Statistical Methods for Analysis with Missing Data

Lecture 16: pattern-mixture models (continued), sensitivity analysis

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Previous Lecture

Introduction to

- The fundamental issue of non-identifiability
- General strategy for identification
- Pattern-mixture models

Today's Lecture

- Common identifying assumptions for pattern-mixture models
  - Reading: Chapter 6 of the lecture notes of Davidian and Tsiatis
- Itemwise conditionally independent nonresponse
- Properties of classes of full-data distributions
- Sensitivity analysis
  - Reading: Chapter 7 of the lecture notes of Davidian and Tsiatis

Outline

Recap from Previous Lecture

Common Identifying Assumptions for Pattern-Mixture Models

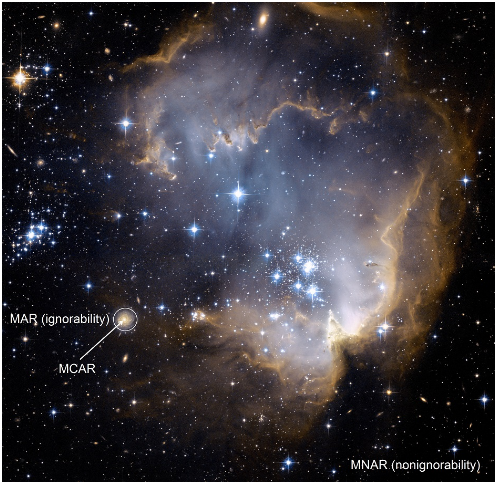
Itemwise Conditionally Independent Nonresponse

Properties of Classes of Full-Data Distributions

Sensitivity Analysis

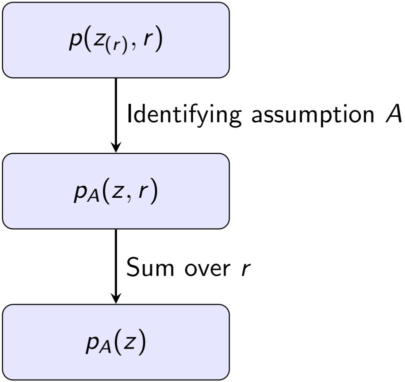
Moving Away from MAR?

- In which direction do we go??
- Remember: there is a universe of missing-data assumptions:



Identification Strategies

- Inference with missing data is impossible without identification assumptions
- Identification strategies generally follow this structure:



Pattern-Mixture Models

- Pattern-mixture models (Little, JASA 1993) provide a transparent way of specifying missing data assumptions
- The pattern-mixture model factorization explicitly reveals:

$$p(z) = \sum_{r \in \{0,1\}^K} \overbrace{p(z_{(\bar{r})} \mid z_{(r)}, r)}^{\text{needs identifying assumption}} \underbrace{p(z_{(r)} \mid r) p(r)}_{\text{can be estimated from data}}$$

- Explicitly shows what needs identifying assumptions and what can be obtained from data alone
- Identifying assumptions explicitly or implicitly amount to constructing $\{p(z_{(\bar{r})} \mid z_{(r)}, r)\}_r$ from $\{p(z_{(r)}, r)\}_r$

Dropout in Longitudinal Study

If missingness only comes from subjects dropping out

- Missingness patterns are uniquely summarized by the dropout time

$$D = 1 + \sum_{j=1}^{T} R_j$$

- The *observed data* are obtained as realizations of

$$(Z_{(D)}, D)$$

- If $D = d$, $Z_{(d)} = (Z_1, \ldots, Z_{d-1})$ and $Z_{(\bar{d})} = (Z_d, \ldots, Z_T)$
- Pattern-mixture model requires modeling the observed-data distribution:
  - $p(D = d)$: simply take empirical frequency
  - $p(z_{(d)} \mid D = d)$: depends on variables' types

Identifying Assumptions for PMMs Under Dropout

- In general, how to obtain $\{p(z_{(\bar{d})} \mid z_{(d)}, d)\}_d$ from $\{p(z_{(d)}, d)\}_d$?
- Note that

$$p(z_{(\bar{d})} \mid z_{(d)}, d) = \prod_{\ell=d}^{T} p(z_\ell \mid z_{(\ell)}, d)$$

- Example: for $T = 3$, we need to identify:
  - If $D = 3$,

$$p(z_{(\bar{3})} \mid z_{(3)}, D = 3) = p(z_3 \mid z_1, z_2, D = 3)$$

  - If $D = 2$,

$$p(z_{(\bar{2})} \mid z_{(2)}, D = 2) = p(z_2 \mid z_1, D = 2)p(z_3 \mid z_1, z_2, D = 2)$$

Identifying Assumptions for PMMs Under Dropout

- In general, how to obtain $\{p(z_{(\bar{d})} \mid z_{(d)}, d)\}_d$ from $\{p(z_{(d)}, d)\}_d$?
- Note that

$$p(z_{(\bar{d})} \mid z_{(d)}, d) = \prod_{\ell=d}^{T} p(z_\ell \mid z_{(\ell)}, d)$$

- Example: for $T = 3$, we need to identify:
  - If $D = 3$,

$$p(z_{(\bar{3})} \mid z_{(3)}, D = 3) = p(z_3 \mid z_1, z_2, D = 3)$$

  - If $D = 2$,

$$p(z_{(\bar{2})} \mid z_{(2)}, D = 2) = p(z_2 \mid z_1, D = 2)p(z_3 \mid z_1, z_2, D = 2)$$

