Probabilistic latent variable models for distinguishing between cause and effect

Joris Mooij

joint work with
Oliver Stegle Dominik Janzing Kun Zhang Bernhard Schölkopf

Max Planck Institute for Biological Cybernetics
Tübingen, Germany
joris.mooij@tuebingen.mpg.de

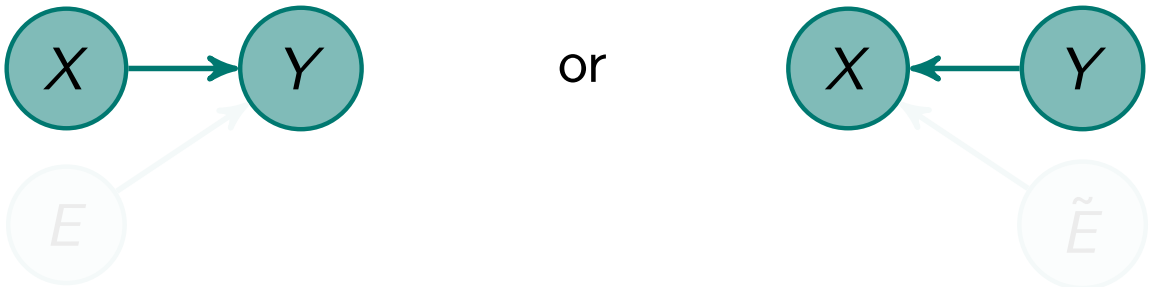
Tübingen, October 7th, 2010

The core problem

Given N observations of two variables X, Y

i	x_i	y_i
1	3.4	2.5
2	7.2	5.2
3	2.3	1.3
⋮	⋮	⋮
N	4.5	2.1

determine whether:



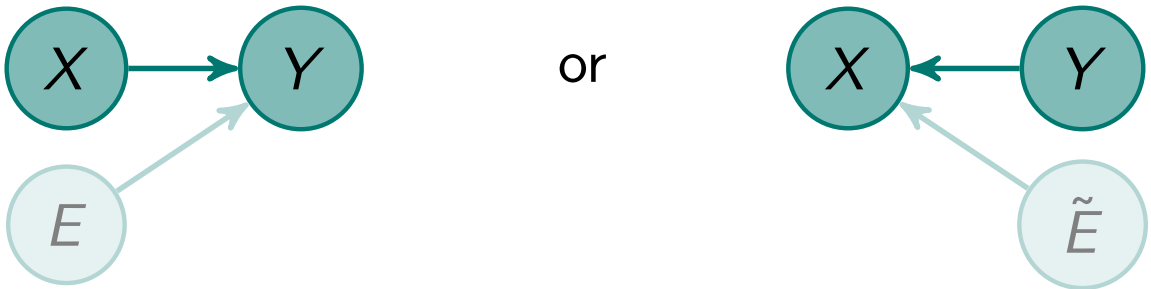
"X causes Y" "Y causes X"

The core problem

Given N observations of two variables X, Y

i	x_i	y_i
1	3.4	2.5
2	7.2	5.2
3	2.3	1.3
⋮	⋮	⋮
N	4.5	2.1

determine whether:

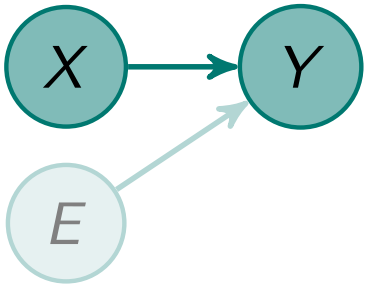


"X causes Y" "Y causes X"

Part I

The basics

Basic assumptions



Suppose:

- X, E cause Y;
- X, Y are *observed*;
- E is *unobserved* (latent)

1. **Determinism**: a function f exists such that:

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

(f is called the *causal mechanism*);

2. **No common causes** of X and E are present:

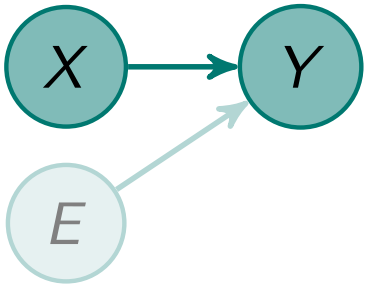
$$X \perp\!\!\!\perp E,$$

(X and E are statistically *independent*).

3. **Independence of distribution of causes and mechanism**:

$$p(X, E) \perp\!\!\!\perp f$$

Basic assumptions



Suppose:

- X, E cause Y;
- X, Y are *observed*;
- E is *unobserved* (latent)

1. **Determinism**: a function f exists such that:

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

(f is called the *causal mechanism*);

2. **No common causes** of X and E are present:

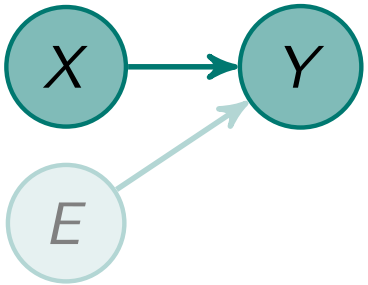
$$X \perp\!\!\!\perp E,$$

(X and E are statistically *independent*).

