Causal discovery and latent-variable models

Aapo Hyvärinen

Dept of Computer Science, University of Helsinki, Finland

UAI CRL workshop, 5th Aug 2022

Abstract

- Estimating causal direction is fundamental problem in science
- Basic bivariate case: Does x cause y, or does y cause x? Generalized as structural equation models (SEM)

Abstract

- Estimating causal direction is fundamental problem in science
- Basic bivariate case: Does x cause y, or does y cause x? Generalized as structural equation models (SEM)
- SEMs are ill-defined, unidentifiable for gaussian data

Abstract

- Estimating causal direction is fundamental problem in science
- Basic bivariate case: Does x cause y, or does y cause x? Generalized as structural equation models (SEM)
- SEMs are ill-defined, unidentifiable for gaussian data
- Approach: SEMs are equivalent to latent-variable models

Abstract

- Estimating causal direction is fundamental problem in science
- Basic bivariate case: Does x cause y, or does y cause x? Generalized as structural equation models (SEM)
- SEMs are ill-defined, unidentifiable for gaussian data
- Approach: SEMs are equivalent to latent-variable models
- Linear non-Gaussian SEM is identifiable since equivalent ICA (Shimizu et al, JMLR 2006)

Abstract

- Estimating causal direction is fundamental problem in science
- Basic bivariate case: Does x cause y, or does y cause x? Generalized as structural equation models (SEM)
- SEMs are ill-defined, unidentifiable for gaussian data
- Approach: SEMs are equivalent to latent-variable models
- Linear non-Gaussian SEM is identifiable since equivalent ICA (Shimizu et al, JMLR 2006)
- Completely nonlinear case identifiable by nonlinear ICA (Monti, Zhang, Hyvärinen, UAI2019)

Linear causality for two variables

- x and y both standardized to zero mean, unit variance
- Goal: distinguish between two statistical models:

$$y = \rho x + d \quad (x \to y) \tag{1}$$

$$x = \rho y + e \quad (y \to x) \tag{2}$$

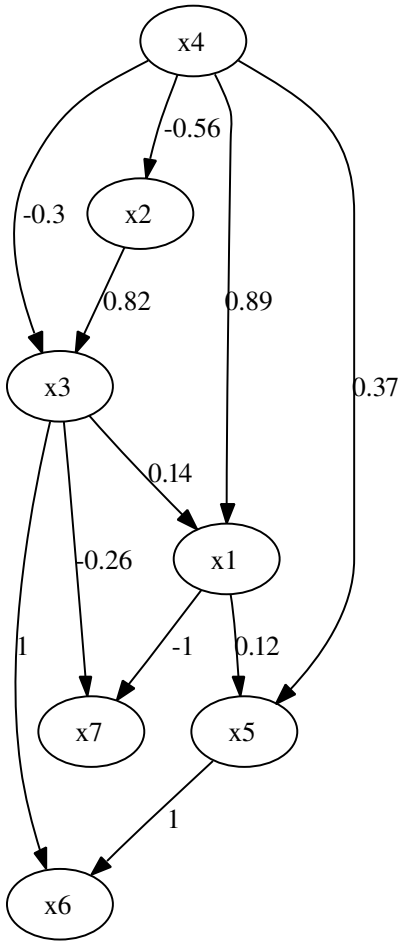
  where disturbances d, e are independent of x, y.
- If variables Gaussian $\Rightarrow$ completely symmetric :
  - Variance explained same for both models
  - Likelihood same for both models (simple function of ρ)

Structural equation models (linear)

- Or: *functional/structural causal model*
- Causal discovery with many variables
- Assume influences are linear, all variables observable:

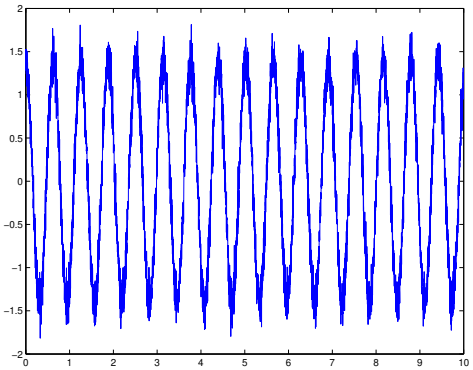
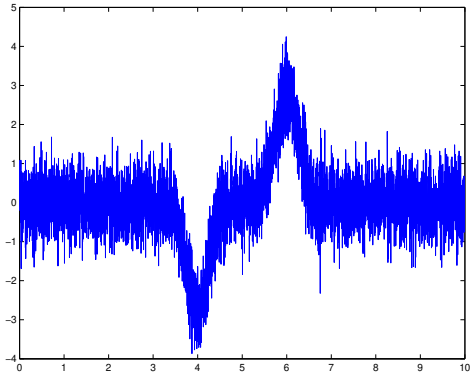
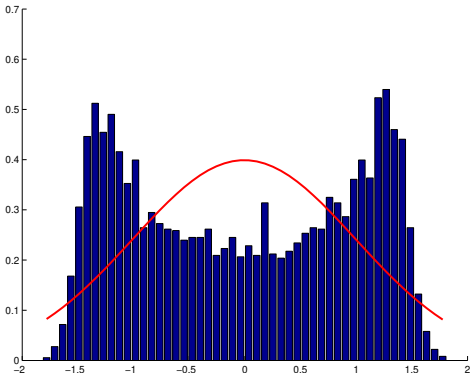
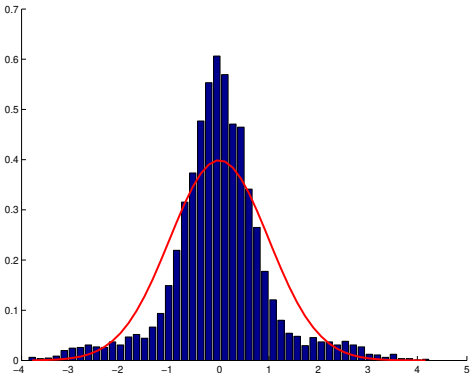
$$x_i = \sum_{j \neq i} b_{ij} x_j + e_i \text{ for all } i$$

- Typical assumption: "Directed Acyclic Graph" (DAG)
- Still, methods based on Gaussianity/covariances fail



Non-Gaussianity comes to rescue

Real-life signals often non-Gaussian

Key trick: ICA as identifiable latent variable model

- Linear independent component analysis (ICA)

$$x_i = \sum_{j=1}^{n} a_{ij} s_j \quad \text{for all } i, j = 1 \dots n \tag{3}$$

  - x_i is i-th observed random variable
  - a_{ij} constant parameters describing "mixing"
  - Assuming independent, non-Gaussian latent "sources" s_j

