Emputation: Identification-Guided Neural Imputation Framework

Yikun Zhang[1,2]

Joint work with *Yanjiao Yang*[3], *Xinwei Shen*[3], and *Yen-Chi Chen*[3]

[1]*Department of Statistics, University of Chicago*
[2]*NSF-Simons AI Institute for the Sky (SkAI Institute)*
[3]*Department of Statistics, University of Washington*

CFE-CMStatistics 2026
December 12, 2026

SkAI
Funded by the U.S. National Science Foundation and the Simons Foundation

THE UNIVERSITY OF CHICAGO

Motivating Example: Alzheimer's Disease

National Alzheimer's Coordinating Center (NACC):

- 4,906 participants with brain autopsy data and a clinical visit in 2 yrs before death.

	Age	BMI	SBP	GDS	R_i
$i = 1$	82	24.1	131	2	1111
$i = 2$	79	?	142	?	1010
$i = 3$	88	27.3	118	?	1110
$i = 4$	75	?	?	4	1001
$i = 5$	91	22.8	126	1	1111
$i = 6$	84	26.5	?	6	1101

BMI: body mass index; SBP: systolic blood pressure; GDS: geriatric depression scale.

Note: these values are illustrative.

- BMI, SBP, and GDS: $\approx$ 40–50% missing.
- Only **1,820** participants (37%) are complete cases.

	Complete ($N = 1{,}820$)	Incomplete ($N = 3{,}086$)
Mean CDR global score	0.97 (*healthier*)	2.23
High AD neuropathology	28.4%	49.5%
APOE (genotype) $\varepsilon 4$ carrier	37.0%	45.4%

Issue: Complete-case analysis excludes 63% of participants and favors healthier individuals, which can lead to *biased* and *inefficient* estimation.

What Makes a Good Imputation?

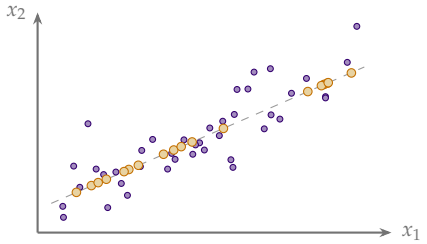
Imputation: train a model $\overset{\text{fill in}}{\Longrightarrow}$ missing values $\Longrightarrow$ complete-case analysis.

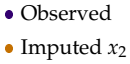
- Study variables $X = (X_1, ..., X_d) \in \mathbb{R}^d$; response pattern $R = (R_1, ..., R_d) \in \{0, 1\}^d$.
- Under pattern $R = r$, observed entries are $X_R \equiv X_r = (X_j : r_j = 1)$.
- Actual observed data: $\{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n$.

What Makes a Good Imputation?

Imputation: train a model $\overset{\text{fill in}}{\Longrightarrow}$ missing values $\Longrightarrow$ complete-case analysis.

- Study variables $X = (X_1, ..., X_d) \in \mathbb{R}^d$; response pattern $R = (R_1, ..., R_d) \in \{0, 1\}^d$.
- Under pattern $R = r$, observed entries are $X_R \equiv X_r = (X_j : r_j = 1)$.
- Actual observed data: $\{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n$.





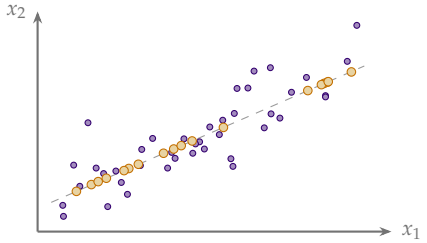
Conditional mean imputation

$\widehat{\mathbb{E}}(X_{\bar{r}} \mid X_r, R = r)$: collapses the variability.

What Makes a Good Imputation?

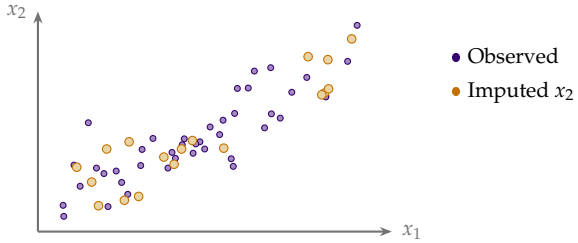
Imputation: train a model $\overset{\text{fill in}}{\Longrightarrow}$ missing values $\Longrightarrow$ complete-case analysis.

- Study variables $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$; response pattern $R = (R_1, \ldots, R_d) \in \{0,1\}^d$.
- Under pattern $R = r$, observed entries are $X_R \equiv X_r = (X_j : r_j = 1)$.
- Actual observed data: $\{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n$.



Conditional mean imputation

$\widehat{\mathbb{E}}(X_{\bar{r}} \mid X_r, R = r)$: collapses the variability.



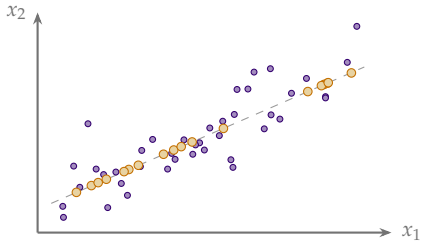
Distributional imputation $\widehat{p}(x_{\bar{r}} \mid x_r, R = r)$:

preserves spread and dependence.

What Makes a Good Imputation?

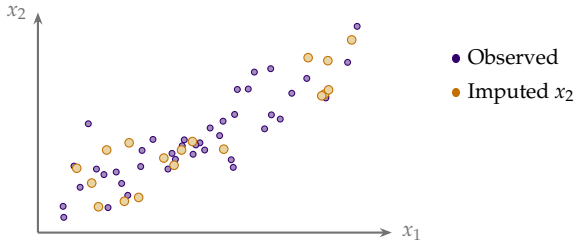
Imputation: train a model $\overset{\text{fill in}}{\Longrightarrow}$ missing values $\Longrightarrow$ complete-case analysis.

- Study variables $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$; response pattern $R = (R_1, \ldots, R_d) \in \{0,1\}^d$.
- Under pattern $R = r$, observed entries are $X_R \equiv X_r = (X_j : r_j = 1)$.
- Actual observed data: $\{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n$.



Conditional mean imputation

$\widehat{\mathbb{E}}(X_{\bar{r}} \mid X_r, R = r)$: collapses the variability.



Distributional imputation $\widehat{p}(x_{\bar{r}} \mid x_r, R = r)$: preserves spread and dependence.

Downstream analysis, especially uncertainty quantification, requires the *whole* conditional distribution (Van Buuren, 2018; Näf et al., 2024).

Imputation Target

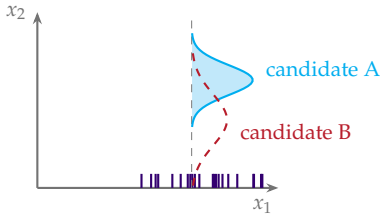
- $\bar{r} = 1_d - r \implies \begin{cases} X_r = (X_j : r_j = 1) & \text{observed variables,} \\ X_{\bar{r}} = (X_j : r_j = 0) & \text{missing variables.} \end{cases}$

Imputation Target

- $\bar{r} = 1_d - r \implies \begin{cases} X_r = (X_j : r_j = 1) & \text{observed variables,} \\ X_{\bar{r}} = (X_j : r_j = 0) & \text{missing variables.} \end{cases}$

Pattern-mixture factorization (Little, 1993):

$$\underbrace{p(x, r)}_{\substack{\text{Joint distribution of } (X,R) \\ \text{identifiable}}} = \underbrace{p(x_r, R = r)}_{\substack{\text{Observed data distribution} \\ \text{identifiable}}} \times \underbrace{p(x_{\bar{r}} \mid x_r, R = r)}_{\substack{\text{Extrapolation distribution} \\ \text{not identifiable}}} .$$



pattern $R = (1, 0)$: same observed x_1, different extrapolation densities

Imputation Target

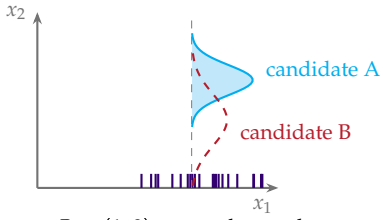
- $\bar{r} = 1_d - r \implies \begin{cases} X_r = (X_j : r_j = 1) & \text{observed variables,} \\ X_{\bar{r}} = (X_j : r_j = 0) & \text{missing variables.} \end{cases}$

Pattern-mixture factorization (Little, 1993):

$$\underbrace{p(x, r)}_{\text{Joint distribution of } (X,R)} = \underbrace{p(x_r, R = r)}_{\substack{\text{Observed data distribution} \\ \text{identifiable}}} \times \underbrace{p(x_{\bar{r}} \mid x_r, R = r)}_{\substack{\text{Extrapolation distribution} \\ \text{not identifiable}}} .$$

- An imputation model imposes an **identifying** assumption that links

  $p(x_{\bar{r}} \mid x_r, R = r) \iff$ observed data quantities.



pattern $R = (1, 0)$: same observed x_1, different extrapolation densities

1. **Missing completely at random (MCAR):** $R \perp\!\!\!\perp X \implies$

$$p(x_{\bar{r}} \mid x_r, R = r) = p(x_{\bar{r}} \mid x_r, R = r') = p(x_{\bar{r}} \mid x_r), \quad \forall r, r' \in \bar{\mathcal{R}},$$

where $\bar{\mathcal{R}} = \{r \in \{0,1\}^d : P(R = r) > 0\}$.

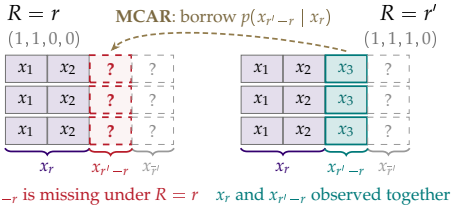
1. **Missing completely at random (MCAR):** $R \perp\!\!\!\perp X \implies$

$$p(x_{\bar{r}} \mid x_r, R = r) = p(x_{\bar{r}} \mid x_r, R = r') = p(x_{\bar{r}} \mid x_r), \quad \forall r, r' \in \bar{\mathcal{R}},$$

where $\bar{\mathcal{R}} = \{r \in \{0,1\}^d : P(R = r) > 0\}$.