3. **Independence of distribution of causes and mechanism**:

$$p(X, E) \perp\!\!\!\perp f$$

Basic assumptions



Suppose:

- X, E cause Y;
- X, Y are *observed*;
- E is *unobserved* (latent)

1. **Determinism**: a function f exists such that:

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

(f is called the *causal mechanism*);

2. **No common causes** of X and E are present:

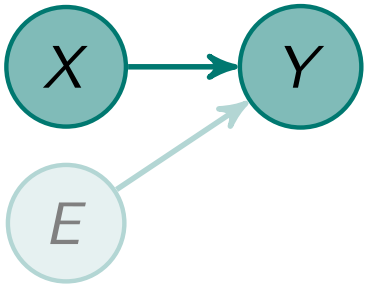
$$X \perp\!\!\!\perp E,$$

(X and E are statistically *independent*).

3. **Independence of distribution of causes and mechanism**:

$$p(X, E) \perp\!\!\!\perp f$$

Basic assumptions



Suppose:

- X, E cause Y;
- X, Y are *observed*;
- E is *unobserved* (latent)

1. **Determinism**: a function f exists such that:

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

(f is called the *causal mechanism*);

2. **No common causes** of X and E are present:

$$X \perp\!\!\!\perp E,$$

(X and E are statistically *independent*).

3. **Independence of distribution of causes and mechanism**:

$$p(X, E) \perp\!\!\!\perp f$$

Previous work: restricting the model class

Several recent approaches restrict the class of possible causal mechanisms:

LINGAM	f is linear	$f(X, E) = \alpha X + \beta E$
AN	Additive Noise	$f(X, E) = F(X) + E$
PNL	Post-Non-Linear model	$f(X, E) = G(F(X) + E)$
HS	Hetero-Schedastic noise	$f(X, E) = F(X) + E \cdot G(X)$

Causal discovery is possible because of the following *identifiability* results:

Theorem (*al.*)

Let $\mathcal{M} \in \{\mathrm{LINGAM}, \mathrm{AN}, \mathrm{PNL}\}$: generically, if a model $X \to Y$ in $\mathcal{M}$ exists for $p(X, Y)$, no model $Y \to X \in \mathcal{M}$ exists for $p(X, Y)$.

The idea is to fit a restricted model in both directions ($X \to Y$ and $Y \to X$) and infer the causal direction to be the one that yields the best fit with the data.

Previous work: restricting the model class

Several recent approaches restrict the class of possible causal mechanisms:

LINGAM	f is linear	$f(X, E) = \alpha X + \beta E$
AN	Additive Noise	$f(X, E) = F(X) + E$
PNL	Post-Non-Linear model	$f(X, E) = G(F(X) + E)$
HS	Hetero-Schedastic noise	$f(X, E) = F(X) + E \cdot G(X)$

Causal discovery is possible because of the following *identifiability* results:

Theorem (*al.*)

Let $\mathcal{M} \in \{\mathrm{LINGAM}, \mathrm{AN}, \mathrm{PNL}\}$: generically, if a model $X \to Y$ in $\mathcal{M}$ exists for $p(X, Y)$, no model $Y \to X \in \mathcal{M}$ exists for $p(X, Y)$.

The idea is to fit a restricted model in both directions ($X \to Y$ and $Y \to X$) and infer the causal direction to be the one that yields the best fit with the data.

Previous work: restricting the model class

Several recent approaches restrict the class of possible causal mechanisms:

LINGAM	f is linear	$f(X, E) = \alpha X + \beta E$
AN	Additive Noise	$f(X, E) = F(X) + E$
PNL	Post-Non-Linear model	$f(X, E) = G(F(X) + E)$
HS	Hetero-Schedastic noise	$f(X, E) = F(X) + E \cdot G(X)$

Causal discovery is possible because of the following *identifiability* results:

Theorem (*al.*)

Let $\mathcal{M} \in \{\mathrm{LINGAM}, \mathrm{AN}, \mathrm{PNL}\}$: generically, if a model $X \to Y$ in $\mathcal{M}$ exists for $p(X, Y)$, no model $Y \to X \in \mathcal{M}$ exists for $p(X, Y)$.

The idea is to fit a restricted model in both directions ($X \to Y$ and $Y \to X$) and infer the causal direction to be the one that yields the best fit with the data.

Previous work: comparing model complexities

A different (recently proposed) approach is based on information theory. It does not restrict the class of possible causal mechanisms.

Theorem (Janzing, Schölkopf)

If $I\big(p(X) : p(Y \mid X)\big) \stackrel{+}{=} 0$, *then*

$$K\big(p(X)\big) + K\big(p(Y \mid X)\big) \stackrel{+}{\leq} K\big(p(Y)\big) + K\big(p(X \mid Y)\big),$$

where $K(\cdot)$ *is the* $\mathit{Kolmogorov\ complexity}$ *and* $I(\cdot : \cdot)$ *is the analogue of mutual information based on Kolmogorov complexity.*

Unfortunately, Kolmogorov complexity is uncomputable, so this result is not directly useful in practice.

Previous work: comparing model complexities

A different (recently proposed) approach is based on information theory. It does not restrict the class of possible causal mechanisms.

