On estimation of functional causal models: Post-nonlinear causal model as an example

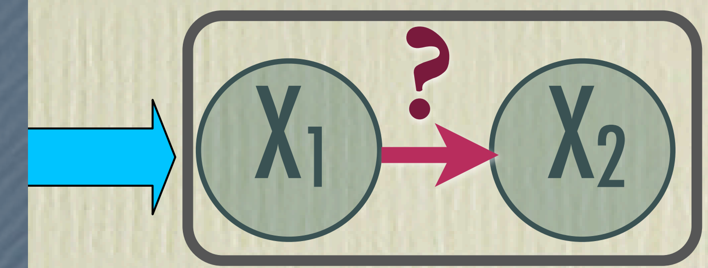
Kun Zhang, Zhikun Wang, Bernhard Schölkopf

Dept. Empirical Inference
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Causal discovery

- Causal discovery: identify causal relations from purely observed data
- In the past decades, under certain assumptions, it was made possible to derive causation from passively observed data
  - statistical data ⇒ causal structure
  - causal Markov assumption
  - faithfulness…

X_1	X_2
-1.1	1.0
2.1	2.0
3.1	4.2
2.3	-0.6
1.3	2.2
-1.8	0.9
.	.
.	.
.	.



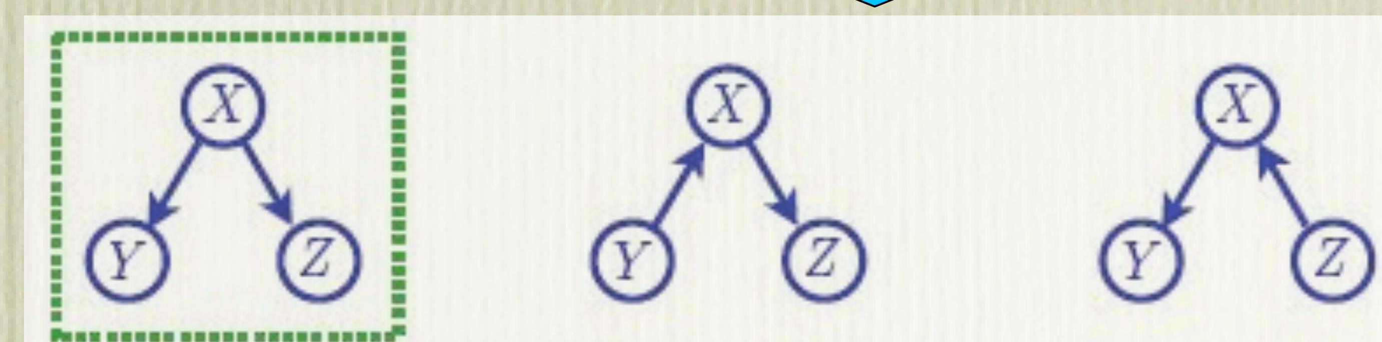
Constraint-based causal discovery

- under Markov condition & faithfulness assumption, uses (conditional) independence <u>constraints</u> to find candidate causal structures

- example: PC algorithm (Spirtes & Glymour, 1991)

$\boxed{Y \perp\!\!\!\perp Z \mid X}$

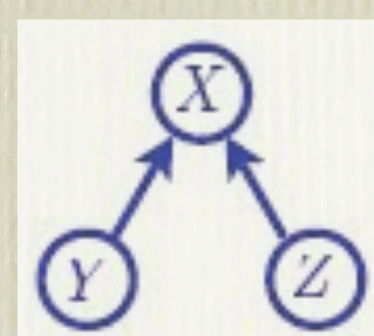
- Markov equivalence class



- pattern Y—X—Z
  - same adjacencies
  - → if all agree on orientation; — if disagree

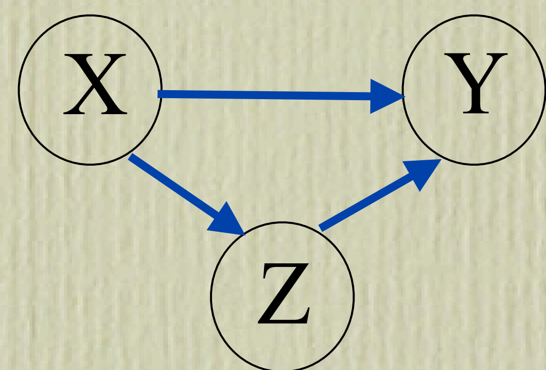
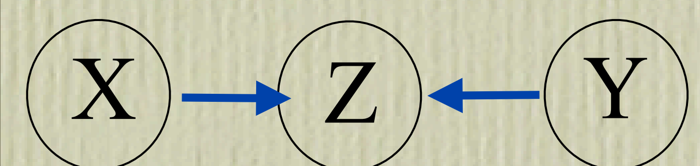
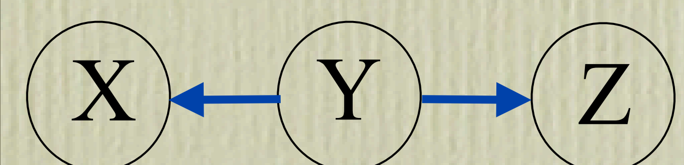
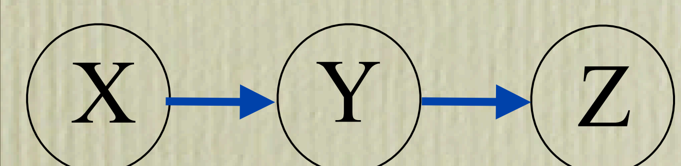
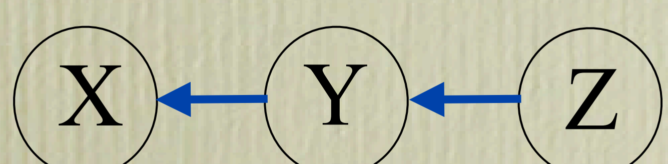
- might be unique: v-structure