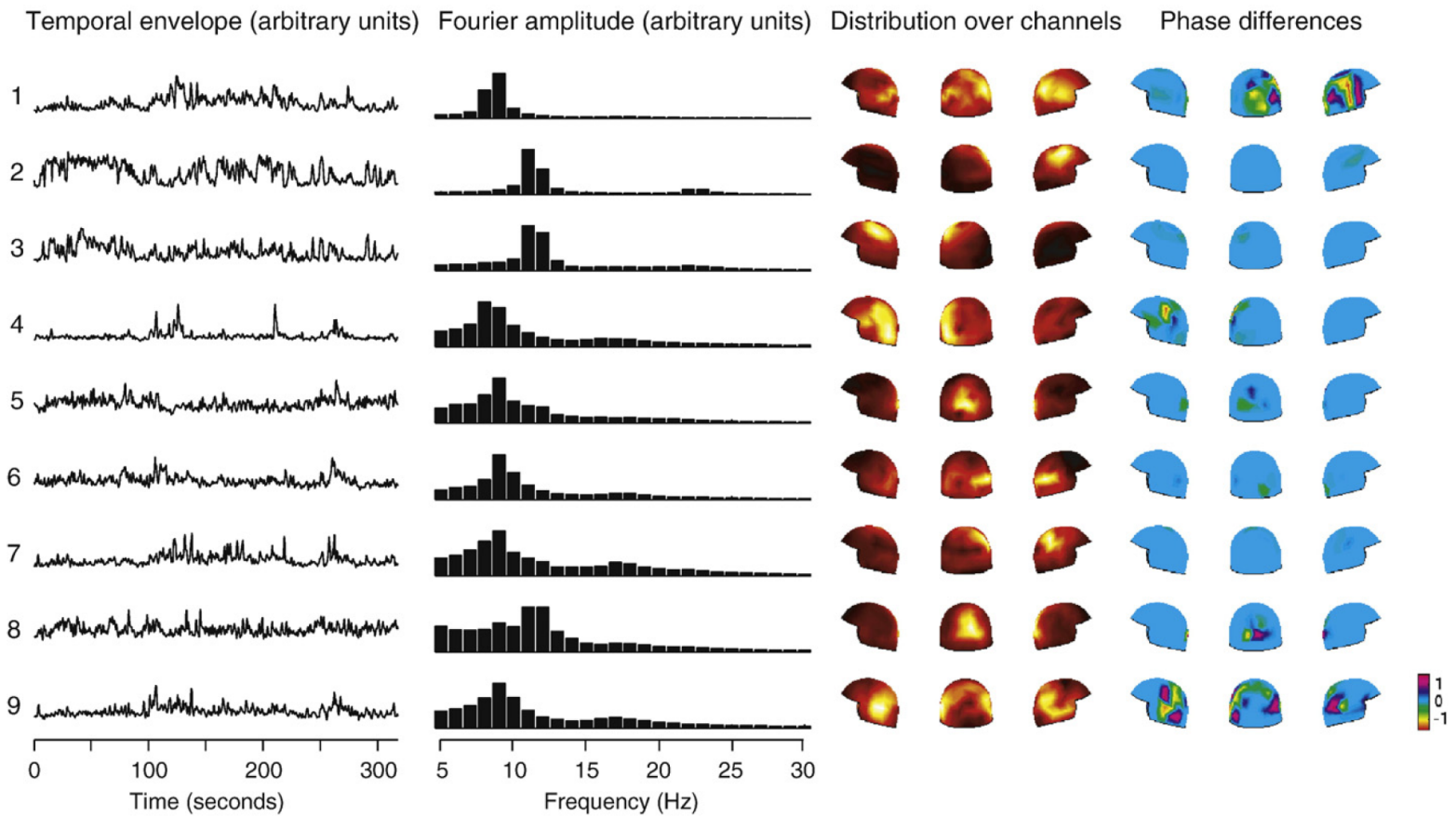
Key trick: ICA as identifiable latent variable model

- Linear independent component analysis (ICA)

$$x_i = \sum_{j=1}^{n} a_{ij} s_j \quad \text{for all } i, j = 1 \dots n \tag{3}$$

  - x_i is i-th observed random variable
  - a_{ij} constant parameters describing "mixing"
  - Assuming independent, non-Gaussian latent "sources" s_j
- ICA is identifiable, i.e. well-defined: (Darmois-Skitovich ∼1950; Comon, 1994)
  - Observing only x_i we can recover both a_{ij} and s_j
  - I.e. original latents can be recovered

Importance of identifiability: Brain source separation



(Hyvärinen, Ramkumar, Parkkonen, Hari, 2010)

Estimation of non-Gaussian SEM by ICA

(Shimizu et al, JMLR2006)

- Transform

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \Leftrightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}$$

Estimation of non-Gaussian SEM by ICA

(Shimizu et al, JMLR2006)

- Transform

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \Leftrightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}$$

- Assume disturbances e_i are non-Gaussian and independent
- Becomes independent component analysis (ICA):
  - $\mathbf{x} = \mathbf{A}\mathbf{e}$ with independent, non-Gaussian e_i

Estimation of non-Gaussian SEM by ICA

(Shimizu et al, JMLR2006)

- Transform
$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \Leftrightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}$$
- Assume disturbances e_i are non-Gaussian and independent
- Becomes independent component analysis (ICA):
  - $\mathbf{x} = \mathbf{A}\mathbf{e}$ with independent, non-Gaussian e_i
- Identifiability of ICA implies identifiability of SEM (almost)