Theorem (Janzing, Schölkopf)

If $I\big(p(X) : p(Y \mid X)\big) \stackrel{+}{=} 0$, *then*

$$K\big(p(X)\big) + K\big(p(Y \mid X)\big) \stackrel{+}{\leq} K\big(p(Y)\big) + K\big(p(X \mid Y)\big),$$

where $K(\cdot)$ *is the* $Kolmogorov\ complexity$ *and* $I(\cdot : \cdot)$ *is the analogue of mutual information based on Kolmogorov complexity.*

Unfortunately, Kolmogorov complexity is uncomputable, so this result is not directly useful in practice.

Previous work: comparing model complexities

A different (recently proposed) approach is based on information theory. It does not restrict the class of possible causal mechanisms.

Theorem (Janzing, Schölkopf)

If $I\big(p(X) : p(Y \mid X)\big) \stackrel{+}{=} 0$, *then*

$$K\big(p(X)\big) + K\big(p(Y \mid X)\big) \stackrel{+}{\leq} K\big(p(Y)\big) + K\big(p(X \mid Y)\big),$$

where $K(\cdot)$ *is the* $Kolmogorov\ complexity$ *and* $I(\cdot : \cdot)$ *is the analogue of mutual information based on Kolmogorov complexity.*

Unfortunately, Kolmogorov complexity is uncomputable, so this result is not directly useful in practice.

This work: non-parametric causal discovery

The goal of this work is to avoid the restrictions on the model class for the causal mechanism.

However, in this way we loose identifiability:

Theorem

Given random variables X, Y, E and a function f such that

$$Y = f(X, E), \qquad X \perp\!\!\!\perp E,$$

we can always construct a function $\tilde{f}$ and a random variable $\tilde{E}$ such that

$$X = \tilde{f}(Y, \tilde{E}), \qquad Y \perp\!\!\!\perp \tilde{E}.$$

The crucial idea is now to compare the *model complexities* and infer the least complex model to be the true causal direction. However, we use other complexity measures than Kolmogorov complexity, so that we can actually compute them.

This work: non-parametric causal discovery

The goal of this work is to avoid the restrictions on the model class for the causal mechanism.

However, in this way we loose identifiability:

Theorem

Given random variables X, Y, E and a function f such that

$$Y = f(X, E), \qquad X \perp\!\!\!\perp E,$$

we can always construct a function $\tilde{f}$ and a random variable $\tilde{E}$ such that

$$X = \tilde{f}(Y, \tilde{E}), \qquad Y \perp\!\!\!\perp \tilde{E}.$$

The crucial idea is now to compare the *model complexities* and infer the least complex model to be the true causal direction. However, we use other complexity measures than Kolmogorov complexity, so that we can actually compute them.

This work: non-parametric causal discovery

The goal of this work is to avoid the restrictions on the model class for the causal mechanism.

However, in this way we loose identifiability:

Theorem

Given random variables X, Y, E and a function f such that

$$Y = f(X, E), \qquad X \perp\!\!\!\perp E,$$

we can always construct a function $\tilde{f}$ and a random variable $\tilde{E}$ such that

$$X = \tilde{f}(Y, \tilde{E}), \qquad Y \perp\!\!\!\perp \tilde{E}.$$

The crucial idea is now to compare the *model complexities* and infer the least complex model to be the true causal direction. However, we use other complexity measures than Kolmogorov complexity, so that we can actually compute them.

Bayesian model selection

Prefer the model with the highest *evidence*:

$$p(D \mid M) = \int p(D \mid \theta, M) p(\theta \mid M)\, d\theta,$$

which is a trade-off between the *likelihood* (goodness-of-fit) and the *prior* (model complexity). $(D = \text{data}, M = \text{model}, \theta = \text{model parameters})$

Basic idea

Causal discovery (for our "core problem") can be done simply by comparing the evidences for two models ($X \to Y$ and $Y \to X$).

Bayesian model selection and causal inference

Bayesian model selection

Prefer the model with the highest *evidence*:

$$p(D \mid M) = \int p(D \mid \theta, M) p(\theta \mid M) \, d\theta,$$

which is a trade-off between the *likelihood* (goodness-of-fit) and the *prior* (model complexity). $(D = \text{data}, M = \text{model}, \theta = \text{model parameters})$

