Relaxing Observability Assumption in Causal Inference with Kernel Methods

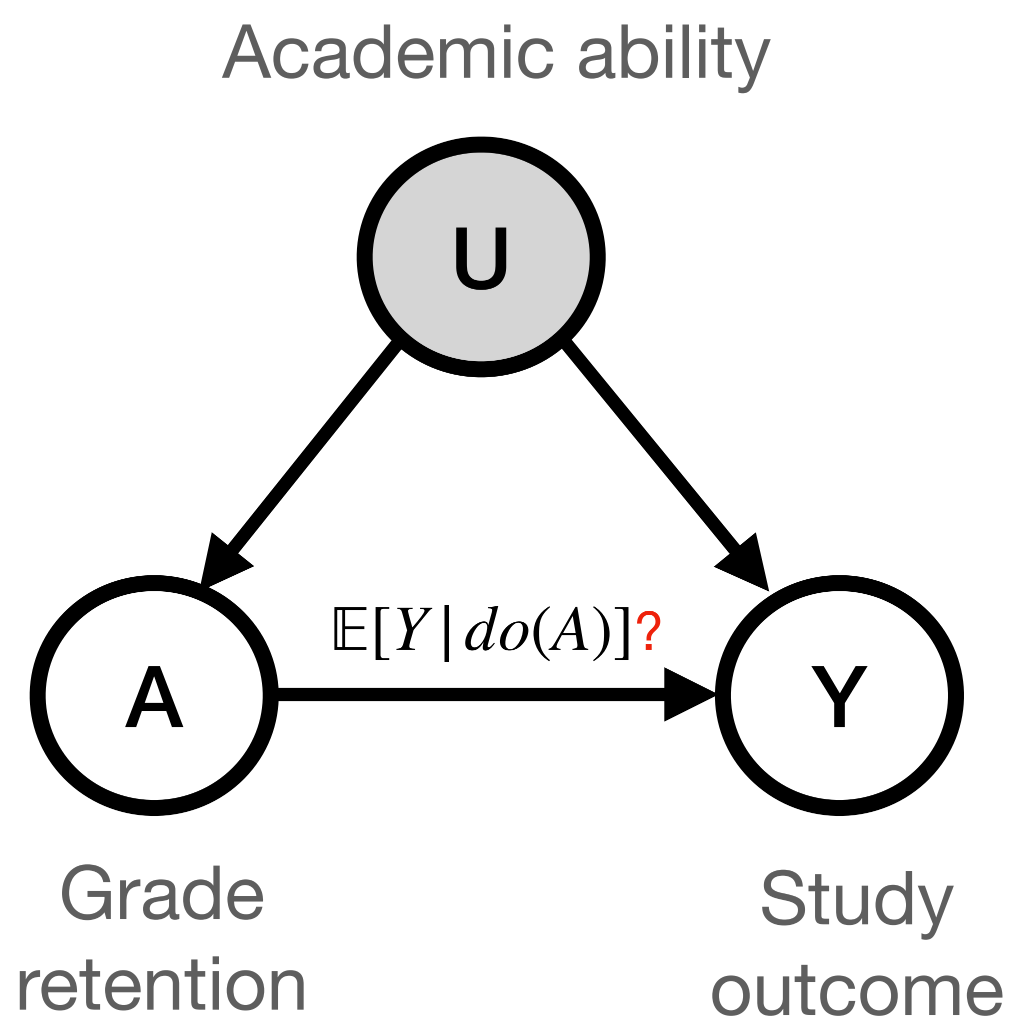
Yuchen Zhu

with Limor Gultchin, Arthur Gretton, Anna Korba, Matt Kusner, Afsaneh Mastouri, Krikamol Muandet, Ricardo Silva

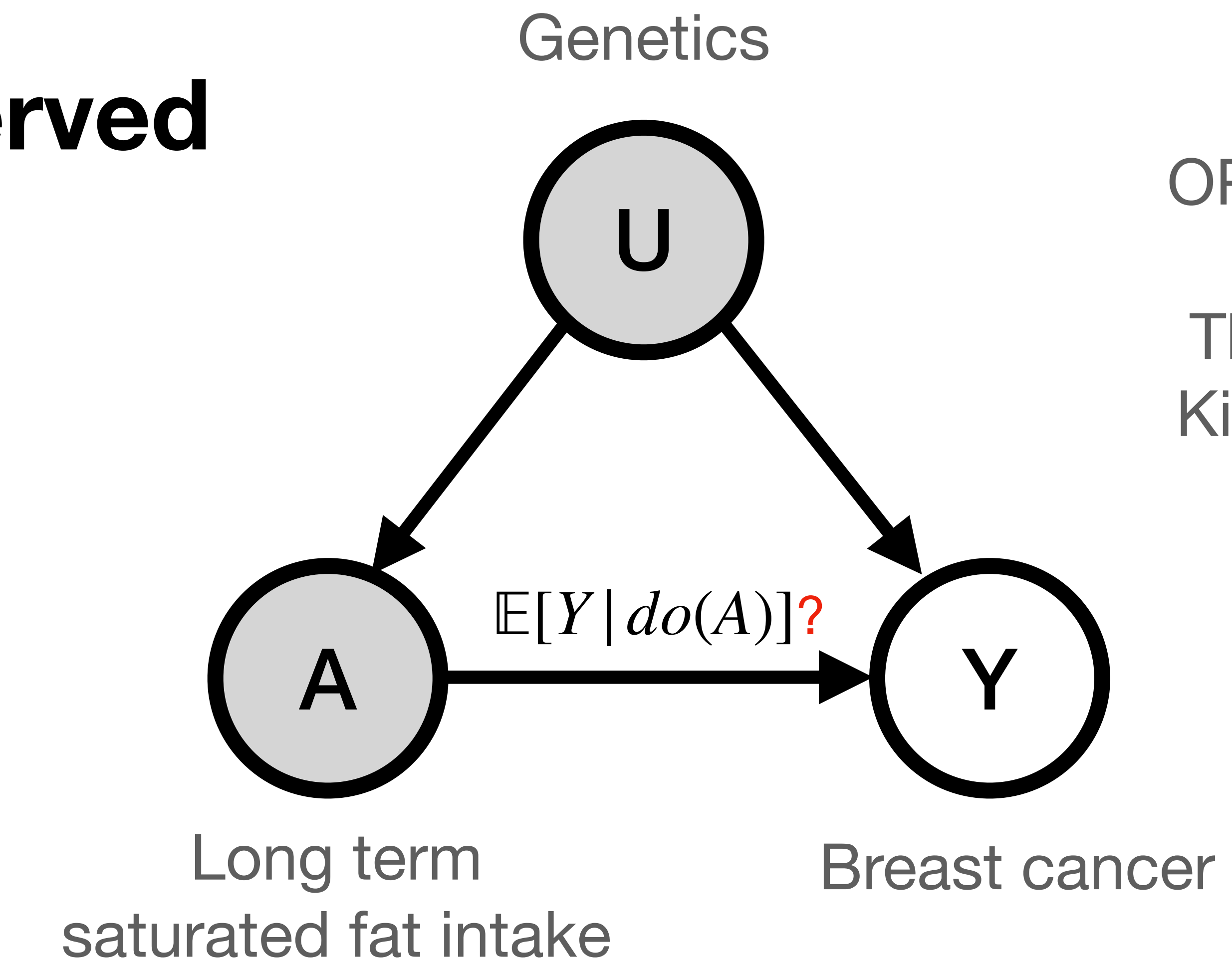
Talk at *When Causal Inference Meets Statistics* Quarterly, 20.04.2023