- If $r' > r$ (*i.e.*, $r'_j \geq r_j$ for all j and $r'_k > r_k$ for at least one k), then

$$p(x_{r'-r} \mid x_r, R = r) \overset{\text{MCAR}}{=} \underbrace{p(x_{r'-r} \mid x_r, R = r')}_{\text{identifiable}}.$$



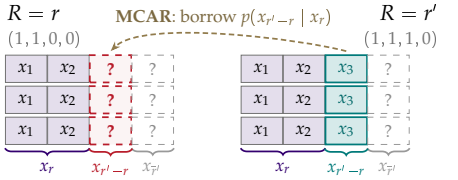
1. **Missing completely at random (MCAR):** $R \perp\!\!\!\perp X \implies$

$$p(x_{\bar{r}} \mid x_r, R = r) = p(x_{\bar{r}} \mid x_r, R = r') = p(x_{\bar{r}} \mid x_r), \quad \forall r, r' \in \bar{\mathcal{R}},$$

where $\bar{\mathcal{R}} = \{r \in \{0,1\}^d : P(R = r) > 0\}$.

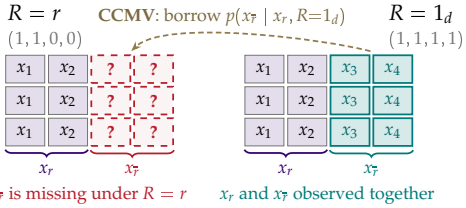
- If $r' > r$ (*i.e.*, $r'_j \geq r_j$ for all j and $r'_k > r_k$ for at least one k), then

$$p(x_{r'-r} \mid x_r, R = r) \overset{\text{MCAR}}{=} \underbrace{p(x_{r'-r} \mid x_r, R = r')}_{\text{identifiable}}.$$



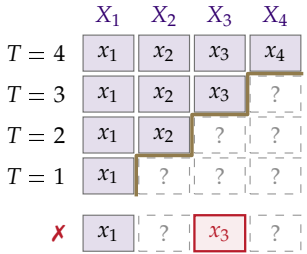
2. **Complete-case missing value (CCMV):**

$$p(x_{\bar{r}} \mid x_r, R = r) \overset{\text{CCMV}}{=} \underbrace{p(x_{\bar{r}} \mid x_r, R = 1_d)}_{\text{identifiable}}.$$



Monotone missingness: for any response pattern r,
$r_j = 0 \implies r_k = 0$ for all $k > j$ (Daniels and Hogan, 2008).

- $(X, R) \in \mathbb{R}^d \times \{0,1\}^d \iff (X_{\leq T}, T)$, $T = |R| \in \{1, ..., d\}$.
- Extrapolation density: $p(x_{>t} \mid x_{\leq t}, t) = \prod_{s=t}^{d-1} p(x_{s+1} \mid x_{\leq s}, t)$.



3. **Complete-case missing value (m-CCMV):**

$$p(x_{s+1} \mid x_{\leq s}, T = t) \overset{\text{m-CCMV}}{=} p(x_{s+1} \mid x_{\leq s}, T = d), \quad s = t, ..., d-1.$$

4. **Available-case missing value (m-ACMV)** $\iff$ missing at random (MAR):

$$p(x_{s+1} \mid x_{\leq s}, T = t) \overset{\text{m-ACMV}}{=} p(x_{s+1} \mid x_{\leq s}, T \geq s+1), \quad s = t, ..., d-1.$$

5. **Nearest-case missing value (m-NCMV):**

$$p(x_{s+1} \mid x_{\leq s}, T = t) \overset{\text{m-NCMV}}{=} p(x_{s+1} \mid x_{\leq s}, T = s+1), \quad s = t, ..., d-1.$$

Imputation as Distributional Learning

Statistics — **Machine learning**

- Explicit identifying assumption.
- Simple working models: EM (Little and Rubin, 2019), hot-deck (Andridge and Little, 2010), MICE (Van Buuren and Oudshoorn, 1999).

✓ Interpretable and principled.

✗ Limited by working models flexibility.

Imputation as Distributional Learning

Statistics

- Explicit identifying assumption.
- Simple working models: EM (Little and Rubin, 2019), hot-deck (Andridge and Little, 2010), MICE (Van Buuren and Oudshoorn, 1999).

✓ Interpretable and principled.

✗ Limited by working models flexibility.

Machine learning

- Pure prediction problem.
- Deep generative models: VAE (Mattei and Frellsen, 2019), GAN (Yoon et al., 2018), diffusion (Zhang et al., 2025).

✓ Expressive and scalable.

✗ Identifying assumptions often implicit.

✗ Mostly evaluated under MCAR by RMSE.

Imputation as Distributional Learning

Emputation

assumption-guided + flexible

Statistics

- Explicit identifying assumption.
- Simple working models: EM (Little and Rubin, 2019), hot-deck (Andridge and Little, 2010), MICE (Van Buuren and Oudshoorn, 1999).

✓ Interpretable and principled.

✗ Limited by working models flexibility.

Machine learning

- Pure prediction problem.
- Deep generative models: VAE (Mattei and Frellsen, 2019), GAN (Yoon et al., 2018), diffusion (Zhang et al., 2025).

✓ Expressive and scalable.

✗ Identifying assumptions often implicit.

✗ Mostly evaluated under MCAR by RMSE.

Goal: leverage the identifying assumption to *guide the training* of deep generative imputation models.

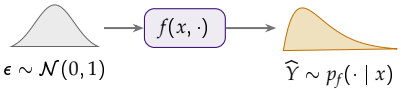
Emputation = Engression + Imputation



Engression: Distributional Regression via the Energy Score

Learn a generative model for the distribution of $Y|X$:

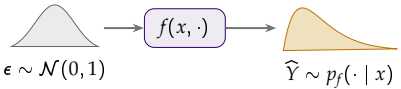
$$Y = f(X, \epsilon), \quad \epsilon \sim \mathcal{N}(0, I_q).$$



Engression: Distributional Regression via the Energy Score

Learn a generative model for the distribution of $Y|X$:

$$Y = f(X, \epsilon), \quad \epsilon \sim \mathcal{N}(0, I_q).$$



- *Energy score* of a distribution P at y (Gneiting and Raftery, 2007):

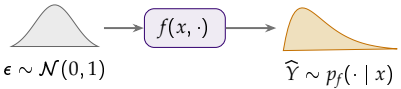
$$\mathrm{ES}(P, y) = \frac{1}{2}\mathbb{E}\|Y - Y'\| - \mathbb{E}\|Y - y\|, \quad Y, Y' \overset{\text{i.i.d.}}{\sim} P.$$

- Strictly proper scoring rule: $\mathbb{E}_{Y\sim P^*}[\mathrm{ES}(P, Y)] \leq \mathbb{E}_{Y\sim P^*}[\mathrm{ES}(P^*, Y)]$.

Engression: Distributional Regression via the Energy Score

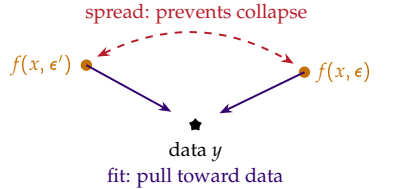
Learn a generative model for the distribution of $Y|X$:

$$Y = f(X, \epsilon), \quad \epsilon \sim \mathcal{N}(0, I_q).$$



- *Energy score* of a distribution P at y (Gneiting and Raftery, 2007):

$$\mathrm{ES}(P, y) = \frac{1}{2}\mathbb{E}\|Y - Y'\| - \mathbb{E}\|Y - y\|, \quad Y, Y' \overset{\text{i.i.d.}}{\sim} P.$$



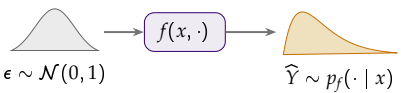
- Strictly proper scoring rule: $\mathbb{E}_{Y\sim P^*}[\mathrm{ES}(P, Y)] \le \mathbb{E}_{Y\sim P^*}[\mathrm{ES}(P^*, Y)]$.
- Population risk = expected negative energy score (Shen and Meinshausen, 2025):

$$L(f) = \mathbb{E}\Big[\underbrace{||f(X,\epsilon) - Y||}_{\text{fit}} - \frac{1}{2}\underbrace{||f(X,\epsilon) - f(X,\epsilon')||}_{\text{spread}}\Big], \qquad \epsilon, \epsilon' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_q).$$

Engression: Distributional Regression via the Energy Score

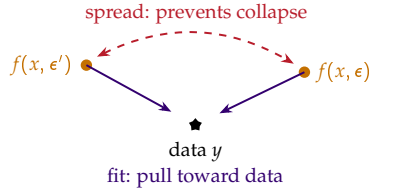
Learn a generative model for the distribution of $Y|X$:

$$Y = f(X, \epsilon), \quad \epsilon \sim \mathcal{N}(0, I_q).$$



- *Energy score* of a distribution P at y (Gneiting and Raftery, 2007):

$$\mathrm{ES}(P, y) = \frac{1}{2}\mathbb{E}\|Y - Y'\| - \mathbb{E}\|Y - y\|, \quad Y, Y' \overset{\text{i.i.d.}}{\sim} P.$$

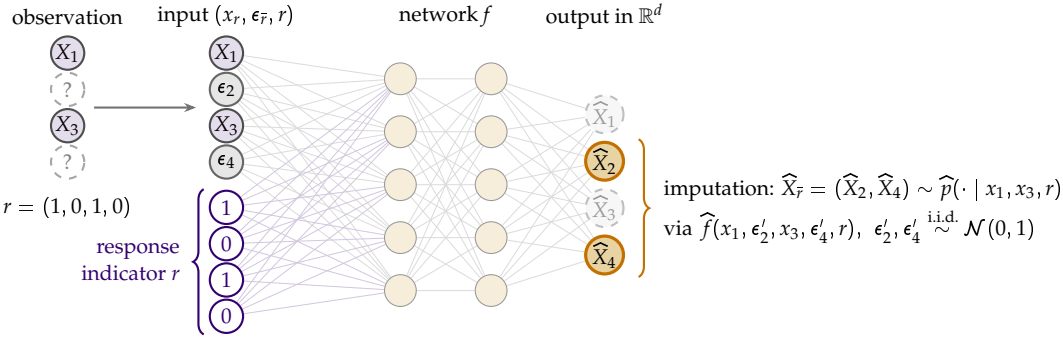


- Strictly proper scoring rule: $\mathbb{E}_{Y\sim P^*}[\mathrm{ES}(P, Y)] \le \mathbb{E}_{Y\sim P^*}[\mathrm{ES}(P^*, Y)]$.
- Population risk = expected negative energy score (Shen and Meinshausen, 2025):

$$L(f) = \mathbb{E}\Big[\underbrace{||f(X,\epsilon) - Y||}_{\text{fit}} - \frac{1}{2}\underbrace{||f(X,\epsilon) - f(X,\epsilon')||}_{\text{spread}}\Big], \qquad \epsilon, \epsilon' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_q).$$

- Learn $\widehat{f} = \arg\min_{f\in\mathcal{F}} \widehat{L}(f)$ through a neural network and sample $\widehat{Y} = \widehat{f}(x, \epsilon)$ with fresh ϵ.

