Nonparametric Pattern-Mixture Models for Inference with Missing Data

Yen-Chi Chen

Department of Statistics
University of Washington

- Joint work with Mauricio Sadinle

- Supported by NSF DMS - 1810960

A regular statistical problem

- We observe IID study variables $X_1, \cdots, X_n \in \mathbb{R}^d$ from a distribution F with a PDF p.
- Our goal is to make inference about a parameter of interest that can be written as a statistical functional

$$\theta = \theta(F).$$

- Common example: the mean vector, the covariance matrix, ...etc.

A regular statistical problem

- We observe IID study variables $X_1, \cdots, X_n \in \mathbb{R}^d$ from a distribution F with a PDF p.
- Our goal is to make inference about a parameter of interest that can be written as a statistical functional

$$\theta = \theta(F).$$

- Common example: the mean vector, the covariance matrix, ...etc.
- A common (nonparametric) estimator: plug-in with the empirical distribution function (EDF)

$$\widehat{\theta}_{\mathsf{naive}} = \theta(\widehat{F}), \quad \widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x).$$

A toy example

ID	X_1	X_2	X_3	X_4
1	15	20	17	32
2	12	15	17	21
3	17	43	35	42
4	11	25	23	43
5	16	37	32	51
6	15	23	32	44
7	21	27	35	53

A toy example

ID	X_1	X_2	X_3	X_4
1	15	20	NA	NA
2	12	NA	NA	NA
3	17	43	35	42
4	11	25	NA	NA
5	16	37	32	51
6	15	23	32	NA
7	21	27	35	NA

Missing data

- When there are missing entries in our data, the problem gets a lot more complicated.
- What we observed is

$$X_{1,\mathsf{obs}}, \cdots, X_{n,\mathsf{obs}}$$

where the original random variable can be decomposed as $X_i = (X_{i,\mathsf{obs}}, X_{i,\mathsf{miss}})$ and $X_{i,\mathsf{miss}}$ is the unobserved part.

Missing data

- When there are missing entries in our data, the problem gets a lot more complicated.
- What we observed is

$$X_{1,\mathsf{obs}}, \cdots, X_{n,\mathsf{obs}}$$

  where the original random variable can be decomposed as $X_i = (X_{i,\mathsf{obs}}, X_{i,\mathsf{miss}})$ and $X_{i,\mathsf{miss}}$ is the unobserved part.
- In this case, we cannot construct the EDF.

Missing data

- When there are missing entries in our data, the problem gets a lot more complicated.
- What we observed is

$$X_{1,\mathsf{obs}}, \cdots, X_{n,\mathsf{obs}}$$

  where the original random variable can be decomposed as $X_i = (X_{i,\mathsf{obs}}, X_{i,\mathsf{miss}})$ and $X_{i,\mathsf{miss}}$ is the unobserved part.
- In this case, we cannot construct the EDF.
- Ignoring observations with missing entries (the complete-case analysis) is a bad idea because the missingness may be dependent with the study variable X.

Monotone missing data

- To simplify the problem, we assume that the missingness is monotone.
- This occurs in many medical research when participants dropout from the study.

Monotone missing data

- To simplify the problem, we assume that the missingness is monotone.
- This occurs in many medical research when participants dropout from the study.
- Let T_i denotes the last observed variable of the i-th individual. Then

$$X_{i,\mathsf{obs}} = X_{i,\leq T_i} = (X_{ij} : j \leq T_i).$$

Monotone missing data

- To simplify the problem, we assume that the missingness is monotone.
- This occurs in many medical research when participants dropout from the study.
- Let T_i denotes the last observed variable of the i-th individual. Then

$$X_{i,\mathsf{obs}} = X_{i,\leq T_i} = (X_{ij} : j \leq T_i).$$

- Thus, the *observed data* can be represented as

$$(X_{1,\leq T_1}, T_1), \cdots, (X_{n,\leq T_n}, T_n).$$

Monotone missing data

- To simplify the problem, we assume that the missingness is monotone.
- This occurs in many medical research when participants dropout from the study.
- Let T_i denotes the last observed variable of the i-th individual. Then

$$X_{i,\mathsf{obs}} = X_{i,\leq T_i} = (X_{ij} : j \leq T_i).$$

- Thus, the *observed data* can be represented as

$$(X_{1,\leq T_1}, T_1), \cdots, (X_{n,\leq T_n}, T_n).$$

- In contrast, we define the *full data*–the hypothetical dataset without missingness:

$$(X_1, T_1), \cdots, (X_n, T_n).$$

Population models

- The population CDF of the study variable $F(x)$ (also called the full-data distribution[1]) can be written as

$$F(x) = \sum_t F(x|T = t)P(T = t)$$

and its PDF can be written as

$$\begin{aligned} p(x) &= \sum_t p(x|T = t)P(T = t) \\ &= \sum_t {\color{red} p(x_{>t}|x_{\leq t}, T = t)} {\color{blue} p(x_{\leq t}|T = t)P(T = t)}. \end{aligned}$$

[1]Sometime the full-data distribution refers to $F(x, t) = F(x|t)P(T = t)$.

Population models

- The population CDF of the study variable $F(x)$ (also called the full-data distribution[1]) can be written as

$$F(x) = \sum_t F(x|T = t)P(T = t)$$

and its PDF can be written as

$$\begin{aligned} p(x) &= \sum_t p(x|T = t)P(T = t) \\ &= \sum_t p(x_{>t}|x_{\leq t}, T = t)p(x_{\leq t}|T = t)P(T = t). \end{aligned}$$

- Extrapolation density: $p(x_{>t}|x_{\leq t}, T = t)$

[1] Sometime the full-data distribution refers to $F(x, t) = F(x|t)P(T = t)$.

Population models

- The population CDF of the study variable $F(x)$ (also called the full-data distribution[1]) can be written as

$$F(x) = \sum_t F(x|T = t)P(T = t)$$

and its PDF can be written as

$$\begin{aligned} p(x) &= \sum_t p(x|T = t)P(T = t) \\ &= \sum_t {\color{red} p(x_{>t}|x_{\leq t}, T = t)} {\color{blue} p(x_{\leq t}|T = t)P(T = t)}. \end{aligned}$$

- Extrapolation density: $p(x_{>t}|x_{\leq t}, T = t)$
- Observed density: $p(x_{\leq t}|T = t)P(T = t)$

[1] Sometime the full-data distribution refers to $F(x, t) = F(x|t)P(T = t)$.

A toy example

Observed density generates what we observed. Extrapolation density describes the density of the unobserved cells.

ID	X_1	X_2	X_3	X_4
1	15	20	NA	NA
2	12	NA	NA	NA
3	17	43	35	42
4	11	25	NA	NA
5	16	37	32	51
6	15	23	32	NA
7	21	27	35	NA

Pattern mixture models

- The factorization:

$$
\begin{aligned}
p(x) &= \sum_t p(x|T=t)P(T=t) \\
&= \sum_t p(x_{>t}|x_{\le t}, T=t)p(x_{\le t}|T=t)P(T=t).
\end{aligned}
$$

is called the *pattern mixture models (PMM)* factorization (Little (1993)).

Pattern mixture models

- The factorization:

$$
\begin{aligned}
p(x) &= \sum_t p(x|T=t)P(T=t) \\
&= \sum_t p(x_{>t}|x_{\le t}, T=t)p(x_{\le t}|T=t)P(T=t).
\end{aligned}
$$

  is called the *pattern mixture models (PMM)* factorization (Little (1993)).
- Extrapolation density $p(x_{>t}|x_{\le t}, T=t)$: cannot be estimated using the observed data; it has to be identified by assumptions.

Pattern mixture models

- The factorization:

$$\begin{aligned} p(x) &= \sum_t p(x|T=t)P(T=t) \\ &= \sum_t p(x_{>t}|x_{\le t}, T=t)p(x_{\le t}|T=t)P(T=t). \end{aligned}$$

  is called the *pattern mixture models (PMM)* factorization (Little (1993)).
- Extrapolation density $p(x_{>t}|x_{\le t}, T=t)$: cannot be estimated using the observed data; it has to be identified by assumptions.
- Observed density $p(x_{\le t}|T=t)P(T=t)$: can be estimated using the observed data.