$\boxed{Y \perp\!\!\!\perp Z}$



Constraint-based method: An inverse problem

- {local causal structures} → {conditional independences}

	$\varnothing$
	$X \perp\!\!\!\perp Y$
	$X \perp\!\!\!\perp Y$
	
	$X \perp\!\!\!\perp Z \mid Y$
	

Constraint-based method: An inverse problem

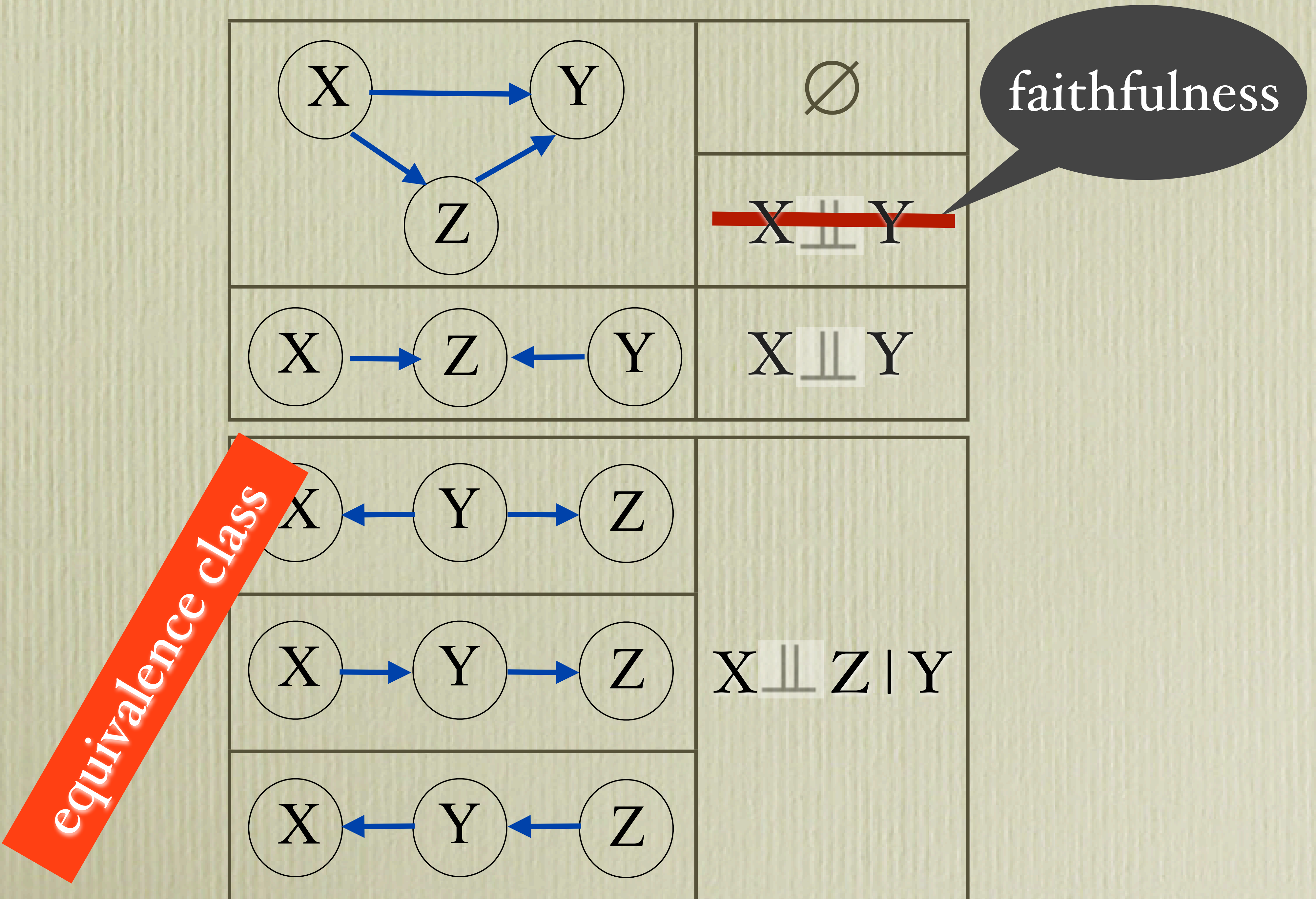
- {local causal structures} → {conditional independences}

$X \rightarrow Y$, $X \rightarrow Z$, $Z \rightarrow Y$	$\varnothing$
	$\cancel{X \perp\!\!\!\perp Y}$ faithfulness
$X \rightarrow Z \leftarrow Y$	$X \perp\!\!\!\perp Y$

$X \leftarrow Y \rightarrow Z$	$X \perp\!\!\!\perp Z \mid Y$
$X \rightarrow Y \rightarrow Z$	
$X \leftarrow Y \leftarrow Z$	

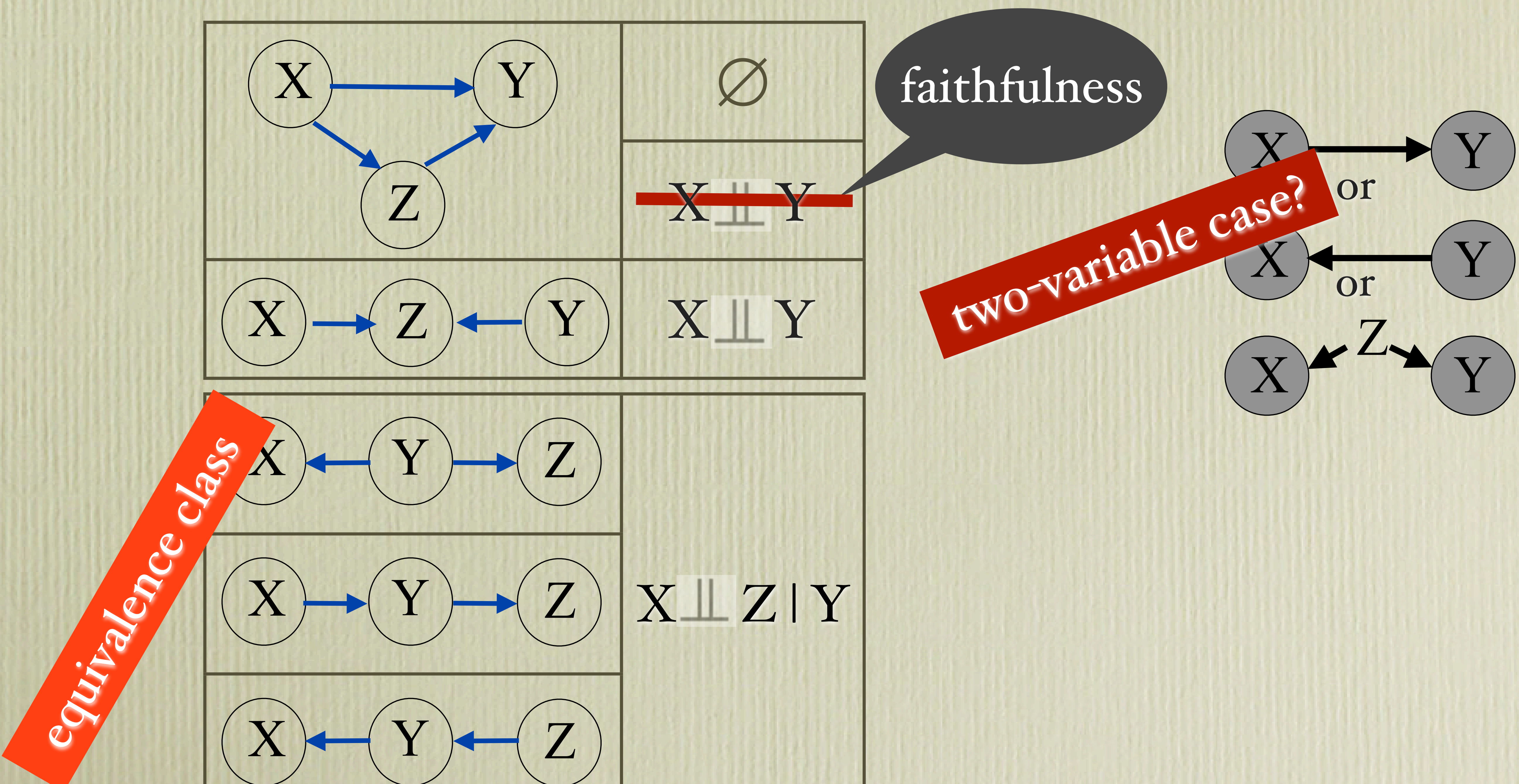
Constraint-based method: An inverse problem