Identifying Assumptions for PMMs Under Dropout

- In general, how to obtain $\{p(z_{(\bar{d})} \mid z_{(d)}, d)\}_d$ from $\{p(z_{(d)}, d)\}_d$?
- Note that

$$p(z_{(\bar{d})} \mid z_{(d)}, d) = \prod_{\ell=d}^{T} p(z_\ell \mid z_{(\ell)}, d)$$

- Example: for $T = 3$, we need to identify:
  - If $D = 3$,

$$p(z_{(\bar{3})} \mid z_{(3)}, D = 3) = p(z_3 \mid z_1, z_2, D = 3)$$

  - If $D = 2$,

$$p(z_{(\bar{2})} \mid z_{(2)}, D = 2) = p(z_2 \mid z_1, D = 2) p(z_3 \mid z_1, z_2, D = 2)$$

Outline

Recap from Previous Lecture

Common Identifying Assumptions for Pattern-Mixture Models

Itemwise Conditionally Independent Nonresponse

Properties of Classes of Full-Data Distributions

Sensitivity Analysis

The Complete-Case Identifying Assumption

Little (JASA 1993) proposed to assume:

$$p_{CC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = T + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Distributions for $D = T + 1$ are identifiable from complete cases
- Example: for $T = 3$, we have:
  - For $D = 3$,

$$p_{CC}(z_3 \mid z_1, z_2, D = 3) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

  - If $D = 2$,

$$p_{CC}(z_2 \mid z_1, D = 2) \equiv p(z_2 \mid z_1, D = 4)$$

$$p_{CC}(z_3 \mid z_1, z_2, D = 2) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

The Complete-Case Identifying Assumption

Little (JASA 1993) proposed to assume:

$$p_{CC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = T + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Distributions for $D = T + 1$ are identifiable from complete cases
- Example: for $T = 3$, we have:
  - For $D = 3$,

$$p_{CC}(z_3 \mid z_1, z_2, D = 3) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

  - If $D = 2$,

$$p_{CC}(z_2 \mid z_1, D = 2) \equiv p(z_2 \mid z_1, D = 4)$$

$$p_{CC}(z_3 \mid z_1, z_2, D = 2) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

The Complete-Case Identifying Assumption

Little (JASA 1993) proposed to assume:

$$p_{CC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = T + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Distributions for $D = T + 1$ are identifiable from complete cases
- Example: for $T = 3$, we have:
  - For $D = 3$,

$$p_{CC}(z_3 \mid z_1, z_2, D = 3) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

  - If $D = 2$,

$$p_{CC}(z_2 \mid z_1, D = 2) \equiv p(z_2 \mid z_1, D = 4)$$

$$p_{CC}(z_3 \mid z_1, z_2, D = 2) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

The Complete-Case Identifying Assumption

Little (JASA 1993) proposed to assume:

$$p_{CC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = T + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Distributions for $D = T + 1$ are identifiable from complete cases
- Example: for $T = 3$, we have:
  - For $D = 3$,

$$p_{CC}(z_3 \mid z_1, z_2, D = 3) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

  - If $D = 2$,

$$p_{CC}(z_2 \mid z_1, D = 2) \equiv p(z_2 \mid z_1, D = 4)$$

$$p_{CC}(z_3 \mid z_1, z_2, D = 2) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

The Complete-Case Identifying Assumption

Little (JASA 1993) proposed to assume:

$$p_{CC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = T + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Distributions for $D = T + 1$ are identifiable from complete cases
- Example: for $T = 3$, we have:
  - For $D = 3$,

$$p_{CC}(z_3 \mid z_1, z_2, D = 3) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

  - If $D = 2$,

$$p_{CC}(z_2 \mid z_1, D = 2) \equiv p(z_2 \mid z_1, D = 4)$$

$$p_{CC}(z_3 \mid z_1, z_2, D = 2) \equiv p(z_3 \mid z_1, z_2, D = 4)$$

The Neighboring-Case Identifying Assumption

The extrapolation distributions could also be obtained from the closest dropout pattern where ℓ is available:

$$p_{NC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = \ell + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among observations with $D = \ell + 1$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that observations with $D = \ell + 1$ are the best for basing extrapolation of the values of Z_ℓ
  - For example, among observations where Z_ℓ is available, those who dropout at time $\ell + 1$ might be the most similar to those that dropout at time ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.

The Neighboring-Case Identifying Assumption

The extrapolation distributions could also be obtained from the closest dropout pattern where ℓ is available:

$$p_{NC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = \ell + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among observations with $D = \ell + 1$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that observations with $D = \ell + 1$ are the best for basing extrapolation of the values of Z_ℓ
  - For example, among observations where Z_ℓ is available, those who dropout at time $\ell + 1$ might be the most similar to those that dropout at time ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.

The Neighboring-Case Identifying Assumption

The extrapolation distributions could also be obtained from the closest dropout pattern where ℓ is available:

$$p_{NC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = \ell + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among observations with $D = \ell + 1$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that observations with $D = \ell + 1$ are the best for basing extrapolation of the values of Z_ℓ
  - For example, among observations where Z_ℓ is available, those who dropout at time $\ell + 1$ might be the most similar to those that dropout at time ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.

The Neighboring-Case Identifying Assumption

The extrapolation distributions could also be obtained from the closest dropout pattern where ℓ is available:

$$p_{NC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = \ell + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among observations with $D = \ell + 1$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that observations with $D = \ell + 1$ are the best for basing extrapolation of the values of Z_ℓ
  - For example, among observations where Z_ℓ is available, those who dropout at time $\ell + 1$ might be the most similar to those that dropout at time ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.