Basic idea

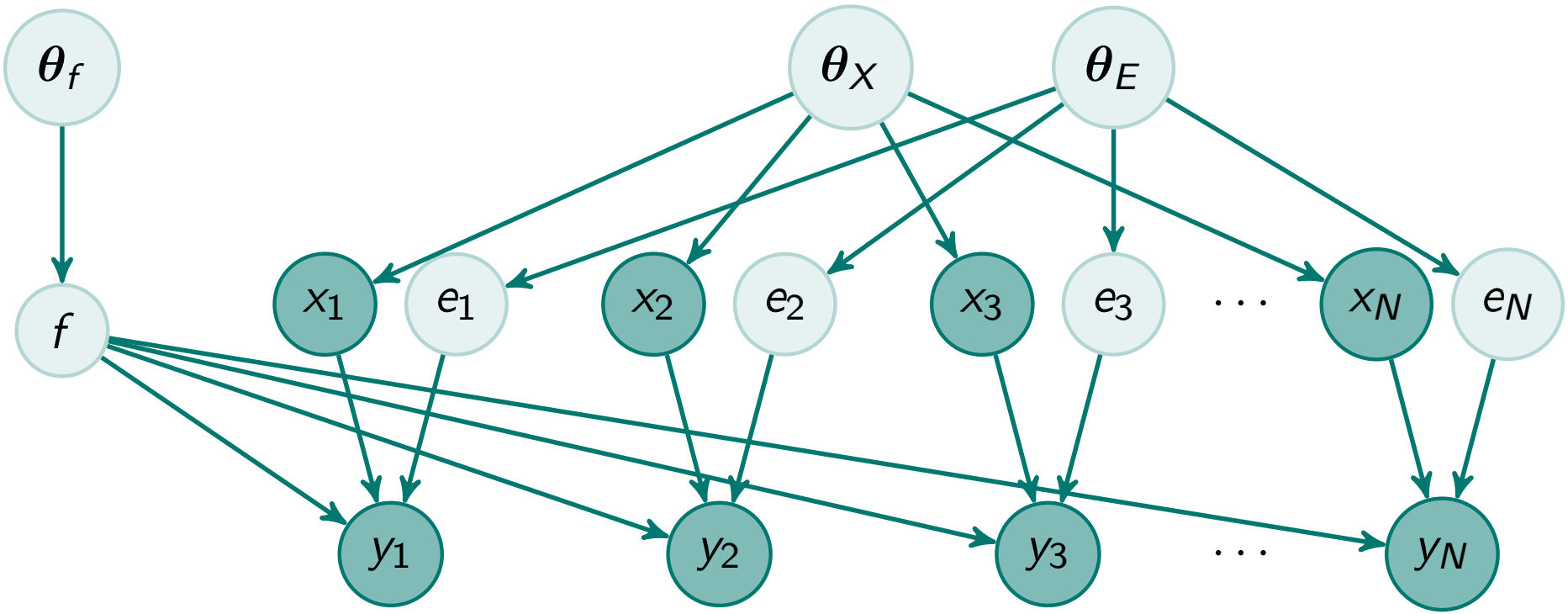
Causal discovery (for our "core problem") can be done simply by comparing the evidences for two models ($X \to Y$ and $Y \to X$).

How we model $X \to Y$

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y} \mid \mathbf{x})$$

$$= \int \left(\prod_{i=1}^{N} p(x_i \mid \boldsymbol{\theta}_X) \right) p(\boldsymbol{\theta}_X)\, d\boldsymbol{\theta}_X$$

$$\cdot \int \left(\prod_{i=1}^{N} \delta\big(y_i - f(x_i, e_i)\big) p(e_i \mid \boldsymbol{\theta}_E) \right) p(f \mid \boldsymbol{\theta}_f)\, d\mathbf{e}\, df \;\; p(\boldsymbol{\theta}_E) d\boldsymbol{\theta}_E \;\; p(\boldsymbol{\theta}_f) d\boldsymbol{\theta}_f$$



Part II

The nasty technical details

Choosing the priors

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ denote the N values for X, Y, and $\mathbf{e} \in \mathbb{R}^N$ the N latent variables.

In order to completely specify the generative model $X \to Y$, we need to choose various priors:

- the prior distribution on the inputs $\mathbf{x}$ (parameterized by θ_X)
- the prior distribution on the latents $\mathbf{e}$ (parameterized by θ_E)
- the prior distribution on the function f (parameterized by θ_f)

Note: even in the discrete, finite case (if X, Y, E can only take a finite number of values), it is not obvious whether the choice of the priors becomes less important as $N \to \infty$: the number of observations ($2N$) is of the same order as the number of unknowns ($N + K$ for some constant K).

Choosing the priors

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ denote the N values for X, Y, and $\mathbf{e} \in \mathbb{R}^N$ the N latent variables.

In order to completely specify the generative model $X \to Y$, we need to choose various priors:

- the prior distribution on the inputs $\mathbf{x}$ (parameterized by θ_X)
- the prior distribution on the latents $\mathbf{e}$ (parameterized by θ_E)
- the prior distribution on the function f (parameterized by θ_f)

Note: even in the discrete, finite case (if X, Y, E can only take a finite number of values), it is not obvious whether the choice of the priors becomes less important as $N \to \infty$: the number of observations ($2N$) is of the same order as the number of unknowns ($N + K$ for some constant K).

Choosing the priors

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ denote the N values for X, Y, and $\mathbf{e} \in \mathbb{R}^N$ the N latent variables.

In order to completely specify the generative model $X \to Y$, we need to choose various priors:

- the prior distribution on the inputs $\mathbf{x}$ (parameterized by θ_X)
- the prior distribution on the latents $\mathbf{e}$ (parameterized by θ_E)
- the prior distribution on the function f (parameterized by θ_f)

Note: even in the discrete, finite case (if X, Y, E can only take a finite number of values), it is not obvious whether the choice of the priors becomes less important as $N \to \infty$: the number of observations ($2N$) is of the same order as the number of unknowns ($N + K$ for some constant K).