Why relax observability assumptions?

Unobserved confounders:

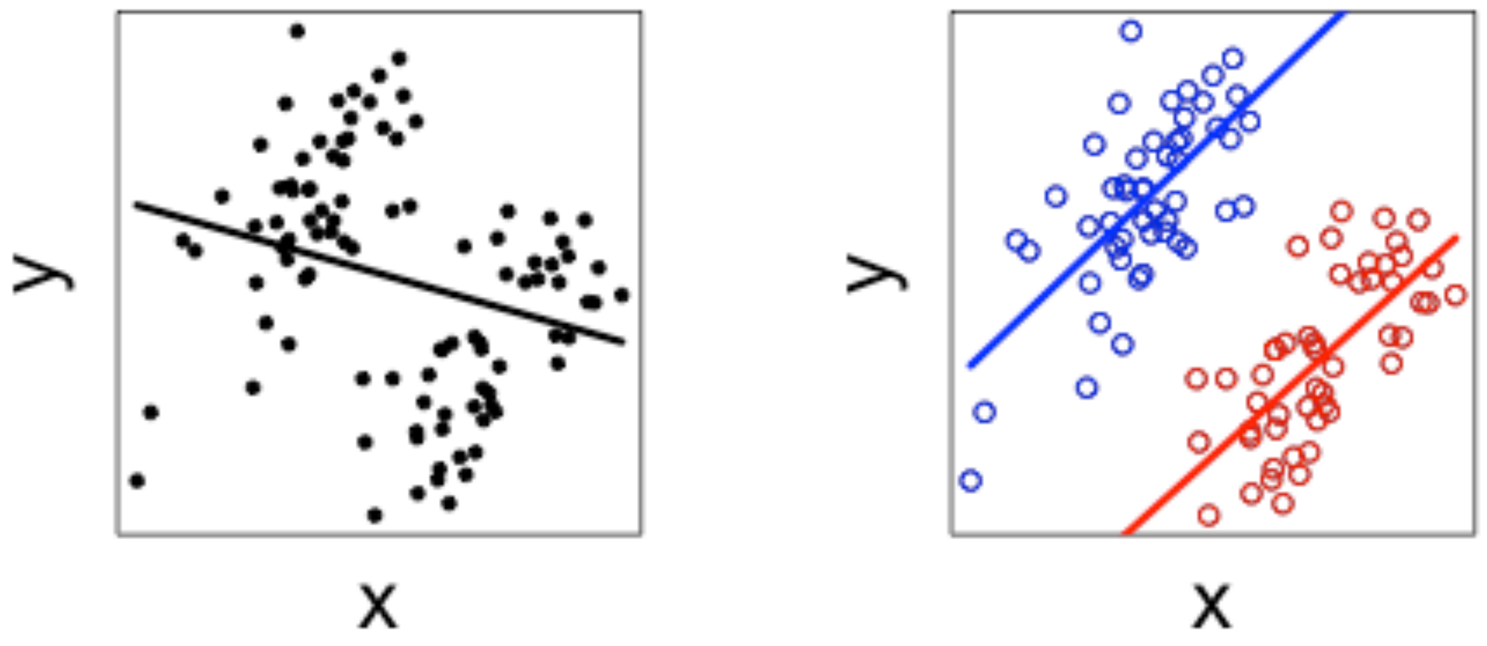


Action observed with error:

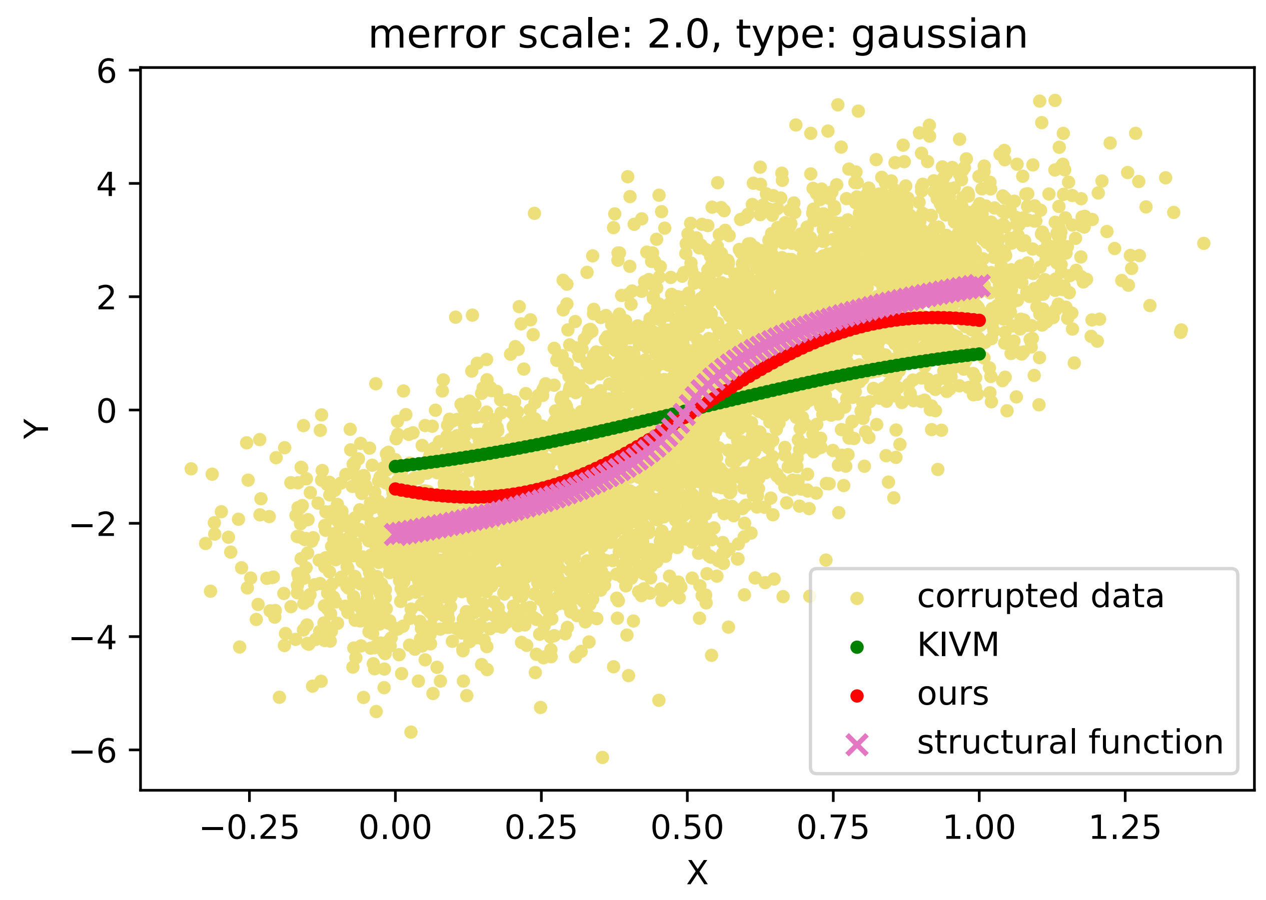


OPEN study: Subar, Thompson, Kipnis, et al. 2001

Simpson's paradox:



Mask interesting relationships:



Kernel Mean Embeddings

$$\mu_{P_X}(x) = \int k(x, y) P_X(y) dy$$

Injective

Characteristic kernel: $P_X \longmapsto \mu_{P_X}(y)$

$$\langle \mu_{P_X}, f \rangle_{H_X} = \mathbb{E}_{P_X}[f(X)]$$

Conditional Kernel Mean Embeddings (CME)

$$\mu_{W|a,x,z} := C_{W|A,X,Z}\left(\phi(a)\otimes\phi(x)\otimes\phi(z)\right)$$

$$\widehat{C}_{W|A,X,Z} = \underset{C\in\mathcal{H}_\Gamma}{\operatorname{argmin}}\ \widehat{E}(C),\ \text{with}$$

$$\widehat{E}(C) = \frac{1}{m}\sum_{i=1}^{m}\|\phi(w_i) - C\phi(a_i, x_i, z_i)\|^2_{\mathcal{H}_\mathcal{W}} + \lambda\ \|C\|^2_{\mathcal{H}_\Gamma}$$

$$\widehat{C}_{W|A,X,Z} = \Phi(W)(\mathcal{K}_{AXZ} + m\ \lambda\ )^{-1}\Phi^T(A,X,Z)$$

Convergence rates are well understood (Singh et al 2019, Mastouri, Zhu, et al 2021)

Connection with Characteristic Functions

Translation invariant: $k(x, y) = k(x - y)$

$$\mu(x) = \int k(x - y)p(y)dy$$

$$\hat{\mu}[\alpha] = \hat{k}[\alpha]\psi[\alpha]$$

Bochner's theorem: $\hat{k}$ is a probability measure.

Connection with Characteristic Functions

KRR estimate of CME:

$$\hat{\mu}^{(s)}_{X|z}(x) = \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z) k(x_j, x)$$

$$\hat{\gamma}^{(s)}_j(z) = (K_Z + s\lambda I)^{-1} K_{Zz}$$

Fourier transform:

$$\tilde{\hat{\mu}}^{(s)}_{X|z}(\alpha) = \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z) e^{-i\alpha x_j} \tilde{k}(\alpha)$$

$$= \tilde{k}(\alpha) \underbrace{\sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z) e^{-j\alpha x_j}}_{=:\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(-\alpha)}$$

Connection with Characteristic Functions

$(x_j, z_j)_{j=1}^{s}$ ⟶ Have $\hat{\mu}^n_{X|z}(y) = \sum_{j=1}^{n} \hat{\gamma}^n_j(z) k(x_j, y)$.

Let $\hat{\psi}^n_{X|z}(\alpha) := \sum_{j=1}^{n} \hat{\gamma}^n_j(z) e^{i\alpha x_j}$.

Where $\hat{\gamma}^n_j(z) = (K_{ZZ} + n\hat{\lambda}^n I)^{-1} K_{Zz}$.

Theorem 1. With real, translation-invariant kernel:

$\hat{\mu}^n_{X|Z} \to^n \mu_{X|Z}$ iff $\hat{\psi}^n_{X|Z} \to^n \psi_{X|Z}$ in IFT of kernel.

