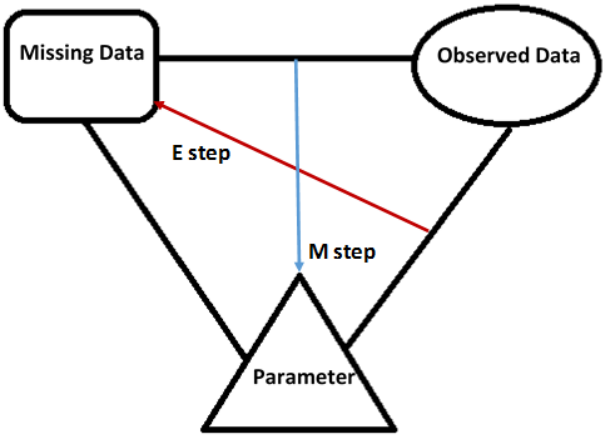

Figure 7.2: The EM solution to the "three-body" problem. Missing values, according to how their appear in the full-data log-likelihood, are replaced by their conditional expectations given observed data and current iterate of the parameter in the E step; the next iterate of parameter is derived based on the "imputed" full-data log-likelihood (the Q function) in the M step.

The Likelihood-Based Analysis Methods

- Under the Bayesian framework, the Bayesian analyst wishes to draw from $[\theta|D_{\text{obs}}]$, and the multiple imputationist wishes to draw from $[D_{\text{mis}}|D_{\text{obs}}]$, both of which may lack an analytic kernel.
- Thus, the following two Gibbs sampler (GS) steps are iterated:

  [GS A]: Draw D_{mis} from $[D_{\text{mis}}|D_{\text{obs}};\theta] \propto L(D|\theta)$;
  [GS B]: Draw θ from $[\theta|D_{\text{obs}}, D_{\text{mis}}] \propto L(D|\theta)q(\theta)$.

  GS A is represented by the red arrow and GS B is represented by the blue arrow.
- After a burn-in period of iterations, the Bayesian analyst collects results from GS B as a posterior sample for θ, and multiple imputationist collects results from GS A and combine them with D_{obs} to form imputed full datasets for aggregated analysis.

The Likelihood-Based Analysis Methods

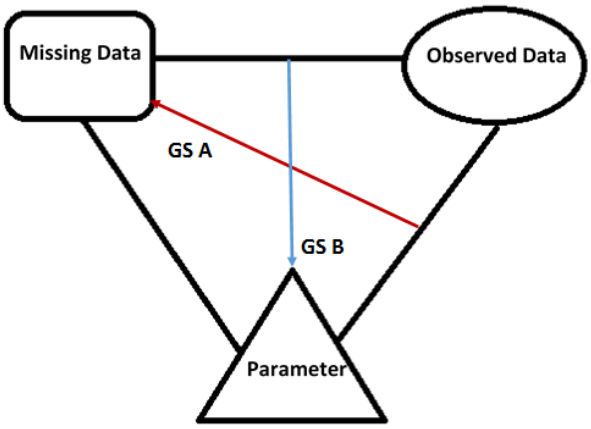

Figure 7.3: The Gibbs sampler solution to the "three-body" problem. GS A: $[D_{\text{mis}}|D_{\text{obs}};\theta]$; GS B: $[\theta|D_{\text{obs}},D_{\text{mis}}]$.

The Likelihood-Based Analysis Methods

- There is a way to break out of the loop if one separates the two processes: the imputation process $(D_{\text{obs}},\theta^*)\rightarrow D_{\text{mis}}$ and the analysis process $(D_{\text{obs}},D_{\text{mis}})\rightarrow\theta$.
- This idea is embodied in the practice of using "uncongenial" imputation models that are mathematically convenient to draw missing values from their posterior predictive distribution than is the analysis model.
- Usually, the "uncongenial" imputation model is set up such that the MLE of model parameters can be easily obtained based on observed data so that the process does not entangle missing data. Then, the posterior predictive distribution of missing data can drawn from by first drawing from the posterior distribution of the parameters and then from the conditional distribution of missing data given observed data and parameters, both sampling steps having closed forms (see examples on pp 5-55 to pp 5-59).

The Likelihood-Based Analysis Methods

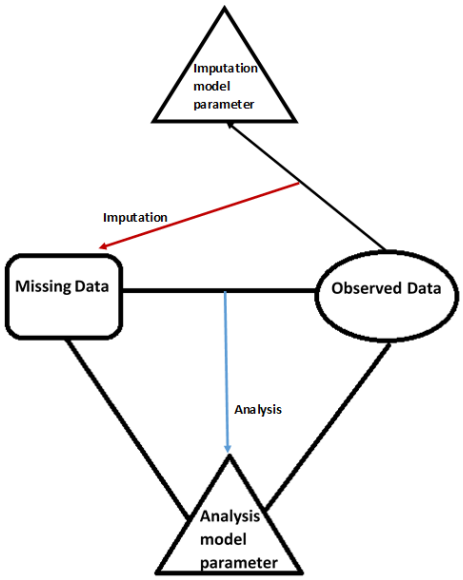

Figure 7.4: Break out of the "three-body" system by using a convenient (uncongenial) imputation model, where model parameters can be easily estimated based on observed data and the posterior predictive distributions have closed forms.