Emputation Model



- Train a *single* neural network $f : \mathbb{R}^d \times \{0, 1\}^d \to \mathbb{R}^d$ for all response patterns $r \in \mathcal{R}$:

$$(X_R, R) = (x_r, r), \ r \neq 1_d \implies \text{replace missing } x_{\bar{r}} \text{ by } \epsilon_{\bar{r}} \sim \mathcal{N}(0, I_{|\bar{r}|}).$$

- (Multiple) imputation = sampling: $\widehat{X}_{\bar{r}} = \left[\widehat{f}(x_r, \epsilon'_{\bar{r}}, r)\right]_{\bar{r}} \in \mathbb{R}^{|r|}$ with fresh $\epsilon'_{\bar{r}} \sim \mathcal{N}(0, I_{|\bar{r}|})$.

Question: How can we guide the training so that $p_{f*}(x_{\bar{r}} \mid x_r, r) = p(x_{\bar{r}} \mid x_r, r)$ under a given missing data (identifying) assumption?

Design a risk that **encodes** the identifying assumption:

$$\mathcal{L}(f) = \mathbb{E}\left\{ \sum_{r \in \mathcal{R}} S_r(R) \left[\underbrace{\left|\left| X_{M_r(R)} - [f(X_r, \epsilon_{\bar{r}}, r)]_{M_r(R)} \right|\right|}_{\text{fit to observed \& masked entries}} - \underbrace{\frac{1}{2} \left|\left| [f(X_r, \epsilon_{\bar{r}}, r)]_{M_r(R)} - [f(X_r, \epsilon'_{\bar{r}}, r)]_{M_r(R)} \right|\right|}_{\text{spread among draws}} \right] \right\}.$$

Empirical emputation risk: $\{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n \implies$

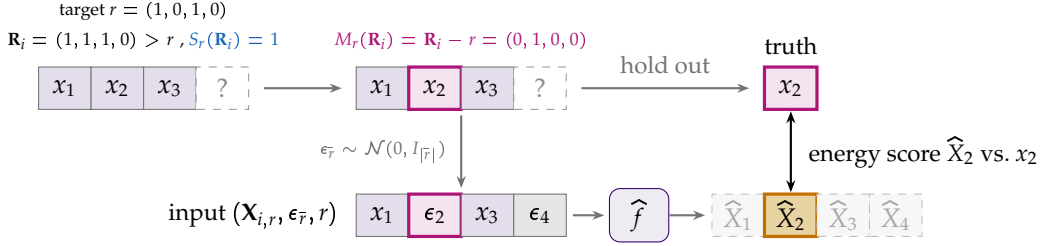
$$\widehat{\mathcal{L}}(f) = \frac{1}{n} \sum_{i=1}^{n} \sum_{r \in \mathcal{R}} S_r(\mathbf{R}_i) \left\{ \frac{1}{B} \sum_{b=1}^{B} \left|\left| \mathbf{X}_{i, M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b)}, r) \right]_{M_r(\mathbf{R}_i)} \right|\right| - \frac{1}{2B(B-1)} \sum_{b,b'=1}^{B} \left|\left| \left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b)}, r) \right]_{M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b')}, r) \right]_{M_r(\mathbf{R}_i)} \right|\right| \right\}, \quad \mathcal{R} = \{ r \neq 1_d : P(R = r) > 0 \}.$$

1. **Selection function** $S_r(\mathbf{R}_i) \geq 0$: whether (and with what weight) unit i contributes to the loss for pattern r.
2. **Masked pattern** $M_r(\mathbf{R}_i) \in \{0,1\}^d$: which *observed* entries of $\mathbf{X}_i \in \mathbb{R}^d$ are treated as targets to be reconstructed.

MCAR: $R \perp\!\!\!\perp X \implies$ for any $r' > r$, $p(x_{r'-r} \mid x_r, R = r) = p(x_{r'-r} \mid x_r, R = r')$.

Learn $p(x_{\bar{r}} \mid x_r, r)$ from every unit with response pattern $\mathbf{R}_i > r$.

- **Selection function:** $S_r(\mathbf{R}_i) = \frac{I(\mathbf{R}_i > r)}{|\mathbf{R}_i - r|}$; **Masked pattern:** $M_r(\mathbf{R}_i) = \mathbf{R}_i - r$.



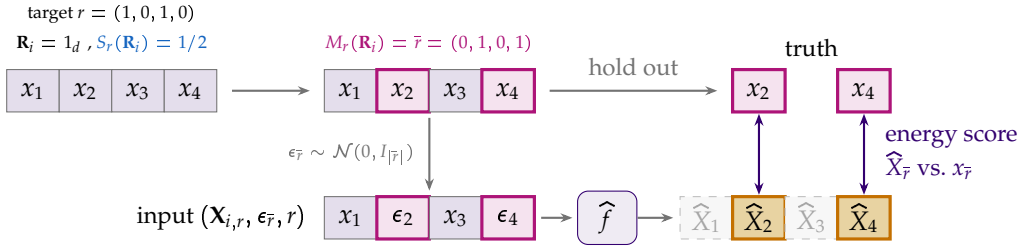
Theorem (Yang et al., 2026): Under regularity conditions, if f^* minimizes $\mathcal{L}(f) = \mathbb{E}[\widehat{\mathcal{L}}(f)]$ under the above $S_r(\mathbf{R}_i)$ and $M_r(\mathbf{R}_i)$, then

$$p_{f^*}(x_{\bar{r}} \mid x_r, r) = p(x_{\bar{r}} \mid x_r) \quad \text{for any } r \in \mathcal{R}.$$

CCMV: $p(x_{\bar{r}} \mid x_r, R = r) \overset{\text{CCMV}}{=} p(x_{\bar{r}} \mid x_r, R = 1_d)$ for every $r \in \mathcal{R}$.

Learn $p(x_{\bar{r}} \mid x_r, r)$ from those complete cases ($\mathbf{R}_i = 1_d$).

- **Selection function:** $S_r(\mathbf{R}_i) = \frac{I(\mathbf{R}_i = 1_d)}{|\bar{r}|}$; **Masked pattern:** $M_r(\mathbf{R}_i) = 1_d - r = \bar{r}$.



Theorem (Yang et al., 2026): Under regularity conditions, if f^* minimizes $\mathcal{L}(f) = \mathbb{E}[\widehat{\mathcal{L}}(f)]$ under the above $S_r(\mathbf{R}_i)$ and $M_r(\mathbf{R}_i)$, then

$$p_{f^*}(x_{\bar{r}} \mid x_r, R = r) = p(x_{\bar{r}} \mid x_r, R = 1_d) \quad \text{for any } r \in \mathcal{R}.$$

$$\widehat{\mathcal{L}}(f) = \frac{1}{n}\sum_{i=1}^{n}\sum_{r\in\mathcal{R}} S_r(\mathbf{R}_i)\left\{\frac{1}{B}\sum_{b=1}^{B}\left\|\mathbf{X}_{i,M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b)}, r)\right]_{M_r(\mathbf{R}_i)}\right\| - \frac{1}{2B(B-1)}\sum_{b,b'=1}^{B}\left\|\left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b)}, r)\right]_{M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b')}, r)\right]_{M_r(\mathbf{R}_i)}\right\|\right\}.$$

- **MCAR:** $S_r(\mathbf{R}_i) = \frac{I(\mathbf{R}_i > r)}{|\mathbf{R}_i - r|}$, $M_r(\mathbf{R}_i) = \mathbf{R}_i - r$. **CCMV:** $S_r(\mathbf{R}_i) = \frac{I(\mathbf{R}_i = 1_d)}{|\bar{r}|}$, $M_r(\mathbf{R}_i) = \bar{r}$.

One example target pattern $r = (1, 0, 1, 0)$: learn $p(x_2, x_4 \mid x_1, x_3, R = r)$.

	Observed units	MCAR risk	CCMV risk
1111	x_1 x_2 x_3 x_4		
1110	x_1 x_2 x_3 ?		
1011	x_1 ? x_3 x_4		
1101	x_1 x_2 ? x_4		
1010	x_1 ? x_3 ?		

Emputation Risk Comparison Between MCAR and CCMV

$$\widehat{\mathcal{L}}(f) = \frac{1}{n}\sum_{i=1}^{n}\sum_{r\in\mathcal{R}} S_r(\mathbf{R}_i)\left\{\frac{1}{B}\sum_{b=1}^{B}\left\|\mathbf{X}_{i,M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b)}, r)\right]_{M_r(\mathbf{R}_i)}\right\| - \frac{1}{2B(B-1)}\sum_{b,b'=1}^{B}\left\|\left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b)}, r)\right]_{M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b')}, r)\right]_{M_r(\mathbf{R}_i)}\right\|\right\}.$$

- **MCAR:** $S_r(\mathbf{R}_i) = \frac{I(\mathbf{R}_i > r)}{|\mathbf{R}_i - r|}$, $M_r(\mathbf{R}_i) = \mathbf{R}_i - r$. **CCMV:** $S_r(\mathbf{R}_i) = \frac{I(\mathbf{R}_i = 1_d)}{|\bar{r}|}$, $M_r(\mathbf{R}_i) = \bar{r}$.