Pattern mixture models

- The factorization:

$$\begin{aligned} p(x) &= \sum_t p(x|T=t)P(T=t) \\ &= \sum_t p(x_{>t}|x_{\le t}, T=t)p(x_{\le t}|T=t)P(T=t). \end{aligned}$$

  is called the *pattern mixture models (PMM)* factorization (Little (1993).
- Extrapolation density $p(x_{>t}|x_{\le t}, T=t)$: cannot be estimated using the observed data; it has to be identified by assumptions.
- Observed density $p(x_{\le t}|T=t)P(T=t)$: can be estimated using the observed data.
- Key of the modeling strategy: try to identify the extrapolation density.

Selection models

- The pattern mixture model is a common approach to handling *missing not at random data*.
- Another common approach is the *selection models*, which uses the following factorization:

$$p(x, T = t) = P(T = t|x)p(x).$$

Selection models

- The pattern mixture model is a common approach to handling *missing not at random data*.
- Another common approach is the *selection models*, which uses the following factorization:

$$p(x, T = t) = P(T = t|x)p(x).$$

- The quantity $P(T = t|x)$ is called the selection probability or missing mechanism (Little and Robin 2002).

Selection models

- The pattern mixture model is a common approach to handling *missing not at random data*.
- Another common approach is the *selection models*, which uses the following factorization:

$$p(x, T = t) = P(T = t|x)p(x).$$

- The quantity $P(T = t|x)$ is called the selection probability or missing mechanism (Little and Robin 2002).
- *Missing completely at random (MCAR):* $P(T = t|x) = P(T = t)$.
- *Missing at random (MAR):* $P(T = t|x) = P(T = t|x_{\leq t})$.
- *Missing not at random (MNAR):* other cases.
- We focus on pattern mixture models in this talk.

Identifying the extrapolation density

- In PMM, we only need to identify the extrapolation density $p(x_{>t} | x_{\leq t}, T = t)$.
- A common strategy is to equate this density to something that is *identifiable/estimatible*.
- Note that we can factorize it as

$$p(x_{>t} | x_{\leq t}, T = t) = \prod_{s=t+1}^{d} p(x_s | x_{<s}, T = t)$$

so it suffices to identify each $p(x_s | x_{<s}, T = t)$ for $s > t$.

Common restrictions

- Here are some common assumptions/restrictions people made.
- Complete-case missing value (CCMV; Little 1993):

$$p(x_s|x_{<s}, T = t) = p(x_s|x_{<s}, T = d).$$

Common restrictions

- Here are some common assumptions/restrictions people made.
- Complete-case missing value (CCMV; Little 1993):

$$p(x_s | x_{<s}, T = t) = p(x_s | x_{<s}, T = d).$$

- Nearsest-case missing value (NCMV; Thijs et al. 2002):

$$p(x_s | x_{<s}, T = t) = p(x_s | x_{<s}, T = s).$$

Common restrictions

- Here are some common assumptions/restrictions people made.
- Complete-case missing value (CCMV; Little 1993):

$$p(x_s|x_{<s}, T = t) = p(x_s|x_{<s}, T = d).$$

- Nearsest-case missing value (NCMV; Thijs et al. 2002):

$$p(x_s|x_{<s}, T = t) = p(x_s|x_{<s}, T = s).$$

- Available-case missing value (ACMV; Molenberghs et al. 1998):

$$p(x_s|x_{<s}, T = t) = p(x_s|x_{<s}, T \geq s).$$

Common restrictions: a toy example

	X_1	X_2	X_3	X_4
T=1	Obs.	Missing	Missing	Missing
T=2	Obs.	Obs.	Missing	Missing
T=3	Obs.	Obs.	Obs.	Missing
T=4	Obs.	Obs.	Obs.	Obs.

Common restrictions: a toy example

	X_1	X_2	X_3	X_4
T=1	Obs.	Missing	Missing	Missing
T=2	Obs.	Obs.	Missing	Missing
T=3	Obs.	Obs.	Obs.	Missing
T=4	Obs.	Obs.	Obs.	Obs.

Common restrictions: a toy example

	X_1	X_2	X_3	X_4
T=1	Obs.	Missing	Missing	Missing
T=2	Obs.	Obs.	Missing	Missing
T=3	Obs.	Obs.	Obs.	Missing
T=4	Obs.	Obs.	Obs.	Obs.

CCMV

Common restrictions: a toy example

	X_1	X_2	X_3	X_4
T=1	Obs.	Missing	Missing	Missing
T=2	Obs.	Obs.	Missing	Missing
T=3	Obs.	Obs.	Obs.	Missing
T=4	Obs.	Obs.	Obs.	Obs.

NCMV

Common restrictions: a toy example

	X_1	X_2	X_3	X_4
T=1	Obs.	Missing	Missing	Missing
T=2	Obs.	Obs.	Missing	Missing
T=3	Obs.	Obs.	Obs.	Missing
T=4	Obs.	Obs.	Obs.	Obs.

ACMV

Donor-based restrictions

- We can generalize these restrictions to a more general 'donor' set by restricting to

$$p(x_s|x_{<s}, T = t) = p(x_s|x_{<s}, T \in \mathscr{A}_{ts}),$$

where $\mathscr{A}_{ts} \subset \{s, s+1, \cdots d\}$ is called the *donor set* of pattern t and variable s.

Donor-based restrictions

- We can generalize these restrictions to a more general 'donor' set by restricting to

$$p(x_s|x_{<s}, T = t) = p(x_s|x_{<s}, T \in \mathscr{A}_{ts}),$$

  where $\mathscr{A}_{ts} \subset \{s, s+1, \cdots d\}$ is called the *donor set* of pattern t and variable s.
- If the set $\{\mathscr{A}_{ts} : t = 1, \cdots, d-1; s = t+1, \cdots\}$ is given, then we can identify the extrapolation density.

Donor-based restrictions

- We can generalize these restrictions to a more general 'donor' set by restricting to

$$p(x_s|x_{<s}, T = t) = p(x_s|x_{<s}, T \in \mathscr{A}_{ts}),$$

where $\mathscr{A}_{ts} \subset \{s, s+1, \cdots d\}$ is called the *donor set* of pattern t and variable s.

- If the set $\{\mathscr{A}_{ts} : t = 1, \cdots, d-1; s = t+1, \cdots\}$ is given, then we can identify the extrapolation density.
- CCMV is the case $\mathscr{A}_{ts} = \{d\}$.
- NCMV is the case $\mathscr{A}_{ts} = \{s\}$.
- ACMV is the case $\mathscr{A}_{ts} = \{s, s+1, \cdots, d\}$.

Donor-based restrictions: a toy example

	X_1	X_2	X_3	X_4
T=1	Obs.	Missing	Missing	Missing
T=2	Obs.	Obs.	Missing	Missing
T=3	Obs.	Obs.	Obs.	Missing
T=4	Obs.	Obs.	Obs.	Obs.

Donor-based restrictions: a toy example

	X_1	X_2	X_3	X_4
T=1	Obs.	Missing	Missing	Missing
T=2	Obs.	Obs.	Missing	Missing
T=3	Obs.	Obs.	Obs.	Missing
T=4	Obs.	Obs.	Obs.	Obs.

Donor 1

Donor-based restrictions: a toy example

	X_1	X_2	X_3	X_4
T=1	Obs.	Missing	Missing	Missing
T=2	Obs.	Obs.	Missing	Missing
T=3	Obs.	Obs.	Obs.	Missing
T=4	Obs.	Obs.	Obs.	Obs.

Donor 2

Estimator under donor-based restrictions

- With a donor-based identifying restriction, we can easily estimate the extrapolation density.
- We can assume a parametric model or use a nonparametric estimator.