Estimation of non-Gaussian SEM by ICA

(Shimizu et al, JMLR2006)

- Transform
$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \Leftrightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}$$
- Assume disturbances e_i are non-Gaussian and independent
- Becomes independent component analysis (ICA):
  - $\mathbf{x} = \mathbf{A}\mathbf{e}$ with independent, non-Gaussian e_i
- Identifiability of ICA implies identifiability of SEM (almost)
- Just one complication: ICA does not estimate order of e_i
  - DAG assumption enables identifiability
- We can (almost) estimate SEM using ICA algorithms

Causal direction with general nonlinearities

- Choose between two regression directions with completely general nonlinearities f_1 or f_2:

$$y = f_1(x, e) \quad (x \to y) \tag{4}$$

$$x = f_2(y, d) \quad (y \to x) \tag{5}$$

with stochastic influences e, d

Causal direction with general nonlinearities

- Choose between two regression directions with completely general nonlinearities f_1 or f_2:

$$y = f_1(x, e) \quad (x \to y) \tag{4}$$

$$x = f_2(y, d) \quad (y \to x) \tag{5}$$

with stochastic influences e, d

- Each can be expressed as nonlinear ICA model, e.g. $x \to y$:

$$x = f_0(e_1) \tag{6}$$

$$y = f_1(x, e_2) = \tilde{f}_1(e_1, e_2) \tag{7}$$

with components e_1, e_2

Causal direction with general nonlinearities

- Choose between two regression directions with completely general nonlinearities f_1 or f_2:

$$y = f_1(x, e) \quad (x \to y) \tag{4}$$

$$x = f_2(y, d) \quad (y \to x) \tag{5}$$

with stochastic influences e, d

- Each can be expressed as nonlinear ICA model, e.g. $x \to y$:

$$x = f_0(e_1) \tag{6}$$

$$y = f_1(x, e_2) = \tilde{f}_1(e_1, e_2) \tag{7}$$

with components e_1, e_2

- But: do we know how to do nonlinear ICA?

Great difficulty of nonlinear ICA

- Darmois (1952) showed "impossibility" of nonlinear ICA:
- For any x_1, x_2, can always construct $z = g(x_1, x_2)$ independent of x_1 as

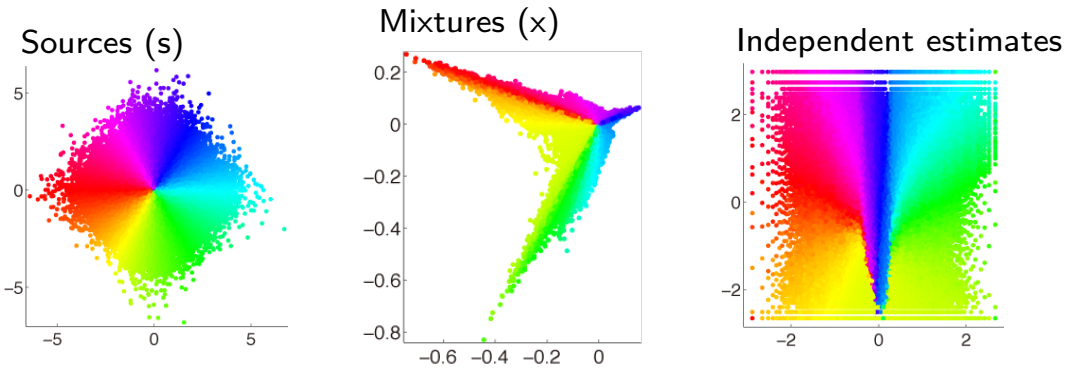
$$g(\xi_1, \xi_2) = P(x_2 < \xi_2 | x_1 = \xi_1) \tag{8}$$

Great difficulty of nonlinear ICA

- Darmois (1952) showed "impossibility" of nonlinear ICA:
- For any x_1, x_2, can always construct $z = g(x_1, x_2)$ independent of x_1 as