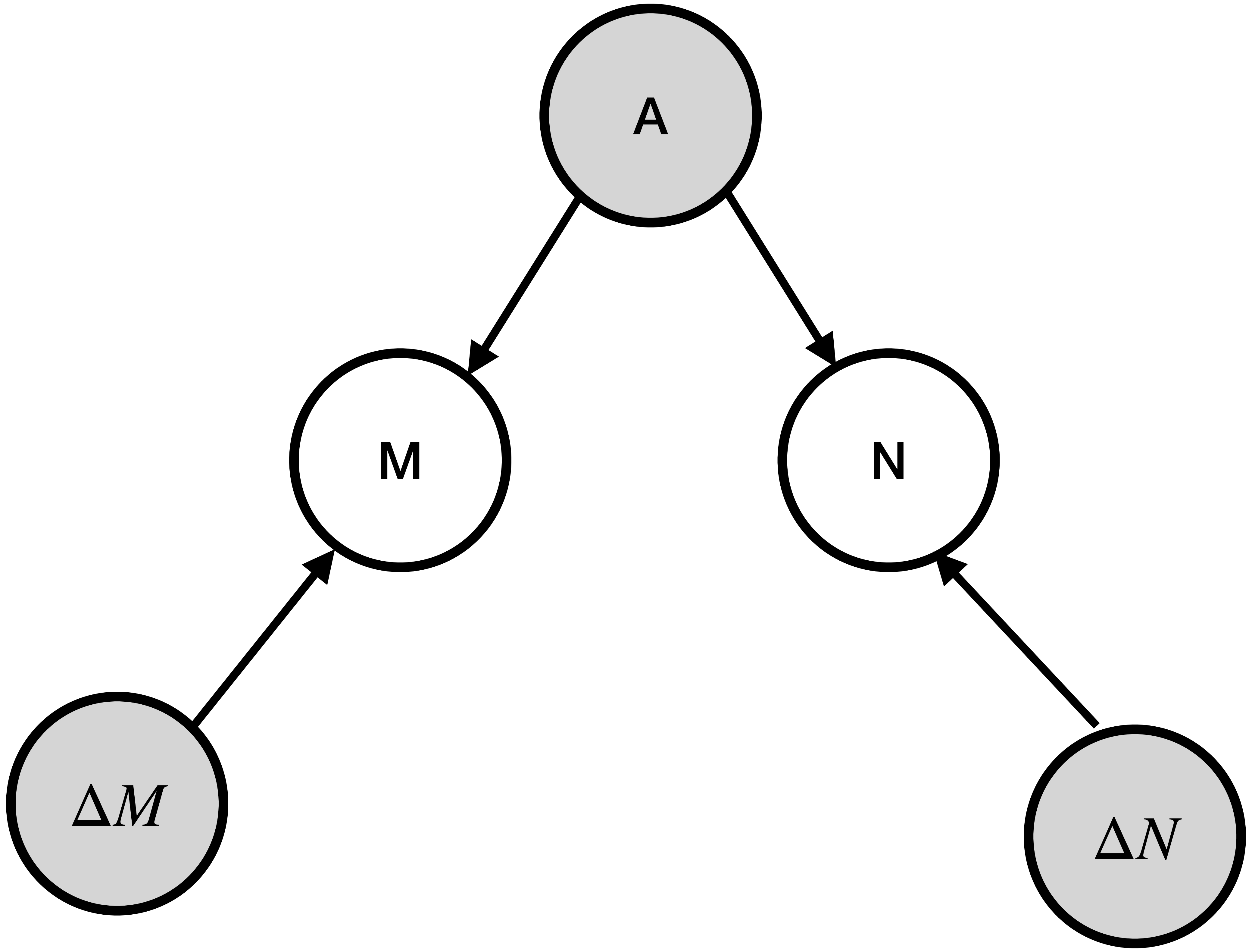
Kotlarski's Lemma

LEMMA 1. *Let* X_1, X_2, X_3 *be three independent real random variables, and let*

$$Z_1 = X_1 - X_3, \; Z_2 = X_2 - X_3 \; .$$

If the characteristic function of the pair (Z_1, Z_2) *does not vanish, then the distribution of* (Z_1, Z_2) *determines the distributions of* X_1, X_2, X_3 *up to a change of the location.*