Estimator under donor-based restrictions

- With a donor-based identifying restriction, we can easily estimate the extrapolation density.
- We can assume a parametric model or use a nonparametric estimator.
- We propose to use the conditional kernel density estimator (CKDE), which can be expressed as

$$\widehat{p}_{A,h}(x_s|x_{<s}, T=t) = \frac{\frac{1}{h}\sum_{i=1}^{n} K\left(\frac{X_{i,s}-x_s}{h}\right) K\left(\frac{X_{i,<s}-x_{<s}}{h}\right) I(T_i \in \mathcal{A}_{ts})}{\sum_{j=1}^{n} K\left(\frac{X_{j,<s}-x_{<s}}{h}\right) I(T_j \in \mathcal{A}_{ts})}$$

$$= \frac{1}{h}\sum_{i=1}^{n} K\left(\frac{X_{i,s}-x_s}{h}\right) W_i(x_{<s}),$$

where

$$W_i(x_{<s}) = \frac{K\left(\frac{X_{i,<s}-x_{<s}}{h}\right) I(T_i \in \mathcal{A}_{ts})}{\sum_{j=1}^{n} K\left(\frac{X_{j,<s}-x_{<s}}{h}\right) I(T_j \in \mathcal{A}_{ts})}.$$

Estimator of the full-data distribution

- With an estimator $\widehat{p}_{A,h}(x_s | x_{<s}, T = t)$, we obtain an estimator of the extrapolation density

$$\widehat{p}_{A,h}(x_{>t} | x_{\leq t}, T = t) = \prod_{s=t+1}^{d} \widehat{p}_{A,h}(x_s | x_{<s}, T = t)$$

which defines a CDF estimator $\widehat{F}_{A,h}(x_{>t} | x_{\leq t}, T = t)$.

Estimator of the full-data distribution

- With an estimator $\widehat{p}_{A,h}(x_s|x_{<s}, T = t)$, we obtain an estimator of the extrapolation density

$$\widehat{p}_{A,h}(x_{>t}|x_{\leq t}, T = t) = \prod_{s=t+1}^{d} \widehat{p}_{A,h}(x_s|x_{<s}, T = t)$$

  which defines a CDF estimator $\widehat{F}_{A,h}(x_{>t}|x_{\leq t}, T = t)$.
- Note that the CDF of the observed density $p(x_{\leq t}|T = t)P(T = t)$ can be estimated by

$$\widehat{F}(x_{\leq t}|T = t)\widehat{P}(T = t) = \frac{1}{n}\sum_{i=1}^{n} I(X_{i,\leq t} \leq x_{\leq t}, T_i = t).$$

Estimator of the full-data distribution

- Putting it altogether, the estimate of $F(x)$ is

$$
\begin{aligned}
\widehat{F}_{A,h}(x) &= \sum_t \widehat{F}_{A,h}(x_{>t} x_{\le t} | T = t) \widehat{P}(T = t) \\
&= \sum_t \int_{-\infty}^{x_{\le t}} \widehat{F}_{A,h}(x_{>t} | x'_{\le t}, T = t) \widehat{F}(dx'_{\le t} | T = t) \widehat{P}(T = t) \\
&= \frac{1}{n} \sum_{i=1}^{n} \widehat{F}_{A,h}(x_{>T_i} | X_{i,\le T_i}, T = T_i) I(X_{i,\le T_i} \le x_{\le T_i}).
\end{aligned}
$$

Estimator of the full-data distribution

- Putting it altogether, the estimate of $F(x)$ is

$$
\begin{aligned}
\widehat{F}_{A,h}(x) &= \sum_t \widehat{F}_{A,h}(x_{>t} x_{\leq t} | T = t)\widehat{P}(T = t) \\
&= \sum_t \int_{-\infty}^{x_{\leq t}} \widehat{F}_{A,h}(x_{>t} | x'_{\leq t}, T = t)\widehat{F}(dx'_{\leq t} | T = t)\widehat{P}(T = t) \\
&= \frac{1}{n} \sum_{i=1}^{n} \widehat{F}_{A,h}(x_{>T_i} | X_{i,\leq T_i}, T = T_i) I(X_{i,\leq T_i} \leq x_{\leq T_i}).
\end{aligned}
$$

- It can be interpreted as a combination of:
    - unobserved variables: kernel CDF estimator.
    - observed variables: EDF.

Estimator of the full-data distribution

- Putting it altogether, the estimate of $F(x)$ is

$$\begin{aligned}
\widehat{F}_{A,h}(x) &= \sum_t \widehat{F}_{A,h}(x_{>t} x_{\leq t} | T = t)\widehat{P}(T = t) \\
&= \sum_t \int_{-\infty}^{x_{\leq t}} \widehat{F}_{A,h}(x_{>t} | x'_{\leq t}, T = t)\widehat{F}(dx'_{\leq t} | T = t)\widehat{P}(T = t) \\
&= \frac{1}{n}\sum_{i=1}^{n} \widehat{F}_{A,h}(x_{>T_i} | X_{i,\leq T_i}, T = T_i) I(X_{i,\leq T_i} \leq x_{\leq T_i}).
\end{aligned}$$

- It can be interpreted as a combination of:
  - unobserved variables: kernel CDF estimator.
  - observed variables: EDF.
- The parameter of interest can be estimated via $\widehat{\theta}_{A,h} = \theta(\widehat{F}_{A,h})$.

Estimator of the full-data distribution

- Although we have a good estimator, computing an estimate of the parameter of interest could be challenging.
- A major problem comes from the fact that the estimated distribution of the unobserved entries $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ does not have a simple form.

Estimator of the full-data distribution

- Although we have a good estimator, computing an estimate of the parameter of interest could be challenging.
- A major problem comes from the fact that the estimated distribution of the unobserved entries $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ does not have a simple form.
- Our solution: instead of analytically computing it, we use a Monte Carlo approximation.

Monte Carlo approximation

Here is a brief description of the Monte Carlo procedure.

- For each i, we generate $X^*_{i,>T_i}$ from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ to replace the missing entries. This is identical to the *imputation* procedure.

Monte Carlo approximation

Here is a brief description of the Monte Carlo procedure.

- For each i, we generate $X^*_{i,>T_i}$ from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ to replace the missing entries. This is identical to the *imputation* procedure.
- After imputing every missing entry, we construct a fully observed (imputed) dataset. Denote the data as

$$\mathcal{X}_n = \{(X^*_{i,>T_i}, X_{i,\leq T_i}) : i = 1, \cdots, n\}.$$

Monte Carlo approximation

Here is a brief description of the Monte Carlo procedure.

- For each i, we generate $X^*_{i,>T_i}$ from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ to replace the missing entries. This is identical to the *imputation* procedure.
- After imputing every missing entry, we construct a fully observed (imputed) dataset. Denote the data as

$$\mathcal{X}_n = \{(X^*_{i,>T_i}, X_{i,\leq T_i}) : i = 1, \cdots, n\}.$$

- To reduce the Monte Carlo errors, we repeat the above imputation procedure V times, leading to $\mathcal{X}_n^{(1)}, \cdots, \mathcal{X}_n^{(V)}$ imputed datasets.

Monte Carlo approximation

Here is a brief description of the Monte Carlo procedure.

- For each i, we generate $X^*_{i,>T_i}$ from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ to replace the missing entries. This is identical to the *imputation* procedure.
- After imputing every missing entry, we construct a fully observed (imputed) dataset. Denote the data as

$$\mathcal{X}_n = \{(X^*_{i,>T_i}, X_{i,\leq T_i}) : i = 1, \cdots, n\}.$$

- To reduce the Monte Carlo errors, we repeat the above imputation procedure V times, leading to $\mathcal{X}_n^{(1)}, \cdots, \mathcal{X}_n^{(V)}$ imputed datasets.
- Combine all datasets to form $\mathcal{X}_n^{[V]} = (\mathcal{X}_n^{(1)}, \cdots, \mathcal{X}_n^{(V)})$ and compute the estimator $\widehat{F}^{[V]}_{A,h}(x)$ using the EDF of $\mathcal{X}_n^{[V]}$.

Monte Carlo approximation

Here is a brief description of the Monte Carlo procedure.