- {local causal structures} → {conditional independences}



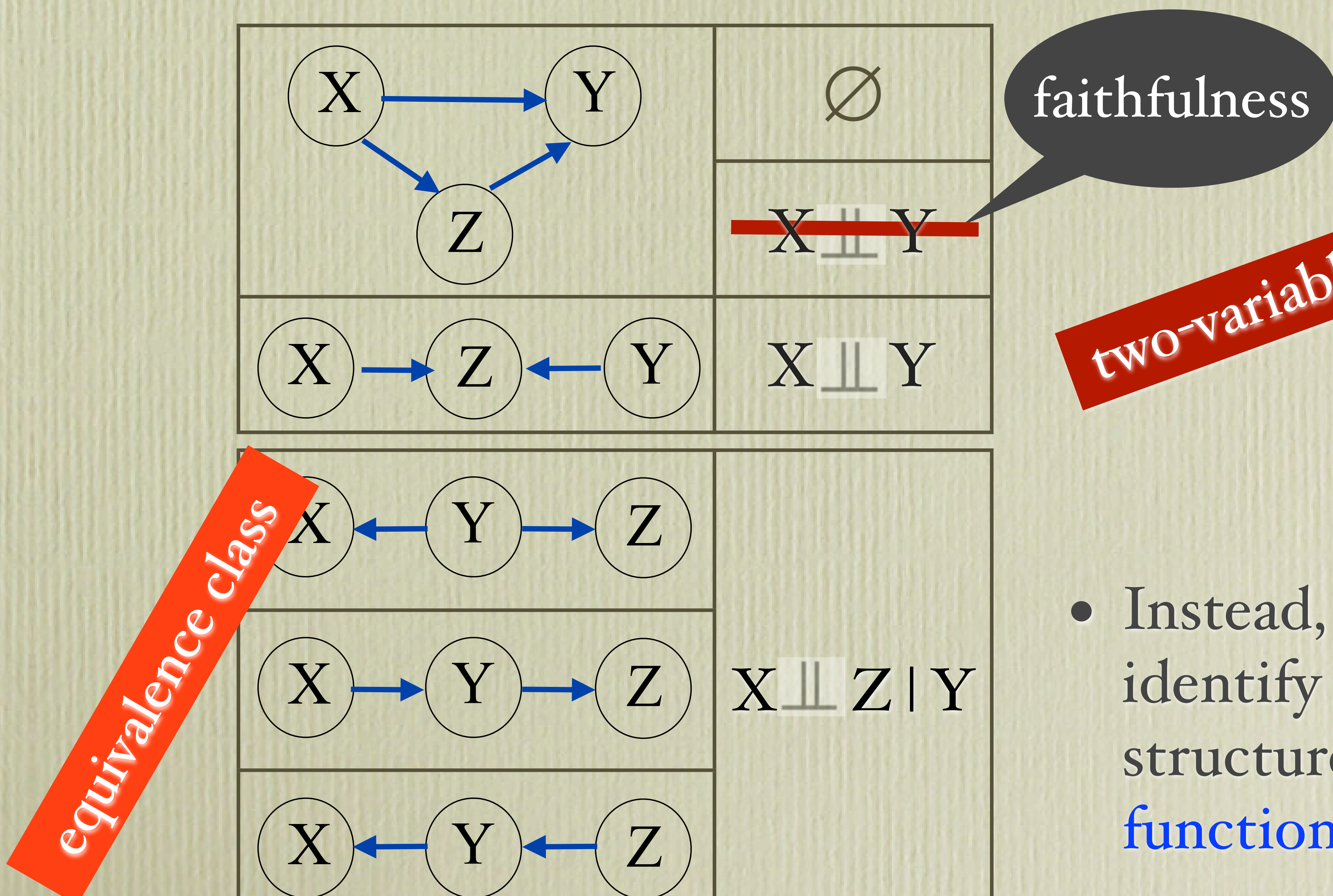
Constraint-based method: An inverse problem

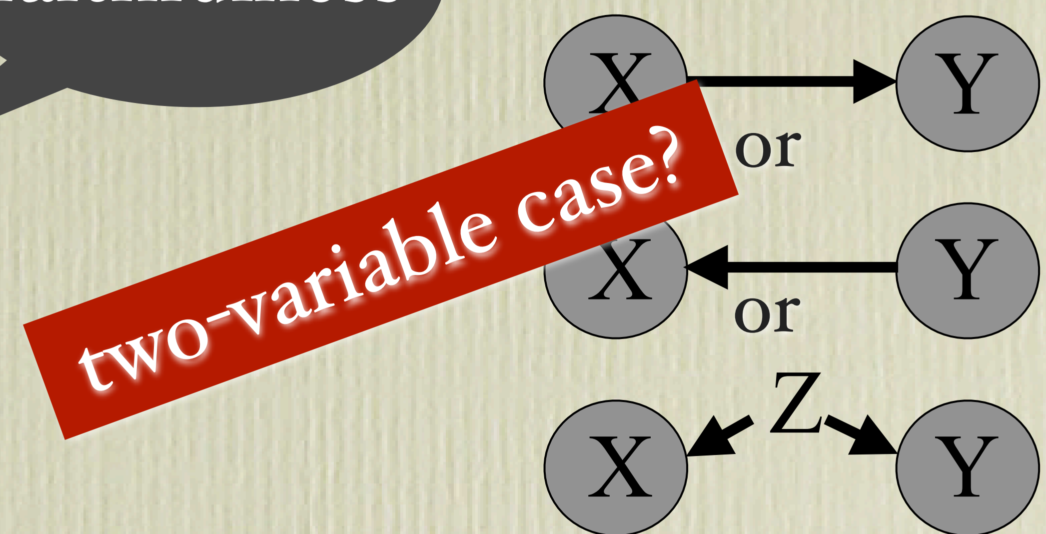
- {local causal structures} → {conditional independences}



Constraint-based method: An inverse problem

- {local causal structures} → {conditional independences}



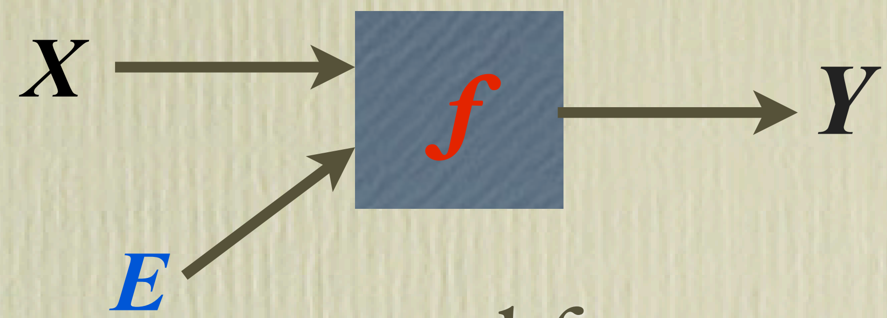


- Instead, try to directly identify local causal structures with functional causal models

Outline

- Causal discovery based on functional causal models (FCMs)
- Estimation of FCMs: Relationship between dependence minimization and maximum likelihood
- Post-nonlinear causal model as an example: By warped Gaussian processes with flexible noise distribution

