On martingales, causality, identifiability and model selection

Alexander Sokol

PhD defense, 14 February, 2014

Agenda

1. A motivating problem
2. Exponential martingales
3. Simplified proofs in the general theory of processes
4. A notion of causality for SDEs
5. Identifiability in ICA
6. Degrees of freedom in nonlinear regression

A motivating problem

Consider the following situation:

- A patient has a disease
- The disease may lead to an event (such as death)
- A treatment can be given, depending on health condition
- The patient may at any time be censored (e.g., recover)

A motivating problem

Consider the following situation:

- A patient has a disease
- The disease may lead to an event (such as death)
- A treatment can be given, depending on health condition
- The patient may at any time be censored (e.g., recover)

We wish to model this scenario and **estimate the causal effect of the treatment on the time to the event.**

A motivating problem

We may use counting processes to set up a modeling framework (Røysland 2010). Assume:

- N^A, N^C, N^D are univariate counting processes
- N^L is a multivariate counting process

A motivating problem

We may use counting processes to set up a modeling framework (Røysland 2010). Assume:

- N^A, N^C, N^D are univariate counting processes
- N^L is a multivariate counting process

and

- $T_A = \inf\{t \geq 0 \mid N_t^A = 1\}$, similarly for T_C and T_D
- $A_t = \int_0^t 1_{(s \leq T_A)} \, \mathrm{d}N_s^A$, similarly for C and D
- $L_t = L_0 + \int_0^t H_s^L \, \mathrm{d}N_s^L$

A motivating problem

We may use counting processes to set up a modeling framework (Røysland 2010). Assume:

- N^A, N^C, N^D are univariate counting processes
- N^L is a multivariate counting process

and

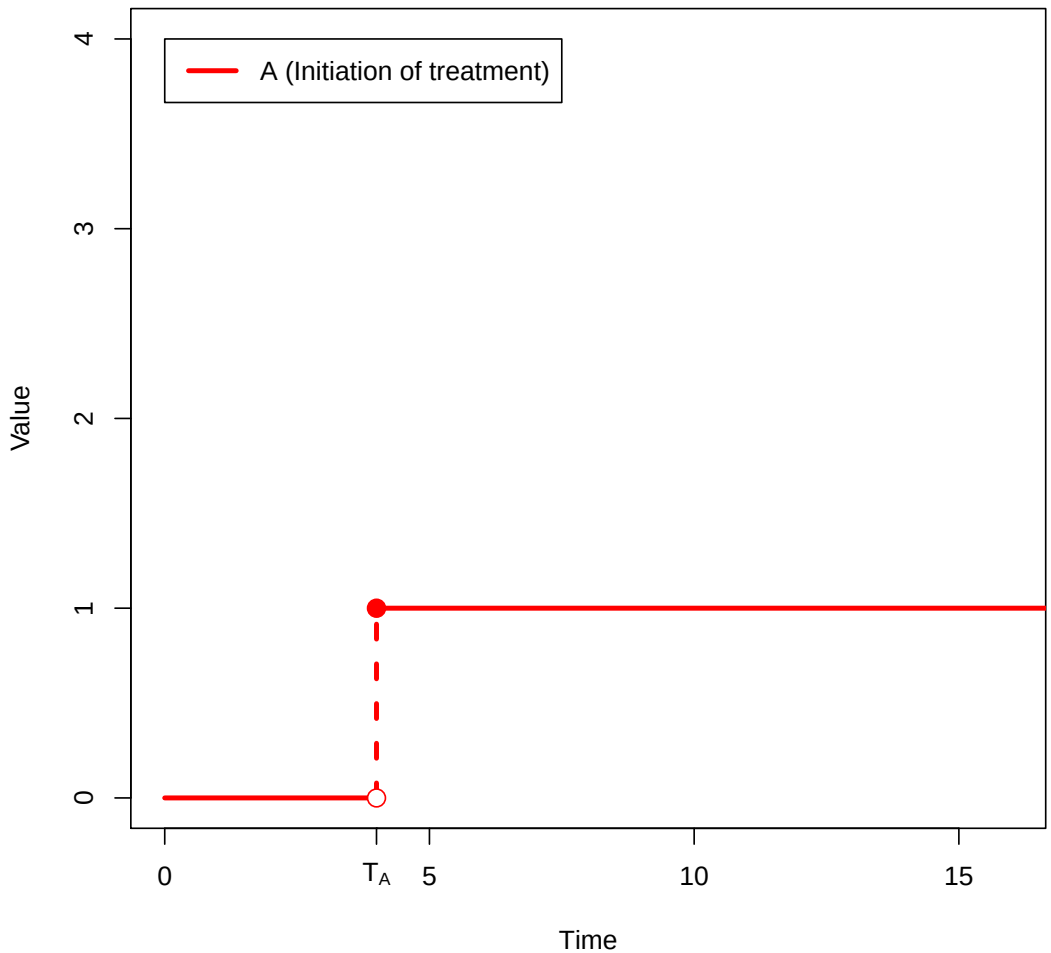
- $T_A = \inf\{t \geq 0 \mid N_t^A = 1\}$, similarly for T_C and T_D
- $A_t = \int_0^t 1_{(s \leq T_A)} \, \mathrm{d}N_s^A$, similarly for C and D
- $L_t = L_0 + \int_0^t H_s^L \, \mathrm{d}N_s^L$

Here:

- A is the counting process for initiation of treatment
- C is the counting process for censoring
- D is the counting process for the event
- L measures the patients multivariate health condition

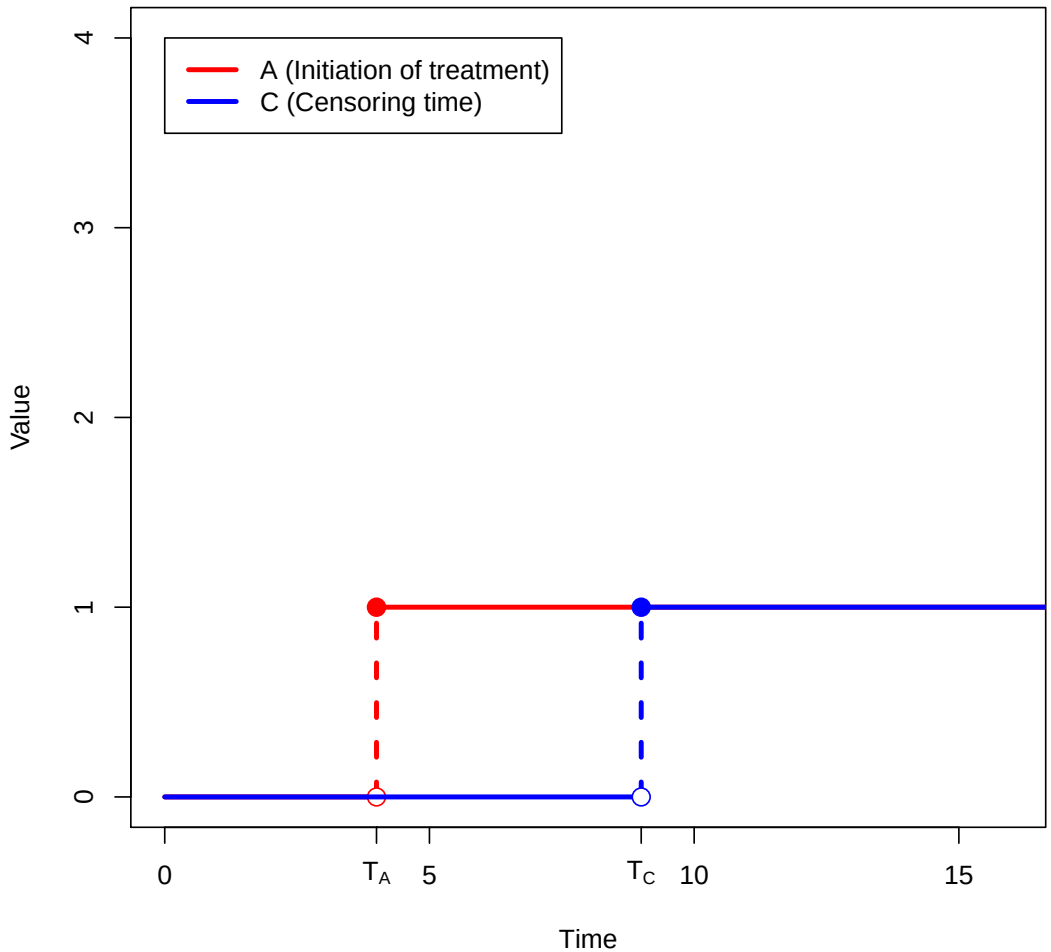
A motivating problem

Illustration of the model setup:



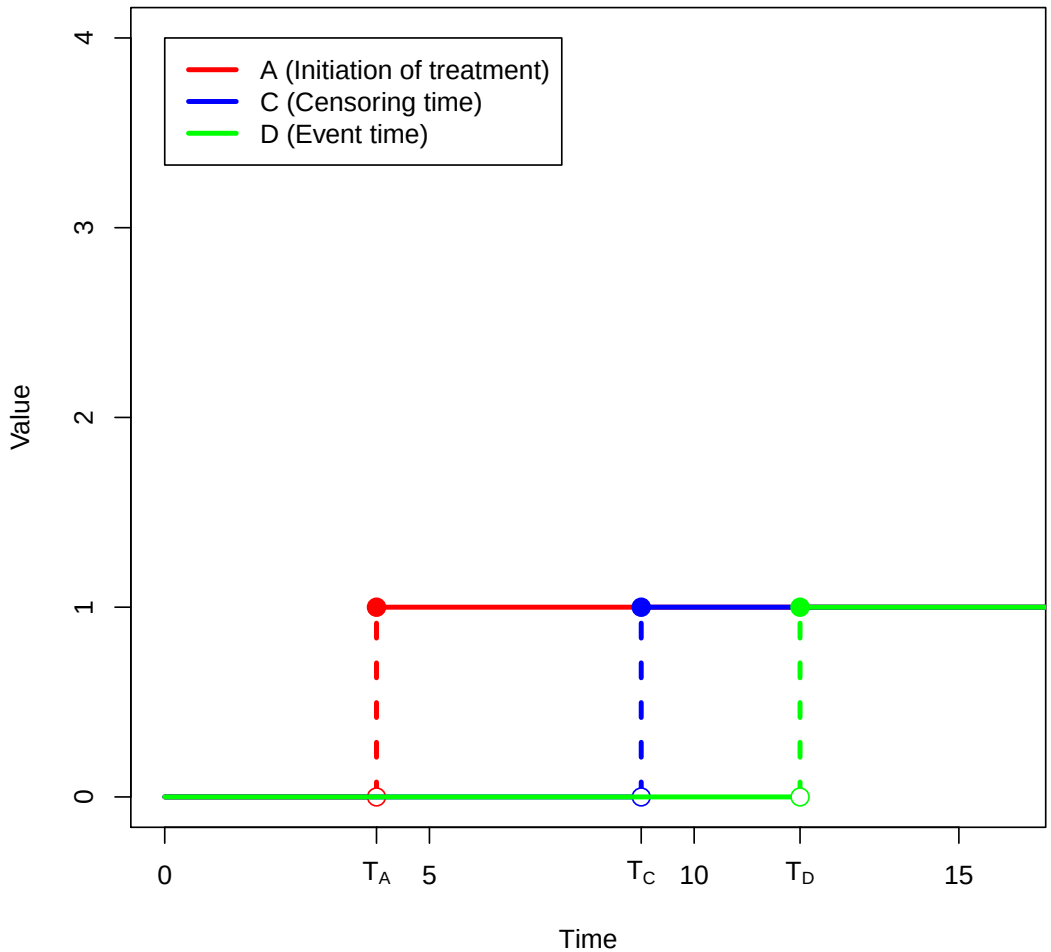
A motivating problem

Illustration of the model setup:



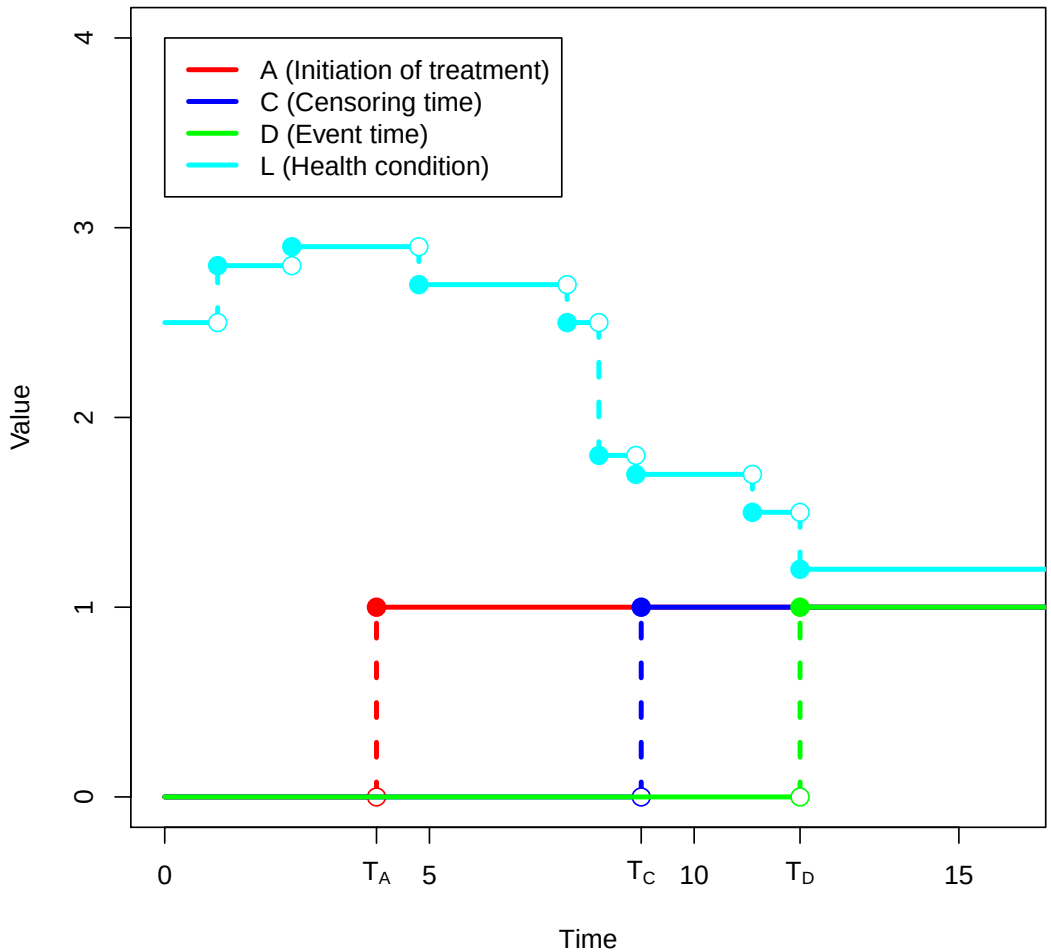
A motivating problem

Illustration of the model setup:



A motivating problem

Illustration of the model setup:



A motivating problem

The behaviour of the model is determined by the intensities of the counting processes N^A, N^C, N^D and N^L. Let N^A have intensity λ^A, etc.

A motivating problem

The behaviour of the model is determined by the intensities of the counting processes N^A, N^C, N^D and N^L. Let N^A have intensity λ^A, etc.

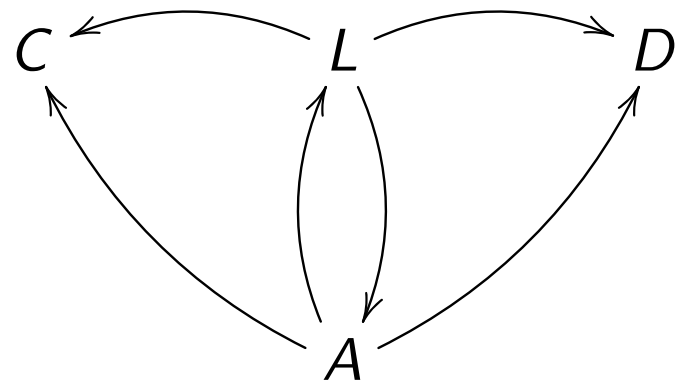
The intensities are stochastic. We allow, say, λ^D, to depend on the processes A and L. This corresponds to that the probability of the event occurring **depends on both treatment status and health condition**.

A motivating problem

The behaviour of the model is determined by the intensities of the counting processes N^A, N^C, N^D and N^L. Let N^A have intensity λ^A, etc.

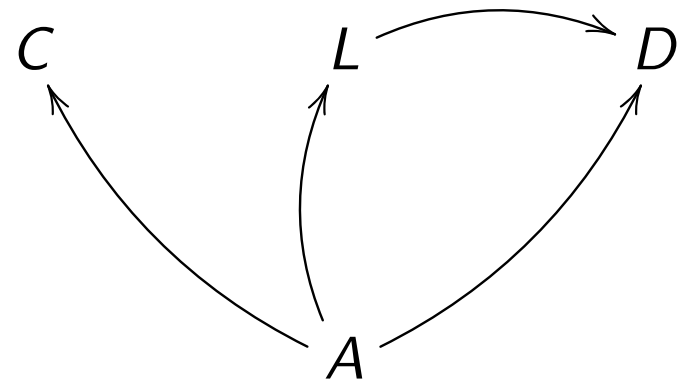
The intensities are stochastic. We allow, say, λ^D, to depend on the processes A and L. This corresponds to that the probability of the event occurring **depends on both treatment status and health condition**.

We put restrictions on the dependencies, corresponding to the local independence graph:



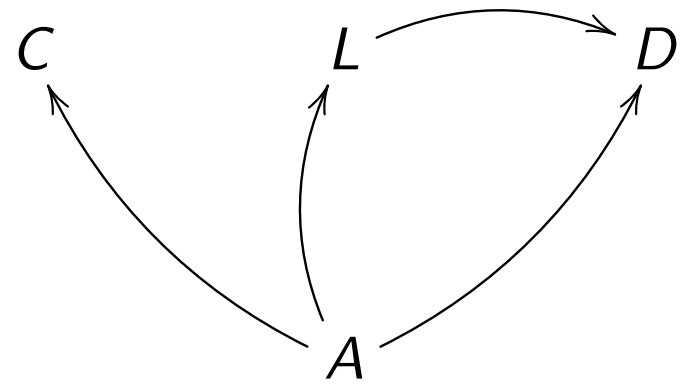
A motivating problem

Røysland proposes estimation in this model for unobserved L by constructing a **randomized trial measure** and carrying out estimation under this measure. A randomized trial measure is a probability measure where the local independence graph is:



A motivating problem

Røysland proposes estimation in this model for unobserved L by constructing a **randomized trial measure** and carrying out estimation under this measure. A randomized trial measure is a probability measure where the local independence graph is:



Røysland shows that this facilitates estimation of the causal effect of A on D in the marginal model where L is unobserved.

A motivating problem

We can now formulate two questions:

A motivating problem

We can now formulate two questions:

1. The estimation methodology outlined above requires the existence of randomized trial measures. **What are sufficient criteria ensuring this existence?**

A motivating problem

We can now formulate two questions:

1. The estimation methodology outlined above requires the existence of randomized trial measures. **What are sufficient criteria ensuring this existence?**
2. Our modeling discussion involved notions of causality. **How do we formalize such notions in a continuous-time framework?**

Exponential martingales

Recall that our motivating question was the existence of randomized trial measures, defined through the local independence graph.

Exponential martingales

Recall that our motivating question was the existence of randomized trial measures, defined through the local independence graph.

The local independence graph is a description of the dependency relationships of the intensities. Thus, we need to construct **counting process distributions with particular intensities.**

Exponential martingales

Recall that our motivating question was the existence of randomized trial measures, defined through the local independence graph.

The local independence graph is a description of the dependency relationships of the intensities. Thus, we need to construct **counting process distributions with particular intensities.**

From a more abstract perspective, our problem is thus:

Problem I. Assume given processes λ and μ, and a multidimensional counting process N with intensity λ. When is it possible to construct from this a multidimensional counting process with intensity μ?

Exponential martingales

To give a solution to this problem, we assume given:

- A filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$
- Predictable d-dimensional processes λ and μ on the probability space
- A counting process N with intensity λ on the probability space

Exponential martingales

To give a solution to this problem, we assume given:

- A filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$
- Predictable d-dimensional processes λ and μ on the probability space
- A counting process N with intensity λ on the probability space

And we define:

- $M_t^i = N_t^i - \int_0^t \lambda_s^i \, \mathrm{d}s$
- $\gamma_t^i = \mu_t^i / \lambda_t^i$
- $H_t^i = \gamma_t^i - 1$

Exponential martingales

To give a solution to this problem, we assume given:

- A filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$
- Predictable d-dimensional processes λ and μ on the probability space
- A counting process N with intensity λ on the probability space

And we define:

- $M_t^i = N_t^i - \int_0^t \lambda_s^i \, \mathrm{d}s$
- $\gamma_t^i = \mu_t^i / \lambda_t^i$
- $H_t^i = \gamma_t^i - 1$

All integrals are vector integrals, meaning that

$$(H \cdot M)_t = \sum_{i=1}^{d} \int_0^t H_s^i \, \mathrm{d}M_s^i.$$

Exponential martingales

Definition. Given a local martingale L with initial value zero, the exponential martingale $\mathcal{E}(L)$ of L is given as the solution to the SDE

$$\mathrm{d}\mathcal{E}(L)_t = \mathcal{E}(L_{t-})\,\mathrm{d}L_t.$$

Exponential martingales

Definition. Given a local martingale L with initial value zero, the exponential martingale $\mathcal{E}(L)$ of L is given as the solution to the SDE

$$\mathrm{d}\mathcal{E}(L)_t = \mathcal{E}(L_{t-})\,\mathrm{d}L_t.$$

The following lemma shows that the martingale property of $\mathcal{E}(H \cdot M)$ can be used to construct a counting process with intensity μ by a change of measure.

Lemma. Assume that $\mathcal{E}(H \cdot M)$ is a martingale. Let $t \geq 0$ and let Q_t have Radon-Nikodym derivative $\mathcal{E}(H \cdot M)_t$. Then N has intensity μ on $[0, t]$ under Q_t.

Exponential martingales

Thus, by the lemma, it suffices to consider:

Problem II. When is $\mathcal{E}(H \cdot M)$ a martingale?

Exponential martingales

Thus, by the lemma, it suffices to consider:

Problem II. When is $\mathcal{E}(H \cdot M)$ a martingale?

We obtained the following sufficient criterion:

Theorem. Assume that there is $\varepsilon > 0$ such that for $0 \leq u \leq t$ with $t - u \leq \varepsilon$, it holds that one of the following two conditions are satisfied:

$$E \exp\left(\sum_{i=1}^{d} \int_u^t (\gamma_s^i \log \gamma_s^i - (\gamma_s^i - 1))\lambda_s^i \, \mathrm{d}s\right) < \infty \quad \text{or}$$

$$E \exp\left(\sum_{i=1}^{d} \int_u^t \lambda_s^i \, \mathrm{d}s + \int_u^t \log_+ \gamma_s^i \, \mathrm{d}N_s^i\right) < \infty.$$

Then $\mathcal{E}(H \cdot M)$ is a martingale.

Exponential martingales

For the case of a homogeneous Poisson process, meaning that $\lambda = 1$, we get the following corollary.

Corollary. Assume that there is $\varepsilon > 0$ such that for $0 \leq u \leq t$ with $t - u \leq \varepsilon$, it holds that one of the following two conditions are satisfied:

$$E \exp\left(\sum_{i=1}^{d} \int_u^t \mu_s^i \log_+ \mu_s^i \, \mathrm{d}s\right) < \infty \quad \text{or}$$

$$E \exp\left(\sum_{i=1}^{d} \int_u^t \log_+ \mu_s^i \, \mathrm{d}N_s^i\right) < \infty.$$

Then $\mathcal{E}(H \cdot M)$ is a martingale.

Exponential martingales

This yields criteria for the existence of:

Exponential martingales

This yields criteria for the existence of:

- Randomized trial measures

Exponential martingales

This yields criteria for the existence of:

- Randomized trial measures
- Counting processes with intensity increasing affinely in N

Exponential martingales

This yields criteria for the existence of:

- Randomized trial measures
- Counting processes with intensity increasing affinely in N
- Counting processes with intensity given as transformations of SDEs

Exponential martingales

This yields criteria for the existence of:

- Randomized trial measures
- Counting processes with intensity increasing affinely in N
- Counting processes with intensity given as transformations of SDEs
- Some self-exciting counting processes

Exponential martingales

We also considered a more classical problem, namely:

Problem III. Given a local martingale M with initial value zero, when is $\mathcal{E}(M)$ a uniformly integrable martingale?