Choosing the priors: the input distribution $p(X \mid \theta_X)$

For $p(X \mid \theta_X)$, we currently use a Gaussian Mixture Model

$$p(X \mid \boldsymbol{\theta}_X) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2)$$

with hyperparameters $\boldsymbol{\theta}_X = (k, \alpha_1, \ldots, \alpha_k, \mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k)$, with an improper Dirichlet prior (with parameters $(-1, -1, \ldots, -1)$) on the component weights $\boldsymbol{\alpha}$ and a flat prior on the component parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$.

Instead of integrating over θ_X, we maximize over θ_X, using a particular penalty for k, the number of mixture components, which is derived using the MML principle. We use an algorithm proposed by Figueiredo and Jain.[1]

[1] Figueiredo & Jain, *Unsupervised learning of finite mixture models*, TPAMI 2002

Choosing the priors: the input distribution $p(X \mid \theta_X)$

For $p(X \mid \theta_X)$, we currently use a Gaussian Mixture Model

$$p(X \mid \boldsymbol{\theta}_X) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2)$$

with hyperparameters $\boldsymbol{\theta}_X = (k, \alpha_1, \ldots, \alpha_k, \mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k)$, with an improper Dirichlet prior (with parameters $(-1, -1, \ldots, -1)$) on the component weights $\boldsymbol{\alpha}$ and a flat prior on the component parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$.

Instead of integrating over $\boldsymbol{\theta}_X$, we maximize over $\boldsymbol{\theta}_X$, using a particular penalty for k, the number of mixture components, which is derived using the MML principle. We use an algorithm proposed by Figueiredo and Jain.[1]

[1] Figueiredo & Jain, *Unsupervised learning of finite mixture models*, TPAMI 2002

Choosing the priors: the noise distribution $p(E \mid \theta_E)$

For $p(E \mid \theta_E)$, we simply use a standard-normal distribution:

$$p(E \mid \theta_E) = \mathcal{N}(0, 1),$$

so there are no hyperparameters.

This may look like a severe restriction on the model, but is not as bad as it seems: in general, there exists a function g such that

$$E = g(\overline{E}), \overline{E} \sim \mathcal{N}(0, 1).$$

Then $Y = f(X, E)$ corresponds with $Y = \overline{f}(X, \overline{E})$, where $\overline{f} := f\big(\cdot, g(\cdot)\big)$.

However, this assumption does introduce a dependency between $p(E)$ and f, which apparently violates our basic assumption $p(X, E) \perp\!\!\!\perp f$.

Choosing the priors: the noise distribution $p(E \mid \theta_E)$

For $p(E \mid \theta_E)$, we simply use a standard-normal distribution:

$$p(E \mid \theta_E) = \mathcal{N}(0, 1),$$

so there are no hyperparameters.

This may look like a severe restriction on the model, but is not as bad as it seems: in general, there exists a function g such that

$$E = g(\overline{E}), \overline{E} \sim \mathcal{N}(0, 1).$$

Then $Y = f(X, E)$ corresponds with $Y = \overline{f}(X, \overline{E})$, where $\overline{f} := f\big(\cdot, g(\cdot)\big)$.

However, this assumption does introduce a dependency between $p(E)$ and f, which apparently violates our basic assumption $p(X, E) \perp\!\!\!\perp f$.

Choosing the priors: the noise distribution $p(E \mid \theta_E)$

For $p(E \mid \theta_E)$, we simply use a standard-normal distribution:

$$p(E \mid \theta_E) = \mathcal{N}(0, 1),$$

so there are no hyperparameters.

This may look like a severe restriction on the model, but is not as bad as it seems: in general, there exists a function g such that

$$E = g(\overline{E}), \overline{E} \sim \mathcal{N}(0, 1).$$

Then $Y = f(X, E)$ corresponds with $Y = \overline{f}(X, \overline{E})$, where $\overline{f} := f\big(\cdot, g(\cdot)\big)$.

However, this assumption does introduce a dependency between $p(E)$ and f, which apparently violates our basic assumption $p(X, E) \perp\!\!\!\perp f$.

Choosing the priors: the function prior $p(f \mid \theta_f)$

For $p(f \mid \theta_f)$, we currently take a Gaussian Process with zero mean function and squared-exponential covariance function:

$$k\big((x,e),(x',e')\big) = \lambda_Y^2 \exp\left(-\frac{(x-x')^2}{2\lambda_X^2}\right) \exp\left(-\frac{(e-e')^2}{2\lambda_E^2}\right)$$

where $\boldsymbol{\theta}_f = (\lambda_X, \lambda_Y, \lambda_E)$ are length-scale parameters.

We currently preprocess the data $\mathbf{x}$, $\mathbf{y}$ such that they have mean 0 and variance 1. For the prior on the length-scale parameters, we use a broad Gamma distribution: $\lambda \sim \Gamma(30, 0.5)$. The only reason for doing this is a numerical one: if the length scales become too large, the kernel matrix $\mathbf{K}_{ij} = k\big((x_i,e_i),(x_j,e_j)\big)$ will become difficult to handle numerically.

Instead of integrating over θ_f, we maximize over θ_f.