The Neighboring-Case Identifying Assumption

The extrapolation distributions could also be obtained from the closest dropout pattern where ℓ is available:

$$p_{NC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D = \ell + 1),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among observations with $D = \ell + 1$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that observations with $D = \ell + 1$ are the best for basing extrapolation of the values of Z_ℓ
  - For example, among observations where Z_ℓ is available, those who dropout at time $\ell + 1$ might be the most similar to those that dropout at time ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.

The Available-Case Identifying Assumption

Here, the extrapolation distributions are obtained from all available cases where ℓ is available:

$$p_{AC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D > \ell),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among all observations with $D > \ell$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that this approach maximizes the use of available information for basing extrapolation of the values of Z_ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.
- HW4: under monotone nonresponse, the AC assumption is equivalent to MAR

The Available-Case Identifying Assumption

Here, the extrapolation distributions are obtained from all available cases where ℓ is available:

$$p_{AC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D > \ell),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among all observations with $D > \ell$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that this approach maximizes the use of available information for basing extrapolation of the values of Z_ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.
- HW4: under monotone nonresponse, the AC assumption is equivalent to MAR

The Available-Case Identifying Assumption

Here, the extrapolation distributions are obtained from all available cases where ℓ is available:

$$p_{AC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D > \ell),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among all observations with $D > \ell$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that this approach maximizes the use of available information for basing extrapolation of the values of Z_ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.
- HW4: under monotone nonresponse, the AC assumption is equivalent to MAR

The Available-Case Identifying Assumption

Here, the extrapolation distributions are obtained from all available cases where ℓ is available:

$$p_{AC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D > \ell),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among all observations with $D > \ell$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that this approach maximizes the use of available information for basing extrapolation of the values of Z_ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.
- HW4: under monotone nonresponse, the AC assumption is equivalent to MAR

The Available-Case Identifying Assumption

Here, the extrapolation distributions are obtained from all available cases where ℓ is available:

$$p_{AC}(z_\ell \mid z_{(\ell)}, D = d) \equiv p(z_\ell \mid z_{(\ell)}, D > \ell),$$

for all $\ell \geq d$, $d = 1, \ldots, T$.

- Among all observations with $D > \ell$ we get to observe z_ℓ and $z_{(\ell)}$
- We could think that this approach maximizes the use of available information for basing extrapolation of the values of Z_ℓ
- HW4: say $T = 3$, write down this restriction for $\ell \geq d$, $d = 1, 2, 3$.
- HW4: under monotone nonresponse, the AC assumption is equivalent to MAR

Outline

Itemwise Conditionally Independent Nonresponse

- Identification assumptions can also be expressed as restrictions on the full-data distribution

- The *itemwise conditionally independent nonresponse* (ICIN)[1] assumption says that

$$Z_j \perp\!\!\!\perp R_j \mid Z_{-j}, R_{-j}, \quad \text{for all} \;\; j = 1, \ldots, K,$$

where $Z_{-j} = (\ldots, Z_{j-1}, Z_{j+1}, \ldots)$, $R_{-j} = (\ldots, R_{j-1}, R_{j+1}, \ldots)$

- Remark: Z_j and R_j being conditionally independent does not imply marginal independence

[1]Sadinle & Reiter (Biometrika 2017): https://doi.org/10.1093/biomet/asw063

Itemwise Conditionally Independent Nonresponse

- Identification assumptions can also be expressed as restrictions on the full-data distribution
- The *itemwise conditionally independent nonresponse* (ICIN)[1] assumption says that

$$Z_j \perp\!\!\!\perp R_j \mid Z_{-j}, R_{-j}, \quad \text{for all} \;\; j = 1, \ldots, K,$$

where $Z_{-j} = (\ldots, Z_{j-1}, Z_{j+1}, \ldots)$, $R_{-j} = (\ldots, R_{j-1}, R_{j+1}, \ldots)$

- Remark: Z_j and R_j being conditionally independent does not imply marginal independence

[1] Sadinle & Reiter (Biometrika 2017): https://doi.org/10.1093/biomet/asw063

ICIN Distribution

- Sadinle & Reiter showed how to construct a full-data distribution that encodes ICIN given an observed-data distribution
- For each missingness pattern $r \in \{0,1\}^K$, given $p(z_{(r)}, r) > 0$, let the function $\eta_r : \mathcal{Z}_{(r)} \mapsto \mathbb{R}$ be defined recursively as

$$\eta_r(z_{(r)}) = \log p(z_{(r)}, r) - \log \int_{\mathcal{Z}_{(\bar{r})}} \exp\left\{ \sum_{\bar{r}' \prec \bar{r}} \eta_{r'}(z_{(r')}) \right\} \mu(dz_{(\bar{r})}).$$

  Then

$$p_{ICIN}(z, r) = \exp\left\{ \sum_{\bar{r}' \preceq \bar{r}} \eta_{r'}(z_{(r')}) \right\}$$
$$= p(z_{(r)}, r) \frac{\exp\left\{ \sum_{\bar{r}' \prec \bar{r}} \eta_{r'}(z_{(r')}) \right\}}{\int_{\mathcal{Z}_{(\bar{r})}} \exp\left\{ \sum_{\bar{r}' \prec \bar{r}} \eta_{r'}(z_{(r')}) \right\} \mu(dz_{(\bar{r})})}$$

- Therefore ICIN can be seen as a restriction for pattern-mixture models!