One example target pattern $r = (1, 0, 1, 0)$: learn $p(x_2, x_4 \mid x_1, x_3, R = r)$.

Observed units		MCAR risk	CCMV risk
1111	x_1 x_2 x_3 x_4	x_1 x_2 x_3 x_4	✓ learns X_2, X_4
1110	x_1 x_2 x_3 ?	x_1 x_2 x_3 ?	✓ learns X_2
1011	x_1 ? x_3 x_4	x_1 ? x_3 x_4	✓ learns X_4
1101	x_1 x_2 ? x_4	✗ X_3 unobserved: no valid input	
1010	x_1 ? x_3 ?	✗ nothing extra to mask	

Emputation Risk Comparison Between MCAR and CCMV

$$\widehat{\mathcal{L}}(f) = \frac{1}{n}\sum_{i=1}^{n}\sum_{r\in\mathcal{R}} S_r(\mathbf{R}_i)\left\{\frac{1}{B}\sum_{b=1}^{B}\left\|\mathbf{X}_{i,M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon^{(b)}_{i,\bar{r}}, r)\right]_{M_r(\mathbf{R}_i)}\right\| - \frac{1}{2B(B-1)}\sum_{b,b'=1}^{B}\left\|\left[f(\mathbf{X}_{i,r}, \epsilon^{(b)}_{i,\bar{r}}, r)\right]_{M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon^{(b')}_{i,\bar{r}}, r)\right]_{M_r(\mathbf{R}_i)}\right\|\right\}.$$

- **MCAR:** $S_r(\mathbf{R}_i) = \frac{I(\mathbf{R}_i > r)}{|\mathbf{R}_i - r|}$, $M_r(\mathbf{R}_i) = \mathbf{R}_i - r$. **CCMV:** $S_r(\mathbf{R}_i) = \frac{I(\mathbf{R}_i = 1_d)}{|\bar{r}|}$, $M_r(\mathbf{R}_i) = \bar{r}$.

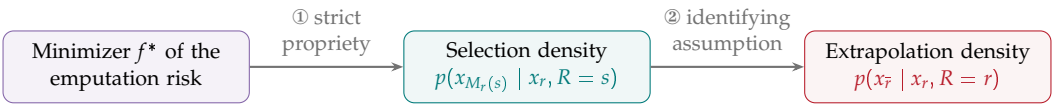
One example target pattern $r = (1, 0, 1, 0)$: learn $p(x_2, x_4 \mid x_1, x_3, R = r)$.

	Observed units	MCAR risk	CCMV risk
1111	x_1 x_2 x_3 x_4	x_1 x_2 x_3 x_4 ✓ learns X_2, X_4	x_1 x_2 x_3 x_4 ✓ learns X_2, X_4
1110	x_1 x_2 x_3 ?	x_1 x_2 x_3 ? ✓ learns X_2	✗ not a complete case
1011	x_1 ? x_3 x_4	x_1 ? x_3 x_4 ✓ learns X_4	✗ not a complete case
1101	x_1 x_2 ? x_4	✗ X_3 unobserved: no valid input	✗ not a complete case
1010	x_1 ? x_3 ?	✗ nothing extra to mask	✗ not a complete case

High-Level Proof Idea

$$\mathcal{L}(f) = \mathbb{E}\left\{\sum_{r\in\mathcal{R}} S_r(R)\left[\left|\left|X_{M_r(R)} - [f(X_r,\epsilon_{\bar{r}},r)]_{M_r(R)}\right|\right| - \frac{1}{2}\left|\left|[f(X_r,\epsilon_{\bar{r}},r)]_{M_r(R)} - [f(X_r,\epsilon'_{\bar{r}},r)]_{M_r(R)}\right|\right|\right]\right\}$$

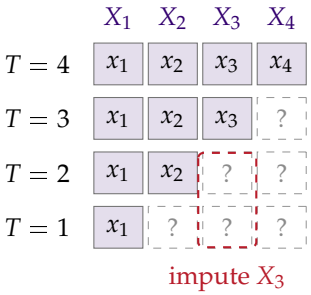
$$= \sum_{r\in\mathcal{R}}\sum_{s\in\bar{\mathcal{R}}} S_r(s)\cdot P(R=s)\cdot \mathbb{E}_{X_r|R=s}\left\{\mathbb{E}_{X_{M_r(s)}\sim p(\cdot|X_r,R=s)}\left[-\mathrm{ES}\left([p_f(\cdot\mid X_r,r)]_{M_r(s)}, X_{M_r(s)}\right)\right]\right\}$$



1. **Strict propriety:** $f^* = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{X_{M_r(s)}\sim p(\cdot|X_r,R=s)}\left[-\mathrm{ES}\left(\left[p_f(\cdot\mid X_r,r)\right]_{M_r(s)}, X_{M_r(s)}\right)\right]$
   $\iff \left[p_f(\cdot\mid x_r,r)\right]_{M_r(s)} = p(x_{M_r(s)}\mid x_r, R=s)$ under the selection $S_r(s), \forall s\in\bar{\mathcal{R}}$.
2. **Identifying assumption:** link $p(x_{M_r(s)}\mid x_r, R=s)$ with $p(x_{\bar{r}}\mid x_r, R=r)$.
   - **MCAR:** $p(x_{M_r(s)}\mid x_r, R=s) = p(x_{M_r(s)}\mid x_r, R=r)$ and $M_r(s) = s - r$.
   - **CCMV:** $p(x_{\bar{r}}\mid x_r, R=1_d) = p(x_{\bar{r}}\mid x_r, R=r)$.

Regularity conditions: a rich model class $\mathcal{F}$, positivity, and finite first moments (▸ details).

Emputation for Monotone Missingness

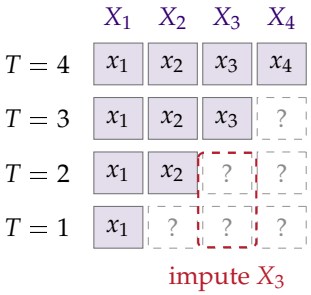


Step $s = 2$: learn $p(x_3 \mid x_1, x_2)$

- Observed vector $(X_{\leq T}, T)$ with $T = |R|$
- Extrapolation density:

$$p(x_{>t} \mid x_{\leq t}, t) = \prod_{s=t}^{d-1} p(x_{s+1} \mid x_{\leq s}, t).$$

Emputation for Monotone Missingness



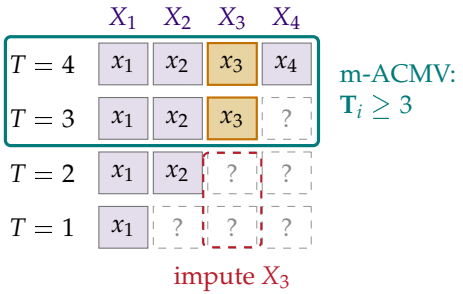
Step $s = 2$: learn $p(x_3 \mid x_1, x_2)$

- Observed vector $(X_{\leq T}, T)$ with $T = |R|$
- Extrapolation density:

$$p(x_{>t} \mid x_{\leq t}, t) = \prod_{s=t}^{d-1} p(x_{s+1} \mid x_{\leq s}, t).$$

- Masked pattern: $M_s(\mathbf{R}_i) \equiv M_s(\mathbf{T}_i) = s + 1$.

Emputation for Monotone Missingness



Step $s = 2$: learn $p(x_3 \mid x_1, x_2)$

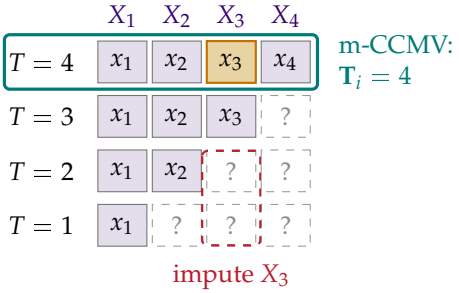
- Observed vector $(X_{\leq T}, T)$ with $T = |R|$
- Extrapolation density:

$$p(x_{>t} \mid x_{\leq t}, t) = \prod_{s=t}^{d-1} p(x_{s+1} \mid x_{\leq s}, t).$$

- Masked pattern: $M_s(\mathbf{R}_i) \equiv M_s(\mathbf{T}_i) = s + 1$.
- Selection function $S_s(\mathbf{T}_i)$ differs:

Assumption	Identify $p(x_{s+1} \mid x_{\leq s}, t)$	$S_s(\mathbf{T}_i)$
m-ACMV	$p(x_{s+1} \mid x_{\leq s}, T \geq s+1)$	$I(\mathbf{T}_i \geq s+1)$

Emputation for Monotone Missingness



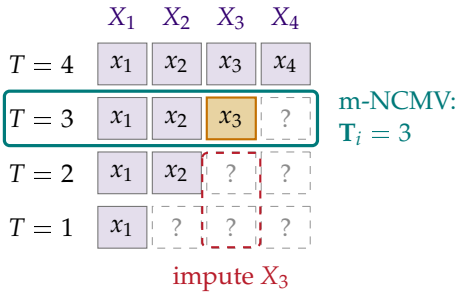
Step $s = 2$: learn $p(x_3 \mid x_1, x_2)$

- Observed vector $(X_{\leq T}, T)$ with $T = |R|$
- Extrapolation density:

$$p(x_{>t} \mid x_{\leq t}, t) = \prod_{s=t}^{d-1} p(x_{s+1} \mid x_{\leq s}, t).$$

- Masked pattern: $M_s(\mathbf{R}_i) \equiv M_s(\mathbf{T}_i) = s + 1$.
- Selection function $S_s(\mathbf{T}_i)$ differs:

Assumption	Identify $p(x_{s+1} \mid x_{\leq s}, t)$	$S_s(\mathbf{T}_i)$
m-ACMV	$p(x_{s+1} \mid x_{\leq s}, T \geq s+1)$	$I(\mathbf{T}_i \geq s+1)$
m-CCMV	$p(x_{s+1} \mid x_{\leq s}, T = d)$	$I(\mathbf{T}_i = d)$