Contents

The Inverse Weighting Technique and Doubly Robustness

- The likelihood-based methods for MAR data basically ignore the missing data mechanism when making inference about a parameter from the sampling distribution $p_Y(y;\theta)$.
- On the contrary, the inverse probability weighting (IPW) uses information on the missingness mechanism to construct M estimators. This is in violation of the *Likelihood Principle*, which states that all information about a parameter is contained in its likelihood function and hence dictates that inference be the same for two models giving rise to the same likelihood (or proportional likelihoods).
- To see the violation, we use the old example where $Y = (X, Y_2)$, Y_2 is possibly missing, the parameter of interest is $\theta := EY_2$, and the distribution of Y is left nonparametric.

The Inverse Weighting Technique and Doubly Robustness

- Write the selection probability function as

$$\pi(X) = \Pr(R = 1|Y),$$

  and suppose we have two models with known selection probabilities $\pi = \pi_0, \pi_0^*$, respectively, where π_0 and π_0^* are two known functions.
- Because of likelihood factorization under MAR, these two models have proportional likelihoods with regard to θ. So by the Principle, inference should be the same.
- However, the IPW estimators for the two models are clearly different, as they are given by

$$n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_0(X_i)}Y_{2_i} \text{ and } n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_0^*(X_i)}Y_{2_i},$$

  respectively.

The Inverse Weighting Technique and Doubly Robustness

- Robins and Ritov (1997) made a strong case in challenging the once axiomatic Principle from an angle of the *curse of dimensionality*.
- They showed that in scenarios like this, no nonparametric estimator for θ that performs reasonably well in finite samples exist unless one engages information on the missingness mechanism and thus violates the Principle.
- The argument went as follows. Any estimator that conforms to the Principle would inevitably estimate the function $\mu(X) := E[Y_2|X]$. Because of the curse of dimensionality, however, nonparametric estimation of μ is infeasible in small- to medium-sized samples when X is high-dimensional and contains multiple continuous components.
- On the other hand, the IPW gives nonparametric estimators that perform well in finite samples.

The Inverse Weighting Technique and Doubly Robustness

- Therefore, IPW-based methods are preferable to the likelihood-based methods when the missing mechanism is *a priori* known, or known to be simple, e.g., missing by design and when the joint distribution of the full data is hard to model.
- Another reason to use IPW-based methods to adjust for confounders while maintaining the causal interpretation of treatment-related parameters (Robins et al., 2000; Hernán, et al, 2001).
- The doubly robust (DR) estimator improves the IPW estimator in terms of both robustness and efficiency.
- DR estimation theory for binary missing and monotone MAR data is well developed. However, for non-monotone MAR data and NMAR data, both IPW and DR techniques still await further exploration.

Contents

Further Readings of Recent Publications

- **Non-monotone missingness**:
  - Sun and Tchetgen Tchetgen (2017; IPW for non-monotone MAR):
- **NMAR**:
  - Miao, W. and Tchetgen Tchetgen (2017; shadow variable for NMAR)
  - Sun et al. (2016; instrumental variables for NMAR)
- **Multiple robustness**:
  - Han and Wang (2013; basic theory)
  - Han (2016; multiple robustness in longitudinal data)
- **Causal inference**:
  - Hernán and Robins (2017) *Causal inference*.
  - Imbens and Rubin (2015) *Causal inference in statistics, social, and biomedical sciences.*
  - van der Laan and Rose (2011). *Targeted learning: causal inference for observational and experimental data*.
  - Pearl (2009) *Causality*.

References

- Han, P. (2016). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. Biometrika.
- Han, P. & Wang, L. (2013). Estimation with missing data: beyond double robustness. Biometrika.
- Hernán, M. A., Brumback, B., & Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association, 96, 440-448.
- Hernán MA, Robins JM (2017). Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming.
- Imbens, G. W. & Rubin, D. B. (2015). Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.
- Pearl, J. (2009). Causality. Cambridge University Press.
- Miao, W. & Tchetgen, E. J. T. (2017). On varieties of doubly robust estimators under missingness not at random with a shadow variable. Biometrika.
- Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.
- Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J., & Tchetgen, E. T. (2016). Semiparametric Estimation with Data Missing Not at Random Using an Instrumental Variable. arXiv preprint arXiv:1607.03197.
- Sun, B. & Tchetgen Tchetgen, E. J. (2017). On Inverse Probability Weighting for Nonmonotone Missing at Random Data. Journal of the American Statistical Association, In press.
- van der Laan, M. J. & Rose, S. (2011). Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Science & Business Media.