- For each i, we generate $X^*_{i,>T_i}$ from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ to replace the missing entries. This is identical to the *imputation* procedure.
- After imputing every missing entry, we construct a fully observed (imputed) dataset. Denote the data as

$$\mathcal{X}_n = \{(X^*_{i,>T_i}, X_{i,\leq T_i}) : i = 1, \cdots, n\}.$$

- To reduce the Monte Carlo errors, we repeat the above imputation procedure V times, leading to $\mathcal{X}_n^{(1)}, \cdots, \mathcal{X}_n^{(V)}$ imputed datasets.
- Combine all datasets to form $\mathcal{X}_n^{[V]} = (\mathcal{X}_n^{(1)}, \cdots, \mathcal{X}_n^{(V)})$ and compute the estimator $\widehat{F}^{[V]}_{A,h}(x)$ using the EDF of $\mathcal{X}_n^{[V]}$.
- Compute the estimator of the parameter of interest $\widehat{\theta}^{[V]}_{A,h} = \theta(\widehat{F}^{[V]}_{A,h})$.

Monte Carlo approximation: a toy example - 1

ID	X_1	X_2	X_3	X_4
1	15	20	NA	NA
2	12	NA	NA	NA
3	17	43	35	42
4	11	25	NA	NA
5	16	37	32	51
6	15	23	32	NA
7	21	27	35	NA

Monte Carlo approximation: a toy example - 1

ID	X_1	X_2	X_3	X_4
1	15	20	30*	NA
2	12	NA	NA	NA
3	17	43	35	42
4	11	25	NA	NA
5	16	37	32	51
6	15	23	32	NA
7	21	27	35	NA

Monte Carlo approximation: a toy example - 1

ID	X_1	X_2	X_3	X_4
1	15	20	30*	43*
2	12	NA	NA	NA
3	17	43	35	42
4	11	25	NA	NA
5	16	37	32	51
6	15	23	32	NA
7	21	27	35	NA

Monte Carlo approximation: a toy example - 1

ID	X_1	X_2	X_3	X_4
1	15	20	30*	43*
2	12	31*	NA	NA
3	17	43	35	42
4	11	25	NA	NA
5	16	37	32	51
6	15	23	32	NA
7	21	27	35	NA

ID	X_1	X_2	X_3	X_4
1	15	20	30*	43*
2	12	31*	32*	NA
3	17	43	35	42
4	11	25	NA	NA
5	16	37	32	51
6	15	23	32	NA
7	21	27	35	NA

Monte Carlo approximation: a toy example - 1

ID	X_1	X_2	X_3	X_4
1	15	20	30*	43*
2	12	31*	32*	42*
3	17	43	35	42
4	11	25	NA	NA
5	16	37	32	51
6	15	23	32	NA
7	21	27	35	NA

Monte Carlo approximation: a toy example - 1

ID	X_1	X_2	X_3	X_4
1	15	20	30*	43*
2	12	31*	32*	42*
3	17	43	35	42
4	11	25	34*	41*
5	16	37	32	51
6	15	23	32	NA
7	21	27	35	NA

Monte Carlo approximation: a toy example - 1

ID	X_1	X_2	X_3	X_4
1	15	20	30*	43*
2	12	31*	32*	42*
3	17	43	35	42
4	11	25	34*	41*
5	16	37	32	51
6	15	23	32	49*
7	21	27	35	45*

ID	X_1	X_2	X_3	X_4
1	15	20	32*	41*
2	12	30*	29*	45*
3	17	43	35	42
4	11	25	34*	46*
5	16	37	32	51
6	15	23	32	42*
7	21	27	35	43*

Monte Carlo approximation: a toy example - 1

ID	X_1	X_2	X_3	X_4
1	15	20	33*	43*
2	12	25*	36*	42*
3	17	43	35	42
4	11	25	33*	41*
5	16	37	32	51
6	15	23	32	49*
7	21	27	35	52*

Monte Carlo approximation: a toy example - 2

ID	X_1	X_2	X_3	X_4
1	15	20	30*	43*
2	12	31*	32*	42*
3	17	43	35	42
4	11	25	34*	41*
5	16	37	32	51
6	15	23	32	49*
7	21	27	35	45*

ID	X_1	X_2	X_3	X_4
1	15	20	32*	41*
2	12	30*	29*	45*
3	17	43	35	42
4	11	25	34*	46*
5	16	37	32	51
6	15	23	32	42*
7	21	27	35	43*

...

ID	X_1	X_2	X_3	X_4
1	15	20	33*	43*
2	12	25*	36*	42*
3	17	43	35	42
4	11	25	33*	41*
5	16	37	32	51
6	15	23	32	49*
7	21	27	35	52*

We then combine these datasets to form a combine data and compute its EDF $\widehat{F}^{[V]}_{A,h}(x)$ and the corresponding estimator $\widehat{\theta}^{[V]}_{A,h} = \theta(\widehat{F}^{[V]}_{A,h})$.

Multiple imputation as Monte Carlo approximation

- This procedure is essentially a *multiple imputation* procedure (Rubin 1987).

Multiple imputation as Monte Carlo approximation

- This procedure is essentially a *multiple imputation* procedure (Rubin 1987).
- We can view our estimator as an estimator based on multiple imputation and the imputation distribution is based on the *estimated extrapolation density*.

Multiple imputation as Monte Carlo approximation

- This procedure is essentially a *multiple imputation* procedure (Rubin 1987).
- We can view our estimator as an estimator based on multiple imputation and the imputation distribution is based on the *estimated extrapolation density*.
- In fact, you can alway interpret the multiple imputation as a Monte Carlo approximation to the EDF formed by imposing an imputation distribution over the unobserved variables.

Multiple imputation as Monte Carlo approximation

- This procedure is essentially a *multiple imputation* procedure (Rubin 1987).
- We can view our estimator as an estimator based on multiple imputation and the imputation distribution is based on the *estimated extrapolation density*.
- In fact, you can alway interpret the multiple imputation as a Monte Carlo approximation to the EDF formed by imposing an imputation distribution over the unobserved variables.
- The imputation distribution is the extrapolation distribution in PMM.

Nonparametric Saturation

- In the missing data literature, an estimator of the full-data distribution $F(x, t)$ satisfies *nonparametric saturation* (NPS Robins, 1997)[2] if the implied observed data distribution agrees with the EDF of the observed data.

[2] Also known as nonparametric identification, just identification.

Nonparametric Saturation

- In the missing data literature, an estimator of the full-data distribution $F(x, t)$ satisfies *nonparametric saturation* (NPS Robins, 1997)[2] if the implied observed data distribution agrees with the EDF of the observed data.
- Namely, an estimator $\widehat{F}_0(x, t)$ has NPS if

$$\widehat{F}_0(x_{\leq t}, t) = \int \widehat{F}_0(x, t)\mu(dx_{>t}) = \widehat{F}(x_{\leq t}, t).$$

[2] Also known as nonparametric identification, just identification.

Nonparametric Saturation

- In the missing data literature, an estimator of the full-data distribution $F(x, t)$ satisfies *nonparametric saturation* (NPS Robins, 1997)[2] if the implied observed data distribution agrees with the EDF of the observed data.
- Namely, an estimator $\widehat{F}_0(x, t)$ has NPS if

$$\widehat{F}_0(x_{\le t}, t) = \int \widehat{F}_0(x, t)\mu(dx_{>t}) = \widehat{F}(x_{\le t}, t).$$

- The NPS can be viewed as a *self-consistent* property–the estimated full-data distribution agrees with the distribution of the observed data.

[2] Also known as nonparametric identification, just identification.

Nonparametric Saturation

- In the missing data literature, an estimator of the full-data distribution $F(x, t)$ satisfies *nonparametric saturation* (NPS Robins, 1997)[2] if the implied observed data distribution agrees with the EDF of the observed data.
- Namely, an estimator $\widehat{F}_0(x, t)$ has NPS if

$$\widehat{F}_0(x_{\leq t}, t) = \int \widehat{F}_0(x, t)\mu(dx_{>t}) = \widehat{F}(x_{\leq t}, t).$$

- The NPS can be viewed as a *self-consistent* property–the estimated full-data distribution agrees with the distribution of the observed data.