ICIN Distribution

- Sadinle & Reiter showed how to construct a full-data distribution that encodes ICIN given an observed-data distribution
- For each missingness pattern $r \in \{0,1\}^K$, given $p(z_{(r)}, r) > 0$, let the function $\eta_r : \mathcal{Z}_{(r)} \mapsto \mathbb{R}$ be defined recursively as

$$\eta_r(z_{(r)}) = \log p(z_{(r)}, r) - \log \int_{\mathcal{Z}_{(\bar{r})}} \exp\left\{ \sum_{\bar{r}' \prec \bar{r}} \eta_{r'}(z_{(r')}) \right\} \mu(dz_{(\bar{r})}).$$

  Then

$$p_{ICIN}(z, r) = \exp\left\{ \sum_{\bar{r}' \preceq \bar{r}} \eta_{r'}(z_{(r')}) \right\}$$

$$= p(z_{(r)}, r) \frac{\exp\left\{ \sum_{\bar{r}' \prec \bar{r}} \eta_{r'}(z_{(r')}) \right\}}{\int_{\mathcal{Z}_{(\bar{r})}} \exp\left\{ \sum_{\bar{r}' \prec \bar{r}} \eta_{r'}(z_{(r')}) \right\} \mu(dz_{(\bar{r})})}$$

- Therefore ICIN can be seen as a restriction for pattern-mixture models!

Outline

Properties of Full-Data Distributions

- Assumptions covered so far: MCAR, MAR, CC, NC, AC, ICIN
- Can we talk about properties of the implied classes of full-data distributions?[2]
  - Observational equivalence
  - Full-data identifiability
  - (Observed-data) identifiability
  - Nonparametric identifiability

[2] Taken from Sadinle & Reiter (forthcoming in Biometrika): https://arxiv.org/pdf/1902.06043.pdf

Observational Equivalence

- Two full-data distributions are said to be *observationally equivalent* if their implied observed-data distributions are the same
- This is, say I have two full-data distributions:

$$p_A(z_{(\bar{r})}, z_{(r)}, r) \quad \text{and} \quad p_B(z_{(\bar{r})}, z_{(r)}, r).$$

If

$$\int p_A(z_{(\bar{r})}, z_{(r)}, r)\, dz_{(\bar{r})} = \int p_B(z_{(\bar{r})}, z_{(r)}, r)\, dz_{(\bar{r})}$$

for all $(z_{(r)}, r)$, then they are *observationally equivalent*

- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are observationally equivalent (under dropout)
- This is an important feature in *sensitivity analysis*, because differences in inferences will be due to the different identifying assumptions and not due to different fits to the observed data!

Observational Equivalence

- Two full-data distributions are said to be *observationally equivalent* if their implied observed-data distributions are the same
- This is, say I have two full-data distributions:

$$p_A(z_{(\bar{r})}, z_{(r)}, r) \quad \text{and} \quad p_B(z_{(\bar{r})}, z_{(r)}, r).$$

  If

$$\int p_A(z_{(\bar{r})}, z_{(r)}, r)\; dz_{(\bar{r})} = \int p_B(z_{(\bar{r})}, z_{(r)}, r)\; dz_{(\bar{r})}$$

  for all $(z_{(r)}, r)$, then they are *observationally equivalent*

- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are observationally equivalent (under dropout)
- This is an important feature in *sensitivity analysis*, because differences in inferences will be due to the different identifying assumptions and not due to different fits to the observed data!

Observational Equivalence

- Two full-data distributions are said to be *observationally equivalent* if their implied observed-data distributions are the same

- This is, say I have two full-data distributions:

$$p_A(z_{(\bar{r})}, z_{(r)}, r) \quad \text{and} \quad p_B(z_{(\bar{r})}, z_{(r)}, r).$$

  If

$$\int p_A(z_{(\bar{r})}, z_{(r)}, r)\; dz_{(\bar{r})} = \int p_B(z_{(\bar{r})}, z_{(r)}, r)\; dz_{(\bar{r})}$$

  for all $(z_{(r)}, r)$, then they are *observationally equivalent*

- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are observationally equivalent (under dropout)

- This is an important feature in *sensitivity analysis*, because differences in inferences will be due to the different identifying assumptions and not due to different fits to the observed data!

Observational Equivalence

- Two full-data distributions are said to be *observationally equivalent* if their implied observed-data distributions are the same

- This is, say I have two full-data distributions:

$$p_A(z_{(\bar{r})}, z_{(r)}, r) \quad \text{and} \quad p_B(z_{(\bar{r})}, z_{(r)}, r).$$

  If

$$\int p_A(z_{(\bar{r})}, z_{(r)}, r)\ dz_{(\bar{r})} = \int p_B(z_{(\bar{r})}, z_{(r)}, r)\ dz_{(\bar{r})}$$

  for all $(z_{(r)}, r)$, then they are *observationally equivalent*

- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are observationally equivalent (under dropout)

- This is an important feature in *sensitivity analysis*, because differences in inferences will be due to the different identifying assumptions and not due to different fits to the observed data!

Observational Equivalence

- Two full-data distributions are said to be *observationally equivalent* if their implied observed-data distributions are the same
- This is, say I have two full-data distributions:

$$p_A(z_{(\bar{r})}, z_{(r)}, r) \quad \text{and} \quad p_B(z_{(\bar{r})}, z_{(r)}, r).$$

If

$$\int p_A(z_{(\bar{r})}, z_{(r)}, r)\ dz_{(\bar{r})} = \int p_B(z_{(\bar{r})}, z_{(r)}, r)\ dz_{(\bar{r})}$$

for all $(z_{(r)}, r)$, then they are *observationally equivalent*

- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are observationally equivalent (under dropout)
- This is an important feature in *sensitivity analysis*, because differences in inferences will be due to the different identifying assumptions and not due to different fits to the observed data!