Exponential martingales

We also considered a more classical problem, namely:

Problem III. Given a local martingale M with initial value zero, when is $\mathcal{E}(M)$ a uniformly integrable martingale?

A classical result is:

Theorem (Novikov, 1972). Assume that M is continuous. If $E\exp(\frac{1}{2}[M]_\infty)$ is finite, then $\mathcal{E}(M)$ is a uniformly integrable martingale.

Exponential martingales

We also considered a more classical problem, namely:

Problem III. Given a local martingale M with initial value zero, when is $\mathcal{E}(M)$ a uniformly integrable martingale?

A classical result is:

Theorem (Novikov, 1972). Assume that M is continuous. If $E\exp(\frac{1}{2}[M]_\infty)$ is finite, then $\mathcal{E}(M)$ is a uniformly integrable martingale.

What happens when M is allowed to have jumps, with $\Delta M \geq -1$?

Exponential martingales

We also considered a more classical problem, namely:

Problem III. Given a local martingale M with initial value zero, when is $\mathcal{E}(M)$ a uniformly integrable martingale?

A classical result is:

Theorem (Novikov, 1972). Assume that M is continuous. If $E\exp(\frac{1}{2}[M]_\infty)$ is finite, then $\mathcal{E}(M)$ is a uniformly integrable martingale.

What happens when M is allowed to have jumps, with $\Delta M \geq -1$?

Theorem (Protter & Shimbo, 2008). If $E\exp(\frac{1}{2}\langle M^c\rangle_\infty + \langle M^d\rangle_\infty)$ is finite, then $\mathcal{E}(M)$ is a uniformly integrable martingale.

Exponential martingales

We showed the following. For $a > -1$ with $a \neq 0$, define

$$\alpha(a) = \frac{(1+a)\log(1+a) - a}{a^2}$$
$$\beta(a) = \frac{(1+a)\log(1+a) - a}{(1+a)a^2},$$

and extend to $[-1, \infty)$ by continuity.

Exponential martingales

We showed the following. For $a > -1$ with $a \neq 0$, define

$$\alpha(a) = \frac{(1+a)\log(1+a) - a}{a^2}$$
$$\beta(a) = \frac{(1+a)\log(1+a) - a}{(1+a)a^2},$$

and extend to $[-1, \infty)$ by continuity.

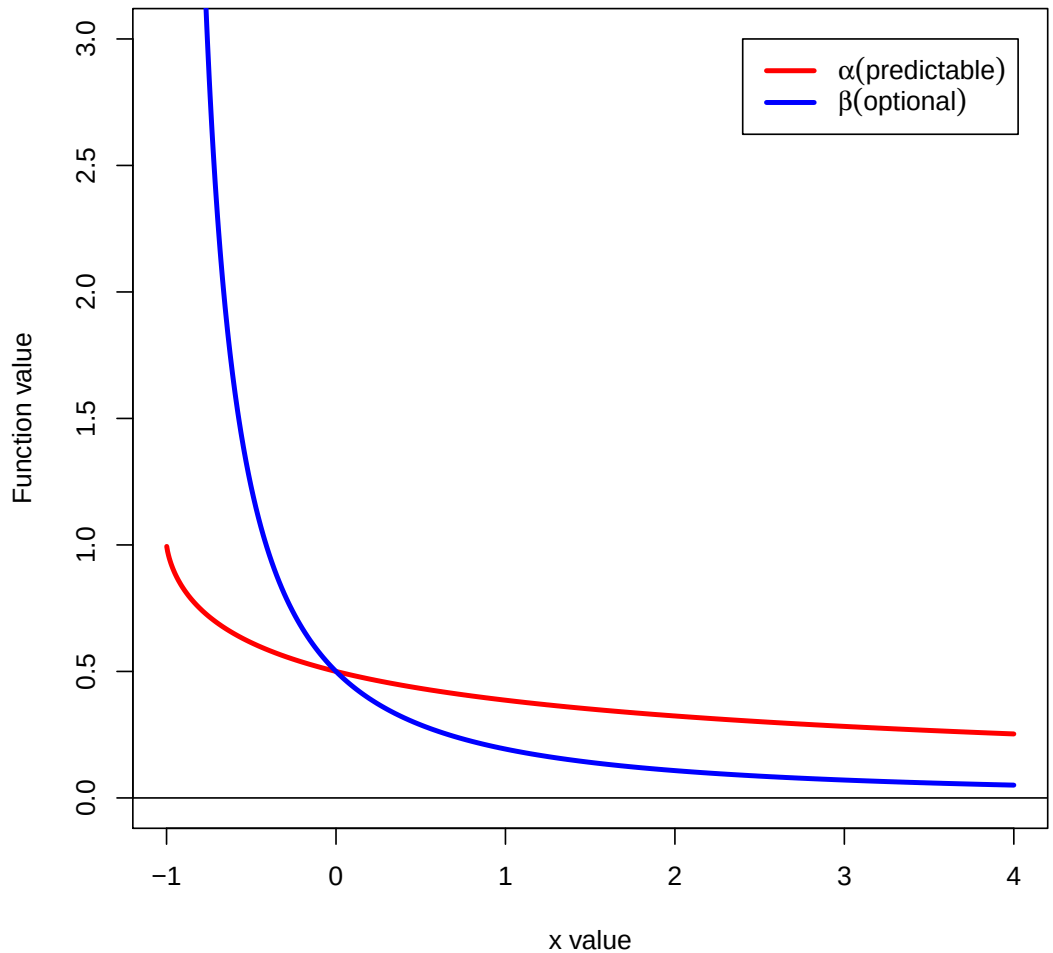
Theorem. Let $a \geq -1$ and assume that $\Delta M 1_{(\Delta M \neq 0)} \geq a$. It holds that

$$E \exp(\tfrac{1}{2}\langle M^c \rangle_\infty + \alpha(a)\langle M^d \rangle_\infty) < \infty \Rightarrow \mathcal{E}(M) \text{ is an UI MG}$$
$$E \exp(\tfrac{1}{2}[M^c]_\infty + \beta(a)[M^d]_\infty) < \infty \Rightarrow \mathcal{E}(M) \text{ is an UI MG.}$$

All constants are optimal. Note that $\beta(-1) = \infty$.

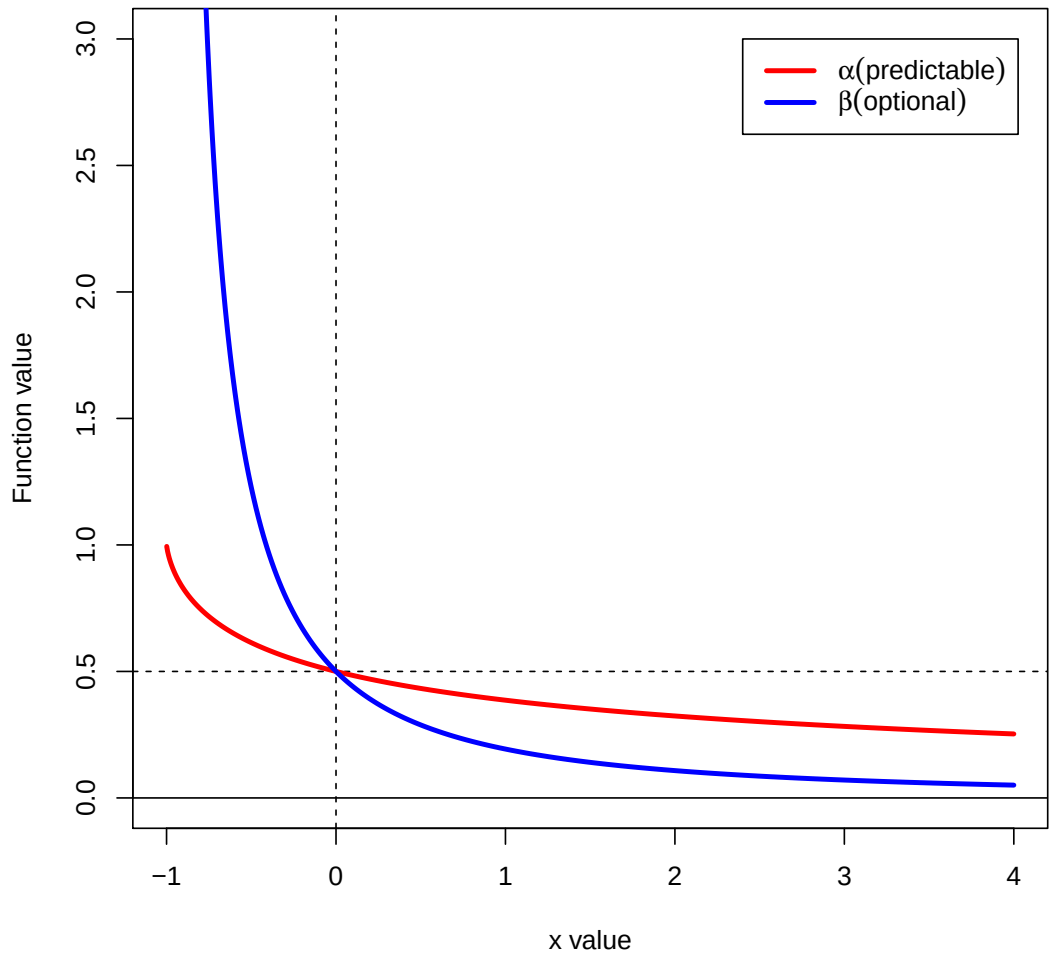
Exponential martingales

Plot of α and β:



Exponential martingales

Plot of α and β:



Exponential martingales

As $\alpha(0) = \beta(0) = \frac{1}{2}$, we obtain:

Corollary. Assume that $\Delta M \geq 0$. If $E \exp(\frac{1}{2}\langle M \rangle_\infty)$ or $E \exp(\frac{1}{2}[M]_\infty)$ is finite, then $\mathcal{E}(M)$ is a uniformly integrable martingale.

Exponential martingales

As $\alpha(0) = \beta(0) = \frac{1}{2}$, we obtain:

Corollary. Assume that $\Delta M \geq 0$. If $E \exp(\frac{1}{2}\langle M \rangle_\infty)$ or $E \exp(\frac{1}{2}[M]_\infty)$ is finite, then $\mathcal{E}(M)$ is a uniformly integrable martingale.

Thus, having nonnegative jumps seems to lead to equal importance of the predictable and optional quadratic variation.

Exponential martingales

As $\alpha(0) = \beta(0) = \frac{1}{2}$, we obtain:

Corollary. Assume that $\Delta M \geq 0$. If $E \exp(\frac{1}{2}\langle M \rangle_\infty)$ or $E \exp(\frac{1}{2}[M]_\infty)$ is finite, then $\mathcal{E}(M)$ is a uniformly integrable martingale.

Thus, having nonnegative jumps seems to lead to equal importance of the predictable and optional quadratic variation.

Based on this observation, we proved, extending a result by Krylov (2009):

Theorem. Assume that $\Delta M \geq 0$. Fix $0 \leq \gamma \leq 1$. Assume that

$$\liminf_{\varepsilon \to 0} \varepsilon \log E \exp\left((1-\varepsilon) \frac{1}{2} \left(\gamma [M]_\infty + (1-\gamma) \langle M \rangle_\infty \right) \right) < \infty,$$

then $\mathcal{E}(M)$ is a uniformly integrable martingale.

Simplified proofs in the general theory of processes

A detour. Having considered some problems applying local martingales, compensators and the quadratic variation, we asked: **Can simplified proofs of the existence of these processes could be found?**

Simplified proofs in the general theory of processes

A detour. Having considered some problems applying local martingales, compensators and the quadratic variation, we asked: **Can simplified proofs of the existence of these processes could be found?**

Beiglböck et al (2012) gave a simplified proof of the Doob-Meyer theorem using the following lemma:

Lemma. Let (X_n) be a uniformly integrable sequence of variables. Then, there exist a limit variable X and convex weights such that

$$\sum_{i=n}^{K_i} \lambda_i^n X_i \xrightarrow{\mathcal{L}^1} X.$$

Simplified proofs in the general theory of processes

Applying an $\mathcal{L}^2$ version of the lemma, we gave simple proofs of:

Simplified proofs in the general theory of processes

Applying an $\mathcal{L}^2$ version of the lemma, we gave simple proofs of:

- **The existence of the dual predictable projection:** For an adapted and locally integrable finite variation process A, there exists a predictable and locally integrable finite variation process $\Pi_p^* A$ such that $A - \Pi_p^* A$ is a local martingale.