Causality is about data-generating process



- Functional causal model Y = f(X, E): Effect generated from cause with independent noise
- Why useful?
    - structural constraints on f guarantee identifiability of the causal model, i.e., asymmetry in X and Y
    - in practice f can usually be approximated with a well-constrained form !
- How to distinguish cause from effect:
    - Fit the model for both directions, and see which direction gives **independence between the assumed cause and noise**

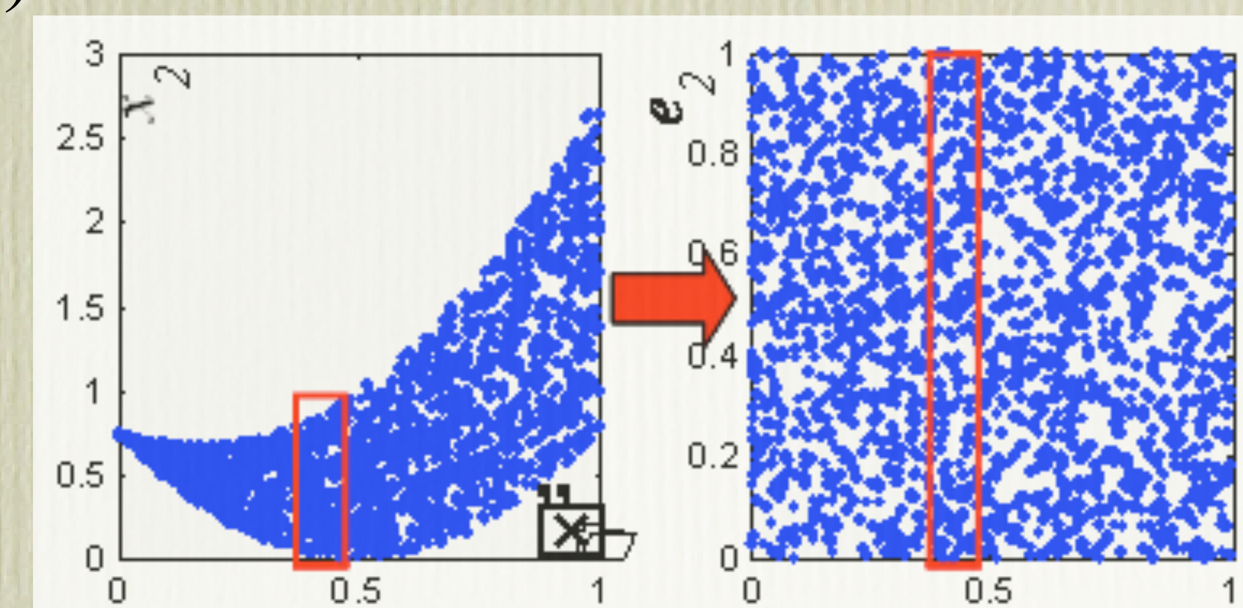
FCM cannot distinguish cause from effect without constraints on *f*

- Without constraints on f, for given (X, Y), both $Y = f_1(X, E)$ with $E \perp\!\!\!\perp X$ and $X = f_2(X, E_1)$ with $E_1 \perp\!\!\!\perp Y$ are possible
  - E.g., with a Gram-Schmidt-orthogonalization procedure (Darmois, '51, Hyvärinen & Pajunen, '99)

$$\tilde{x} = \mathrm{cdf}(x_1), \quad \text{so } \tilde{x} \sim \mathbf{U}(0,1);$$

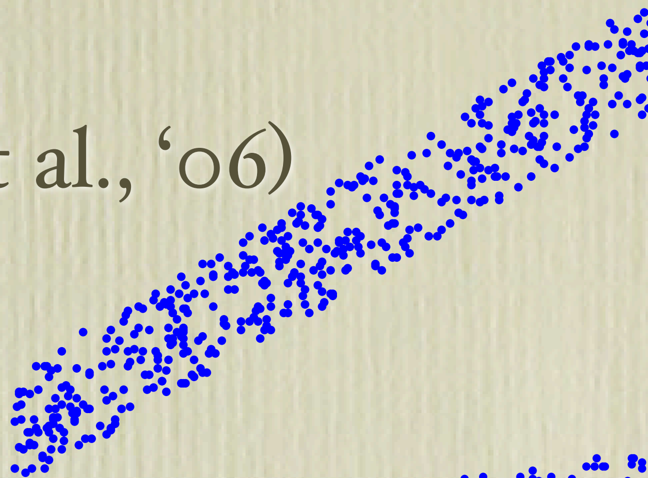
$$e = \mathrm{cdf}(y \mid \tilde{x}) = \int_{-\infty}^{x_2} p_{\tilde{x},y}(\tilde{x},t)\,dt.$$

Then $(x, y) \Rightarrow (\tilde{x}, e)$, with $E \perp\!\!\!\perp X$



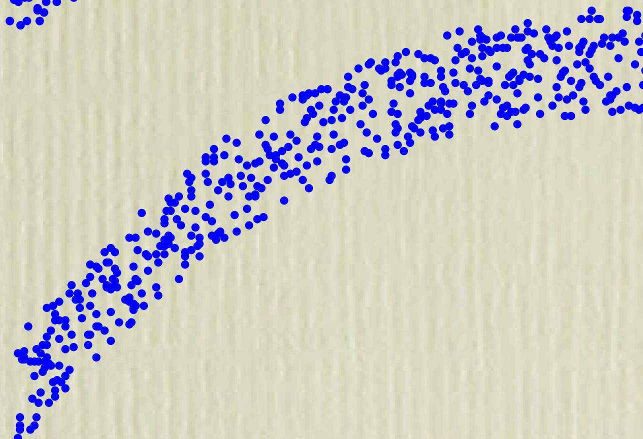
(Generally) identifiable FCMs with independent noise

- linear non-Gaussian acyclic causal model (Shimizu et al., '06)



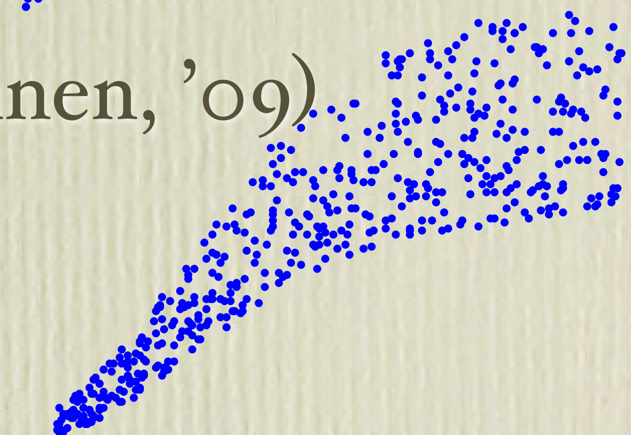
$$Y = aX + E$$

- additive noise model (Hoyer et al., '09)



$$Y = f(X) + E$$

- post-nonlinear (PNL) causal model (Zhang & Hyvärinen, '09)



$$Y = f_2\,(\,f_1(X) + E\,)$$

Some papers estimate the models by maximum likelihood; some propose to minimize the dependence between X and E: What is the difference?