Observational Equivalence

- Two full-data distributions are said to be *observationally equivalent* if their implied observed-data distributions are the same
- This is, say I have two full-data distributions:

$$p_A(z_{(\bar{r})}, z_{(r)}, r) \quad \text{and} \quad p_B(z_{(\bar{r})}, z_{(r)}, r).$$

  If

$$\int p_A(z_{(\bar{r})}, z_{(r)}, r)\ dz_{(\bar{r})} = \int p_B(z_{(\bar{r})}, z_{(r)}, r)\ dz_{(\bar{r})}$$

  for all $(z_{(r)}, r)$, then they are *observationally equivalent*
- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are observationally equivalent (under dropout)
- This is an important feature in *sensitivity analysis*, because differences in inferences will be due to the different identifying assumptions and not due to different fits to the observed data!

Full-Data Identifiability

- $\mathcal{C}_\Theta$: class of full-data distributions
- Θ: parameter space, either finite- or infinite-dimensional
- Say we were able to observe Z regardless of the value of R
- Identifiability of $\mathcal{C}_\Theta$ in the usual sense (e.g., Lehmann & Casella 1998, p. 24) here is referred to as *full-data identifiability*
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *full-data identifiable* if there exists a bijection from Θ to $\mathcal{C}_\Theta$
- Full-data identifiability is an elementary requirement which simply says that the class is properly parameterized

Full-Data Identifiability

- $\mathcal{C}_\Theta$: class of full-data distributions
- Θ: parameter space, either finite- or infinite-dimensional
- Say we were able to observe Z regardless of the value of R
- Identifiability of $\mathcal{C}_\Theta$ in the usual sense (e.g., Lehmann & Casella 1998, p. 24) here is referred to as *full-data identifiability*
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *full-data identifiable* if there exists a bijection from Θ to $\mathcal{C}_\Theta$
- Full-data identifiability is an elementary requirement which simply says that the class is properly parameterized

Full-Data Identifiability

- $\mathcal{C}_\Theta$: class of full-data distributions
- Θ: parameter space, either finite- or infinite-dimensional
- Say we were able to observe Z regardless of the value of R
- Identifiability of $\mathcal{C}_\Theta$ in the usual sense (e.g., Lehmann & Casella 1998, p. 24) here is referred to as *full-data identifiability*
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *full-data identifiable* if there exists a bijection from Θ to $\mathcal{C}_\Theta$
- Full-data identifiability is an elementary requirement which simply says that the class is properly parameterized

Full-Data Identifiability

- $\mathcal{C}_\Theta$: class of full-data distributions
- Θ: parameter space, either finite- or infinite-dimensional
- Say we were able to observe Z regardless of the value of R
- Identifiability of $\mathcal{C}_\Theta$ in the usual sense (e.g., Lehmann & Casella 1998, p. 24) here is referred to as *full-data identifiability*
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *full-data identifiable* if there exists a bijection from Θ to $\mathcal{C}_\Theta$
- Full-data identifiability is an elementary requirement which simply says that the class is properly parameterized

Full-Data Identifiability

- $\mathcal{C}_\Theta$: class of full-data distributions
- Θ: parameter space, either finite- or infinite-dimensional
- Say we were able to observe Z regardless of the value of R
- Identifiability of $\mathcal{C}_\Theta$ in the usual sense (e.g., Lehmann & Casella 1998, p. 24) here is referred to as *full-data identifiability*
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *full-data identifiable* if there exists a bijection from Θ to $\mathcal{C}_\Theta$
- Full-data identifiability is an elementary requirement which simply says that the class is properly parameterized

Full-Data Identifiability

- $\mathcal{C}_\Theta$: class of full-data distributions
- Θ: parameter space, either finite- or infinite-dimensional
- Say we were able to observe Z regardless of the value of R
- Identifiability of $\mathcal{C}_\Theta$ in the usual sense (e.g., Lehmann & Casella 1998, p. 24) here is referred to as *full-data identifiability*
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *full-data identifiable* if there exists a bijection from Θ to $\mathcal{C}_\Theta$
- Full-data identifiability is an elementary requirement which simply says that the class is properly parameterized

(Observed-Data) Identifiability

- obs($\mathcal{C}_\Theta$): the class of observed-data distributions implied by $\mathcal{C}_\Theta$
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *identifiable* if there exist bijections from Θ to $\mathcal{C}_\Theta$ and from obs($\mathcal{C}_\Theta$) to $\mathcal{C}_\Theta$
  - First bijection: full-data identifiability for $\mathcal{C}_\Theta$
  - Second bijection: we need a unique way to go back and forth from obs($\mathcal{C}_\Theta$) to $\mathcal{C}_\Theta$
  - These imply a third bijection between obs($\mathcal{C}_\Theta$) and Θ: the common notion of identifiability applied to obs($\mathcal{C}_\Theta$)

(Observed-Data) Identifiability

- $\text{obs}(\mathcal{C}_\Theta)$: the class of observed-data distributions implied by $\mathcal{C}_\Theta$
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *identifiable* if there exist bijections from Θ to $\mathcal{C}_\Theta$ and from $\text{obs}(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$
  - First bijection: full-data identifiability for $\mathcal{C}_\Theta$
  - Second bijection: we need a unique way to go back and forth from $\text{obs}(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$
  - These imply a third bijection between $\text{obs}(\mathcal{C}_\Theta)$ and Θ: the common notion of identifiability applied to $\text{obs}(\mathcal{C}_\Theta)$

(Observed-Data) Identifiability

- $\text{obs}(\mathcal{C}_\Theta)$: the class of observed-data distributions implied by $\mathcal{C}_\Theta$
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *identifiable* if there exist bijections from Θ to $\mathcal{C}_\Theta$ and from $\text{obs}(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$
  - First bijection: full-data identifiability for $\mathcal{C}_\Theta$
  - Second bijection: we need a unique way to go back and forth from $\text{obs}(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$
  - These imply a third bijection between $\text{obs}(\mathcal{C}_\Theta)$ and Θ: the common notion of identifiability applied to $\text{obs}(\mathcal{C}_\Theta)$