Kotlarski's Lemma

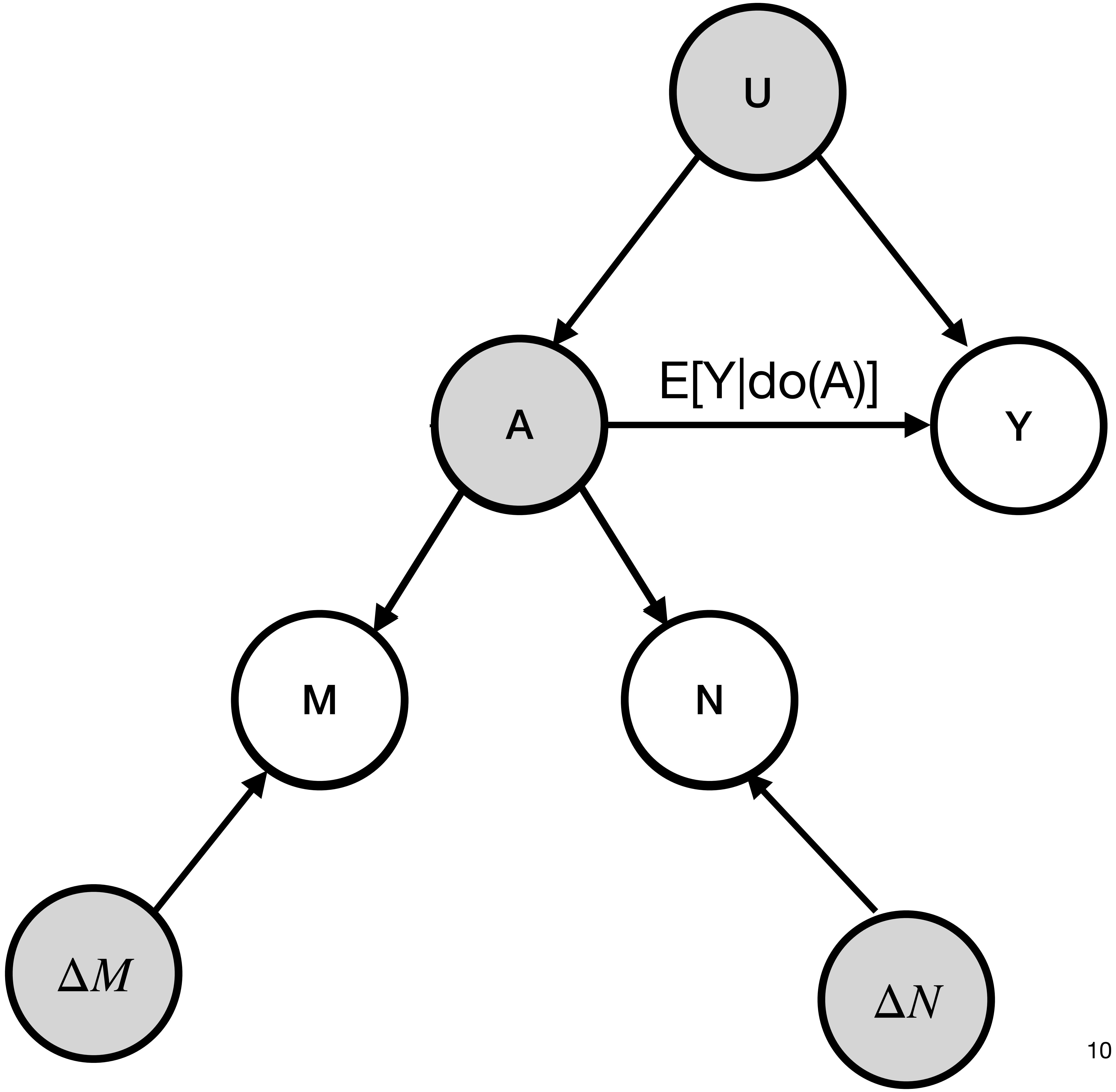


$$M = A + \Delta M$$

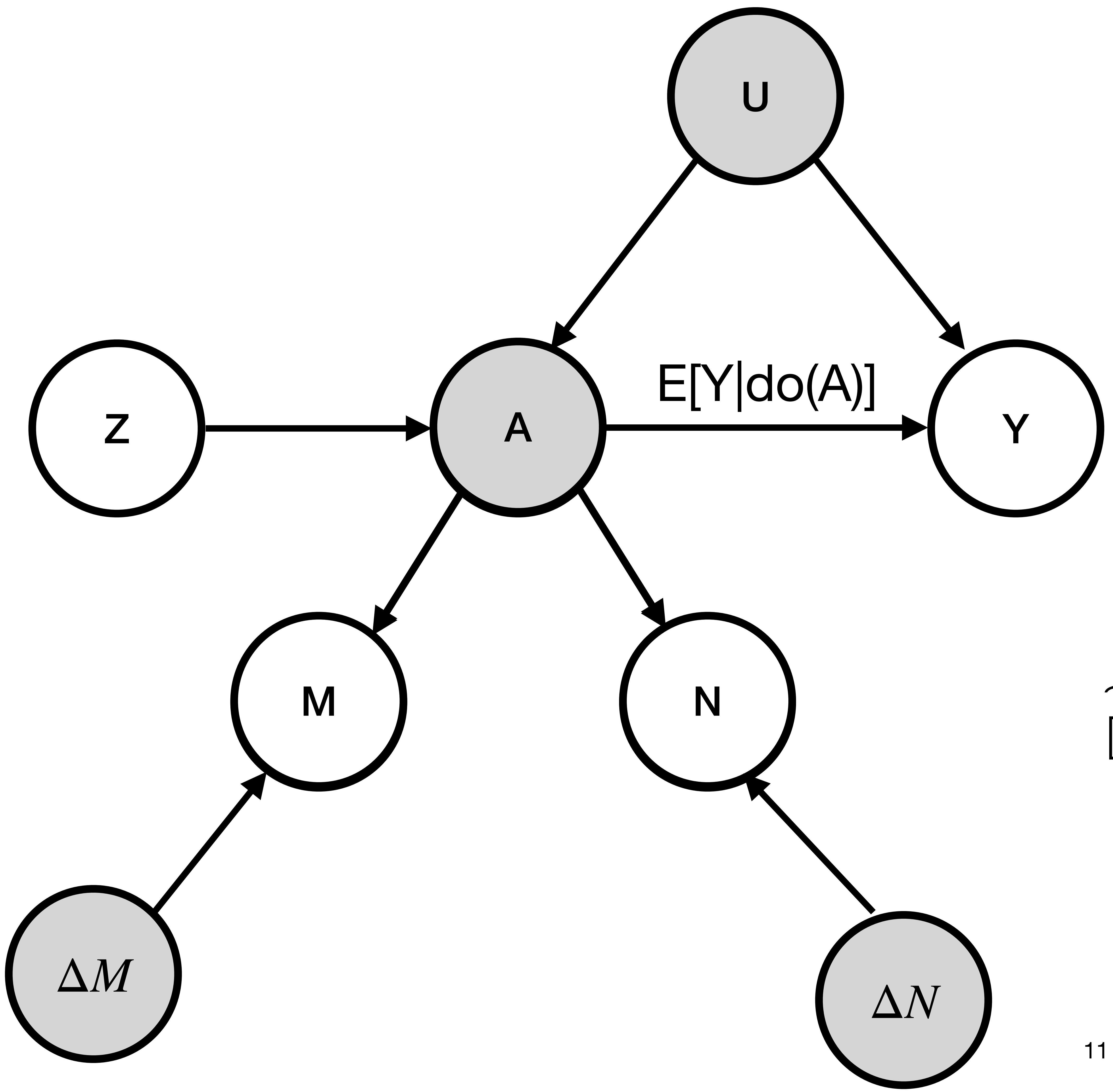
$$N = A + \Delta N$$

$$\overbrace{\mathbb{E}_{\mathscr{P}_A}\left[e^{i\alpha A}\right]}^{\psi_{\mathscr{P}_A}(\alpha):=} = \exp\left(\int_0^\alpha i\frac{\mathbb{E}\left[Me^{i\nu N}\right]}{\mathbb{E}\left[e^{i\nu N}\right]}d\nu\right)$$