Mutual information minimization vs. maximum likelihood ?

- Model:

$$\boxed{Y = f(X, E; \boldsymbol{\theta}_1);\ E \perp\!\!\!\perp X,\ E \sim p(E; \boldsymbol{\theta}_2),\ f \in \mathcal{F}\ \text{(appropriately constrained)}}$$

Maximum likelihood

$$\text{maximizing} \quad l_{X \to Y}(\boldsymbol{\theta}) = \sum_{i=1}^{T} \log P_{\mathcal{F}}(\mathbf{x}_i, \mathbf{y}_i)$$

$$= \sum_{i=1}^{T} \log p(X = \mathbf{x}_i) + \sum_{i=1}^{T} \log p(E = \hat{\mathbf{e}}_i; \boldsymbol{\theta}_2) - \sum_{i=1}^{T} \log \left| \frac{\partial f}{\partial E} \Big|_{E=\hat{\mathbf{e}}_i} \right|.$$

Independence criterion

$$\text{minimizing} \quad I(X, \hat{E}; \boldsymbol{\theta}) = -\frac{1}{T}\sum_{i=1}^{T} \log p(X = \mathbf{x}_i) - \frac{1}{T}\sum_{i=1}^{T} \log p(X = \mathbf{x}_i, Y = \mathbf{y}_i) +$$

$$\frac{1}{T}\sum_{i=1}^{T} \log p(\hat{E} = \hat{\mathbf{e}}_i; \boldsymbol{\theta}_2) + \frac{1}{T}\sum_{i=1}^{T} \log \left| \frac{\partial f}{\partial E} \Big|_{E=\hat{\mathbf{e}}_i} \right|$$

$$\boxed{\frac{1}{T} l_{X \to Y}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{i=1}^{T} \log p(X = \mathbf{x}_i, Y = \mathbf{y}_i) - I(X, \hat{E}; \boldsymbol{\theta}).}$$

- Maximum likelihood is equivalent to mutual information minimization
- However, convenient to do model selection with maximum likelihood !

Loss that might be caused by a wrongly specified noise distribution

$$Y = f(X, E; \boldsymbol{\theta}_1);\ E \perp\!\!\!\perp X,\ E \sim p(E; \boldsymbol{\theta}_2),\ f \in \mathcal{F}\ \text{(appropriately constrained)}$$

- Maximum likelihood (or mutual information minimization) aims to maximize

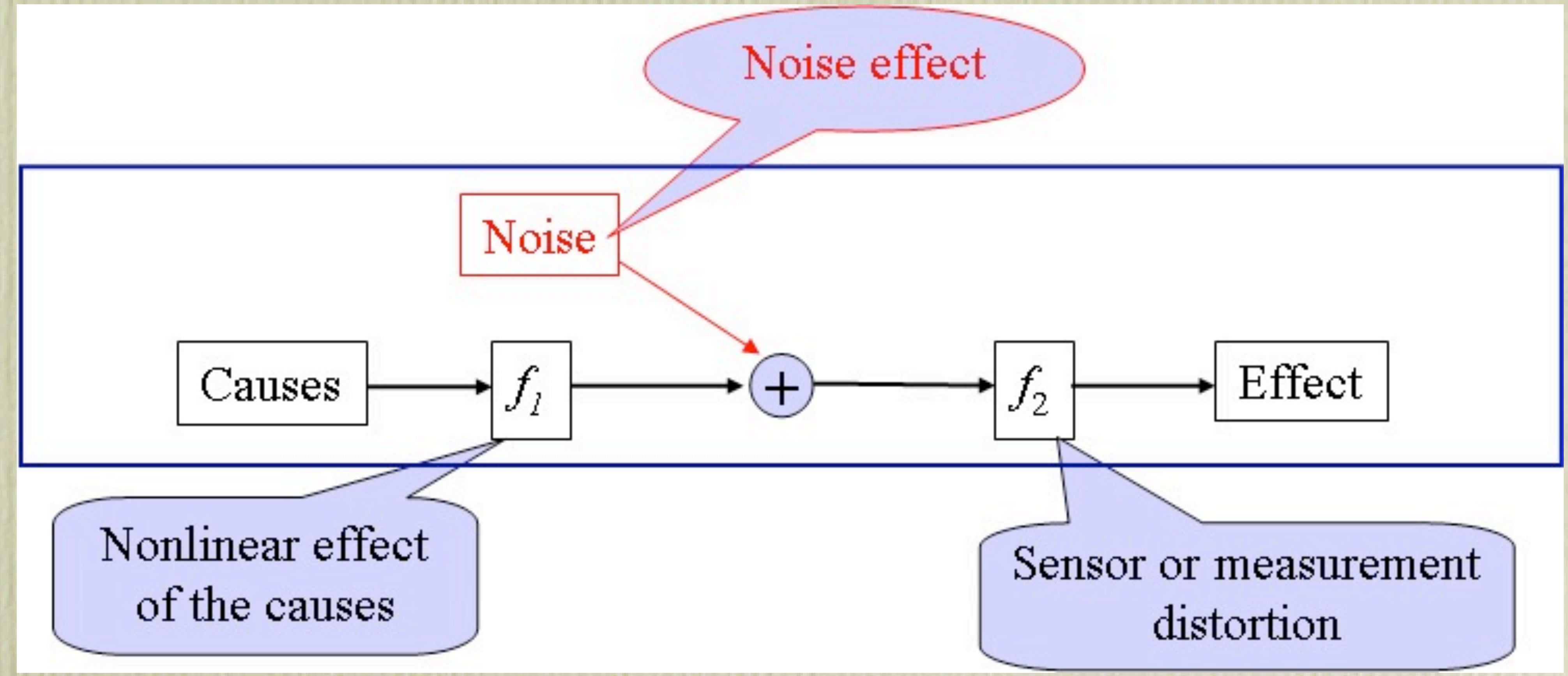
$$J_{X \to Y} = \sum_{i=1}^{T} \log p(E = \hat{\mathbf{e}}_i) - \sum_{i=1}^{T} \log \left| \frac{\partial f}{\partial E} \Big|_{E=\hat{\mathbf{e}}_i} \right|$$

- If f has additive noise, $\partial f / \partial E \equiv 1$; reduces to ordinary regression problem

- In general, if $p(E)$ is wrongly specified (e.g., simply set to Gaussian), the estimated f might not be statistically consistent
  - The estimated f might have to sacrifice to make the estimated noise closer to the specified $p(E)$ such that term 1 becomes bigger; a trade-off of the two terms, is maximized

Post-nonlinear causal model: An example

- Without prior knowledge, the assumed model is expected to be
  - general enough: adapted to approximate the true generating process
  - identifiable: asymmetry in causes and effects



- post-nonlinear (PNL) causal model: $Y = f_2(f_1(X) + E)$
- PNL causal model is generally identifiable (Zhang & Hyvärinen, '09)