(Observed-Data) Identifiability

- $\text{obs}(\mathcal{C}_\Theta)$: the class of observed-data distributions implied by $\mathcal{C}_\Theta$
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *identifiable* if there exist bijections from Θ to $\mathcal{C}_\Theta$ and from $\text{obs}(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$
  - First bijection: full-data identifiability for $\mathcal{C}_\Theta$
  - Second bijection: we need a unique way to go back and forth from $\text{obs}(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$
  - These imply a third bijection between $\text{obs}(\mathcal{C}_\Theta)$ and Θ: the common notion of identifiability applied to $\text{obs}(\mathcal{C}_\Theta)$

(Observed-Data) Identifiability

- $\text{obs}(\mathcal{C}_\Theta)$: the class of observed-data distributions implied by $\mathcal{C}_\Theta$
- A class of full-data distributions $\mathcal{C}_\Theta$ is said to be *identifiable* if there exist bijections from Θ to $\mathcal{C}_\Theta$ and from $\text{obs}(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$
  - First bijection: full-data identifiability for $\mathcal{C}_\Theta$
  - Second bijection: we need a unique way to go back and forth from $\text{obs}(\mathcal{C}_\Theta)$ to $\mathcal{C}_\Theta$
  - These imply a third bijection between $\text{obs}(\mathcal{C}_\Theta)$ and Θ: the common notion of identifiability applied to $\text{obs}(\mathcal{C}_\Theta)$

Nonparametric Identifiability

- $\mathcal{G}$: all possible observed-data distributions
- Say obs($\mathcal{C}_\Theta$) is a proper subset of $\mathcal{G}$: $\mathcal{C}_\Theta$ imposes parametric restrictions on what could be nonparametrically recovered from observed data alone
- $\mathcal{C}_\Theta$ is said to be *nonparametrically identifiable* if it is identifiable and obs($\mathcal{C}_\Theta$) $= \mathcal{G}$
- Also known as nonparametric saturation or just-identification (Robins 1997, Vansteelandt et al. 2006, Hoonhout & Ridder 2018)
- Bijection between Θ and obs($\mathcal{C}_\Theta$) $= \mathcal{G}$: we can think of a nonparametrically identifiable class as being indexed by the set of all observed-data distributions $\mathcal{G}$

Nonparametric Identifiability

- $\mathcal{G}$: all possible observed-data distributions
- Say obs$(\mathcal{C}_\Theta)$ is a proper subset of $\mathcal{G}$: $\mathcal{C}_\Theta$ imposes parametric restrictions on what could be nonparametrically recovered from observed data alone
- $\mathcal{C}_\Theta$ is said to be *nonparametrically identifiable* if it is identifiable and obs$(\mathcal{C}_\Theta) = \mathcal{G}$
- Also known as nonparametric saturation or just-identification (Robins 1997, Vansteelandt et al. 2006, Hoonhout & Ridder 2018)
- Bijection between Θ and obs$(\mathcal{C}_\Theta) = \mathcal{G}$: we can think of a nonparametrically identifiable class as being indexed by the set of all observed-data distributions $\mathcal{G}$

Nonparametric Identifiability

- $\mathcal{G}$: all possible observed-data distributions
- Say $\text{obs}(\mathcal{C}_\Theta)$ is a proper subset of $\mathcal{G}$: $\mathcal{C}_\Theta$ imposes parametric restrictions on what could be nonparametrically recovered from observed data alone
- $\mathcal{C}_\Theta$ is said to be *nonparametrically identifiable* if it is identifiable and $\text{obs}(\mathcal{C}_\Theta) = \mathcal{G}$
- Also known as nonparametric saturation or just-identification (Robins 1997, Vansteelandt et al. 2006, Hoonhout & Ridder 2018)
- Bijection between Θ and $\text{obs}(\mathcal{C}_\Theta) = \mathcal{G}$: we can think of a nonparametrically identifiable class as being indexed by the set of all observed-data distributions $\mathcal{G}$

Nonparametric Identifiability

- $\mathcal{G}$: all possible observed-data distributions
- Say obs($\mathcal{C}_\Theta$) is a proper subset of $\mathcal{G}$: $\mathcal{C}_\Theta$ imposes parametric restrictions on what could be nonparametrically recovered from observed data alone
- $\mathcal{C}_\Theta$ is said to be *nonparametrically identifiable* if it is identifiable and obs($\mathcal{C}_\Theta$) $= \mathcal{G}$
- Also known as nonparametric saturation or just-identification (Robins 1997, Vansteelandt et al. 2006, Hoonhout & Ridder 2018)
- Bijection between Θ and obs($\mathcal{C}_\Theta$) $= \mathcal{G}$: we can think of a nonparametrically identifiable class as being indexed by the set of all observed-data distributions $\mathcal{G}$

Nonparametric Identifiability

- $\mathcal{G}$: all possible observed-data distributions
- Say $\text{obs}(\mathcal{C}_\Theta)$ is a proper subset of $\mathcal{G}$: $\mathcal{C}_\Theta$ imposes parametric restrictions on what could be nonparametrically recovered from observed data alone
- $\mathcal{C}_\Theta$ is said to be *nonparametrically identifiable* if it is identifiable and $\text{obs}(\mathcal{C}_\Theta) = \mathcal{G}$
- Also known as nonparametric saturation or just-identification (Robins 1997, Vansteelandt et al. 2006, Hoonhout & Ridder 2018)
- Bijection between Θ and $\text{obs}(\mathcal{C}_\Theta) = \mathcal{G}$: we can think of a nonparametrically identifiable class as being indexed by the set of all observed-data distributions $\mathcal{G}$