Choosing the priors: the function prior $p(f \mid \theta_f)$

For $p(f \mid \theta_f)$, we currently take a Gaussian Process with zero mean function and squared-exponential covariance function:

$$k\big((x,e),(x',e')\big) = \lambda_Y^2 \exp\left(-\frac{(x-x')^2}{2\lambda_X^2}\right) \exp\left(-\frac{(e-e')^2}{2\lambda_E^2}\right)$$

where $\theta_f = (\lambda_X, \lambda_Y, \lambda_E)$ are length-scale parameters.

We currently preprocess the data $\mathbf{x}$, $\mathbf{y}$ such that they have mean 0 and variance 1. For the prior on the length-scale parameters, we use a broad Gamma distribution: $\lambda \sim \Gamma(30, 0.5)$. The only reason for doing this is a numerical one: if the length scales become too large, the kernel matrix $\mathbf{K}_{ij} = k\big((x_i,e_i),(x_j,e_j)\big)$ will become difficult to handle numerically.

Instead of integrating over θ_f, we maximize over θ_f.

Back to our integral

By maximizing over the hyperparameters instead of integrating over them, we have reduced the computational problem to solving:

$$p(\mathbf{x}, \mathbf{y} \,|\, X \to Y) \approx \max_{\boldsymbol{\theta}_X} \left(\prod_{i=1}^{N} p(x_i \,|\, \boldsymbol{\theta}_X) \right) p(\boldsymbol{\theta}_X)$$

$$\cdot \max_{\boldsymbol{\theta}_f} p(\boldsymbol{\theta}_f) \int \left(\prod_{i=1}^{N} \delta\big(y_i - f(x_i, e_i)\big) p(e_i) \right) p(f \,|\, \boldsymbol{\theta}_f) \, d\mathbf{e} \, df$$

The first maximization problem is solved numerically by using a modified EM algorithm written by Figueiredo and Jain, and gives

$$\max_{\boldsymbol{\theta}_X} \left(\prod_{i=1}^{N} p(x_i \,|\, \boldsymbol{\theta}_X) \right) p(\boldsymbol{\theta}_X) = \exp\big(- \mathcal{L}_{\mathrm{MML}}(\mathbf{x}) \big)$$

From now on, we focus on the second part.

Back to our integral

By maximizing over the hyperparameters instead of integrating over them, we have reduced the computational problem to solving:

$$p(\mathbf{x}, \mathbf{y} \,|\, X \to Y) \approx \max_{\boldsymbol{\theta}_X} \left(\prod_{i=1}^{N} p(x_i \,|\, \boldsymbol{\theta}_X) \right) p(\boldsymbol{\theta}_X)$$
$$\cdot \max_{\boldsymbol{\theta}_f} p(\boldsymbol{\theta}_f) \int \left(\prod_{i=1}^{N} \delta\big(y_i - f(x_i, e_i)\big) p(e_i) \right) p(f \,|\, \boldsymbol{\theta}_f) \, d\mathbf{e} \, df$$

The first maximization problem is solved numerically by using a modified EM algorithm written by Figueiredo and Jain, and gives

$$\max_{\boldsymbol{\theta}_X} \left(\prod_{i=1}^{N} p(x_i \,|\, \boldsymbol{\theta}_X) \right) p(\boldsymbol{\theta}_X) = \exp\big(- \mathcal{L}_{\mathrm{MML}}(\mathbf{x}) \big)$$

From now on, we focus on the second part.

Integrating over the latent variables $\mathbf{e}$

We first integrate out the latent variables $\mathbf{e}$, using the Dirac delta function calculus:

$$\int \left(\prod_{i=1}^{N} \delta\big(y_i - f(x_i, e_i)\big) p(e_i) \right) p(f \mid \boldsymbol{\theta}_f)\, d\mathbf{e}\, df = \int \frac{p(\mathbf{e}_0(f))}{J\big(\mathbf{e}_0(f)\big)}\; p(f \mid \boldsymbol{\theta}_f)\, df$$

where

$$J\big(\mathbf{e}_0(f)\big) = \det \left| \frac{\partial f}{\partial \mathbf{e}}\big(\mathbf{x}, \mathbf{e}_0(f)\big) \right| = \prod_{i=1}^{N} \left| \frac{\partial f}{\partial e}\big(x_i, (\mathbf{e}_0(f))_i\big) \right|$$

is the absolute value of the determinant of the Jacobian and $\mathbf{e}_0(f)$ is the unique vector satisfying $f\big(x_i, \big(\mathbf{e}_0(f)\big)_i\big) = y_i$.

Here we assumed that for each x, the function $f_x : e \mapsto f(x, e)$ is invertible. Although this is not a restriction on the model class, this assumption is not compatible with our Gaussian Process prior on f.