Application in causal inference with corrupted treatments



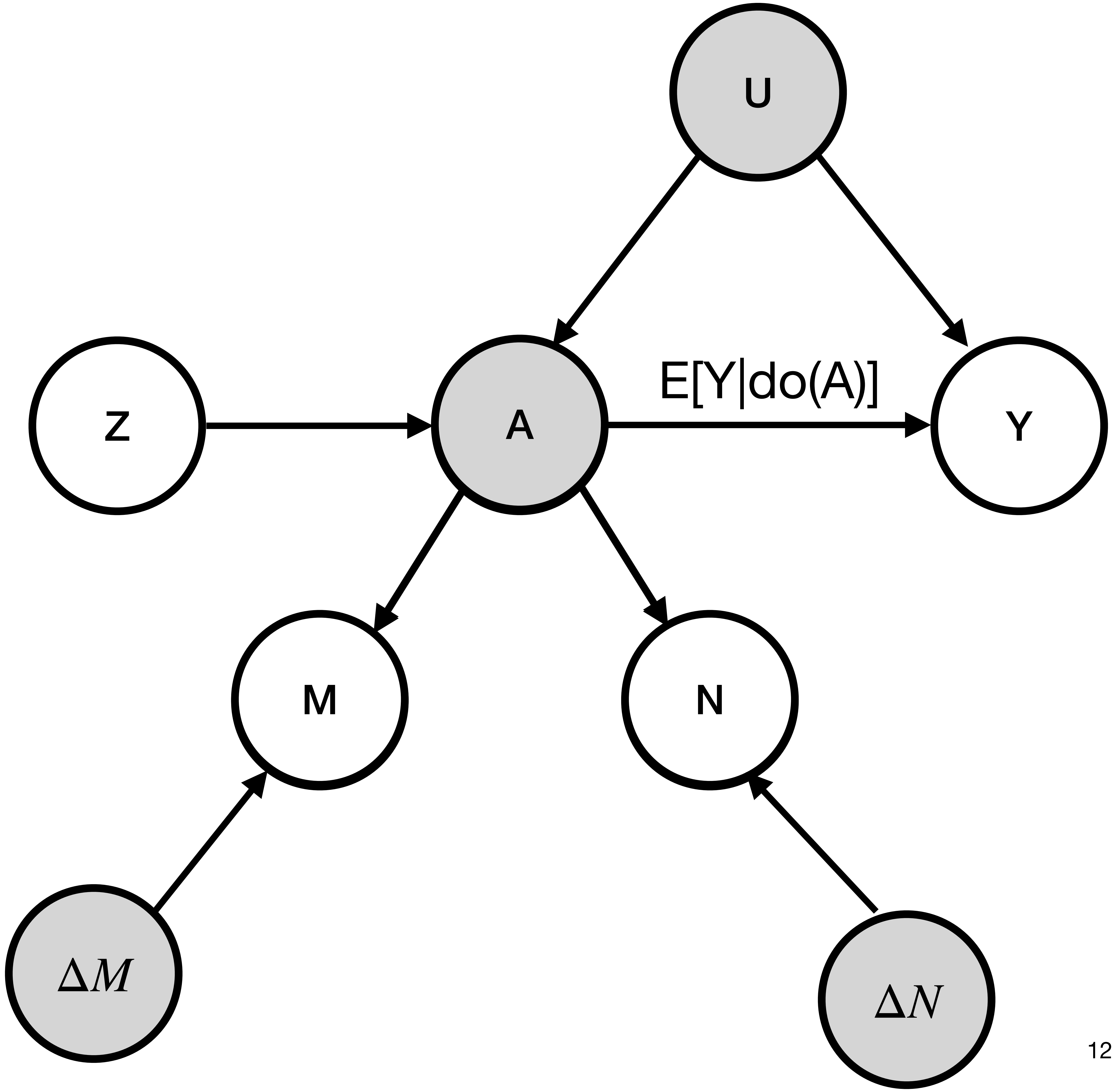
Application in causal inference with corrupted treatments



How to compute the right hand side?

$$\underbrace{\mathbb{E}_{\mathscr{P}_{A|z}}\left[e^{i\alpha A} \,|\, z\right]}_{\psi_{\mathscr{P}_{A|z}}(\alpha):=} = \exp\left(\int_0^\alpha i \frac{\mathbb{E}\left[M e^{i\nu N} \,|\, z\right]}{\mathbb{E}\left[e^{i\nu N} \,|\, z\right]} d\nu \right)$$

Application in causal inference with corrupted treatments



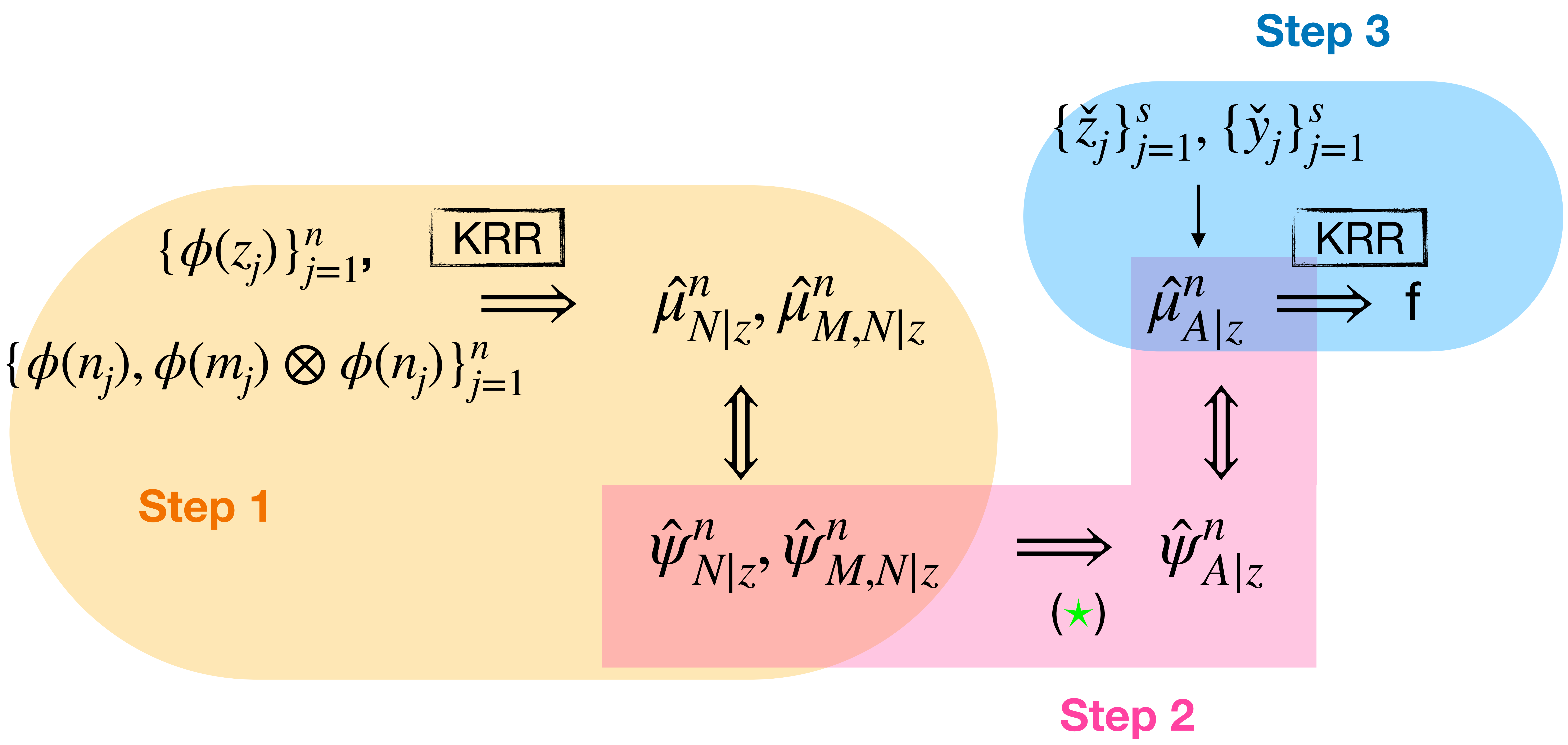
To obtain $\hat{\psi}^n_{A|z}$:

$$\overbrace{\mathbb{E}_{\mathscr{P}_{A|z}}[e^{i\alpha X}](\alpha)}^{\psi_{A|z}(\alpha)} = \exp\left(\int_0^\alpha i\frac{\overbrace{\mathbb{E}[Me^{i\nu N}\,|\,z]}^{\frac{\partial}{\partial v}\psi_{M,N|z}(v,\nu)\big|_{v=0}}}{\underbrace{\mathbb{E}[e^{i\nu N}\,|\,z]}_{\psi_{N|z}(\nu)}}d\nu\right) \quad (1)$$

1. **Differentiate wrt α to remove integral.**
2. **Replace with sample estimates.**

$$\frac{\frac{d}{d\alpha}\hat{\psi}^n_{A|z}(\alpha)}{\hat{\psi}^n_{A|z}(\alpha)} = \frac{\frac{\partial}{\partial v}\hat{\psi}^n_{M,N|z}(v,\alpha)\Big|_{v=0}}{\hat{\psi}^n_{N|z}(\alpha)} \quad (2)$$

Measurement Error KIV (MEKIV)



Zhu et al, UAI 2022, Causal Inference with Treatment Measurement Error: A Nonparametric IV Approach.

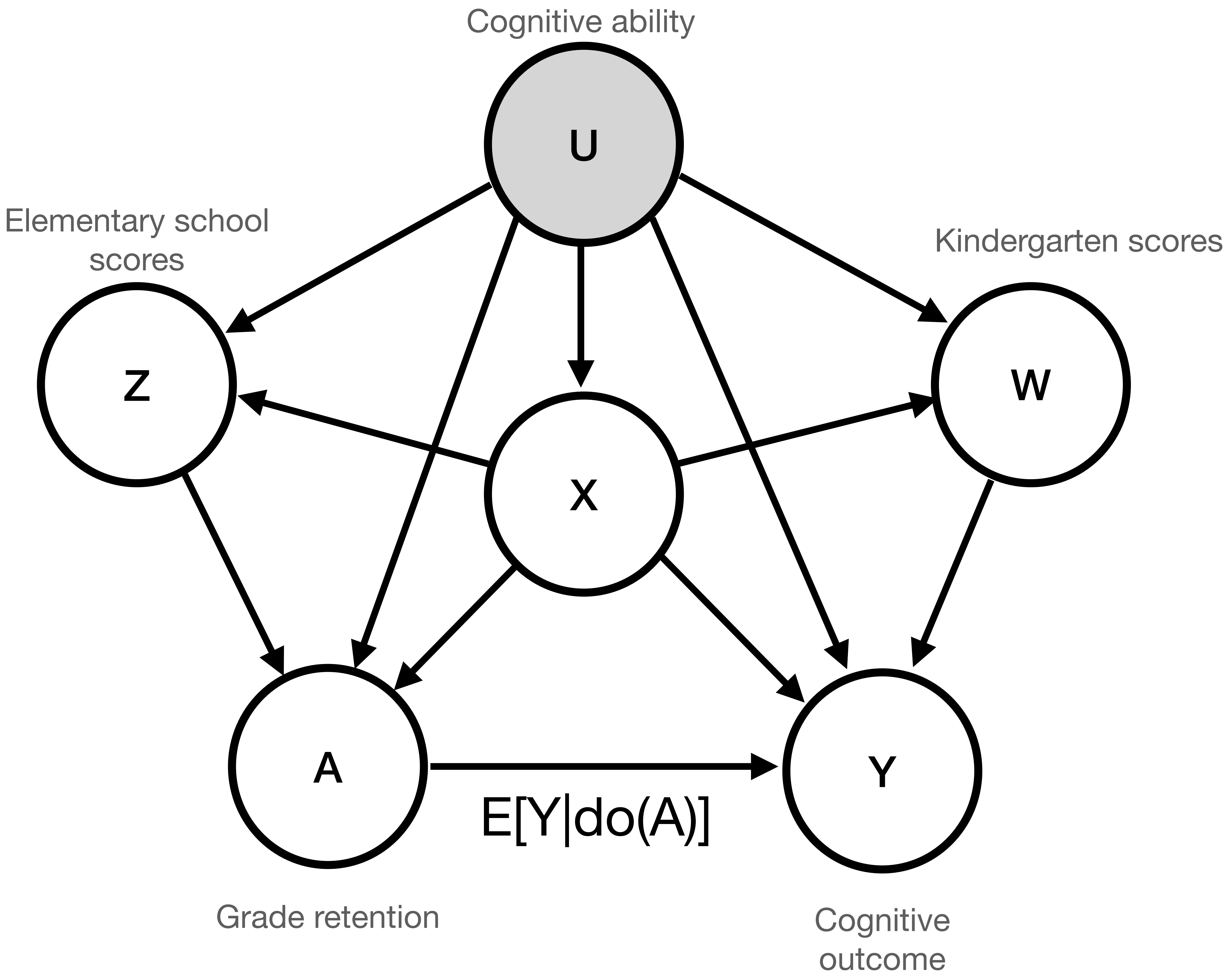
Advantages of MEKIV

- No distributional assumptions. Further relaxation: Evdokimov and White 2011.
- Very little hyper parameter tuning.
- Models the distributions using mean embeddings and not the full densities.