Nonparametric Identifiability

- Two nonparametrically identifiable classes are necessarily observationally equivalent
- Nonparametric identification additionally guarantees that these restrictions do not constrain the observed-data distribution, and therefore cannot be rejected based on the observed data
- Nonparametric identification is therefore a basic desirable property, particularly useful for comparing inferences under different missing data assumptions
- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are nonparametric identified
- Fun fact: MAR and ICIN lead to nonparametric identification

Nonparametric Identifiability

- Two nonparametrically identifiable classes are necessarily observationally equivalent
- Nonparametric identification additionally guarantees that these restrictions do not constrain the observed-data distribution, and therefore cannot be rejected based on the observed data
- Nonparametric identification is therefore a basic desirable property, particularly useful for comparing inferences under different missing data assumptions
- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are nonparametric identified
- Fun fact: MAR and ICIN lead to nonparametric identification

Nonparametric Identifiability

- Two nonparametrically identifiable classes are necessarily observationally equivalent
- Nonparametric identification additionally guarantees that these restrictions do not constrain the observed-data distribution, and therefore cannot be rejected based on the observed data
- Nonparametric identification is therefore a basic desirable property, particularly useful for comparing inferences under different missing data assumptions
- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are nonparametric identified
- Fun fact: MAR and ICIN lead to nonparametric identification

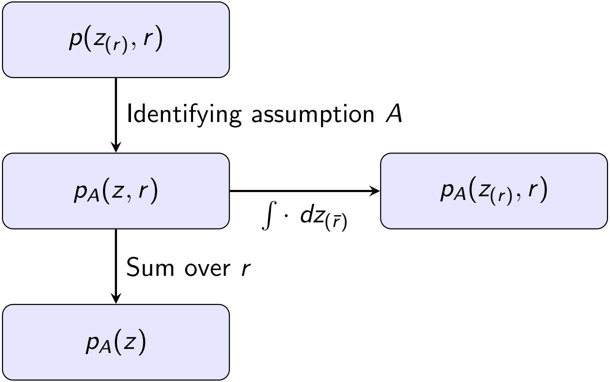
Nonparametric Identifiability

- Two nonparametrically identifiable classes are necessarily observationally equivalent
- Nonparametric identification additionally guarantees that these restrictions do not constrain the observed-data distribution, and therefore cannot be rejected based on the observed data
- Nonparametric identification is therefore a basic desirable property, particularly useful for comparing inferences under different missing data assumptions
- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are nonparametric identified
- Fun fact: MAR and ICIN lead to nonparametric identification

Nonparametric Identifiability

- Two nonparametrically identifiable classes are necessarily observationally equivalent
- Nonparametric identification additionally guarantees that these restrictions do not constrain the observed-data distribution, and therefore cannot be rejected based on the observed data
- Nonparametric identification is therefore a basic desirable property, particularly useful for comparing inferences under different missing data assumptions
- HW4: the full-data distributions obtained under the CC, NC, and AC assumptions are nonparametric identified
- Fun fact: MAR and ICIN lead to nonparametric identification

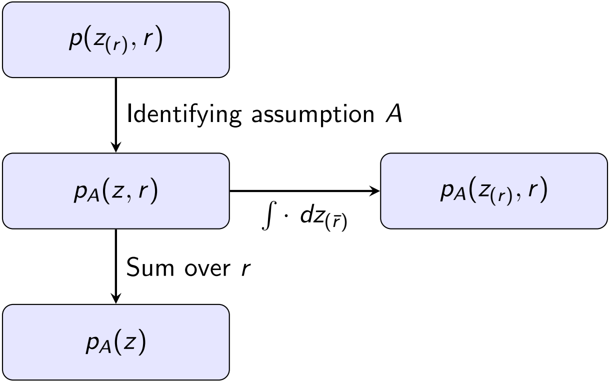
Nonparametric Identifiability



If $p_A(z_{(r)}, r) = p(z_{(r)}, r)$, then A leads to nonparametric identification:

- A does not impose restrictions on observed-data distribution
- A cannot be rejected from observed data alone

Nonparametric Identifiability



If $p_A(z_{(r)}, r) = p(z_{(r)}, r)$, then A leads to nonparametric identification:

- A does not impose restrictions on observed-data distribution
- A cannot be rejected from observed data alone

Outline

Sensitivity Analysis

Based on Scharfstein et al. (Biometrics 2018)[3]:

- *Local sensitivity analysis*
  - Make a missing-data assumption A, explore changes in inferences implied by departures from A in a "neighborhood" around A
    - See Chapter 7 of the lecture notes of Davidian and Tsiatisfor more on this
- *Global sensitivity analysis*
  - Make completely different missing-data assumptions, explore inferences under all such assumptions
    - We will illustrate global sensitivity analysis with an example presented by Sadinle & Reiter (2017)

[3] https://doi.org/10.1111/biom.12729

The Slovenian Plebiscite Data Revisited

- Slovenians voted for independence from Yugoslavia in a plebiscite in 1991
- The Slovenian public opinion survey included these questions:
  1. Independence: Are you in favor of Slovenian independence?
  2. Secession: Are you in favor of Slovenia's secession from Yugoslavia?
  3. Attendance: Will you attend the plebiscite?
- Rubin, Stern and Vehovar (1995) analyzed these three questions under MAR
- Plebiscite results give us the proportions of non-attendants and attendants in favor of independence

The Slovenian Plebiscite Data Revisited

- Slovenians voted for independence from Yugoslavia in a plebiscite in 1991

- The Slovenian public opinion survey included these questions:

  1. Independence: Are you in favor of Slovenian independence?
  2. Secession: Are you in favor of Slovenia's secession from Yugoslavia?
  3. Attendance: Will you attend the plebiscite?