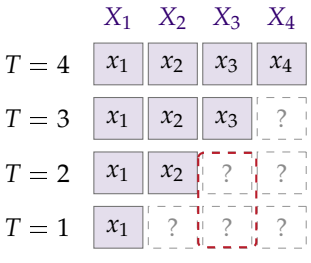
Step $s = 2$: learn $p(x_3 \mid x_1, x_2)$

- Observed vector $(X_{\leq T}, T)$ with $T = |R|$
- Extrapolation density:

$$p(x_{>t} \mid x_{\leq t}, t) = \prod_{s=t}^{d-1} p(x_{s+1} \mid x_{\leq s}, t).$$

- Masked pattern: $M_s(\mathbf{R}_i) \equiv M_s(\mathbf{T}_i) = s + 1$.
- Selection function $S_s(\mathbf{T}_i)$ differs:

Assumption	Identify $p(x_{s+1} \mid x_{\leq s}, t)$	$S_s(\mathbf{T}_i)$
m-ACMV	$p(x_{s+1} \mid x_{\leq s}, T \geq s+1)$	$I(\mathbf{T}_i \geq s+1)$
m-CCMV	$p(x_{s+1} \mid x_{\leq s}, T = d)$	$I(\mathbf{T}_i = d)$
m-NCMV	$p(x_{s+1} \mid x_{\leq s}, T = s+1)$	$I(\mathbf{T}_i = s+1)$



Step $s = 2$: learn $p(x_3 \mid x_1, x_2)$

- Observed vector $(X_{\leq T}, T)$ with $T = |R|$
- Extrapolation density:

$$p(x_{>t} \mid x_{\leq t}, t) = \prod_{s=t}^{d-1} p(x_{s+1} \mid x_{\leq s}, t).$$

- Masked pattern: $M_s(\mathbf{R}_i) \equiv M_s(\mathbf{T}_i) = s + 1$.
- Selection function $S_s(\mathbf{T}_i)$ differs:

Assumption	Identify $p(x_{s+1} \mid x_{\leq s}, t)$	$S_s(\mathbf{T}_i)$
m-ACMV	$p(x_{s+1} \mid x_{\leq s}, T \geq s+1)$	$I(\mathbf{T}_i \geq s+1)$
m-CCMV	$p(x_{s+1} \mid x_{\leq s}, T = d)$	$I(\mathbf{T}_i = d)$
m-NCMV	$p(x_{s+1} \mid x_{\leq s}, T = s+1)$	$I(\mathbf{T}_i = s+1)$

$$\widehat{\mathcal{L}}(g) = \frac{1}{n} \sum_{i=1}^{n} \sum_{s=1}^{d-1} S_s(\mathbf{T}_i) \left\{ \frac{1}{B} \sum_{b=1}^{B} \left| \mathbf{X}_{i, M_s(\mathbf{T}_i)} - g(\mathbf{X}_{i,\leq s}, \epsilon_{i,>s}^{(b)}, s) \right| - \frac{1}{2B(B-1)} \sum_{b,b'=1}^{B} \left| g(\mathbf{X}_{i,\leq s}, \epsilon_{i,>s}^{(b)}, s) - g(\mathbf{X}_{i,\leq s}, \epsilon_{i,>s}^{(b')}, s) \right| \right\}.$$

$$\widehat{\mathcal{L}}(f) = \frac{1}{n}\sum_{i=1}^{n}\sum_{r\in\mathcal{R}} S_r(\mathbf{R}_i)\left\{\frac{1}{B}\sum_{b=1}^{B}\left|\left|\mathbf{X}_{i,M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b)}, r)\right]_{M_r(\mathbf{R}_i)}\right|\right| - \frac{1}{2B(B-1)}\sum_{b\neq b'}^{B}\left|\left|\left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b)}, r)\right]_{M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r}, \epsilon_{i,\bar{r}}^{(b')}, r)\right]_{M_r(\mathbf{R}_i)}\right|\right|\right\}.$$

	Assumption	Masked pattern $M_r(\mathbf{R}_i)$	Selection function $S_r(\mathbf{R}_i)$
Nonmonotone	MCAR	$\mathbf{R}_i - r$	$\lvert\mathbf{R}_i - r\rvert^{-1} I(r < \mathbf{R}_i)$
	CCMV	$\bar{r}$	$\lvert\bar{r}\rvert^{-1} I(\mathbf{R}_i = 1_d)$
	Tree graph	$\mathsf{PA}(r) - r$	$\lvert\mathsf{PA}(r) - r\rvert^{-1} I(\mathbf{R}_i = \mathsf{PA}(r))$
Monotone	m-ACMV	$\lvert r\rvert + 1$	$I(\lvert\mathbf{R}_i\rvert \geq \lvert r\rvert + 1)$
	m-CCMV	$\lvert r\rvert + 1$	$I(\lvert\mathbf{R}_i\rvert = d)$
	m-NCMV	$\lvert r\rvert + 1$	$I(\lvert\mathbf{R}_i\rvert = \lvert r\rvert + 1)$

A Unified Emputation Framework

$$\widehat{\mathcal{L}}(f) = \frac{1}{n}\sum_{i=1}^{n}\sum_{r\in\mathcal{R}} S_r(\mathbf{R}_i)\left\{\frac{1}{B}\sum_{b=1}^{B}\left\|\mathbf{X}_{i,M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r},\epsilon_{i,\bar{r}}^{(b)},r)\right]_{M_r(\mathbf{R}_i)}\right\| - \frac{1}{2B(B-1)}\sum_{b\neq b'}^{B}\left\|\left[f(\mathbf{X}_{i,r},\epsilon_{i,\bar{r}}^{(b)},r)\right]_{M_r(\mathbf{R}_i)} - \left[f(\mathbf{X}_{i,r},\epsilon_{i,\bar{r}}^{(b')},r)\right]_{M_r(\mathbf{R}_i)}\right\|\right\}.$$

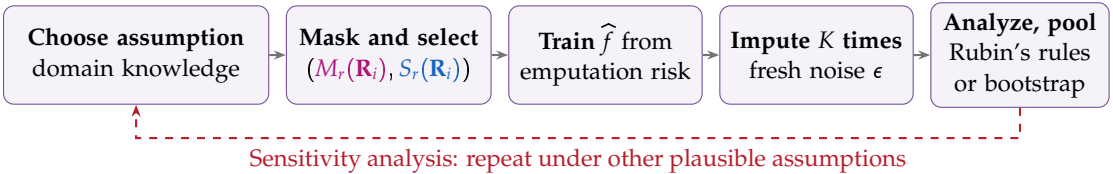
	Assumption	Masked pattern $M_r(\mathbf{R}_i)$	Selection function $S_r(\mathbf{R}_i)$
Nonmonotone	MCAR	$\mathbf{R}_i - r$	$\lvert\mathbf{R}_i - r\rvert^{-1} I(r < \mathbf{R}_i)$
	CCMV	$\bar{r}$	$\lvert\bar{r}\rvert^{-1} I(\mathbf{R}_i = 1_d)$
	Tree graph	$\mathsf{PA}(r) - r$	$\lvert\mathsf{PA}(r) - r\rvert^{-1} I(\mathbf{R}_i = \mathsf{PA}(r))$
Monotone	m-ACMV	$\lvert r\rvert + 1$	$I(\lvert\mathbf{R}_i\rvert \geq \lvert r\rvert + 1)$
	m-CCMV	$\lvert r\rvert + 1$	$I(\lvert\mathbf{R}_i\rvert = d)$
	m-NCMV	$\lvert r\rvert + 1$	$I(\lvert\mathbf{R}_i\rvert = \lvert r\rvert + 1)$

- Encode the identifying assumption through $M_r(\mathbf{R}_i)$ and $S_r(\mathbf{R}_i)$.
- $f^* = \arg\min_{f\in\mathcal{F}} \mathbb{E}\left[\widehat{\mathcal{L}}(f)\right]$ recovers the identified extrapolation density.

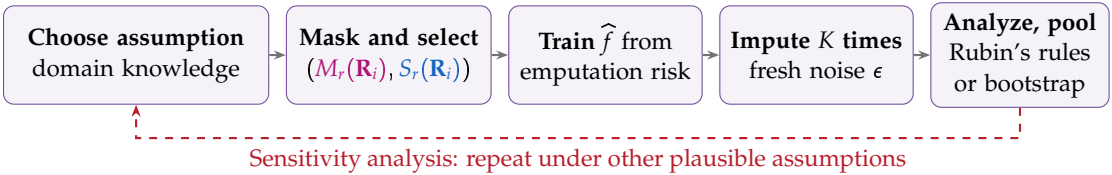
Recipe for a new assumption: specify how it identifies $p(x_{\bar{r}} \mid x_r, R = r)$:

- Reference units give $S_r(\mathbf{R}_i)$, the revealed coordinates give $M_r(\mathbf{R}_i)$.

Emputation in Practice

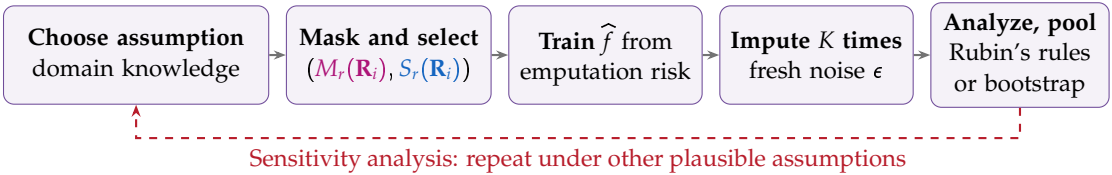


Emputation in Practice



- **Assumption choice:** Except for MCAR, none of these assumptions can be rejected by the observed data (Robins, 1997; Vansteelandt et al., 2006).
  - CCMV $\iff$ a logit discrete choice selection model (Tchetgen et al., 2018): complete cases are a credible reference.
  - m-ACMV = MAR for dropout.
  - m-NCMV: adjacent dropouts are alike.