Theorem (Chen and Sadinle (2019))

The proposed estimator $\widehat{F}_{A,h}(x, t)$ satisfies the NPS.

[2] Also known as nonparametric identification, just identification.

Convergence rates

- Recall that $\theta = \theta(F)$ is the true parameter of interest and we use the estimator $\widehat{\theta}_{A,h} = \theta(\widehat{F}_{A,h})$.

- Recall that $\theta = \theta(F)$ is the true parameter of interest and we use the estimator $\widehat{\theta}_{A,h} = \theta(\widehat{F}_{A,h})$.
- Their difference can be decomposed into three components:

$$\widehat{\theta}_{A,h} - \theta = \widehat{\theta}_{A,h} - \bar{\theta}_{A,h} + \bar{\theta}_{A,h} - \theta_A + \theta_A - \theta$$

and under good conditions (including $\frac{\log n}{nh^d} \to 0$), we have the following results.

Convergence rates

- Recall that $\theta = \theta(F)$ is the true parameter of interest and we use the estimator $\widehat{\theta}_{A,h} = \theta(\widehat{F}_{A,h})$.
- Their difference can be decomposed into three components:

$$\widehat{\theta}_{A,h} - \theta = \widehat{\theta}_{A,h} - \bar{\theta}_{A,h} + \bar{\theta}_{A,h} - \theta_A + \theta_A - \theta$$

  and under good conditions (including $\frac{\log n}{nh^d} \to 0$), we have the following results.
- $\widehat{\theta}_{A,h} - \bar{\theta}_{A,h} = O_P\left(\sqrt{\frac{1}{n}}\right)$: the stochastic variation.
- $\bar{\theta}_{A,h} - \theta_A = O(h^2)$: the bias of the smoothing.
- $\theta_A - \theta$: *the bias of identifying restriction*. It will be 0 if our identifying restriction leads to the correct extrapolation density.

Asymptotic normality

Theorem (Chen and Sadinle (2019))

Under regularity conditions, when $\frac{\log n}{nh^d} \to 0$ *and* $h \to 0$,

$$\sqrt{n}(\widehat{F}_{A,h}(x) - \bar{F}_{A,h}(x))$$

converges to a Gaussian process where

$$\bar{F}_{A,h}(x) = \sum_t \int_{x'_{\leq t}=-\infty}^{x'_{\leq t}=x_{\leq t}} \bar{F}_{A,h}(x_{>t}|x'_{\leq t}, T = t)F(dx'_{\leq t}|T = t)P(T = t),$$

$$\bar{F}_{A,h}(x_{>t}|x_{\leq t}, T = t) \approx \mathbb{E}(\widehat{F}_{A,h}(x_{>t}|x_{\leq t}, T = t)).$$

- $\bar{F}_{A,h}(x)$ behaves like the expected quantity of the estimator $\widehat{F}_{A,h}(x)$.
- $\bar{\theta}_{A,h} = \theta(\bar{F}_{A,h})$.

Bootstrap method

- Sampling with replacement from the original data (including missing entries) to obtain a bootstrap sample.
- Use the bootstrap sample to estimate the conditional density.

Bootstrap method

- Sampling with replacement from the original data (including missing entries) to obtain a bootstrap sample.
- Use the bootstrap sample to estimate the conditional density.
- Perform the Monte Carlo procedure (multiple imputation) for V times. Compute the estimator $\widehat{\theta}_{A,h}^{[V]*}$.

Bootstrap method

- Sampling with replacement from the original data (including missing entries) to obtain a bootstrap sample.
- Use the bootstrap sample to estimate the conditional density.
- Perform the Monte Carlo procedure (multiple imputation) for V times. Compute the estimator $\widehat{\theta}_{A,h}^{[V]*}$.
- Repeat the above procedure B times, leading to B bootstrap estimates

$$\widehat{\theta}_{A,h}^{[V]*(1)}, \cdots, \widehat{\theta}_{A,h}^{[V]*(B)}.$$

Bootstrap method

- Sampling with replacement from the original data (including missing entries) to obtain a bootstrap sample.
- Use the bootstrap sample to estimate the conditional density.
- Perform the Monte Carlo procedure (multiple imputation) for V times. Compute the estimator $\widehat{\theta}_{A,h}^{[V]*}$.
- Repeat the above procedure B times, leading to B bootstrap estimates
$$\widehat{\theta}_{A,h}^{[V]*(1)}, \cdots, \widehat{\theta}_{A,h}^{[V]*(B)}.$$
- Compute the upper and the lower limits $(\ell_{1-\alpha}, u_{1-\alpha})$ of the confidence interval using the quantiles. Namely, $\ell_{B,1-\alpha} = \widehat{G}^{-1}(\alpha/2)$ and $u_{B,1-\alpha} = \widehat{G}^{-1}(1-\alpha/2)$ where
$$\widehat{G}(s) = \frac{1}{B}\sum_{b=1}^{B} I(\widehat{\theta}_{A,h}^{[V]*(b)}).$$

Confidence interval

Let $u_{1-\alpha}$ and $\ell_{1-\alpha}$ be the upper and lower bound from the bootstrap approach when the number of bootstrap replicates $B \to \infty$ and $V \to \infty$.

Theorem (Chen and Sadinle (2019))

Under regularity conditions, when $\frac{\log n}{nh^d} \to 0$ *and* $h \to 0$*,*

$$P(\ell_{1-\alpha} \leq \bar{\theta}_{A,h} \leq u_{1-\alpha}) \to 1 - \alpha.$$

Confidence interval

Let $u_{1-\alpha}$ and $\ell_{1-\alpha}$ be the upper and lower bound from the bootstrap approach when the number of bootstrap replicates $B \to \infty$ and $V \to \infty$.

Theorem (Chen and Sadinle (2019))

Under regularity conditions, when $\frac{\log n}{nh^d} \to 0$ *and* $h \to 0$,

$$P(\ell_{1-\alpha} \leq \bar{\theta}_{A,h} \leq u_{1-\alpha}) \to 1 - \alpha.$$

- Namely, the bootstrap confidence interval is valid for $\bar{\theta}_{A,h} = \theta(F_{A,h})$.

Confidence interval

Let $u_{1-\alpha}$ and $\ell_{1-\alpha}$ be the upper and lower bound from the bootstrap approach when the number of bootstrap replicates $B \to \infty$ and $V \to \infty$.

Theorem (Chen and Sadinle (2019))

Under regularity conditions, when $\frac{\log n}{nh^d} \to 0$ *and* $h \to 0$*,*

$$P(\ell_{1-\alpha} \leq \bar{\theta}_{A,h} \leq u_{1-\alpha}) \to 1 - \alpha.$$

- Namely, the bootstrap confidence interval is valid for $\bar{\theta}_{A,h} = \theta(F_{A,h})$.
- Note that

$$\bar{\theta}_{A,h} - \theta = \bar{\theta}_{A,h} - \theta_A + \theta_A - \theta$$

consists of the bias from smoothing and the bias from identifying restriction.

$$
\begin{tikzcd}
& \text{Actual} & & \text{Bootstrap} & \\
F_{\mathsf{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\mathsf{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^{*} \arrow[r, Rightarrow] & \widehat{F}^{*}_{\mathsf{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\mathsf{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\mathsf{obs,h}} \arrow[d, dotted] & & \widehat{F}^{*}_{\mathsf{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^{*}_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^{*}_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^{*}_{A,h}
\end{tikzcd}
$$

Actual — Bootstrap

$$
\begin{tikzcd}
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & \boxed{S} \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^*_{A,h}
\end{tikzcd}
$$

Original data

Actual Bootstrap

$$
\begin{tikzcd}
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \boxed{\widehat{F}_{\text{obs}}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^*_{A,h}
\end{tikzcd}
$$

EDF on the observed variables

Bootstrap Diagram (Efron 1994)