Simplified proofs in the general theory of processes

Applying an $\mathcal{L}^2$ version of the lemma, we gave simple proofs of:

- **The existence of the dual predictable projection:** For an adapted and locally integrable finite variation process A, there exists a predictable and locally integrable finite variation process $\Pi_p^* A$ such that $A - \Pi_p^* A$ is a local martingale.
- **The existence of the quadratic variation process:** For a local martingale M, there exists an adapted and increasing process $[M]$ such that $M^2 - [M]$ is a local martingale and $(\Delta M)^2 = \Delta[M]$.

Simplified proofs in the general theory of processes

Applying an $\mathcal{L}^2$ version of the lemma, we gave simple proofs of:

- **The existence of the dual predictable projection:** For an adapted and locally integrable finite variation process A, there exists a predictable and locally integrable finite variation process $\Pi_p^* A$ such that $A - \Pi_p^* A$ is a local martingale.
- **The existence of the quadratic variation process:** For a local martingale M, there exists an adapted and increasing process $[M]$ such that $M^2 - [M]$ is a local martingale and $(\Delta M)^2 = \Delta[M]$.

The first proof is a minor variation of the arguments presented in Beiglböck et al (2012).

Simplified proofs in the general theory of processes

Outline of the proof of the existence of the quadratic variation:

Simplified proofs in the general theory of processes

Outline of the proof of the existence of the quadratic variation:

By the fundamental theorem of local martingales, it suffices to consider the case of a bounded local martingale M.

Simplified proofs in the general theory of processes

Outline of the proof of the existence of the quadratic variation:

By the fundamental theorem of local martingales, it suffices to consider the case of a bounded local martingale M.

Write $M_t^2 = N_t^n + Q_t^n$, where $t_k^n = k2^{-n}$ and

$$N_t^n = 2 \sum_{k: t_{k-1}^n < t} M_{t_{k-1}^n}^t (M_{t_k^n}^t - M_{t_{k-1}^n}^t).$$

Simplified proofs in the general theory of processes

Outline of the proof of the existence of the quadratic variation:

By the fundamental theorem of local martingales, it suffices to consider the case of a bounded local martingale M.

Write $M_t^2 = N_t^n + Q_t^n$, where $t_k^n = k2^{-n}$ and

$$N_t^n = 2 \sum_{k: t_{k-1}^n < t} M_{t_{k-1}^n}^t (M_{t_k^n}^t - M_{t_{k-1}^n}^t).$$

Observe that (N_t^n) is a sequence of local martingales with (N_∞^n) bounded in $\mathcal{L}^2$. Thus, by the lemma, there exists a limiting martingale N.

Simplified proofs in the general theory of processes

Outline of the proof of the existence of the quadratic variation:

By the fundamental theorem of local martingales, it suffices to consider the case of a bounded local martingale M.

Write $M_t^2 = N_t^n + Q_t^n$, where $t_k^n = k2^{-n}$ and

$$N_t^n = 2 \sum_{k: t_{k-1}^n < t} M_{t_{k-1}^n}^t (M_{t_k^n}^t - M_{t_{k-1}^n}^t).$$

Observe that (N_t^n) is a sequence of local martingales with (N_∞^n) bounded in $\mathcal{L}^2$. Thus, by the lemma, there exists a limiting martingale N.

Put $[M] = M^2 - N$ and verify using strong convergence.

A notion of causality for SDEs

We return to our main line of problems. Recall that through our main motivating problem, we can to ask the question: **How do we formalize causality in a continuous-time framework?**

A notion of causality for SDEs

We return to our main line of problems. Recall that through our main motivating problem, we can to ask the question: **How do we formalize causality in a continuous-time framework?**

Instead of considering counting processes, where causality already has been discussed through local independence (e.g. Didelez 2008), we considered causality for stochastic differential equations (SDEs) of the type

$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t,$$

where $a : \mathbb{R}^p \to \mathbb{M}(p, d)$ and Z is a d-dimensional semimartingale.

A notion of causality for SDEs

We return to our main line of problems. Recall that through our main motivating problem, we can to ask the question: **How do we formalize causality in a continuous-time framework?**

Instead of considering counting processes, where causality already has been discussed through local independence (e.g. Didelez 2008), we considered causality for stochastic differential equations (SDEs) of the type

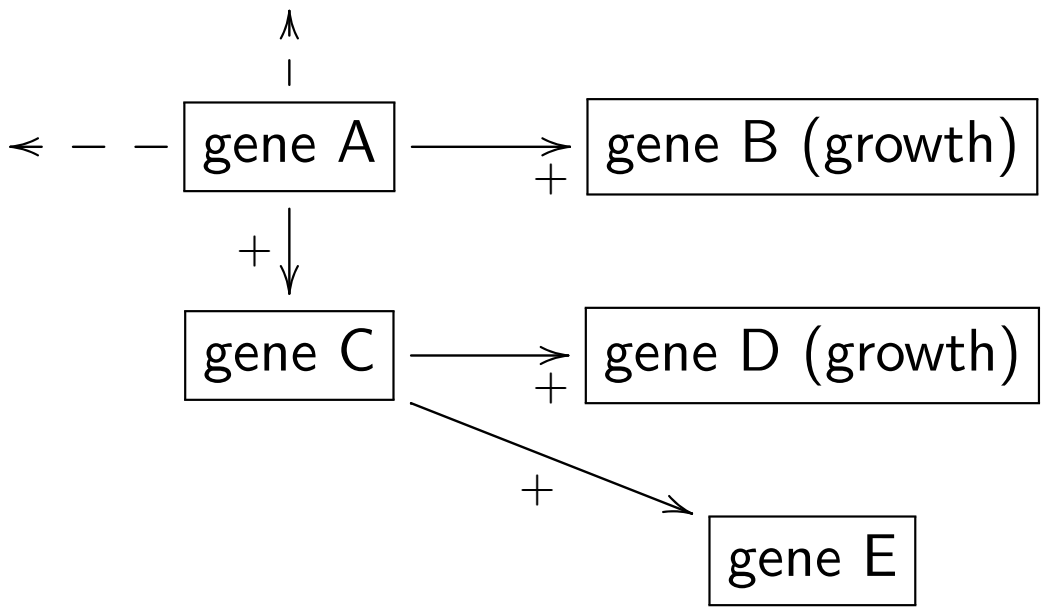
$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t,$$

where $a : \mathbb{R}^p \to \mathbb{M}(p, d)$ and Z is a d-dimensional semimartingale.

We begin by considering an example.

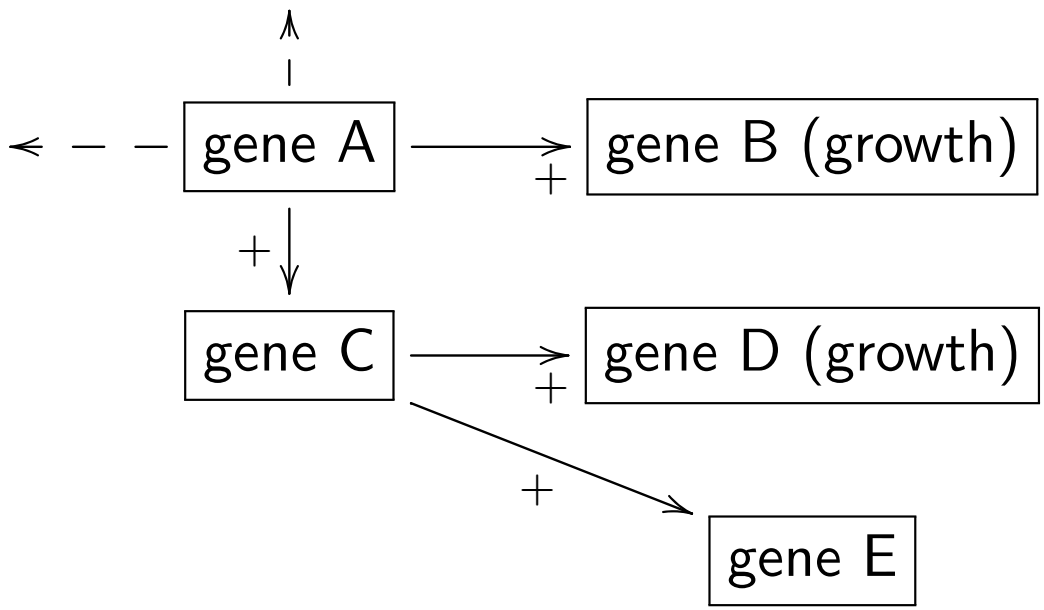
A notion of causality for SDEs

Example. We consider a plant whose growth depends on the expression of some known particular genes. Furthermore, the expression level of the genes are causally dependent on each other.



A notion of causality for SDEs

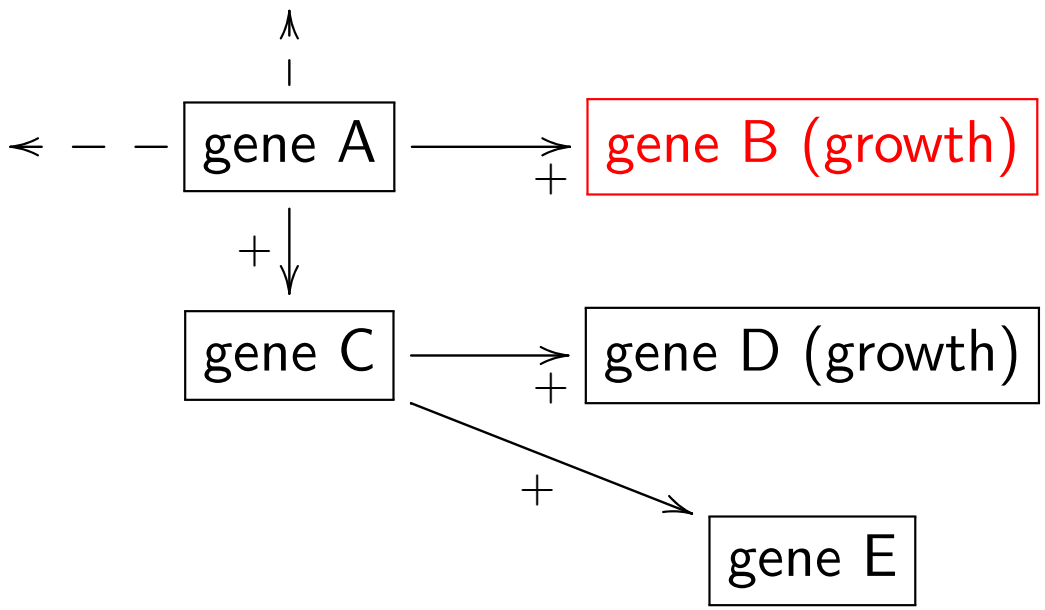
Example. We consider a plant whose growth depends on the expression of some known particular genes. Furthermore, the expression level of the genes are causally dependent on each other.



What happens when we intervene by changing the activity level of some gene?

A notion of causality for SDEs

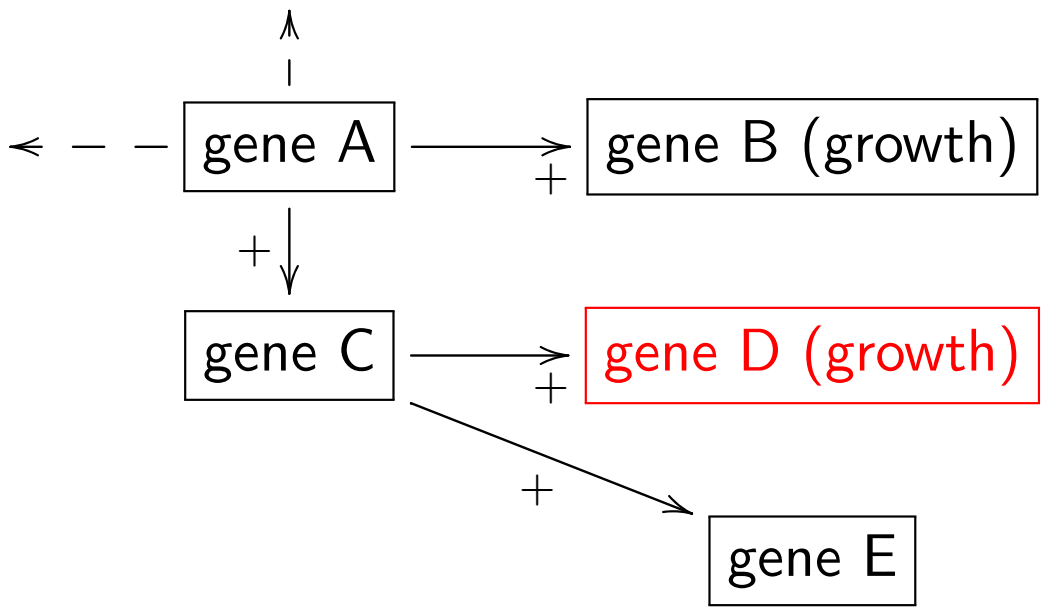
Example. We consider a plant whose growth depends on the expression of some known particular genes. Furthermore, the expression level of the genes are causally dependent on each other.