- Rubin, Stern and Vehovar (1995) analyzed these three questions under MAR

- Plebiscite results give us the proportions of non-attendants and attendants in favor of independence

The Slovenian Plebiscite Data Revisited

Table 1. SPO Survey Results (n = 2,074)

		Independence		
Secession	*Attendance*	*Yes*	*No*	*Don't know*
Yes	Yes	1,191	8	21
	No	8	0	4
	Don't Know	107	3	9
No	Yes	158	68	29
	No	7	14	3
	Don't Know	18	43	31
Don't Know	Yes	90	2	109
	No	1	2	25
	Don't Know	19	8	96

Taken from Rubin, Stern and Vehovar (JASA, 1995)

Slovenian Data: a Bayesian Approach

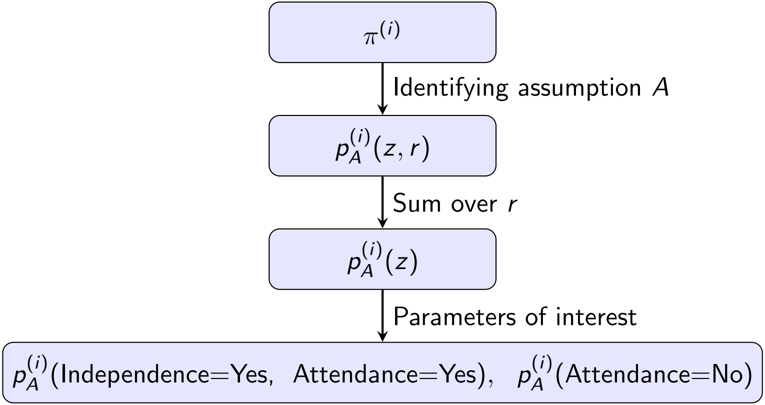
- Let $\pi_{z_{(r)},r} = p(z_{(r)}, r), \quad \pi = \{\pi_{z_{(r)},r}\}$
- Draw $\pi^{(i)}$ from posterior of π given the Slovenian data (problems 9 and 10 of HW3)
- Do, for $i = 1, \ldots, m$:

Slovenian Data: a Bayesian Approach

- Let $\pi_{z_{(r)},r} = p(z_{(r)}, r)$, $\quad \pi = \{\pi_{z_{(r)},r}\}$
- Draw $\pi^{(i)}$ from posterior of π given the Slovenian data (problems 9 and 10 of HW3)
- Do, for $i = 1, \ldots, m$:

Slovenian Data: a Bayesian Approach

- Let $\pi_{z_{(r)},r} = p(z_{(r)}, r), \quad \pi = \{\pi_{z_{(r)},r}\}$
- Draw $\pi^{(i)}$ from posterior of π given the Slovenian data (problems 9 and 10 of HW3)
- Do, for $i = 1, \ldots, m$:



Global Sensitivity Analysis for The Slovenian Data

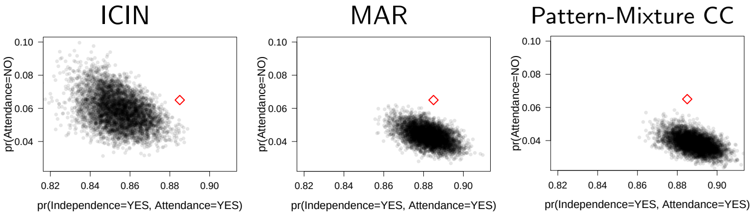


Figure: Samples from joint posterior distributions of p(Independence = Yes, Attendance = Yes) and p(Attendance = No). Red diamond: results of plebiscite.

- All of these identifying assumptions lead to nonparametric identifiability!

Global Sensitivity Analysis for The Slovenian Data

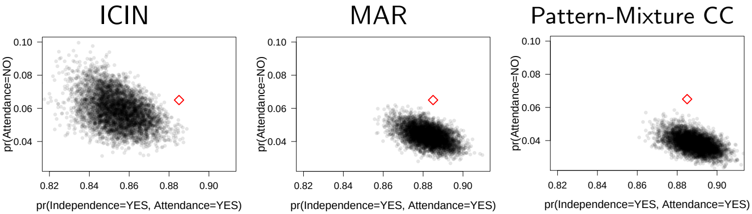


Figure: Samples from joint posterior distributions of p(Independence = Yes, Attendance = Yes) and p(Attendance = No). Red diamond: results of plebiscite.

- All of these identifying assumptions lead to nonparametric identifiability!

Summary

Main take-aways from today's lecture:

- Pattern-mixture models provide an alternative way of thinking about missing data
- Ensuring identifiability is key when moving away from MAR
- Nonparametric identifiability is an important and desirable property
- Sensitivity analyses are recommended

Summary

Main take-aways from this class:

- The fundamental problem of inference with missing data: *it is impossible without extra, usually untestable, assumptions on how missingness arises*
- Willing to assume MAR?: plethora of approaches readily available:
  - Frequentist likelihood-based inference (EM algorithm)
  - Bayesian inference (Data Augmentation / Gibbs samplers)
  - Multiple imputation (although worry about uncongeniality)
  - Inverse-probability weighting (also, double robust procedures)

  (they also could handle nonignorable missingness – current area of research)
- Avoid ad-hoc approaches such as single imputation and complete-case analysis

Thank you for your patience and participation!

THE END