Estimating PNL model by Warped Gaussian processes with non-Gaussian noise

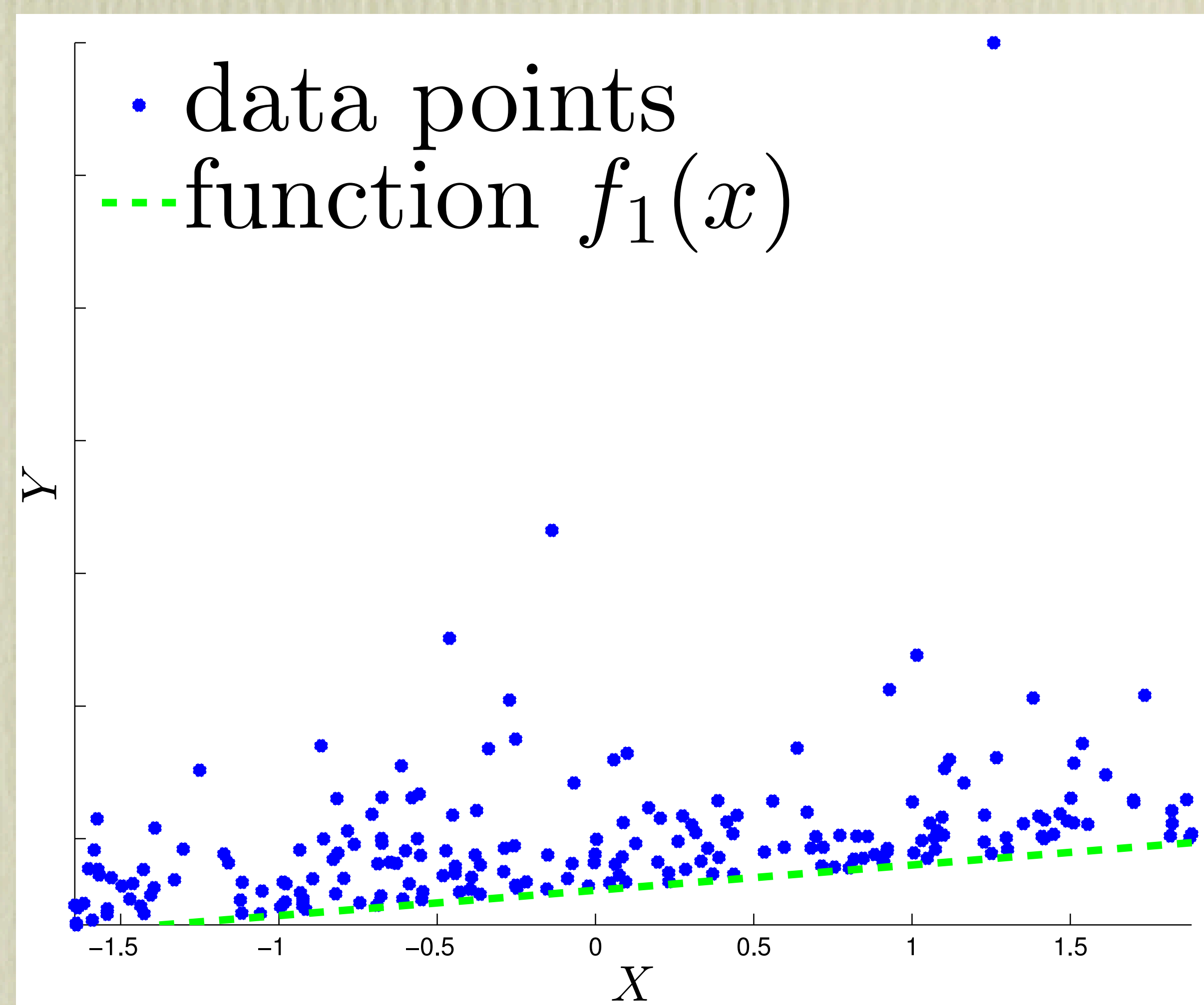
$$Y = f_2(f_1(X) + E)$$

- Previously estimated by minimizing mutual information between X and $\hat{E}$: difficult to do model selection
- A maximum likelihood perspective
  - Using a Gaussian process (GP) prior for $f_1 \Rightarrow$ warped Gaussian processes (Snelson et al., 2004)
  - Further use a flexible noise distribution (mixture of Gaussians) for E; otherwise the estimated f may be inconsistent
  - Represent f_2 with some basis functions
  - Use MCMC for Bayesian inference on f_1, f_2, and E

Simulation: Settings and results

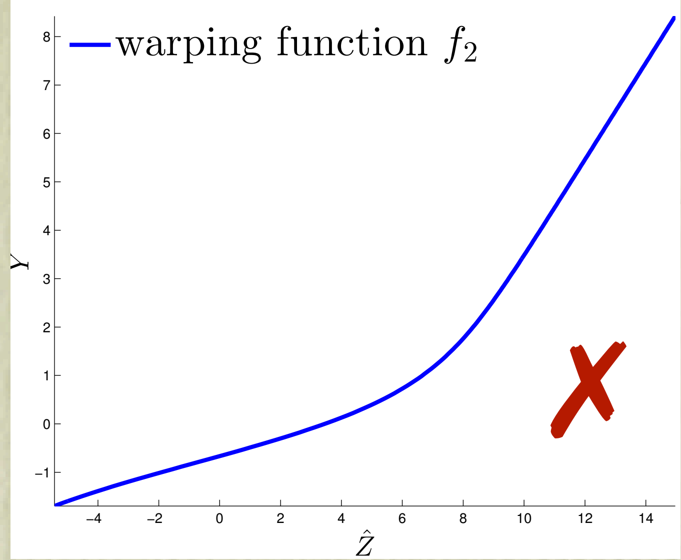
$$Y = f_2(f_1(X) + E)$$

- To illustrate different behaviors of estimated PNL causal model by
  - warped GPs with MoG noise (WGP-MoG)
  - warped GPs with Gaussian noise (WGP-Gaussian)
  - mutual information minimization approach with MLPs and MoG noise (PNL-MLP)
- Generated data: Z = 2X+E, Y = Z, i.e., f_1 = 2X, $f_2(Z)$ = Z, and E is log-normal