Increasing the activity of gene B: **growth effect.**

A notion of causality for SDEs

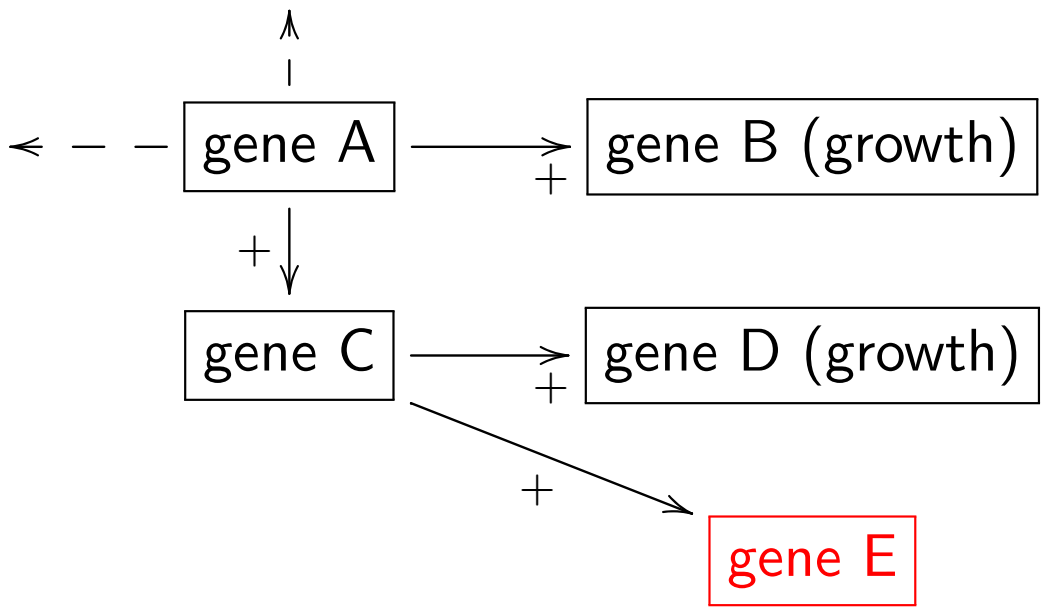
Example. We consider a plant whose growth depends on the expression of some known particular genes. Furthermore, the expression level of the genes are causally dependent on each other.



Increasing the activity of gene D: **growth effect.**

A notion of causality for SDEs

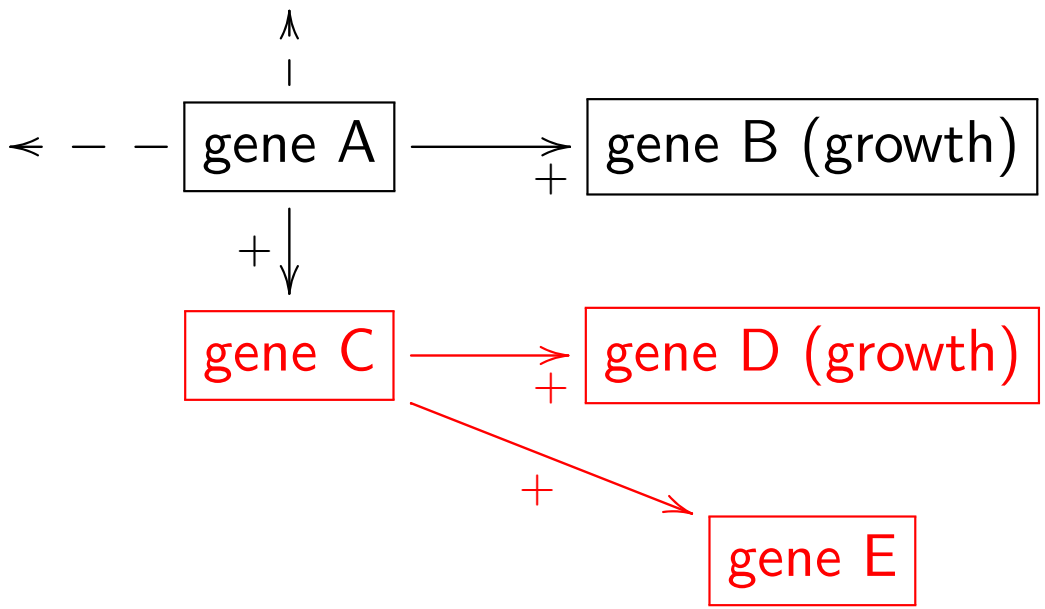
Example. We consider a plant whose growth depends on the expression of some known particular genes. Furthermore, the expression level of the genes are causally dependent on each other.



Increasing the activity of gene E: **no effect.** (in spite of positive correlation with growth)

A notion of causality for SDEs

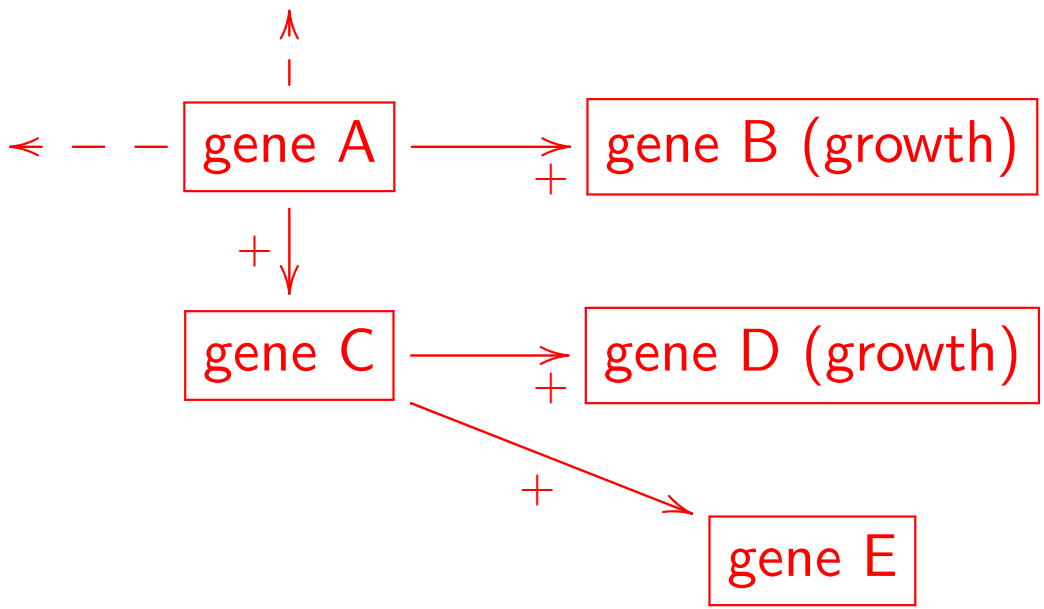
Example. We consider a plant whose growth depends on the expression of some known particular genes. Furthermore, the expression level of the genes are causally dependent on each other.

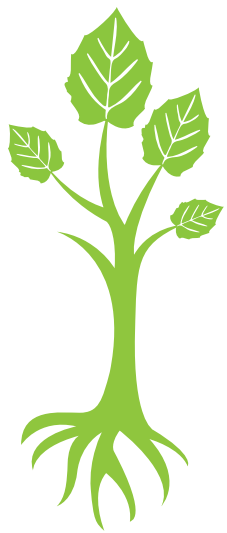


Increasing the activity of gene C: **growth effect.** (through gene D)

A notion of causality for SDEs

Example. We consider a plant whose growth depends on the expression of some known particular genes. Furthermore, the expression level of the genes are causally dependent on each other.





Increasing the activity of gene A: **large growth effect.** (through genes B and D)

A notion of causality for SDEs

In reality, expression levels of genes change over time. Consider a network of p genes. One simple model of the activity of these genes over time is through an Ornstein-Uhlenbeck SDE of the type

$$\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t.$$

A notion of causality for SDEs

In reality, expression levels of genes change over time. Consider a network of p genes. One simple model of the activity of these genes over time is through an Ornstein-Uhlenbeck SDE of the type

$$\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t.$$

In this case, the matrix B controls the association of the p genes to each other.

A notion of causality for SDEs

In reality, expression levels of genes change over time. Consider a network of p genes. One simple model of the activity of these genes over time is through an Ornstein-Uhlenbeck SDE of the type

$$\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t.$$

In this case, the matrix B controls the association of the p genes to each other.

If we could endow this model with a notion of causality, we might be able to use it to **predict the effect of interventions in the system.** This is our motivation for introducing a notion of causality for SDEs.

A notion of causality for SDEs

Formally, our setup consists of the following:

A notion of causality for SDEs

Formally, our setup consists of the following:

- A p-dimensional initial condition X_0

A notion of causality for SDEs

Formally, our setup consists of the following:

- A p-dimensional initial condition X_0
- A d-dimensional semimartingale Z

A notion of causality for SDEs

Formally, our setup consists of the following:

- A p-dimensional initial condition X_0
- A d-dimensional semimartingale Z
- A continuous function $a : \mathbb{R}^p \to \mathbb{M}(p, d)$.

A notion of causality for SDEs

Formally, our setup consists of the following:

- A p-dimensional initial condition X_0
- A d-dimensional semimartingale Z
- A continuous function $a : \mathbb{R}^p \to \mathbb{M}(p, d)$.

We consider the SDE

$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t.$$

A notion of causality for SDEs

As the argument t in X_t frequently denotes time, it is natural to interpret our SDE in terms of causality where variables from previous timepoints are the causes of variables for future timepoints.

A notion of causality for SDEs

As the argument t in X_t frequently denotes time, it is natural to interpret our SDE in terms of causality where variables from previous timepoints are the causes of variables for future timepoints.

In particular, we may use the formula

$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t$$

to make the approximation

$$X_{t+\Delta} = X_t + a(X_t)(Z_{t+\Delta} - Z_t)$$

and consider this as describing how interventions in X_t will influence $X_{t+\Delta}$.

A notion of causality for SDEs

Definition. Consider some $m \leq p$ and $\zeta \in \mathbb{R}$. The postintervention SDE arising from making the intervention $X^m := \zeta$ in the p-dimensional SDE

$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t$$

is the p-dimensional SDE

$$\mathrm{d}Y_t = b(Y_{t-})\,\mathrm{d}Z_t$$

where $b : \mathbb{R}^p \to \mathbb{M}(p, d)$, $b_{ij}(y) = a_{ij}(y)$ for $i \neq m$ and $b_{mj}(y) = 0$, $Y_0^i = X_0^i$ for $i \neq m$ and $Y_0^m = \zeta$.