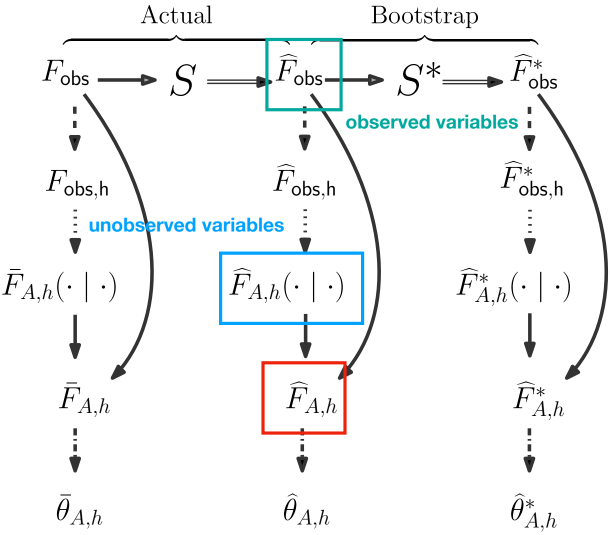
$$
\begin{tikzcd}
 & \text{Actual} & & \text{Bootstrap} & \\
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^{*} \arrow[r, Rightarrow] & \widehat{F}^{*}_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \boxed{\widehat{F}_{\text{obs,h}}} \arrow[d, dotted] & & \widehat{F}^{*}_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^{*}_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^{*}_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^{*}_{A,h}
\end{tikzcd}
$$

Kernel smoothing

$$
\begin{tikzcd}
& \text{Actual} & & \text{Bootstrap} & \\
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \boxed{\widehat{F}_{A,h}(\cdot \mid \cdot)} \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^*_{A,h}
\end{tikzcd}
$$

The estimated extrapolation distribution via smoothing & the identifying restriction

Bootstrap Diagram (Efron 1994)



Estimator of the full-data distribution

Bootstrap Diagram (Efron 1994)

$$
\begin{tikzcd}
 & \text{Actual} & & \text{Bootstrap} & \\
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \boxed{\hat{\theta}_{A,h}} & & \hat{\theta}^*_{A,h}
\end{tikzcd}
$$

Estimator of the parameter of interest

Actual Bootstrap

$$
\begin{tikzcd}
F_{\text{obs}} \arrow[r] \arrow[ddd, bend left] \arrow[d, dashed] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[ddd, bend left] \arrow[d, dashed] & \boxed{S^*} \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[ddd, bend left] \arrow[d, dashed] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^*_{A,h}
\end{tikzcd}
$$

Bootstrap sample

$$
\overbrace{\hphantom{F_{\text{obs}} \longrightarrow S \Longrightarrow}}^{\text{Actual}} \quad \overbrace{\hphantom{\widehat{F}_{\text{obs}} \longrightarrow S^* \Longrightarrow}}^{\text{Bootstrap}}
$$

$$
\begin{tikzcd}
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^* \arrow[r, Rightarrow] & \boxed{\widehat{F}^*_{\text{obs}}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^*_{A,h}
\end{tikzcd}
$$

Bootstrap EDF

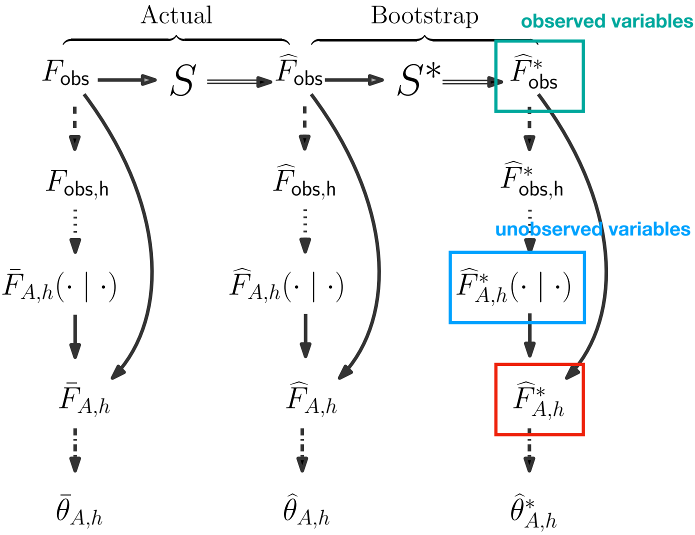
$$
\begin{tikzcd}
 & \text{Actual} & & \text{Bootstrap} & \\
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[dddd, bend left, end anchor=north east] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[dddd, bend left, end anchor=north east] & S^{*} \arrow[r, Rightarrow] & \widehat{F}^{*}_{\text{obs}} \arrow[d, dashed] \arrow[dddd, bend left, end anchor=north east] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \boxed{\widehat{F}^{*}_{\text{obs,h}}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^{*}_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^{*}_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^{*}_{A,h}
\end{tikzcd}
$$

Kernel smoothing on bootstrap sample

$$
\begin{tikzcd}
 & \text{Actual} & & \text{Bootstrap} & \\
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \boxed{\widehat{F}^*_{A,h}(\cdot \mid \cdot)} \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^*_{A,h}
\end{tikzcd}
$$

Bootstrap extrapolation distribution via smoothing & the identifying restriction

Bootstrap Diagram (Efron 1994)



Bootstrap estimator of the full data distribution

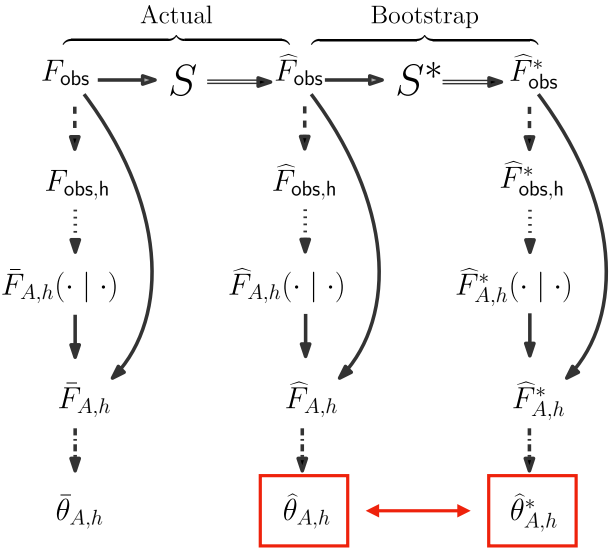
Bootstrap Diagram (Efron 1994)

Actual Bootstrap

$$
\begin{tikzcd}
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \boxed{\hat{\theta}^*_{A,h}}
\end{tikzcd}
$$

Bootstrap estimate of the parameter of interest

Bootstrap Diagram (Efron 1994)



This difference is how we do resampling inference

Bootstrap Diagram (Efron 1994)

$$
\begin{tikzcd}
 & \text{Actual} & & \text{Bootstrap} & \\
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\boxed{\bar{\theta}_{A,h}} \arrow[rr, leftrightarrow] & & \boxed{\hat{\theta}_{A,h}} & & \hat{\theta}^*_{A,h}
\end{tikzcd}
$$

It can be viewed as a plug-in estimate of this difference

$$\begin{tikzcd}
 & \text{Actual} & & \text{Bootstrap} & \\
\boxed{F_{\mathsf{obs}}} \arrow[r] \arrow[d, dashed] \arrow[dddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\mathsf{obs}} \arrow[r] \arrow[d, dashed] \arrow[dddd, bend left] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\mathsf{obs}} \arrow[d, dashed] \arrow[dddd, bend left] \\
F_{\mathsf{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\mathsf{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\mathsf{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^*_{A,h}
\end{tikzcd}$$