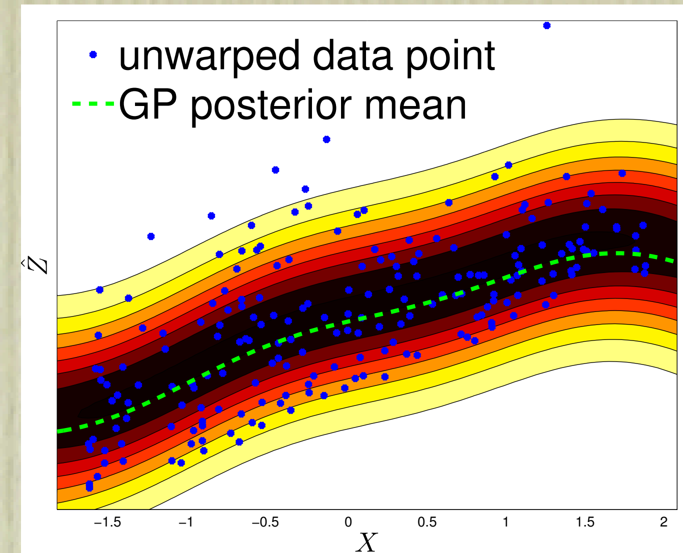


$$Y = f_2(f_1(X) + E)$$

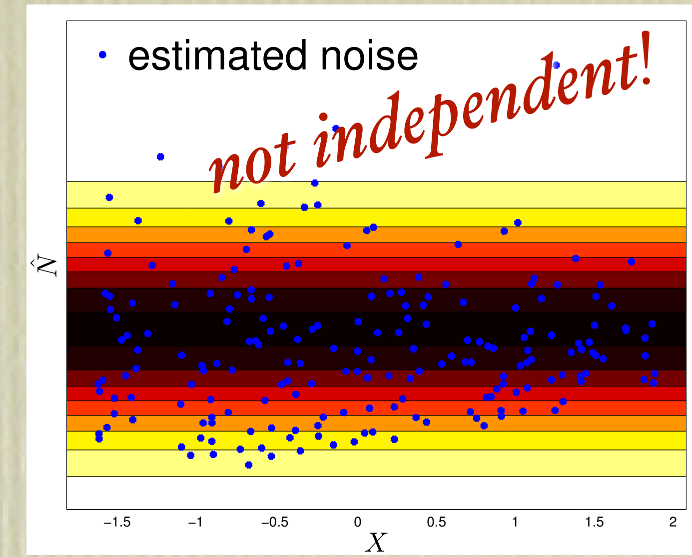
- WGP-Gaussian



(b) Estimated PNL function f_2

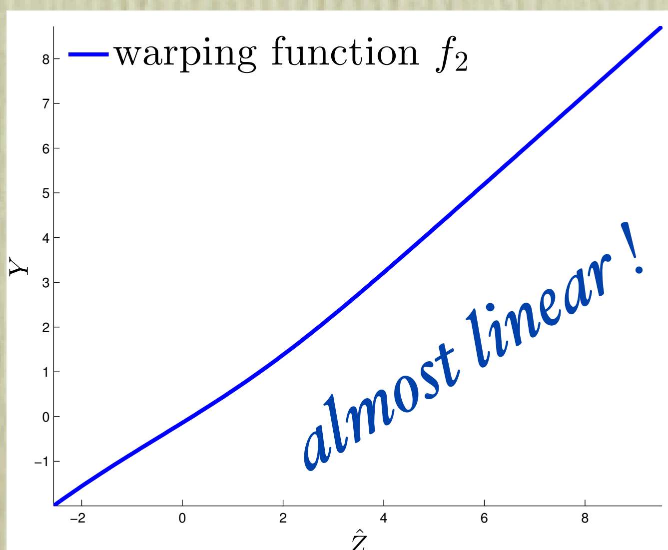


(c) Estimated f_1

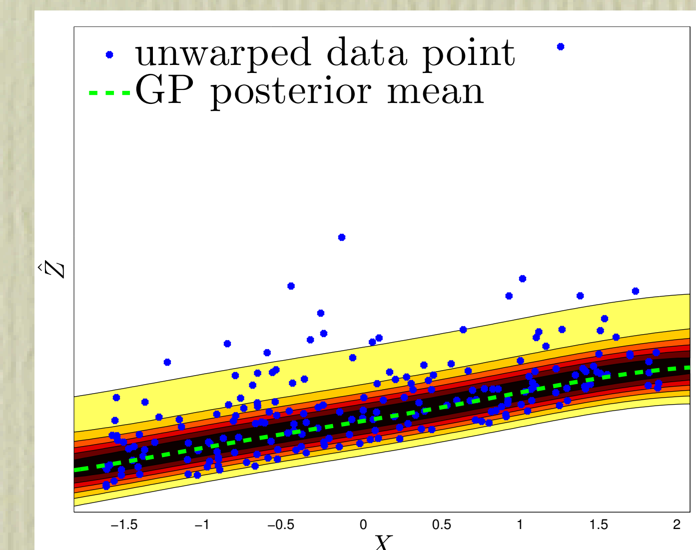


(d) X and estimated noise

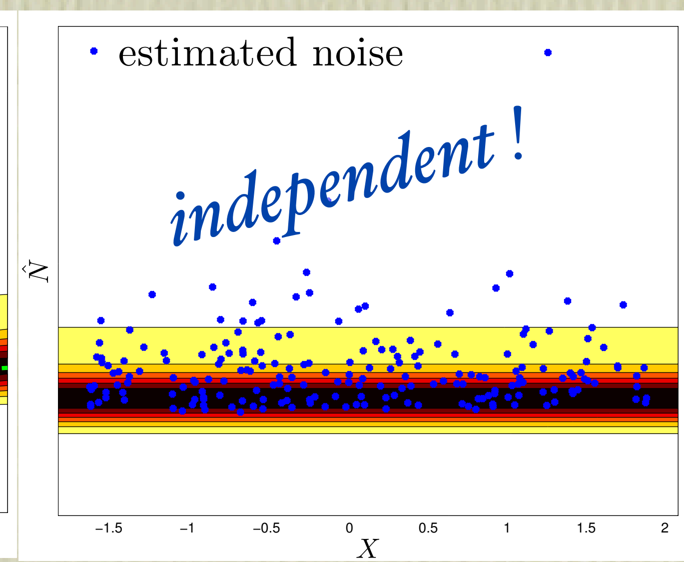
- WGP-MoG



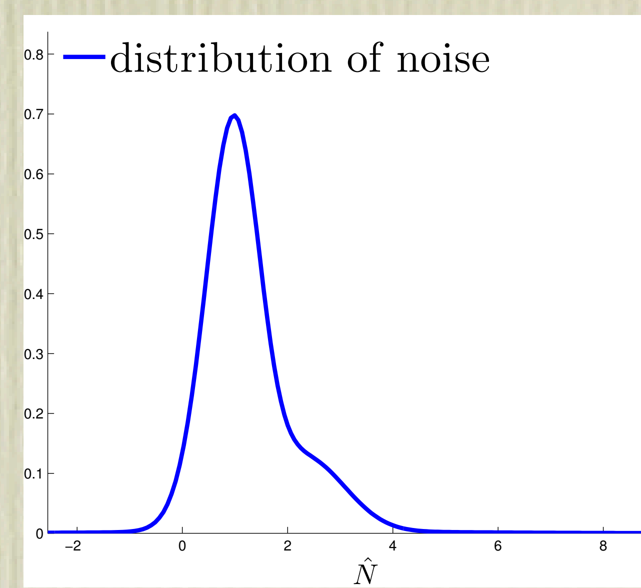
(a) Estimated PNL function f_2



(b) Estimated f_1

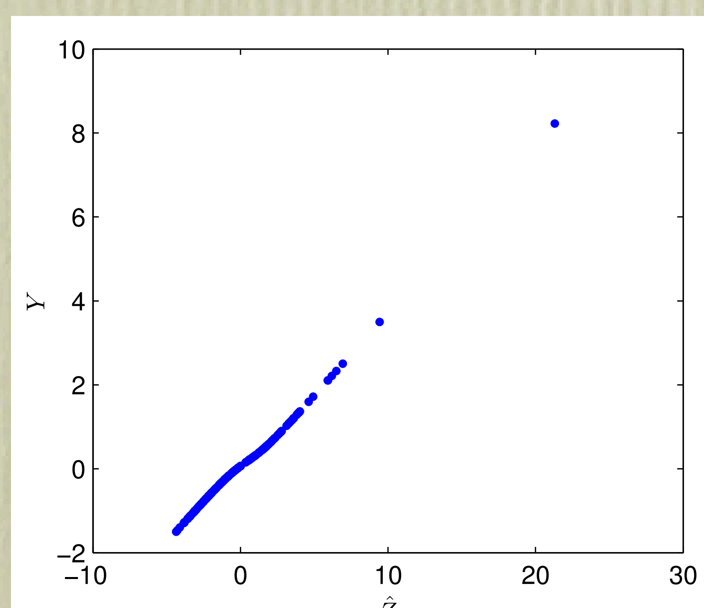


(c) X and estimated noise

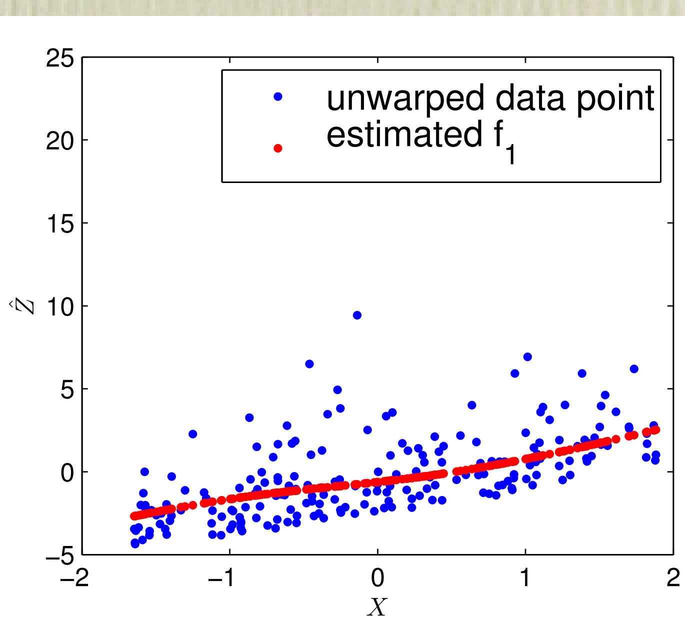


(d) MoG Noise Distribution

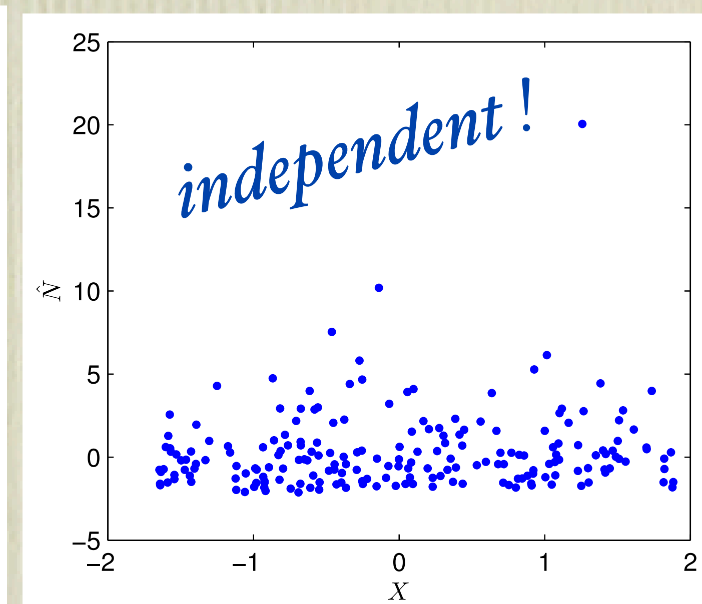
- PNL-MLP



(a) Estimated PNL function f_2



(b) Estimated f_1



(c) X and estimated noise

On real data

- Apply different approaches for causal direction determination on 77 cause-effect pairs, on which ground truth is known based on background info

Additive noise model

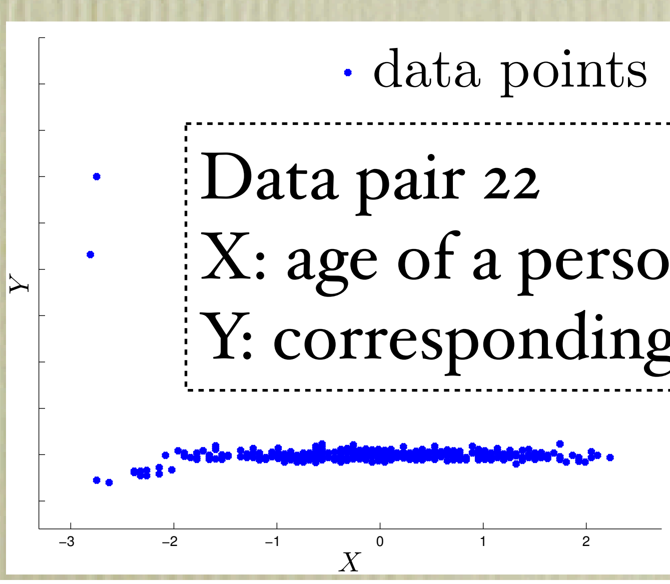
Gaussian process latent variable model