A notion of causality for SDEs

Definition. Consider some $m \leq p$ and $\zeta \in \mathbb{R}$. The postintervention SDE arising from making the intervention $X^m := \zeta$ in the p-dimensional SDE

$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t$$

is the p-dimensional SDE

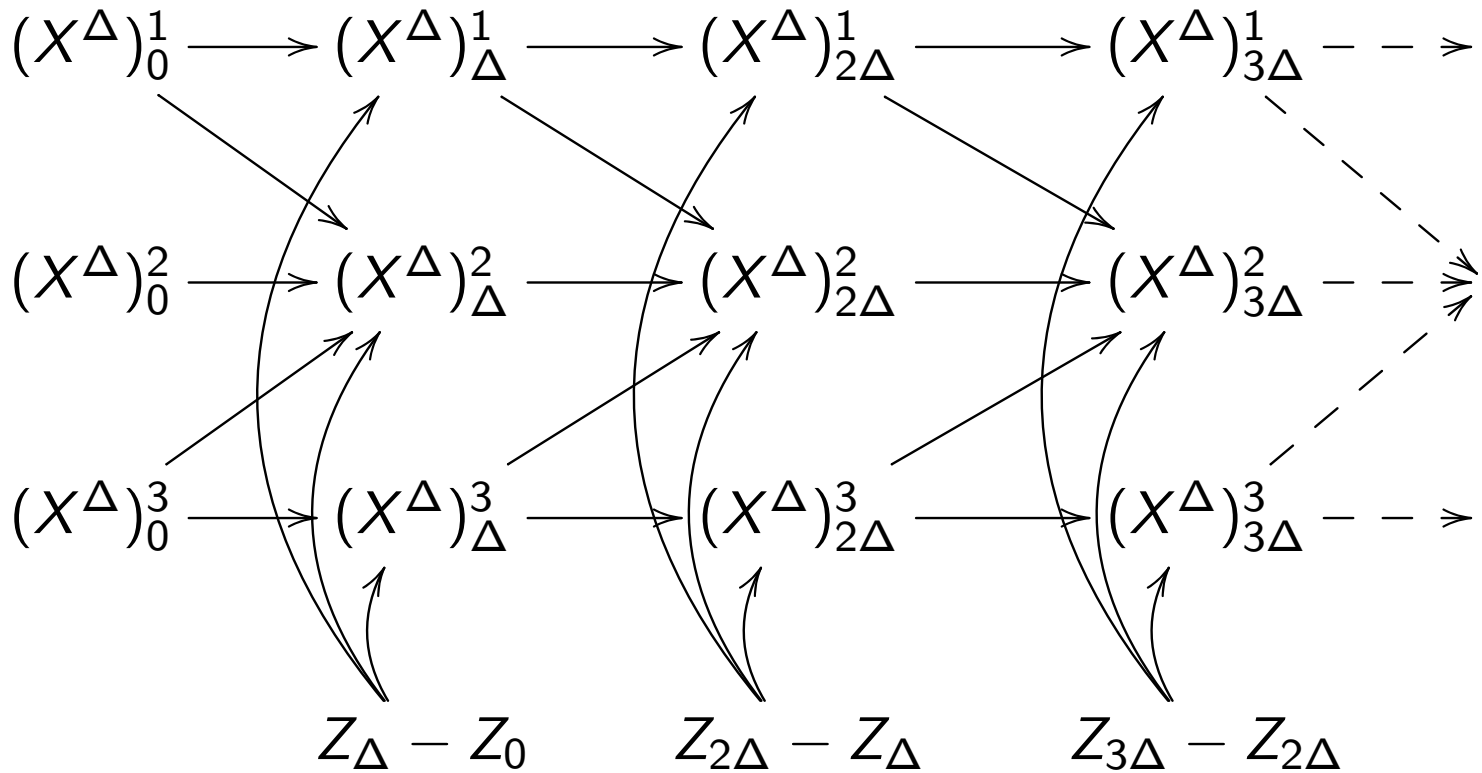
$$\mathrm{d}Y_t = b(Y_{t-})\,\mathrm{d}Z_t$$

where $b : \mathbb{R}^p \to \mathbb{M}(p, d)$, $b_{ij}(y) = a_{ij}(y)$ for $i \neq m$ and $b_{mj}(y) = 0$, $Y_0^i = X_0^i$ for $i \neq m$ and $Y_0^m = \zeta$.

This **directly defines** interventions in SDEs without specifying the underlying notion of causality.

A notion of causality for SDEs

However, a limiting argument shows that our notion of intervention is **in agreement with intervening in the DAG (directed acyclic graph) of the discretized SDE given as below**, where arrows corresponds to entries of $a(x)$ independent of particular coordinates of x.



A notion of causality for SDEs

We now have a notion of intervention for SDEs, and thus a notion of causality for SDEs, at our disposal.

A notion of causality for SDEs

We now have a notion of intervention for SDEs, and thus a notion of causality for SDEs, at our disposal.

Question: Assume that we observe some SDE. Are postintervention distributions idenfiable from the observational distribution?

A notion of causality for SDEs

We now have a notion of intervention for SDEs, and thus a notion of causality for SDEs, at our disposal.

Question: Assume that we observe some SDE. Are postintervention distributions idenfiable from the observational distribution?

In Pearl's classical DAG-based notion of causality, this is not the case.

A notion of causality for SDEs

We now have a notion of intervention for SDEs, and thus a notion of causality for SDEs, at our disposal.

Question: Assume that we observe some SDE. Are postintervention distributions idenfiable from the observational distribution?

In Pearl's classical DAG-based notion of causality, this is not the case.

In our SDE case, the fact that distinct SDEs can have the same distributions (for example, through varying the diffusion matrix) could be a source of complications.

A notion of causality for SDEs

Theorem. Consider the two SDEs

$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t$$

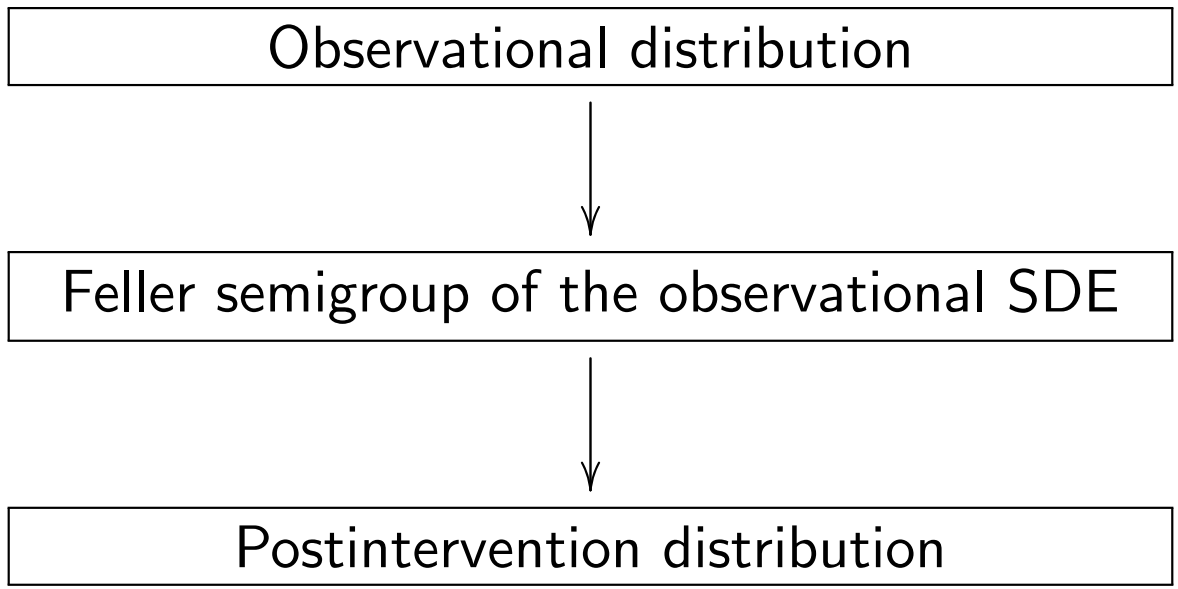
and

$$\mathrm{d}X_t = \tilde{a}(X_{t-})\,\mathrm{d}\tilde{Z}_t,$$

where Z and $\tilde{Z}$ are Lévy processes of dimension d and $\tilde{d}$, respectively, and a and $\tilde{a}$ are Lipschitz and bounded. Then the results of intervention in the SDEs are the same **whenever the Feller semigroups and the initial distributions are the same for the SDEs.**

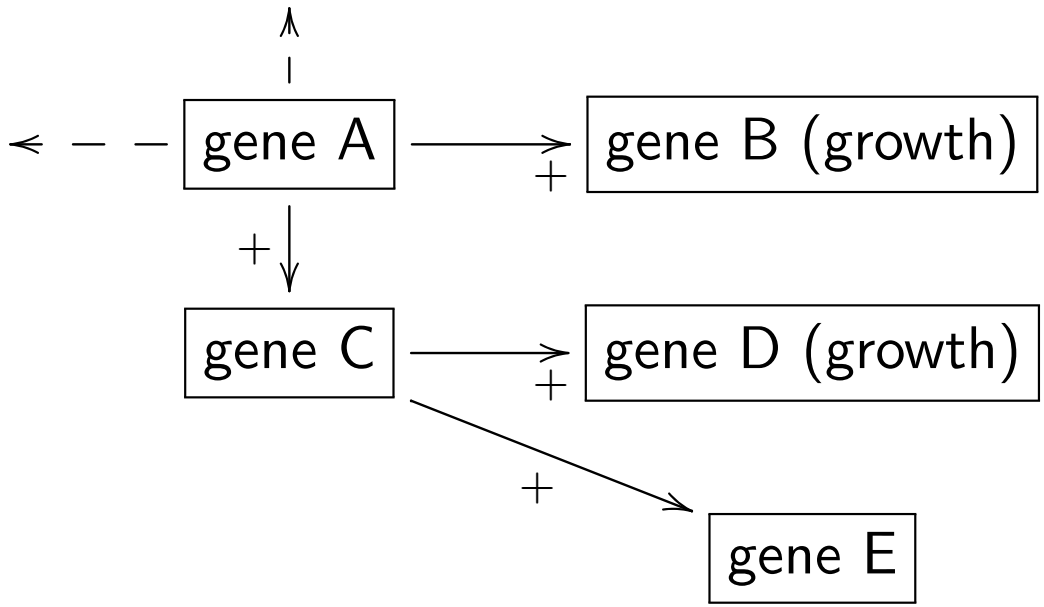
A notion of causality for SDEs

This enables practical inference of postintervention distributions through the following line of inference:



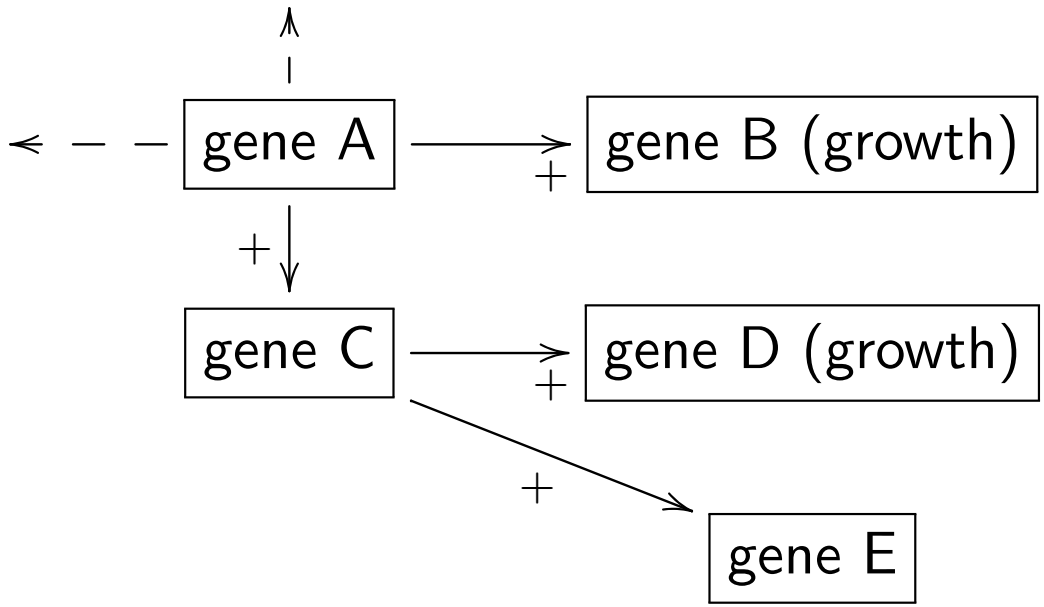
A notion of causality for SDEs

Consider again our previous plant growth example:



A notion of causality for SDEs

Consider again our previous plant growth example:



Our results imply that for time-dependent SDE observations with Lévy noise, **all intervention effects can be identified.**

Identifiability in ICA

A detour (again). Having come into contact with the literature on DAG-based causality, we also considered a problem related to this.

Identifiability in ICA

A detour (again). Having come into contact with the literature on DAG-based causality, we also considered a problem related to this.