Emputation in Practice

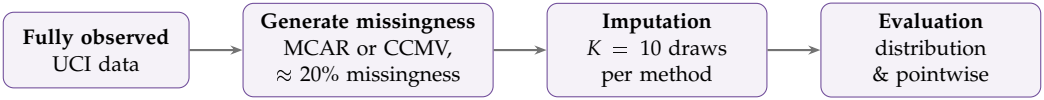


- **Assumption choice:** Except for MCAR, none of these assumptions can be rejected by the observed data (Robins, 1997; Vansteelandt et al., 2006).
  - CCMV $\iff$ a logit discrete choice selection model (Tchetgen et al., 2018): complete cases are a credible reference.
  - m-ACMV = MAR for dropout.
  - m-NCMV: adjacent dropouts are alike.
- **Recommendation:** fit a few plausible emputation risks ($\equiv$ identifying assumptions) and compare downstream conclusions.

Simulation Studies and Application to Alzheimer's Disease Data



Simulation Study Design



Data	n	d
Concrete	1030	9
Wine	4898	12
CCPP	9568	5

- **Emputation** trained separately with MCAR and CCMV risks: *matched* vs. *mismatched* to the true missingness mechanism.
- **Competitors:** mean, EM (Gaussian), MICE with PMM (Van Buuren and Oudshoorn, 1999), missForest (MF; Stekhoven and Bühlmann 2012), GAIN (Yoon et al., 2018).

- **Metrics** (100 Monte Carlo repetitions):
  - *Pointwise:* root mean squared error (RMSE) and mean absolute error (MAE).
  - *Distributional:* mean absolute difference in correlations (MADC), energy distance with $\beta = 1$, MMD^2 with a Gaussian kernel.

Pointwise Accuracy Is Not Enough

RMSE at 20% missingness (**best**, second best)

Data	DGP	Emp-MCAR	Emp-CCMV	missForest	MICE	EM	GAIN	Mean
Concrete	MCAR	0.6188	0.6293	**0.5571**	0.7118	0.7326	0.9768	1.0013
	CCMV	0.6239	**0.6238**	0.7168	0.7495	0.8046	1.0759	1.0385
Wine	MCAR	0.7735	0.7896	**0.7123**	0.9325	0.9577	1.0483	1.0013
	CCMV	0.7343	**0.7339**	0.8219	0.9714	0.9753	1.0702	0.9085
CCPP	MCAR	0.6071	0.6069	**0.5433**	0.6274	0.6421	0.8304	1.0004
	CCMV	0.5659	0.5539	**0.5530**	0.6055	0.6247	0.8219	1.1072

Pointwise Accuracy Is Not Enough

RMSE at 20% missingness (**best**, second best)

Data	DGP	Emp-MCAR	Emp-CCMV	missForest	MICE	EM	GAIN	Mean
Concrete	MCAR	0.6188	0.6293	**0.5571**	0.7118	0.7326	0.9768	1.0013
	CCMV	0.6239	**0.6238**	0.7168	0.7495	0.8046	1.0759	1.0385
Wine	MCAR	0.7735	0.7896	**0.7123**	0.9325	0.9577	1.0483	1.0013
	CCMV	0.7343	**0.7339**	0.8219	0.9714	0.9753	1.0702	0.9085
CCPP	MCAR	0.6071	0.6069	**0.5433**	0.6274	0.6421	0.8304	1.0004
	CCMV	0.5659	0.5539	**0.5530**	0.6055	0.6247	0.8219	1.1072

- missForest is fit to minimize pointwise errors, yielding the lowest RMSE in 4 of 6 settings.

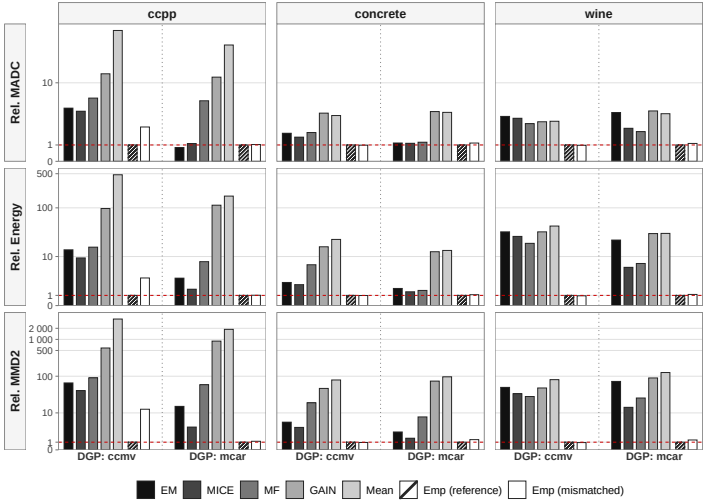
Pointwise Accuracy Is Not Enough

RMSE at 20% missingness (**best**, second best)

Data	DGP	Emp-MCAR	Emp-CCMV	missForest	MICE	EM	GAIN	Mean
Concrete	MCAR	0.6188	0.6293	**0.5571**	0.7118	0.7326	0.9768	1.0013
	CCMV	0.6239	**0.6238**	0.7168	0.7495	0.8046	1.0759	1.0385
Wine	MCAR	0.7735	0.7896	**0.7123**	0.9325	0.9577	1.0483	1.0013
	CCMV	0.7343	**0.7339**	0.8219	0.9714	0.9753	1.0702	0.9085
CCPP	MCAR	0.6071	0.6069	**0.5433**	0.6274	0.6421	0.8304	1.0004
	CCMV	0.5659	0.5539	**0.5530**	0.6055	0.6247	0.8219	1.1072

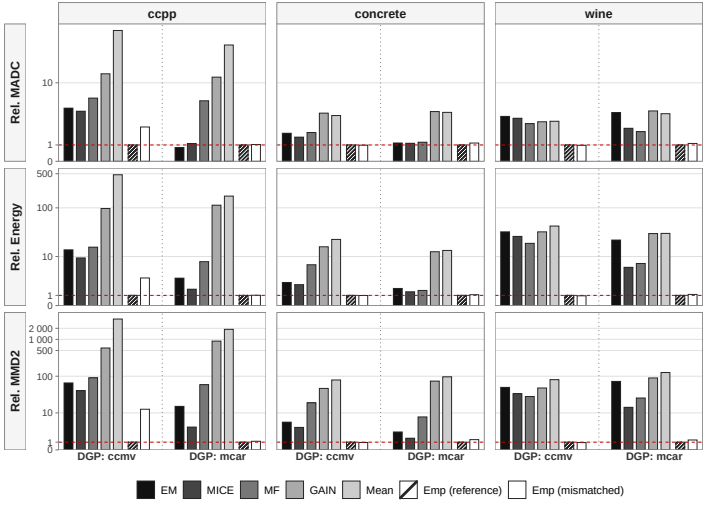
- missForest is fit to minimize pointwise errors, yielding the lowest RMSE in 4 of 6 settings.
- … yet, good point predictions $\neq$ good imputations!!

Distributional Evaluation at ≈20% Missingness



- Each metric is divided by that of **matched** emputation (hatched = 1).

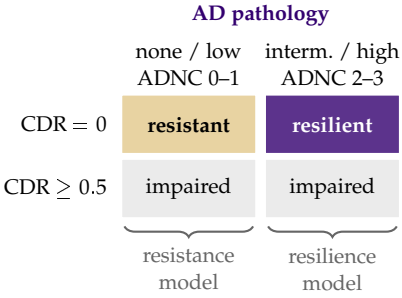
Distributional Evaluation at ≈20% Missingness



- Each metric is divided by that of **matched** emputation (hatched = 1).
- Matched emputation has the smallest energy distance, MMD^2, and MADC in almost every setting.
- Mismatched emputation remain competitive.
- Other methods (*e.g.*, missForest) show larger distributional discrepancies.

NACC Application: Cognitive Resilience and Resistance

NACC: $n = 4{,}906$ autopsied participants, visit closest to death (≤ 2 years).



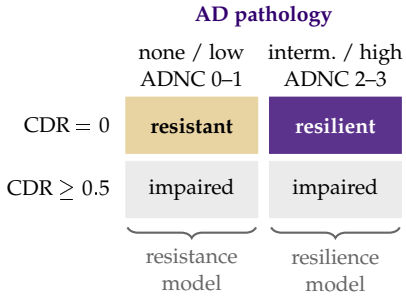
Resistance: normal cognition (CDR $= 0$) with none or low AD pathology.

Resilience: normal cognition (CDR $= 0$) *despite* intermediate or high AD pathology.

ADNC: AD Neuropathologic Change; CDR: Clinical Dementia Rating; GDS: Geriatric Depression Scale.

NACC Application: Cognitive Resilience and Resistance

NACC: $n = 4{,}906$ autopsied participants, visit closest to death (≤ 2 years).



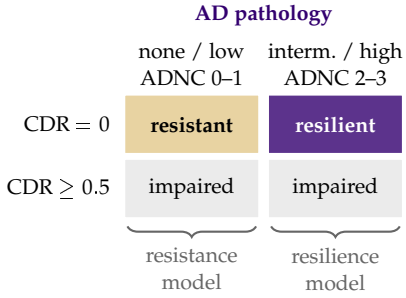
- **Resistance:** normal cognition (CDR $= 0$) with none or low AD pathology.
- **Resilience:** normal cognition (CDR $= 0$) *despite* intermediate or high AD pathology.
- **Models:** separate logistic regressions of $I(\text{CDR} = 0)$ within the two pathology stratum.

- Predictors: age, sex, race, education, APOE $\varepsilon 4/\varepsilon 2$, BMI, SBP, GDS, diuretic use.

ADNC: AD Neuropathologic Change; CDR: Clinical Dementia Rating; GDS: Geriatric Depression Scale.

NACC Application: Cognitive Resilience and Resistance

NACC: $n = 4{,}906$ autopsied participants, visit closest to death (≤ 2 years).



Resistance: normal cognition (CDR $= 0$) with none or low AD pathology.

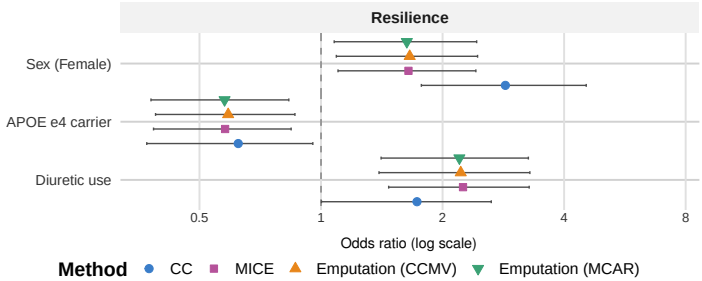
Resilience: normal cognition (CDR $= 0$) *despite* intermediate or high AD pathology.