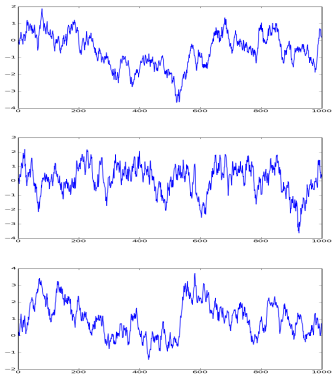
$$g(\xi_1, \xi_2) = P(x_2 < \xi_2 | x_1 = \xi_1) \tag{8}$$

- Independence alone too weak for identifiability:
  We could take x_1 as independent component which is absurd

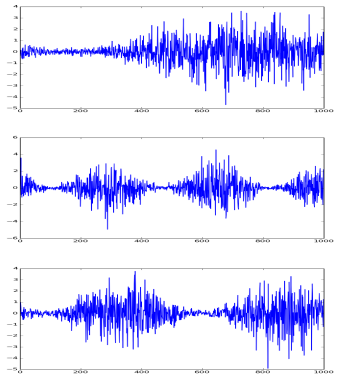


Temporal structure or multimodality enable nonlinear ICA

- Theory above considered i.i.d. sampled random variables
- What if we have time series? some "auxiliary" data?



Autocorrelations
(Harmeling et al 2003)



Nonstationarity
(Hyvärinen and Morioka, NIPS2016)

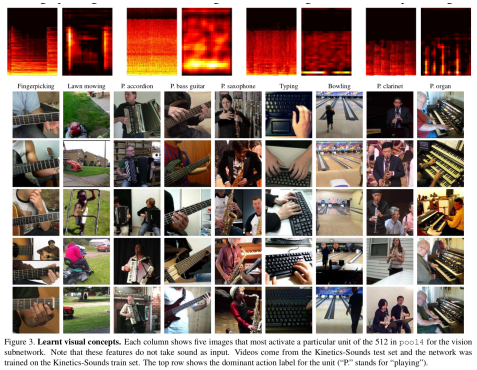
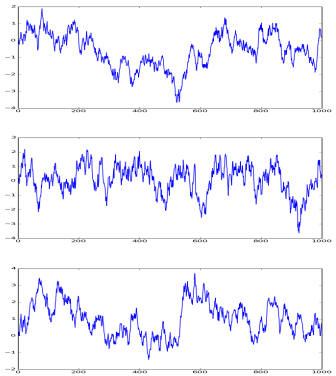


Figure 3. **Learnt visual concepts.** Each column shows five images that most activate a particular unit of the 512 in pool4 for the vision subnetwork. Note that these features do not take sound as input. Videos come from the Kinetics-Sounds test set and the network was trained on the Kinetics-Sounds train set. The top row shows the dominant action label for the unit ("P." stands for "playing").

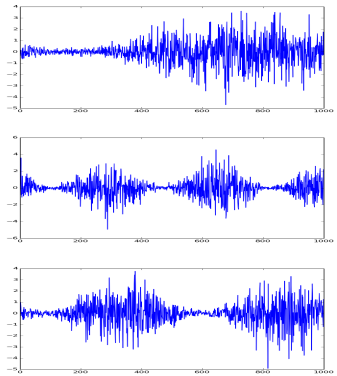
Dependencies btw audio and video
(Arandjelovic&Zisserman 2017; Hyvärinen et al, 2019)

Temporal structure or multimodality enable nonlinear ICA

- Theory above considered i.i.d. sampled random variables
- What if we have time series? some "auxiliary" data?



Autocorrelations
(Harmeling et al 2003)



Nonstationarity
(Hyvärinen and Morioka, NIPS2016)

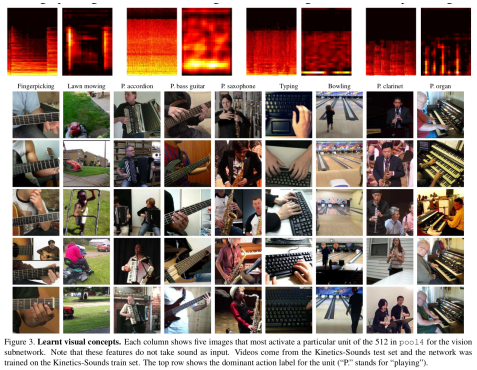


Figure 3. **Learnt visual concepts.** Each column shows five images that most activate a particular unit of the 512 in pool4 for the vision subnetwork. Note that these features do not take sound as input. Videos come from the Kinetics-Sounds test set and the network was trained on the Kinetics-Sounds train set. The top row shows the dominant action label for the unit ("P." stands for "playing").