In the DAG-based model for causal discovery, we:

- Consider the distribution of p variables $X^1, \ldots, X^p$.
- Assume that the distribution of these variables have conditional independence properties consistent with some DAG.
- Wish to identify the DAG, corresponding to the **causal network.**

Identifiability in ICA

A detour (again). Having come into contact with the literature on DAG-based causality, we also considered a problem related to this.

In the DAG-based model for causal discovery, we:

- Consider the distribution of p variables $X^1, \ldots, X^p$.
- Assume that the distribution of these variables have conditional independence properties consistent with some DAG.
- Wish to identify the DAG, corresponding to the **causal network.**

The central problem is that the DAG is not uniquely identifiable from the distribution.

Identifiability in ICA

However, Shimizu et al (2006) showed the following. Assume that the variables $X^1, \ldots, X^p$ are linearly related in the sense that

$$X = CX + \varepsilon$$

where $C \in \mathbb{M}(p, p)$ is acyclic in the sense that PCP^t is strictly lower triangular for some permutation matrix P. Also assume that the error variable ε has independent, non-degenerate and **non-Gaussian** coordinates of mean zero.

Identifiability in ICA

However, Shimizu et al (2006) showed the following. Assume that the variables $X^1, \ldots, X^p$ are linearly related in the sense that

$$X = CX + \varepsilon$$

where $C \in \mathbb{M}(p, p)$ is acyclic in the sense that PCP^t is strictly lower triangular for some permutation matrix P. Also assume that the error variable ε has independent, non-degenerate and **non-Gaussian** coordinates of mean zero.

Then, the DAG can be recovered from the distribution of X.

Identifiability in ICA

This identifiability result is derived from a result on ICA (Independent Component Analysis).

Identifiability in ICA

This identifiability result is derived from a result on ICA (Independent Component Analysis). The ICA model considers p variables $X^1, \ldots, X^p$ satisfying

$$X = A\varepsilon,$$

where ε has independent non-degenerate coordinates of mean zero. Both A and the distribution of ε are assumed unknown and are to be estimated. X is observed.

Identifiability in ICA

This identifiability result is derived from a result on ICA (Independent Component Analysis). The ICA model considers p variables $X^1, \ldots, X^p$ satisfying

$$X = A\varepsilon,$$

where ε has independent non-degenerate coordinates of mean zero. Both A and the distribution of ε are assumed unknown and are to be estimated. X is observed.

From the results of Comon (1994), it follows that:

Identifiability in ICA

This identifiability result is derived from a result on ICA (Independent Component Analysis). The ICA model considers p variables $X^1, \ldots, X^p$ satisfying

$$X = A\varepsilon,$$

where ε has independent non-degenerate coordinates of mean zero. Both A and the distribution of ε are assumed unknown and are to be estimated. X is observed.

From the results of Comon (1994), it follows that:

1. If the true distribution of ε has **only Gaussian coordinates**, A is identifiable up to transpose products.

Identifiability in ICA

This identifiability result is derived from a result on ICA (Independent Component Analysis). The ICA model considers p variables $X^1, \ldots, X^p$ satisfying

$$X = A\varepsilon,$$

where ε has independent non-degenerate coordinates of mean zero. Both A and the distribution of ε are assumed unknown and are to be estimated. X is observed.

From the results of Comon (1994), it follows that:

1. If the true distribution of ε has **only Gaussian coordinates**, A is identifiable up to transpose products.
2. If the true distribution of ε has **no Gaussian coordinates**, then A is identifiable up to scaling and permutation.

Identifiability in ICA

This presents the following practical conundrum:

Identifiability in ICA

This presents the following practical conundrum:

It is infeasible to claim that the error distribution is exactly Gaussian. Therefore, it is reasonable to expect that we will always be in **the latter of the two scenarios.**

Identifiability in ICA

This presents the following practical conundrum:

It is infeasible to claim that the error distribution is exactly Gaussian. Therefore, it is reasonable to expect that we will always be in **the latter of the two scenarios.**

Nonetheless, we might think that in practice, if the error distribution is close to Gaussian, the behaviour of the model will more closely resemble **the former of the two scenarios.**

Identifiability in ICA

This presents the following practical conundrum:

It is infeasible to claim that the error distribution is exactly Gaussian. Therefore, it is reasonable to expect that we will always be in **the latter of the two scenarios.**

Nonetheless, we might think that in practice, if the error distribution is close to Gaussian, the behaviour of the model will more closely resemble **the former of the two scenarios.**

Question: What actually happens to identifiability when we sample finitely many times and the error distribution is close to Gaussian while not being Gaussian?

Identifiability in ICA

This issue may be framed as a discrepancy between continuous and discontinuous behaviour:

Identifiability in ICA

This issue may be framed as a discrepancy between continuous and discontinuous behaviour:

- When considering the entire **family of distributions**, the model behaves **discontinuously**: Identifiability properties can change drastically even for small changes in the true distribution of ε.

Identifiability in ICA

This issue may be framed as a discrepancy between continuous and discontinuous behaviour:

- When considering the entire **family of distributions**, the model behaves **discontinuously**: Identifiability properties can change drastically even for small changes in the true distribution of ε.
- When considering a **single distribution at a time**, the model behaves **continuously**: When the distribution of ε varies a small amount, the distribution of $A\varepsilon$ also varies only a small amount.

Identifiability in ICA

This issue may be framed as a discrepancy between continuous and discontinuous behaviour:

- When considering the entire **family of distributions**, the model behaves **discontinuously**: Identifiability properties can change drastically even for small changes in the true distribution of ε.
- When considering a **single distribution at a time**, the model behaves **continuously**: When the distribution of ε varies a small amount, the distribution of $A\varepsilon$ also varies only a small amount.

We set out to understand this phenomenon better.

Identifiability in ICA

To obtain results, we considered estimation only of the mixing matrix A, assuming the error distribution fixed. We considered error distributions which are independent and identical contaminated Gaussian distributions.

Identifiability in ICA

To obtain results, we considered estimation only of the mixing matrix A, assuming the error distribution fixed. We considered error distributions which are independent and identical contaminated Gaussian distributions.

Theorem. Consider n samples $X_1, \ldots, X_n$ from the p-dimensional true distribution. Assume that the common error distribution P_n is $P_n = \beta_n \xi + (1 - \beta_n)\mathcal{N}$, where $\mathcal{N}$ is the standard normal distribution. Letting n tend to infinity, it holds that:

Identifiability in ICA

To obtain results, we considered estimation only of the mixing matrix A, assuming the error distribution fixed. We considered error distributions which are independent and identical contaminated Gaussian distributions.

Theorem. Consider n samples $X_1, \ldots, X_n$ from the p-dimensional true distribution. Assume that the common error distribution P_n is $P_n = \beta_n \xi + (1 - \beta_n)\mathcal{N}$, where $\mathcal{N}$ is the standard normal distribution. Letting n tend to infinity, it holds that:

- If $\beta_n \simeq n^{-\rho}$ for $\rho > 1/2$ (fast approach to Gaussianity), then the asymptotic behaviour of the model is similar to the Gaussian scenario.

Identifiability in ICA

To obtain results, we considered estimation only of the mixing matrix A, assuming the error distribution fixed. We considered error distributions which are independent and identical contaminated Gaussian distributions.

Theorem. Consider n samples $X_1, \ldots, X_n$ from the p-dimensional true distribution. Assume that the common error distribution P_n is $P_n = \beta_n \xi + (1 - \beta_n)\mathcal{N}$, where $\mathcal{N}$ is the standard normal distribution. Letting n tend to infinity, it holds that:

- If $\beta_n \simeq n^{-\rho}$ for $\rho > 1/2$ (fast approach to Gaussianity), then the asymptotic behaviour of the model is similar to the Gaussian scenario.
- If $\beta_n \simeq n^{-\rho}$ for $\rho < 1/2$ (slow approach to Gaussianity), then the asymptotic behaviour of the model is similar to the non-Gaussian scenario.

Identifiability in ICA

To obtain results, we considered estimation only of the mixing matrix A, assuming the error distribution fixed. We considered error distributions which are independent and identical contaminated Gaussian distributions.

Theorem. Consider n samples $X_1, \ldots, X_n$ from the p-dimensional true distribution. Assume that the common error distribution P_n is $P_n = \beta_n \xi + (1 - \beta_n)\mathcal{N}$, where $\mathcal{N}$ is the standard normal distribution. Letting n tend to infinity, it holds that:

- If $\beta_n \simeq n^{-\rho}$ for $\rho > 1/2$ (fast approach to Gaussianity), then the asymptotic behaviour of the model is similar to the Gaussian scenario.
- If $\beta_n \simeq n^{-\rho}$ for $\rho < 1/2$ (slow approach to Gaussianity), then the asymptotic behaviour of the model is similar to the non-Gaussian scenario.

(Subject to very favorable interpretations)

Degrees of freedom in nonlinear regression

Recall that one of our motivating examples for the development of a notion of causality for SDEs was an Ornstein-Uhlenbeck SDE for modeling gene expression networks:

$$\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t.$$

Degrees of freedom in nonlinear regression

Recall that one of our motivating examples for the development of a notion of causality for SDEs was an Ornstein-Uhlenbeck SDE for modeling gene expression networks:

$$\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t.$$

Having introduced our notion of causality for such SDEs, we next turned to the question of estimating the **causal network** in such an SDE.

Degrees of freedom in nonlinear regression

Recall that one of our motivating examples for the development of a notion of causality for SDEs was an Ornstein-Uhlenbeck SDE for modeling gene expression networks:

$$\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t.$$

Having introduced our notion of causality for such SDEs, we next turned to the question of estimating the **causal network** in such an SDE.

For the OU SDE, the causal network is determined by the zeroes of the mean reversion speed matrix B ($B_{ij} = 0$ means the absence of a causal link from j to i).

Degrees of freedom in nonlinear regression

Recall that one of our motivating examples for the development of a notion of causality for SDEs was an Ornstein-Uhlenbeck SDE for modeling gene expression networks:

$$\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t.$$

Having introduced our notion of causality for such SDEs, we next turned to the question of estimating the **causal network** in such an SDE.

For the OU SDE, the causal network is determined by the zeroes of the mean reversion speed matrix B ($B_{ij} = 0$ means the absence of a causal link from j to i).

We will take particular interest in sparse networks, meaning that we wish to obtain **sparse estimates of** B (many entries equal to zero).

Degrees of freedom in nonlinear regression

For simplicity, we consider the Ornstein-Uhlenbeck SDE

$$\mathrm{d}X_t = BX_t\,\mathrm{d}t + \mathrm{d}W_t$$

with mean reversion level zero and diffusion matrix I_p.

Degrees of freedom in nonlinear regression

For simplicity, we consider the Ornstein-Uhlenbeck SDE

$$\mathrm{d}X_t = BX_t\,\mathrm{d}t + \mathrm{d}W_t$$

with mean reversion level zero and diffusion matrix I_p.

We assume that we observe this SDE discretely at times $t_k = \Delta k$ for $k = 0, \ldots, n$. A natural loss function for the estimation of B is then $R : \mathbb{M}(p, p) \to [0, \infty)$ given by

$$R(B) = \sum_{k=1}^{n} \|X_{t_k} - \exp(\Delta B) X_{t_{k-1}}\|_2^2$$

Degrees of freedom in nonlinear regression

For simplicity, we consider the Ornstein-Uhlenbeck SDE

$$\mathrm{d}X_t = BX_t\,\mathrm{d}t + \mathrm{d}W_t$$

with mean reversion level zero and diffusion matrix I_p.