The CDF of the observed variables

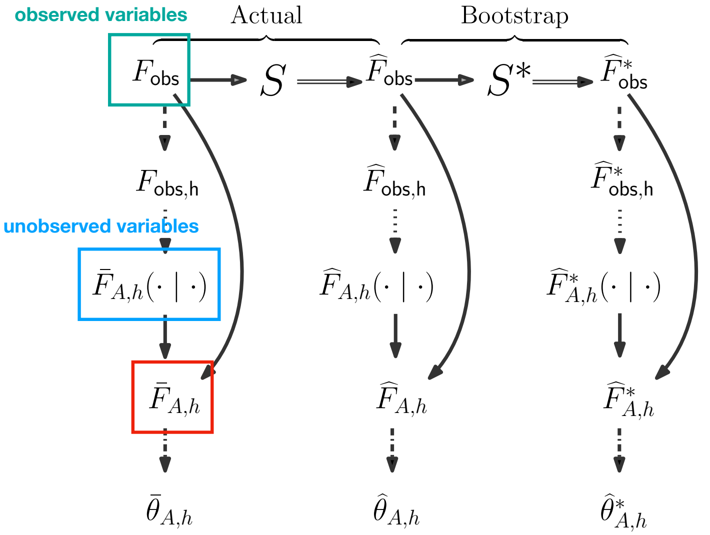
$$
\begin{tikzcd}
 & \overbrace{\hspace{6em}}^{\text{Actual}} & & \overbrace{\hspace{6em}}^{\text{Bootstrap}} & \\
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^{*} \arrow[r, Rightarrow] & \widehat{F}^{*}_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
\boxed{F_{\text{obs,h}}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^{*}_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^{*}_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^{*}_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^{*}_{A,h}
\end{tikzcd}
$$

The kernel-smoothed version of the CDF

$$\begin{tikzcd}
 & \text{Actual} & & \text{Bootstrap} & \\
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[ddd, bend left] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[d, dashed] \arrow[ddd, bend left] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\boxed{\bar{F}_{A,h}(\cdot \mid \cdot)} \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\bar{\theta}_{A,h} & & \hat{\theta}_{A,h} & & \hat{\theta}^*_{A,h}
\end{tikzcd}$$

The extrapolation distribution from smoothed CDF
& the identifying restriction

Bootstrap Diagram (Efron 1994)



The full-data distribution

Bootstrap Diagram (Efron 1994)

$$
\begin{tikzcd}
\overbrace{\hspace{3cm}}^{\text{Actual}} & & \overbrace{\hspace{3cm}}^{\text{Bootstrap}} \\[-2em]
F_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[dddd, bend left=40] & S \arrow[r, Rightarrow] & \widehat{F}_{\text{obs}} \arrow[r] \arrow[d, dashed] \arrow[dddd, bend left=40] & S^* \arrow[r, Rightarrow] & \widehat{F}^*_{\text{obs}} \arrow[d, dashed] \arrow[dddd, bend left=40] \\
F_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}_{\text{obs,h}} \arrow[d, dotted] & & \widehat{F}^*_{\text{obs,h}} \arrow[d, dotted] \\
\bar{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}_{A,h}(\cdot \mid \cdot) \arrow[d] & & \widehat{F}^*_{A,h}(\cdot \mid \cdot) \arrow[d] \\
\bar{F}_{A,h} \arrow[d, dashed] & & \widehat{F}_{A,h} \arrow[d, dashed] & & \widehat{F}^*_{A,h} \arrow[d, dashed] \\
\boxed{\bar{\theta}_{A,h}} & & \hat{\theta}_{A,h} & & \hat{\theta}^*_{A,h}
\end{tikzcd}
$$

The mapped parameter of interest

- Since the bias $\theta_A - \theta$ is hard to know in practice, the sensitivity analysis is a common procedure to evaluate the stability of an estimator.

A structural sensitivity analysis

- Since the bias $\theta_A - \theta$ is hard to know in practice, the sensitivity analysis is a common procedure to evaluate the stability of an estimator.
- In the class of donor-based identifying restrictions, we may perform the sensitivity analysis by perturbing a given restriction within the class.

A structural sensitivity analysis

- Since the bias $\theta_A - \theta$ is hard to know in practice, the sensitivity analysis is a common procedure to evaluate the stability of an estimator.
- In the class of donor-based identifying restrictions, we may perform the sensitivity analysis by perturbing a given restriction within the class.
- For instance, the NCMV requires $\mathscr{A}_{ts} = \{s\}$. We may consider perturbing it via considering the 'k-NCMV' restrictions

$$\mathscr{A}_{ts}^{\mathsf{k-NC}} = \{\tau : \tau \geq s, |\tau - s| \leq k - 1\} = \{s, s+1, \cdots, s+k-1\}.$$

- Since the bias $\theta_A - \theta$ is hard to know in practice, the sensitivity analysis is a common procedure to evaluate the stability of an estimator.
- In the class of donor-based identifying restrictions, we may perform the sensitivity analysis by perturbing a given restriction within the class.
- For instance, the NCMV requires $\mathscr{A}_{ts} = \{s\}$. We may consider perturbing it via considering the 'k-NCMV' restrictions

$$\mathscr{A}_{ts}^{\mathsf{k-NC}} = \{\tau : \tau \geq s, |\tau - s| \leq k - 1\} = \{s, s+1, \cdots, s+k-1\}.$$

- When $k = 1$ this reduces to NCMV and when $k = d$, this becomes ACMV.

- Our method is not limited to a nonparametric estimator; one can use a parametric density estimator as well.
- All we need is an estimator of the conditional density, which can be done parametrically or nonparametrically.

[3]When using a parametric model, the sequential imputation reduces to the parametric sequential imputation described in p.60 of Liu (2008).

- Our method is not limited to a nonparametric estimator; one can use a parametric density estimator as well.
- All we need is an estimator of the conditional density, which can be done parametrically or nonparametrically.
- In our framework, the modeling strategy on the distribution and the identifying restrictions are *decoupled*–one can choose any distribution estimator and any donor-based identifying restriction.

[3] When using a parametric model, the sequential imputation reduces to the parametric sequential imputation described in p.60 of Liu (2008).

Decoupling modeling procedure and identifying restriction

- Our method is not limited to a nonparametric estimator; one can use a parametric density estimator as well.
- All we need is an estimator of the conditional density, which can be done parametrically or nonparametrically.
- In our framework, the modeling strategy on the distribution and the identifying restrictions are *decoupled*–one can choose any distribution estimator and any donor-based identifying restriction.
- The Monte Carlo approximation (multiple imputation) and the bootstrap can be done in a similar manner[3].

[3] When using a parametric model, the sequential imputation reduces to the parametric sequential imputation described in p.60 of Liu (2008).

The flexibility and transparency of modeling

- When handling missing data, there are three modeling components:
  - Assumptions on missingness.
  - Models on distributions.
  - Formulation of the parameter of interest.
- Many classical methods would require all three components to be dependent.
- Our methods allow them to all be independent.
- Also, our method leads to the *model congenial property* (Meng 1994)[4] as long as we are using a nonparametric estimator on the distribution.

[4]In short, this means the model on missing data and the model used for formulating parameter of interest are consistent.

Conclusion

- We introduce a class called the donor-based identifying restrictions for handling missing data.

Conclusion

- We introduce a class called the donor-based identifying restrictions for handling missing data.
- We proposed a nonparametric estimator of the full-data distribution but a similar idea can be applied to a parametric model. This estimator is nonparamteric saturated and model congenial.

Conclusion

- We introduce a class called the donor-based identifying restrictions for handling missing data.
- We proposed a nonparametric estimator of the full-data distribution but a similar idea can be applied to a parametric model. This estimator is nonparamteric saturated and model congenial.
- Even if we cannot directly compute the estimator, we may use a Monte Carlo approximation in the form of multiple imputation to approximate it.

Conclusion

- We introduce a class called the donor-based identifying restrictions for handling missing data.
- We proposed a nonparametric estimator of the full-data distribution but a similar idea can be applied to a parametric model. This estimator is nonparamteric saturated and model congenial.
- Even if we cannot directly compute the estimator, we may use a Monte Carlo approximation in the form of multiple imputation to approximate it.
- In a sense, our work provides an alternative view of multiple imputation–it can be viewed as a Monte Carlo approximation to a PMM estimator.