Dependencies btw audio and video
(Arandjelovic&Zisserman 2017; Hyvärinen et al, 2019)

- Identifiability of nonlinear ICA can be proven
(Sprekeler et al, 2014; Hyvärinen and Morioka, NIPS2016 & AISTATS2017; Khemakhem et al 2020)
Can find original sources!

NonSENS: (Monti, K. Zhang, Hyvärinen, UAI2019) Non-Linear SEM Estimation using Non-Stationarity

- Do nonlinear ICA by nonstationarity (e.g. due to interventions)

NonSENS: (Monti, K. Zhang, Hyvärinen, UAI2019) Non-Linear SEM Estimation using Non-Stationarity

- Do nonlinear ICA by nonstationarity (e.g. due to interventions)
- Further trick: If true causal model is $x \to y$, x is independent of e_2, since

$$x = f_0(e_1) \tag{9}$$

$$y = f_1(x, e_2) = \tilde{f}_1(e_1, e_2) \tag{10}$$

NonSENS: (Monti, K. Zhang, Hyvärinen, UAI2019) Non-Linear SEM Estimation using Non-Stationarity

- Do nonlinear ICA by nonstationarity (e.g. due to interventions)
- Further trick: If true causal model is $x \to y$, x is independent of e_2, since

$$x = f_0(e_1) \tag{9}$$

$$y = f_1(x, e_2) = \tilde{f}_1(e_1, e_2) \tag{10}$$

- Our NonSENS algorithm proceeds as follows:
  1. Recover latent sources, e_1, e_2, by nonlinear ICA
  2. Run various independence tests (we employ HSIC)
  3. Conclude cause is the variable independent of one influence e_j (or output result is inconclusive)

Digression: Causal Autoregressive Flows

(Khemakhem et al, AISTATS2021)

- E.g. "Affine flow":

$$x_j = \exp(\Phi_j(x_1, \ldots, x_{j-1}))z_j + \Psi_j(x_1, \ldots, x_{j-1}) \qquad (11)$$

with two NN's Φ and Ψ

Digression: Causal Autoregressive Flows

(Khemakhem et al, AISTATS2021)

- E.g. "Affine flow":

$$x_j = \exp(\Phi_j(x_1, \ldots, x_{j-1}))z_j + \Psi_j(x_1, \ldots, x_{j-1}) \qquad (11)$$

  with two NN's Φ and Ψ
- A generative model, but immediately similar to SEM
- Ordering in autoregressive model $\sim$ causal ordering in DAG

Digression: Causal Autoregressive Flows

(Khemakhem et al, AISTATS2021)

- E.g. "Affine flow":

$$x_j = \exp(\Phi_j(x_1, \ldots, x_{j-1}))z_j + \Psi_j(x_1, \ldots, x_{j-1}) \qquad (11)$$

  with two NN's Φ and Ψ
- A generative model, but immediately similar to SEM
- Ordering in autoregressive model $\sim$ causal ordering in DAG
- Theorem: with 2 variables, the model is identifiable if
  - z_1, z_2 are Gaussian and statistically independent,
  - Ψ is invertible and nonlinear
- Generalizes (in one sense) additive nonlinear models (Hoyer et al, 2008)
  - Violates independence assumption usually made

Combining feature learning and causality?

- We saw different properties that enable identifiability:
  - non-Gaussianity (in linear case)
  - non-stationarity
  - temporal dependencies
  - multimodality (e.g. audio + video)

Combining feature learning and causality?

- We saw different properties that enable identifiability:
  - non-Gaussianity (in linear case)
  - non-stationarity
  - temporal dependencies
  - multimodality (e.g. audio + video)

→ We could use one to estimate features, another for causal discovery between them

  - E.g. In linear case, combine nonstationarity with autoregressive model (K. Zhang and Hyvärinen, UAI2010)

Combining feature learning and causality?