Summary of techniques and future work

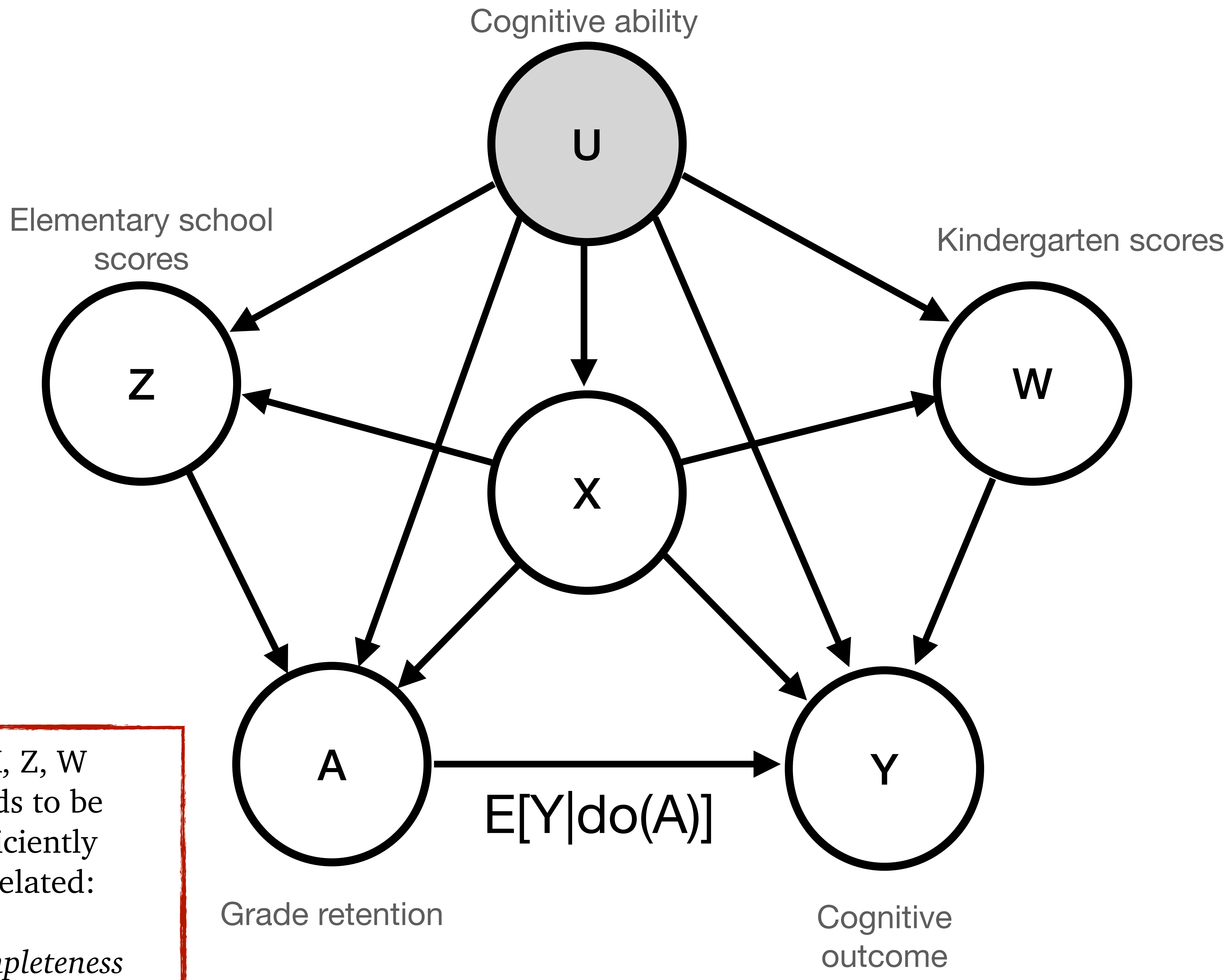
- Kotlarski's Lemma allows us to identify three unseen variables from just two of their linear combinations. Can this be extended further?
- Duality between characteristic functions and mean embeddings.
- Need to relax the additive measurement error assumption.
- Need to relax additive error on outcome assumption.

Proximal Causal Learning Background



Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Proximal Causal Learning Background



U, X, Z, W needs to be sufficiently correlated:

Completeness Condition (Miao et al. 2018)

Average causal effect estimation:

$$\mathbb{E}[Y \,|\, do(A = a)] = \int_{XW} h(a, w, x) p(w, x) dx dw$$

How to get h?

$$\Downarrow$$

$$\mathbb{E}[Y - h(A, W, X) \,|\, A, Z, X] = 0 \quad \text{a.s. } P_{AZX}$$

Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Proximal Maximum Moment Restriction

$$\mathbb{E}[Y - h(A, X, W) \mid A, X, Z] = 0 \quad \text{a.s. } P_{AXZ}$$

CMR

$\Downarrow$

- If E[A|B] = 0,
- Then (for g measurable):
- E[Ag(B)] = E[E[Ag(B)|B]]
- = E[E[A|B]g(B)] = 0

$$\mathbb{E}[(Y - h(A, X, W))g(A, X, Z)] = 0 \quad \textbf{a.s. } P_{AXZ}$$

For all g

Precursor loss:

$$R(h) = \sup_g (\mathbb{E}[(Y - h(A, W, X))g(A, Z, X)])^2$$

$\Updownarrow$

- Restrict g to $\mathscr{H}_{\mathscr{AXZ}}$

PMMR surrogate loss $R_k(h)$ k indexes the kernel.

Mastouri*, **Z.***, et al. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restrictions. *ICML 2021.*

Proximal Maximum Moment Restriction

Precursor loss:

$$R(h) = \sup_g (\mathbb{E}[(Y - h(A, W, X))g(A, Z, X)])^2$$

$$\Downarrow$$

- Restrict g to $\mathscr{H}_{\mathscr{A}\mathscr{X}\mathscr{Z}}$

$$R_k(h) = \sup_{g \in \mathscr{H}_{\mathscr{A}\mathscr{Z}\mathscr{X}}, \quad \|g\| \leq 1} (\mathbb{E}[(Y - h(A, W, X))\langle g, k((A, Z, X), \cdot)\rangle])^2$$

$$= \mathbb{E}[(Y - h(A, W, X))(Y' - h(A', W', X'))k((A, Z, X), (A', Z', X'))]$$

V-statistic: $R_V(h) := \frac{1}{n^2} \sum_{i,j=1}^{n} (y_i - h_i)(y_j - h_j)k_{ij}$ (reweighed ERM!)

Mastouri*, **Z.***, et al. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restrictions. *ICML 2021.*