Integrating over the latent variables $\mathbf{e}$

We first integrate out the latent variables $\mathbf{e}$, using the Dirac delta function calculus:

$$\int \left(\prod_{i=1}^{N} \delta\big(y_i - f(x_i, e_i)\big) p(e_i) \right) p(f \mid \boldsymbol{\theta}_f)\, d\mathbf{e}\, df = \int \frac{p(\mathbf{e}_0(f))}{J\big(\mathbf{e}_0(f)\big)}\; p(f \mid \boldsymbol{\theta}_f)\, df$$

where

$$J\big(\mathbf{e}_0(f)\big) = \det \left| \frac{\partial f}{\partial \mathbf{e}} \big(\mathbf{x}, \mathbf{e}_0(f)\big) \right| = \prod_{i=1}^{N} \left| \frac{\partial f}{\partial e} \big(x_i, (\mathbf{e}_0(f))_i\big) \right|$$

is the absolute value of the determinant of the Jacobian and $\mathbf{e}_0(f)$ is the unique vector satisfying $f\big(x_i, \big(\mathbf{e}_0(f)\big)_i\big) = y_i$.

Here we assumed that for each x, the function $f_x : e \mapsto f(x, e)$ is invertible. Although this is not a restriction on the model class, this assumption is not compatible with our Gaussian Process prior on f.

"Integrating over f"

We then evaluate the remaining integral:

$$\int \frac{p(\mathbf{e}_0(f))p(f \mid \boldsymbol{\theta}_f)}{J\big(\mathbf{e}_0(f)\big)}\, df = \int \mathcal{N}\big(\mathbf{y} \mid \mathbf{0}, \mathbf{K}\big)\mathcal{N}\big(\mathbf{e}_0(f) \mid \mathbf{0}, \mathbf{I}\big)J^{-1}\big(\mathbf{e}_0(f)\big)\, d\mathbf{e}_0(f)$$

where $\mathbf{K} = k\big(\mathbf{x}, \mathbf{e}_0(f)\big)$.

We then approximate this integral by maximizing over $\mathbf{e}_0(f)$ (henceforth simply denoted $\mathbf{e}$):

$$\cdots \approx \max_{\mathbf{e}} \Big(\mathcal{N}\big(\mathbf{y} \mid \mathbf{0}, \mathbf{K}\big)\mathcal{N}\big(\mathbf{e} \mid \mathbf{0}, \mathbf{I}\big)J^{-1}\big(\mathbf{e}\big)\Big)$$

and using the mean predicted partial derivatives of the GP f:

$$\frac{\partial f}{\partial e}(x_i, e_i) = \frac{\partial k}{\partial e}\big((x_i, e_i), (\mathbf{x}, \mathbf{e})\big)\mathbf{K}^{-1}\mathbf{y}$$

for approximating the Jacobian.

"Integrating over f"

We then evaluate the remaining integral:

$$\int \frac{p(\mathbf{e}_0(f))p(f \mid \boldsymbol{\theta}_f)}{J\big(\mathbf{e}_0(f)\big)}\, df = \int \mathcal{N}\big(\mathbf{y} \mid \mathbf{0}, \mathbf{K}\big)\mathcal{N}\big(\mathbf{e}_0(f) \mid \mathbf{0}, \mathbf{I}\big)J^{-1}\big(\mathbf{e}_0(f)\big)\, d\mathbf{e}_0(f)$$

where $\mathbf{K} = k\big(\mathbf{x}, \mathbf{e}_0(f)\big)$.

We then approximate this integral by maximizing over $\mathbf{e}_0(f)$ (henceforth simply denoted $\mathbf{e}$):

$$\cdots \approx \max_{\mathbf{e}} \Big(\mathcal{N}\big(\mathbf{y} \mid \mathbf{0}, \mathbf{K}\big)\mathcal{N}\big(\mathbf{e} \mid \mathbf{0}, \mathbf{I}\big)J^{-1}\big(\mathbf{e}\big)\Big)$$

and using the mean predicted partial derivatives of the GP f:

$$\frac{\partial f}{\partial e}(x_i, e_i) = \frac{\partial k}{\partial e}\big((x_i, e_i), (\mathbf{x}, \mathbf{e})\big)\mathbf{K}^{-1}\mathbf{y}$$

for approximating the Jacobian.

Integrating or optimizing latent variables?

Summarizing, we first integrated out the latent variables and then optimized over control points of the GP.

Alternatively, one might start with integrating out the GP exactly, and then optimize over the latent variables, similarly to what is usually done in GPLVMs (Gaussian Process Latent Variable Models).

However, we believe that for the purpose of causal discovery, the latter approach would not work well. The reason is that when optimizing over $\mathbf{e}$, the result is often quite dependent on $\mathbf{x}$, which violates our basic assumption that $X \perp\!\!\!\perp E$. Indeed, our approach seems more related to (nonlinear) ICA then to PCA (which is related to GPLVMs).

Integrating or optimizing latent variables?

Summarizing, we first integrated out the latent variables and then optimized over control points of the GP.

Alternatively, one might start with integrating out the GP exactly, and then optimize over the latent variables, similarly to what is usually done in GPLVMs (Gaussian Process Latent Variable Models).

However, we believe that for the purpose of causal discovery, the latter approach would not work well. The reason is that when optimizing over $\mathbf{e}$, the result is often quite dependent on $\mathbf{x}$, which violates our basic assumption that $X \perp\!\!\!\perp E$. Indeed, our approach seems more related to (nonlinear) ICA then to PCA (which is related to GPLVMs).

Integrating or optimizing latent variables?

Summarizing, we first integrated out the latent variables and then optimized over control points of the GP.