Conclusion

- We introduce a class called the donor-based identifying restrictions for handling missing data.
- We proposed a nonparametric estimator of the full-data distribution but a similar idea can be applied to a parametric model. This estimator is nonparamteric saturated and model congenial.
- Even if we cannot directly compute the estimator, we may use a Monte Carlo approximation in the form of multiple imputation to approximate it.
- In a sense, our work provides an alternative view of multiple imputation–it can be viewed as a Monte Carlo approximation to a PMM estimator.
- Our estimator has nice asymptotic property but there is an identifying restriction bias we have to be cautious.

Future work

- Generalization to nonmonotone case (work in progress with Mauricio).
- How to interpret the donor-based identifying restriction?
- How to do data analysis with multiple identifying restrictions?
- Missing covariates in regression/causal inference problem.
- Will the bootstrap always include the imputation uncertainty?
- Equivalent selection models and semi-parametric inference.

Thank You!

More details can be found in `https://arxiv.org/abs/1904.11085`.

References

1. Chen, Y. C., & Sadinle, M. (2019). Nonparametric Pattern-Mixture Models for Inference with Missing Data. arXiv preprint arXiv:1904.11085.
2. Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. J. Am. Statist. Assoc., 88(421), 125-134.
3. Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. Hoboken, New Jersey: Wiley, 2nd ed.
4. Little, R. (1995). Modeling the drop-out mechanism in longitudinal studies. Journal of the American Statistical Association, 90(1), 1.
5. Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., & Curran, D. (2002). Strategies to fit pattern-mixture models. Biostatistics, 3(2), 245-265.
6. Molenberghs, G., Michiels, B., Kenward, M. G., & Diggle, P. J. (1998). Monotone missing data and pattern-mixture models. Statistica Neerlandica, 52(2), 153-161.
7. Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys (Vol. 81). John Wiley & Sons.
8. Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. Statistics in Medicine, 16(1), 21-37.
9. Efron, B. (1994). Missing data, imputation, and the bootstrap. Journal of the American Statistical Association, 89(426), 463-475.
10. Liu, J. S. (2008). Monte Carlo strategies in scientific computing. Springer Science & Business Media.
11. Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical Science, 538-558.

The sequential imputation

- Generating $X^*_{i,>T_i}$ from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ can be done via a sequential sampling from the conditional KDE.

The sequential imputation

- Generating $X^*_{i,>T_i}$ from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ can be done via a sequential sampling from the conditional KDE.
- Note that sampling from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ is the same as sampling from its PDF

$$\widehat{p}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i) = \prod_{s=T_i+1}^{d} \widehat{p}_{A,h}(x_s|x_{<s}, T = T_i).$$

The sequential imputation

- Generating $X^*_{i,>T_i}$ from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ can be done via a sequential sampling from the conditional KDE.
- Note that sampling from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ is the same as sampling from its PDF

$$\widehat{p}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i) = \prod_{s=T_i+1}^{d} \widehat{p}_{A,h}(x_s|x_{<s}, T = T_i).$$

- We can sample X_{T_i+1} and then sample X_{T_i+2} conditioned on the previously sampled X_{T_1+1}.

The sequential imputation

- Generating $X^*_{i,>T_i}$ from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ can be done via a sequential sampling from the conditional KDE.
- Note that sampling from $\widehat{F}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i)$ is the same as sampling from its PDF

$$\widehat{p}_{A,h}(x_{>T_i}|X_{i,\leq T_i}, T = T_i) = \prod_{s=T_i+1}^{d} \widehat{p}_{A,h}(x_s|x_{<s}, T = T_i).$$

- We can sample X_{T_i+1} and then sample X_{T_i+2} conditioned on the previously sampled X_{T_1+1}.
- Because

$$\widehat{p}_{A,h}(x_s|x_{<s}, T = T_i) = \frac{1}{h}\sum_{j=1}^{n} K\left(\frac{X_{j,s} - x_s}{h}\right) W_j(x_{<s}),$$

sampling from can be done from a weighted smoothed bootstrap procedure.

Richness of donor-based identifications

One may be wondering how large the donor-based identification class. The following theorem shows that this class contains many, many distinct elements.

Theorem (Chen and Sadinle (2019+); in progress)

Suppose that there are d variables that are subject to monotone missingness. Then there are

$$L_d = \prod_{t=0}^{d-1} (2^{d-t} - 1)$$

numbers of distinct donor-based identifying restrictions.

Richness of donor-based identifications

One may be wondering how large the donor-based identification class. The following theorem shows that this class contains many, many distinct elements.

Theorem (Chen and Sadinle (2019+); in progress)

Suppose that there are d variables that are subject to monotone missingness. Then there are

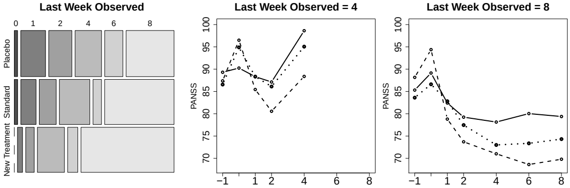
$$L_d = \prod_{t=0}^{d-1} (2^{d-t} - 1)$$

numbers of distinct donor-based identifying restrictions.

- Here are some numbers of L_d:

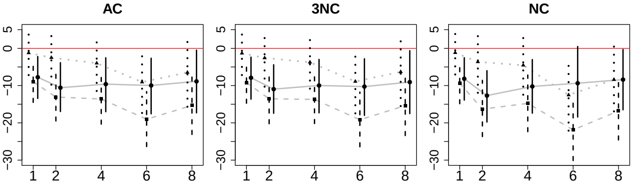
  $L_1 = 1, L_2 = 3, L_3 = 21, L_4 = 315, L_5 = 9765, L_6 = 615195, L_7 > 7 \times 10^7.$

PANSS Datasets - 1



- The purpose of the trial was to evaluate the effectiveness of four different doses of a new treatment (N) compared with placebo (P) and with a standard of care (S) in patients with chronic schizophrenia.
- The Positive and Negative Syndrome Scale for Schizophrenia (PANSS) score X_t was measured on patients one week before, the day of, and on weeks $t = 1, 2, 4, 6,$ and 8 after randomization.
- We are interested in estimating average treatment effects (ATEs) over time $\mu_t^{G_1} - \mu_t^{G_2} = \mathbb{E}(X_t|G_1) - \mathbb{E}(X_t|G_2)$, where

PANSS Datasets - 2



- Dashed lines: $\mu_t^N - \mu_t^P$; dotted lines: $\mu_t^S - \mu_t^P$; and solid lines: $\mu_t^N - \mu_t^S$.
- We use Gaussian kernels in conditional KDE with Silverman's rule (Silverman 1986) for the bandwidth.
- We consider the AC, 3NC and NC identifying restrictions.
- 95% Confidence intervals are constructed using the bootstrap.

Assumptions

(A1) The true full-data distribution function $F(x, t)$ has a density function $f_0(x, t)$ satisfying

1. $\inf_{x \in \mathcal{X}} f_0(x, t) > 0$ for each $t = 1, \cdots, d$.
2. $f_0(x, t) \in \mathbf{UBC}_2$ for each $t = 1, \cdots, d$.

(A2) The statistical functional θ is Hadamard differentiable.

(K1) $K(z)$ has at least second-order bounded derivative and

$$\int z^2 K(z)\mu(dz) < \infty, \qquad \int K^2(z)\mu(dz) < \infty.$$

(K2) Let $\mathcal{K} = \left\{z \mapsto K\left(\frac{z-w}{h}\right) : w \in \mathbb{R}, \bar{h} > h > 0\right\}$, for some fixed constant $\bar{h}$. We assume that $\mathcal{K}$ is a VC-type class. Namely, there exists constants A, v and a constant envelope b_0 such that

$$\sup_Q N(\mathcal{K}, \mathcal{L}^2(Q), b_0\epsilon) \leq \left(\frac{A}{\epsilon}\right)^v,$$

where $N(T, d_T, \epsilon)$ is the ϵ-covering number for a semi-metric set T with metric d_T, and $\mathcal{L}^2(Q)$ is the L_2 norm with respect to the probability measure Q.