Information geometric causal inference

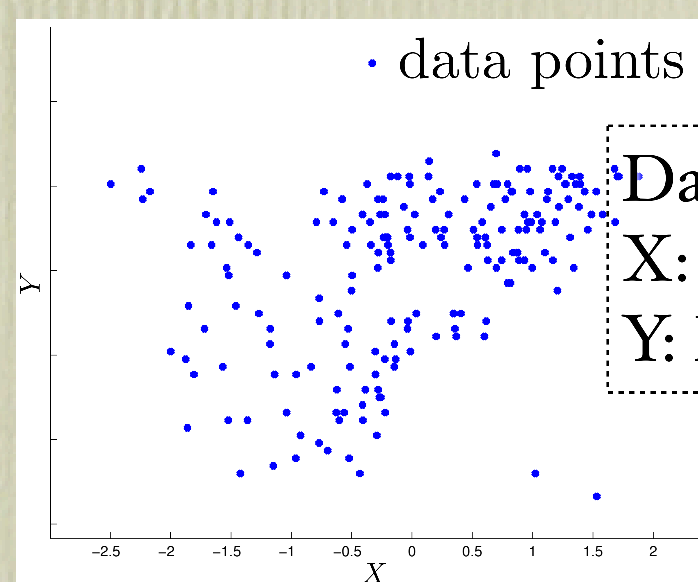
Accuracy of different methods for causal direction determination on the cause-effect pairs.

Method	PNL-MLP	PNL-WGP-Gaussian	PNL-WGP-MoG	ANM	GPI	IGCI
Accuracy (%)	70	67	76 ✔	63	72	73

- On pairs 22 and 57, PNL-WGP-MoG prefers X→Y, which is plausible due to background info, but PNL-WMP-Gaussian prefers Y→X



Data pair 22
X: age of a person,
Y: corresponding height



Data pair 57
X: latitude of the country's capital,
Y: life expectancy

Conclusion

- Functional causal model (FCM) based approach could fully identify causal structure
- For estimating FCMs, maximum likelihood is equivalent to minimizing the mutual info between the assume cause and noise
  - Bayesian model selection is easier to do from the maximum likelihood point of view
- In general, the performance depends on the specified noise distribution; better to learn it from data

Thanks for your listening.