Alternatively, one might start with integrating out the GP exactly, and then optimize over the latent variables, similarly to what is usually done in GPLVMs (Gaussian Process Latent Variable Models).

However, we believe that for the purpose of causal discovery, the latter approach would not work well. The reason is that when optimizing over $\mathbf{e}$, the result is often quite dependent on $\mathbf{x}$, which violates our basic assumption that $X \perp\!\!\!\perp E$. Indeed, our approach seems more related to (nonlinear) ICA then to PCA (which is related to GPLVMs).

The final objective function

Thus our final optimization problem is:

$$\min_{\boldsymbol{\lambda},\mathbf{e}} \left(-\log p(\boldsymbol{\lambda}) - \log \mathcal{N}\left(\mathbf{y} \,|\, \mathbf{0}, \mathbf{K}\right) - \log \mathcal{N}\left(\mathbf{e} \,|\, \mathbf{0}, \mathbf{I}\right) + \sum_{i=1}^{N} \log \left| \frac{\partial f}{\partial e}(x_i, e_i) \right| \right)$$

where

$$\mathbf{K} = k\left(\mathbf{x}, \mathbf{e}\right)$$

and

$$\frac{\partial f}{\partial e}(x_i, e_i) = \frac{\partial k}{\partial e}\left((x_i, e_i), (\mathbf{x}, \mathbf{e})\right)\mathbf{K}^{-1}\mathbf{y}$$

Problems

There are still two major issues here:

1. if $\frac{\partial f}{\partial e}$ becomes 0 for some (x_i, e_i), the objective function becomes $-\infty$;
2. the kernel matrix $\mathbf{K}$ is extremely ill-conditioned.

The final objective function

Thus our final optimization problem is:

$$\min_{\boldsymbol{\lambda},\mathbf{e}} \left(-\log p(\boldsymbol{\lambda}) - \log \mathcal{N}\left(\mathbf{y} \,|\, \mathbf{0}, \mathbf{K}\right) - \log \mathcal{N}\left(\mathbf{e} \,|\, \mathbf{0}, \mathbf{I}\right) + \sum_{i=1}^{N} \log \left| \frac{\partial f}{\partial e}(x_i, e_i) \right| \right)$$

where

$$\mathbf{K} = k\left(\mathbf{x}, \mathbf{e}\right)$$

and

$$\frac{\partial f}{\partial e}(x_i, e_i) = \frac{\partial k}{\partial e}\left((x_i, e_i), (\mathbf{x}, \mathbf{e})\right)\mathbf{K}^{-1}\mathbf{y}$$

Problems

There are still two major issues here:

1. if $\frac{\partial f}{\partial e}$ becomes 0 for some (x_i, e_i), the objective function becomes $-\infty$;
2. the kernel matrix $\mathbf{K}$ is extremely ill-conditioned.

Regularization and initialization

To deal with the first problem, we regularize our objective function as follows:

1. We approximate $|\log| \, x \approx \log \sqrt{x^2 + \epsilon}$ with $\epsilon \ll 1$ (using $\epsilon = 10^{-3}$).
2. We implemented a log barrier that heavily penalized negative values of $\frac{\partial f}{\partial e}(\hat{x}_i, e_i)$. This was done to avoid sign flips of these terms that would violate the invertability assumption.
3. We added a tiny amount of additive $\mathcal{N}(0, \sigma^2)$-noise to each y_i-value, which is equivalent to replacing $\mathbf{K}$ by $\mathbf{K} + \sigma^2 \mathbf{I}$ (using $\sigma = 10^{-5}$).

Further, we initialize $\mathbf{e}$ with additive noise models. The main reason is that in an additive noise model, the $\frac{\partial f}{\partial e}(x_i, e_i)$ are all positive and constant. This initialization effectively leads to a solution that satisfies the invertability assumption that we made.

Regularization and initialization

To deal with the first problem, we regularize our objective function as follows:

1. We approximate $|\log| \, x \approx \log \sqrt{x^2 + \epsilon}$ with $\epsilon \ll 1$ (using $\epsilon = 10^{-3}$).
2. We implemented a log barrier that heavily penalized negative values of $\frac{\partial f}{\partial e}(\hat{x}_i, e_i)$. This was done to avoid sign flips of these terms that would violate the invertability assumption.
3. We added a tiny amount of additive $\mathcal{N}(0, \sigma^2)$-noise to each y_i-value, which is equivalent to replacing $\mathbf{K}$ by $\mathbf{K} + \sigma^2 \mathbf{I}$ (using $\sigma = 10^{-5}$).

Further, we initialize $\mathbf{e}$ with additive noise models. The main reason is that in an additive noise model, the $\frac{\partial f}{\partial e}(x_i, e_i)$ are all positive and constant. This initialization effectively leads to a solution that satisfies the invertability assumption that we made.

References

- See also our NIPS paper *Probabilistic latent variable models for distinguishing between cause and effect*,
  J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, B. Schölkopf,
  Advances in Neural Information Processing Systems 23 (NIPS*2010)
- See also our code package (also distributed as part of the Causality Challenge):
  `http://webdav.tuebingen.mpg.de/causality/nips2010-gpi-code.tar.gz`