Models: separate logistic regressions of $I(\text{CDR} = 0)$ within the two pathology stratum.

- Predictors: age, sex, race, education, APOE $\varepsilon 4/\varepsilon 2$, BMI, SBP, GDS, diuretic use.
- BMI, SBP, and GDS $\approx$40–50% missing; only **1,820** complete cases.
- Compare CC, MICE (PMM), emputation-MCAR, and emputation-CCMV under 500 bootstrap replicates $\times$ 10 imputations.

ADNC: AD Neuropathologic Change; CDR: Clinical Dementia Rating; GDS: Geriatric Depression Scale.

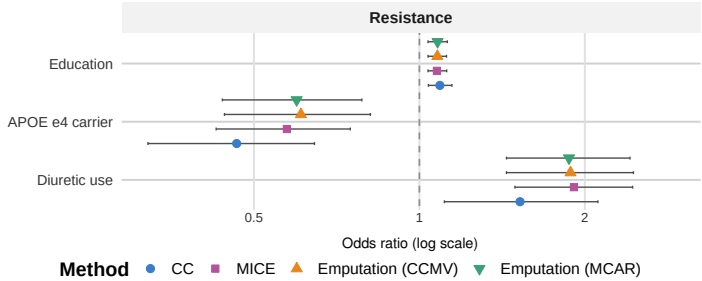
NACC Results: Odds Ratios with 95% Bootstrap CIs



- APOE $\varepsilon 4$: lower odds of resilience under every method.
- Female sex: the CC estimate is larger than those from MICE and emputation.
- Diuretic use: positive under MICE and emputation; CC not significant.

Selected predictors from the full logistic models (all predictors).

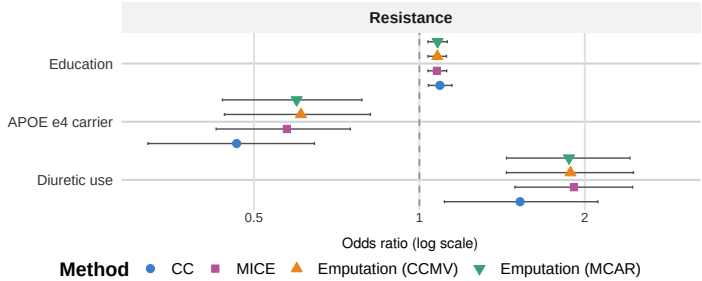
NACC Results: Odds Ratios with 95% Bootstrap CIs



- APOE $\varepsilon 4$: lower odds of resistance under every method.
- Education: higher odds, consistent with cognitive reserve (Stern, 2012).
- Diuretic use: positive under every method.

Selected predictors from the full logistic models (▸ all predictors).

NACC Results: Odds Ratios with 95% Bootstrap CIs



- APOE $\varepsilon 4$: lower odds of resistance under every method.
- Education: higher odds, consistent with cognitive reserve (Stern, 2012).
- Diuretic use: positive under every method.

Finding: Scientific conclusions agree across MICE and emputation, but complete-case analysis can diverge.

Selected predictors from the full logistic models (▸ all predictors).

Conclusion



Main Takeaways of Today's Talk

Emputation: a neural network generative framework with an energy score risk.

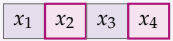
- The identifying assumption decides **what to mask** and **whom to learn from**.

① Energy score



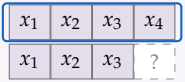
learn *distributions*, not means

② Masked pattern $M_r(\mathbf{R}_i)$



which entries to hold out

③ Selection function $S_r(\mathbf{R}_i)$



which units act as donors

Main Takeaways of Today's Talk

Emputation: a neural network generative framework with an energy score risk.

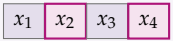
- The identifying assumption decides **what to mask** and **whom to learn from**.

① Energy score



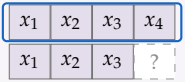
learn *distributions*, not means

② Masked pattern $M_r(\mathbf{R}_i)$



which entries to hold out

③ Selection function $S_r(\mathbf{R}_i)$



which units act as donors

- **Theory:** the population minimizer recovers the identified extrapolation density (MCAR, CCMV, tree graphs, monotone ACMV/CCMV/NCMV).

Main Takeaways of Today's Talk

Emputation: a neural network generative framework with an energy score risk.

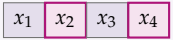
- The identifying assumption decides **what to mask** and **whom to learn from**.

① Energy score



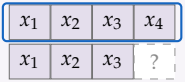
learn *distributions*, not means

② Masked pattern $M_r(\mathbf{R}_i)$



which entries to hold out

③ Selection function $S_r(\mathbf{R}_i)$



which units act as donors

- **Theory:** the population minimizer recovers the identified extrapolation density (MCAR, CCMV, tree graphs, monotone ACMV/CCMV/NCMV).
- **Practice:** one network for all patterns, direct sampling $\Longrightarrow$ multiple imputation.
- **Evidence:** simulations and an Alzheimer's disease application.

Main Takeaways of Today's Talk

Emputation: a neural network generative framework with an energy score risk.

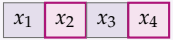
- The identifying assumption decides **what to mask** and **whom to learn from**.

① Energy score



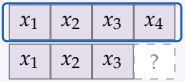
learn *distributions*, not means

② Masked pattern $M_r(\mathbf{R}_i)$



which entries to hold out

③ Selection function $S_r(\mathbf{R}_i)$



which units act as donors

- **Theory:** the population minimizer recovers the identified extrapolation density (MCAR, CCMV, tree graphs, monotone ACMV/CCMV/NCMV).
- **Practice:** one network for all patterns, direct sampling $\Longrightarrow$ multiple imputation.
- **Evidence:** simulations and an Alzheimer's disease application.

Thank you!

[1] Y. Yang, Y. Zhang, X. Shen, and Y.-C. Chen. Emputation: Identification-Guided Neural Imputation Framework. *arXiv preprint*, 2026. https://arxiv.org/abs/2607.05279.

References

R. R. Andridge and R. J. Little. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64, 2010.

Y.-C. Chen. Pattern graphs: a graphical approach to nonmonotone missing data. *The Annals of Statistics*, 50(1): 129–146, 2022.

M. J. Daniels and J. W. Hogan. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall/CRC, 2008.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

R. J. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88 (421):125–134, 1993.

R. J. Little and D. B. Rubin. *Statistical Analysis With Missing Data*. John Wiley & Sons, 2019.

P.-A. Mattei and J. Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, volume 97, pages 4413–4423. PMLR, 2019.

J. Näf, E. Scornet, and J. Josse. What is a good imputation under mar missingness? *arXiv preprint arXiv:2403.19196*, 2024.

J. M. Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16(1):21–37, 1997.

X. Shen and N. Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):653–677, 2025.

References

D. J. Stekhoven and P. Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Y. Stern. Cognitive reserve in ageing and Alzheimer's disease. *The Lancet Neurology*, 11(11):1006–1012, 2012.

D. Suen and Y.-C. Chen. Modeling multivariate missingness with tree graphs and conjugate odds. *arXiv preprint arXiv:2602.16992*, 2026.

G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

E. J. T. Tchetgen, L. Wang, and B. Sun. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4):2069, 2018.

S. Van Buuren. *Flexible imputation of missing data*. chapman and hall/CRC, 2018.

S. Van Buuren and K. Oudshoorn. Flexible multivariate imputation by MICE. *Leiden, The Netherlands: TNO Prevention Center*, pages 1–20, 1999.

S. Vansteelandt, E. Goetghebeur, M. G. Kenward, and G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, pages 953–979, 2006.

Y. Yang, Y. Zhang, X. Shen, and Y.-C. Chen. Emputation: Identification-guided neural imputation framework. *arXiv preprint arXiv:2607.05279*, 2026.

J. Yoon, J. Jordon, and M. Schaar. GAIN: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pages 5689–5698. PMLR, 2018.

H. Zhang, L. Fang, Q. Wu, and P. S. Yu. DiffPuter: An EM-driven diffusion model for missing data imputation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Precise Regularity Conditions

(A1) **Positivity:** $P(R = 1_d \mid X) > 0$ almost surely.

- For tree graphs, $P(R = \mathsf{PA}(r) \mid X_r) > 0$ on the relevant target support for every $r \in \mathcal{R} = \{r \neq 1_d : P(R = r) > 0\}$.

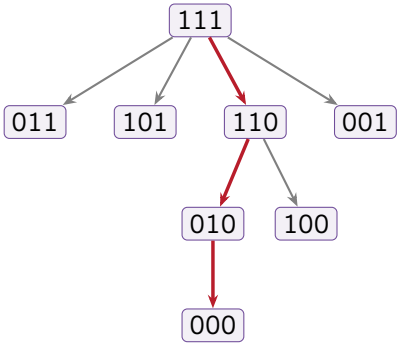
(A2) **Finite moments:** $\mathbb{E}\|X\| < \infty$ and $\mathbb{E}\|f(x_r, \epsilon, r)\| < \infty$ for $f \in \mathcal{F}$, so that the energy score is finite and strictly proper (Gneiting and Raftery, 2007; Székely and Rizzo, 2013).

(A3) **Expressiveness:** $\mathcal{F}$ contains a generator whose induced distribution equals the donor distribution for every $r \in \mathcal{R}$.

(A4) **Identification:** the chosen identifying assumption holds. Otherwise, f^* still recovers the donor distribution, but not the true extrapolation density.