- We saw different properties that enable identifiability:
  - non-Gaussianity (in linear case)
  - non-stationarity
  - temporal dependencies
  - multimodality (e.g. audio + video)

→ We could use one to estimate features, another for causal discovery between them

  - E.g. In linear case, combine nonstationarity with autoregressive model (K. Zhang and Hyvärinen, UAI2010)

- Should we use causal notation for generative (latent variable) models?
  - $\mathbf{x} := \mathbf{A}\mathbf{s}$
  - since the relation is not symmetric, $\mathbf{s}$ is generated first

Combining feature learning and causality?

- We saw different properties that enable identifiability:
  - non-Gaussianity (in linear case)
  - non-stationarity
  - temporal dependencies
  - multimodality (e.g. audio + video)

→ We could use one to estimate features, another for causal discovery between them

  - E.g. In linear case, combine nonstationarity with autoregressive model (K. Zhang and Hyvärinen, UAI2010)

- Should we use causal notation for generative (latent variable) models?
  - $\mathbf{x} := \mathbf{A}\mathbf{s}$
  - since the relation is not symmetric, $\mathbf{s}$ is generated first
- Identifiable ⇒ Causal ? Causal ⇒ Identifiable ?

Conclusion

- Finding direction of effect is fundamental in causal discovery

Conclusion

- Finding direction of effect is fundamental in causal discovery
- Linear Gaussian case is hopeless: $y = ax + d$ and $x = ay + e$ give equal fit according to any measure

Conclusion

- Finding direction of effect is fundamental in causal discovery
- Linear Gaussian case is hopeless: $y = ax + d$ and $x = ay + e$ give equal fit according to any measure
- In linear case, direction was first found by non-Gaussianity (Shimizu et al, JMLR2006; Dodge and Rousson, 2001)
- Transform SEM to ICA
  - identifiable & algorithms available

Conclusion

- Finding direction of effect is fundamental in causal discovery
- Linear Gaussian case is hopeless: $y = ax + d$ and $x = ay + e$ give equal fit according to any measure
- In linear case, direction was first found by non-Gaussianity (Shimizu et al, JMLR2006; Dodge and Rousson, 2001)
- Transform SEM to ICA
  - identifiable & algorithms available
- Completely nonlinear case:
  - Basic ICA unidentifiable with i.i.d. sampling
  - Nonstationarity ($\approx$ interventions) enables identifiability

  $\Rightarrow$ NonSENS method (Monti, K. Zhang, Hyvärinen, UAI2019)

Conclusion

- Finding direction of effect is fundamental in causal discovery
- Linear Gaussian case is hopeless: $y = ax + d$ and $x = ay + e$ give equal fit according to any measure
- In linear case, direction was first found by non-Gaussianity (Shimizu et al, JMLR2006; Dodge and Rousson, 2001)
- Transform SEM to ICA
  - identifiable & algorithms available
- Completely nonlinear case:
  - Basic ICA unidentifiable with i.i.d. sampling
  - Nonstationarity ($\approx$ interventions) enables identifiability

  $\Rightarrow$ NonSENS method (Monti, K. Zhang, Hyvärinen, UAI2019)
- We show connection to autoregressive flows (Khemakhem et al, AISTATS2021)

Conclusion

- Finding direction of effect is fundamental in causal discovery
- Linear Gaussian case is hopeless: $y = ax + d$ and $x = ay + e$ give equal fit according to any measure
- In linear case, direction was first found by non-Gaussianity (Shimizu et al, JMLR2006; Dodge and Rousson, 2001)
- Transform SEM to ICA
  - identifiable & algorithms available
- Completely nonlinear case:
  - Basic ICA unidentifiable with i.i.d. sampling
  - Nonstationarity ($\approx$ interventions) enables identifiability

  $\Rightarrow$ NonSENS method (Monti, K. Zhang, Hyvärinen, UAI2019)
- We show connection to autoregressive flows (Khemakhem et al, AISTATS2021)
- Future work: Combine feature learning and causal discovery