▸ Back to the proof sketch

Beyond CCMV: Tree Graphs



An example tree graph
arrows: parent → child it informs; red: path of 000

- Each $r \in \mathcal{R}$ has a parent $\mathsf{PA}(r) > r$ in a tree graph rooted at 1_d (Chen, 2022; Suen and Chen, 2026).
  - CCMV: $\mathsf{PA}(r) \equiv 1_d$ for every $r \in \mathcal{R}$.
- Identifying assumption:

$$p(x_{\bar{r}} \mid x_r, R = r) \overset{\text{Tree graph}}{=} p(x_{\bar{r}} \mid x_r, R = \mathsf{PA}(r)).$$

- Along the path $r = \gamma_0 \leftarrow \gamma_1 \leftarrow \cdots \leftarrow \gamma_K = 1_d$:

$$p(x_{\bar{r}} \mid x_r, r) = \prod_{k=0}^{K-1} \underbrace{p(x_{\gamma_{k+1}-\gamma_k} \mid x_{\gamma_k}, R = \gamma_{k+1})}_{\text{identifiable}}.$$

- Learn one-step laws from the parent's units:

$$M_r(\mathbf{R}_i) = \mathsf{PA}(r) - r, \quad S_r(\mathbf{R}_i) = \frac{I(\mathbf{R}_i = \mathsf{PA}(r))}{|\mathsf{PA}(r) - r|}.$$

- Imputation path: $000 \to 010 \to 110 \to 111$.

- **MCAR:** keep 20% of rows complete; in the others, each entry is missing with probability ρ ($\rho = 0.25$ or 0.5).
- **CCMV** via a pattern graph (Chen, 2022): selection odds

$$\frac{p(R = r \mid X)}{p(R = 1_d \mid X)} = \exp(\alpha_r + \beta_r^\top X_r), \qquad \beta_r = \mathbf{1};\ \alpha_r = -4\ (20\%)\ \text{ or }\ -0.5\ (40\%).$$

- **MAR:** sequential construction from Yoon et al. (2018), with 20% complete rows.
- **Energy distance:**

$$\widehat{\mathrm{ED}}(P, Q) = \tfrac{2}{n_X n_Y} \sum_{i,j} \|X_i - Y_j\| - \tfrac{1}{n_X^2} \sum_{i,j} \|X_i - X_j\| - \tfrac{1}{n_Y^2} \sum_{i,j} \|Y_i - Y_j\|, \quad X_i \sim P, Y_i \sim Q.$$

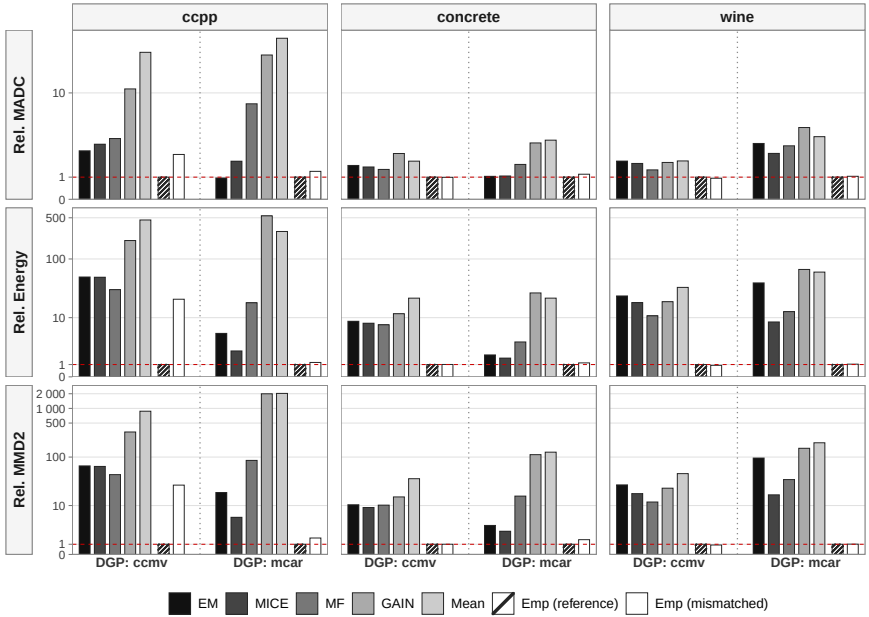
- **MADC** (mean absolute difference of pairwise correlations):

$$\mathrm{MADC} = \tfrac{2}{d(d-1)} \sum_{j<l} \left| C_{jl}^{\text{true}} - \bar{C}_{jl}^{\text{imp}} \right|, \quad \bar{C}^{\text{imp}} = \tfrac{1}{K} \textstyle\sum_{k=1}^{K} C_k^{\text{imp}} \text{ (average over imputations)}.$$

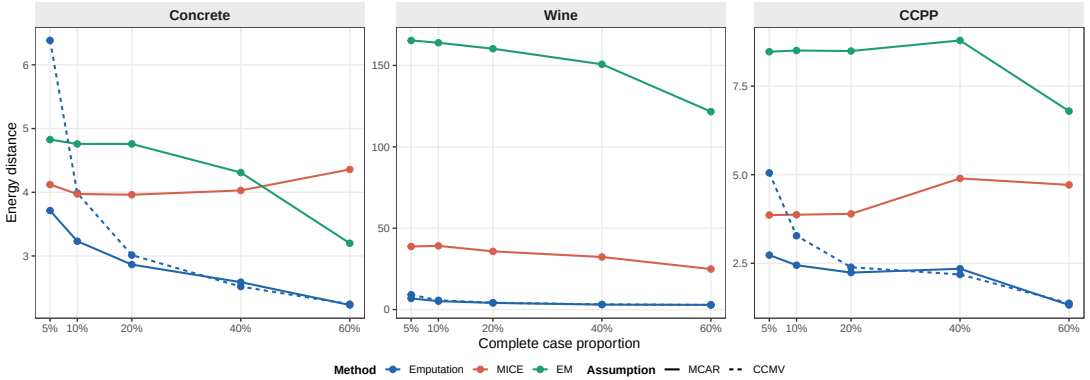
- **MMD^2** with a Gaussian kernel:

$$\widehat{\mathrm{MMD}}^2(P, Q) = \tfrac{1}{n_X(n_X-1)} \sum_{i \neq j} k(X_i, X_j) + \tfrac{1}{n_Y(n_Y-1)} \sum_{i \neq j} k(Y_i, Y_j) - \tfrac{2}{n_X n_Y} \sum_{i,j} k(X_i, Y_j).$$

Distributional Evaluation at ≈40% Missingness

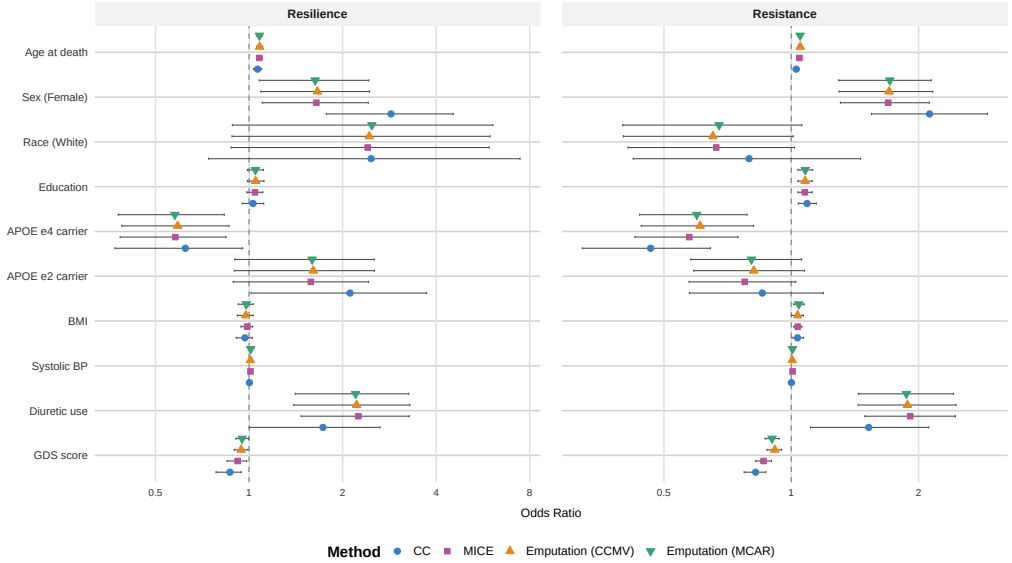


How Many Complete Cases Does Emputation Need?



- Analyze under MAR, with complete-case proportions varied from 5% to 60%.
- Emputation (MCAR or CCMV risk) generally beats MICE and EM.
- A complete-case proportion of $\geq 20\%$ is a useful guide for CCMV emputation in these simulations.

NACC Results for All Predictors



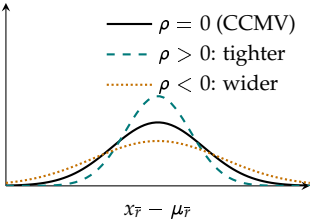
Full resilience and resistance models: 500 bootstrap replicates $\times$ 10 imputations. (back)

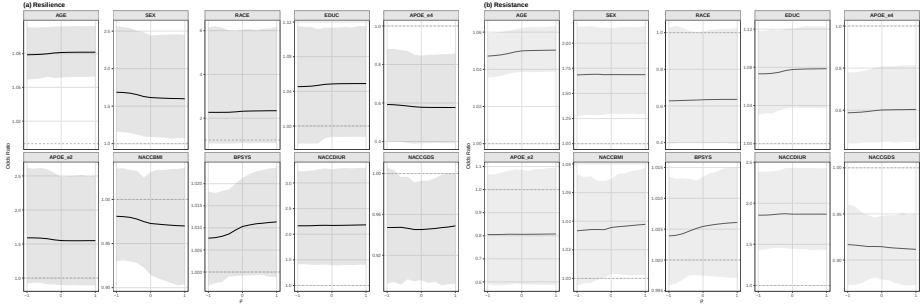
NACC Data: Sensitivity to the Missing Data Assumption

- MCAR and CCMV emputations give similar odds ratios.
- **Departure from CCMV** by exponential tilting:

$$p_\rho(x_{\bar{r}} \mid x_r, r) \propto p(x_{\bar{r}} \mid x_r, R = 1_d) \exp\left(-\rho \left|\left|x_{\bar{r}} - \mu_{\bar{r}}\right|\right|^2\right),$$

  - $\rho = 0$ is CCMV and $\mu_{\bar{r}}$ is the CCMV conditional mean.





- **Stable:** age, sex, APOE $\varepsilon 4$, and diuretic use keep their direction and significance.
- **More sensitive:** BMI (resistance) and GDS (resilience).