We assume that we observe this SDE discretely at times $t_k = \Delta k$ for $k = 0, \ldots, n$. A natural loss function for the estimation of B is then $R : \mathbb{M}(p, p) \to [0, \infty)$ given by

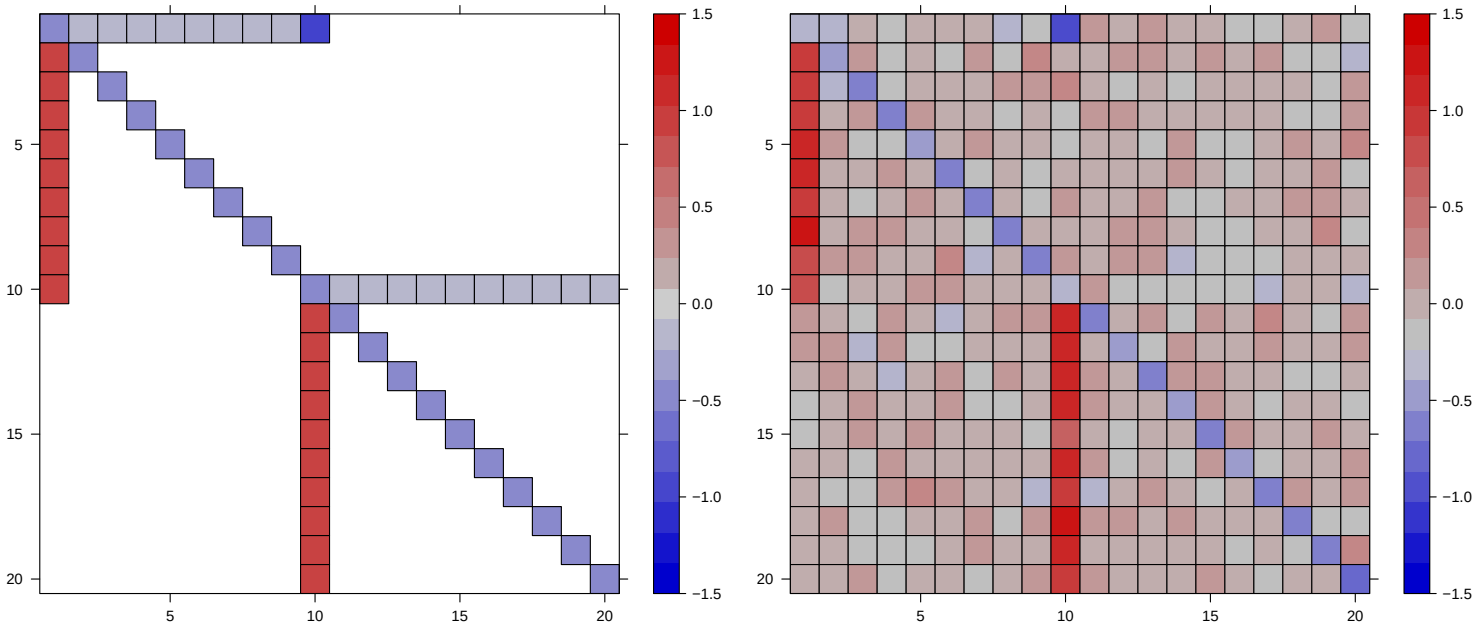
$$R(B) = \sum_{k=1}^{n} \|X_{t_k} - \exp(\Delta B) X_{t_{k-1}}\|_2^2$$

and sparse estimates can be obtained by, for $\lambda \geq 0$,

$$\hat{B}_\lambda(X_{t_0}, \ldots, X_{t_n}) = \operatorname*{argmin}_{B \in \mathbb{M}(p,p)} R(B) + \lambda \|B\|_1.$$

Degrees of freedom in nonlinear regression

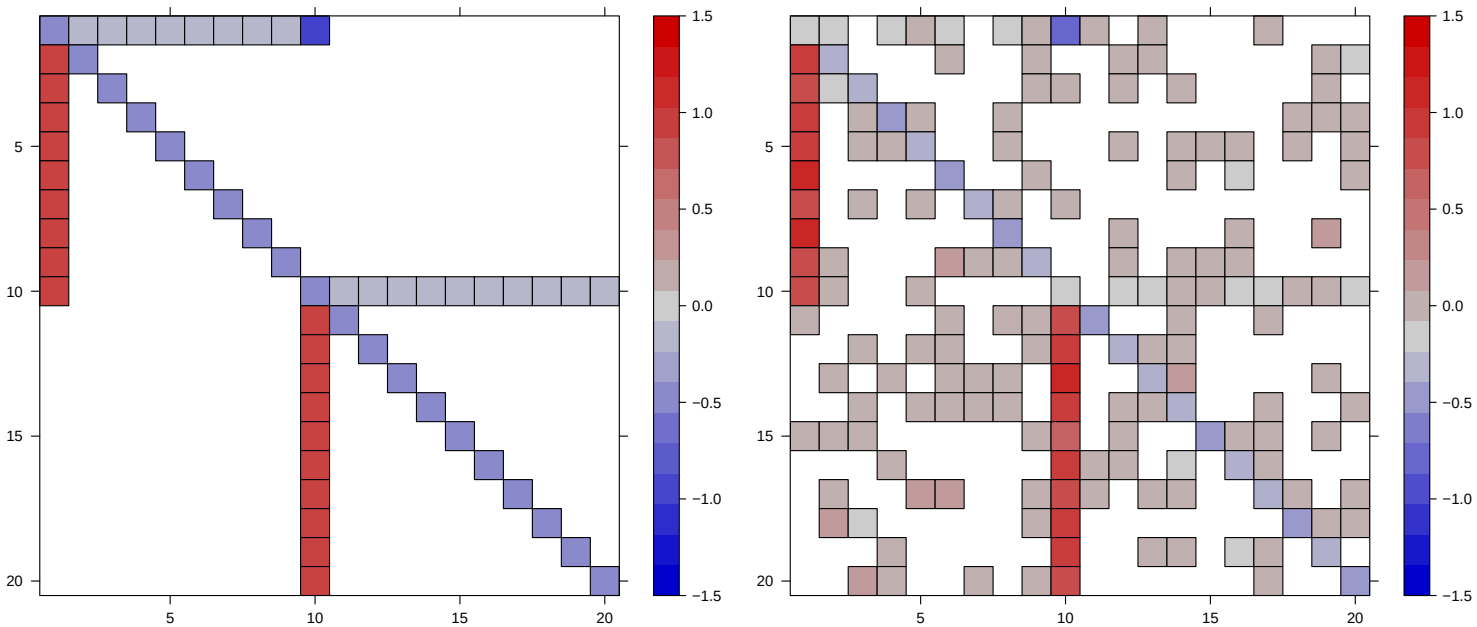
Some numerical results from simulated data:



(Left matrix: True B, right matrix: $\hat{B}_\lambda$ with $\lambda = 0.000$).

Degrees of freedom in nonlinear regression

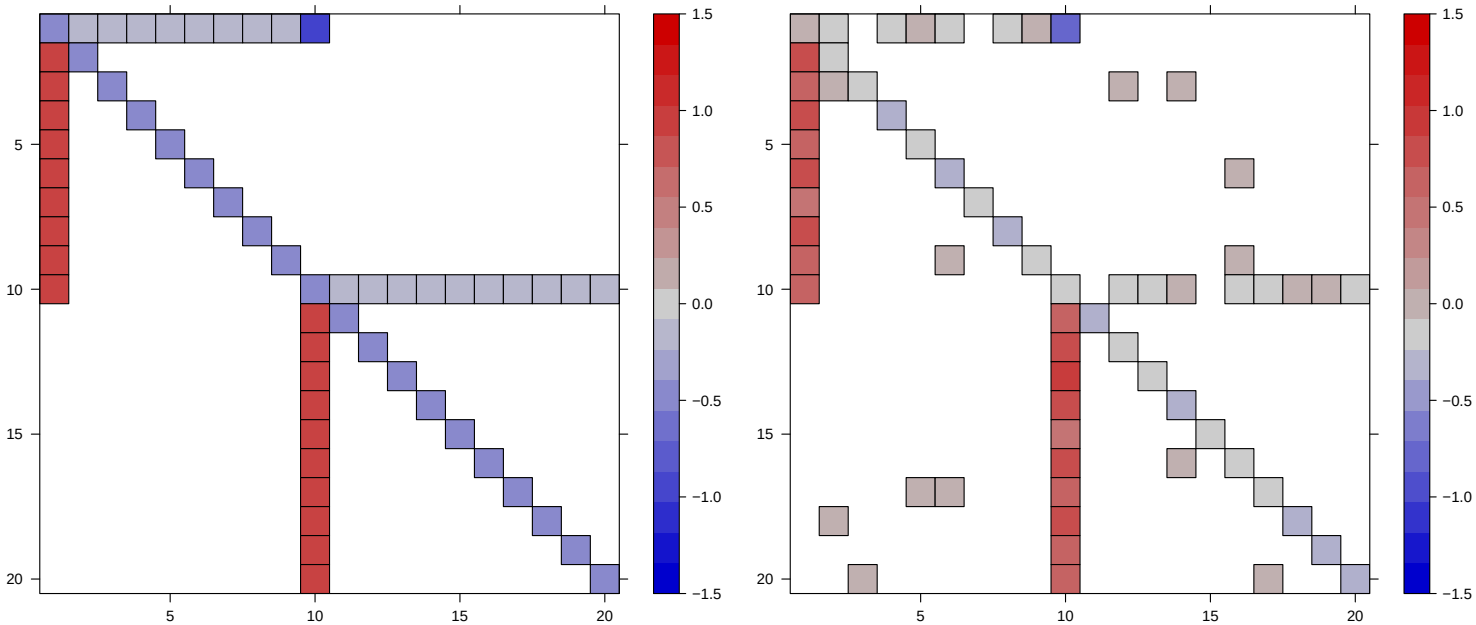
Some numerical results from simulated data:



(Left matrix: True B, right matrix: $\hat{B}_\lambda$ with $\lambda = 0.050$).

Degrees of freedom in nonlinear regression

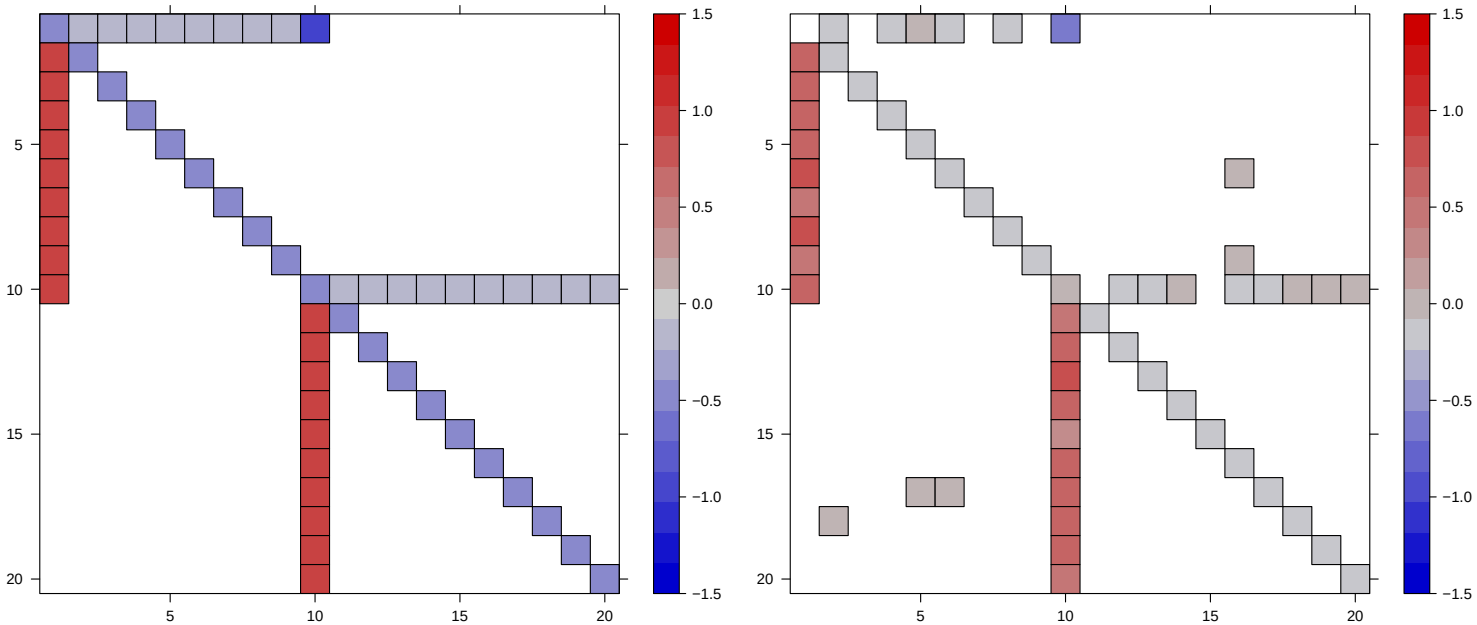
Some numerical results from simulated data:



(Left matrix: True B, right matrix: $\hat{B}_\lambda$ with $\lambda = 0.200$).

Degrees of freedom in nonlinear regression

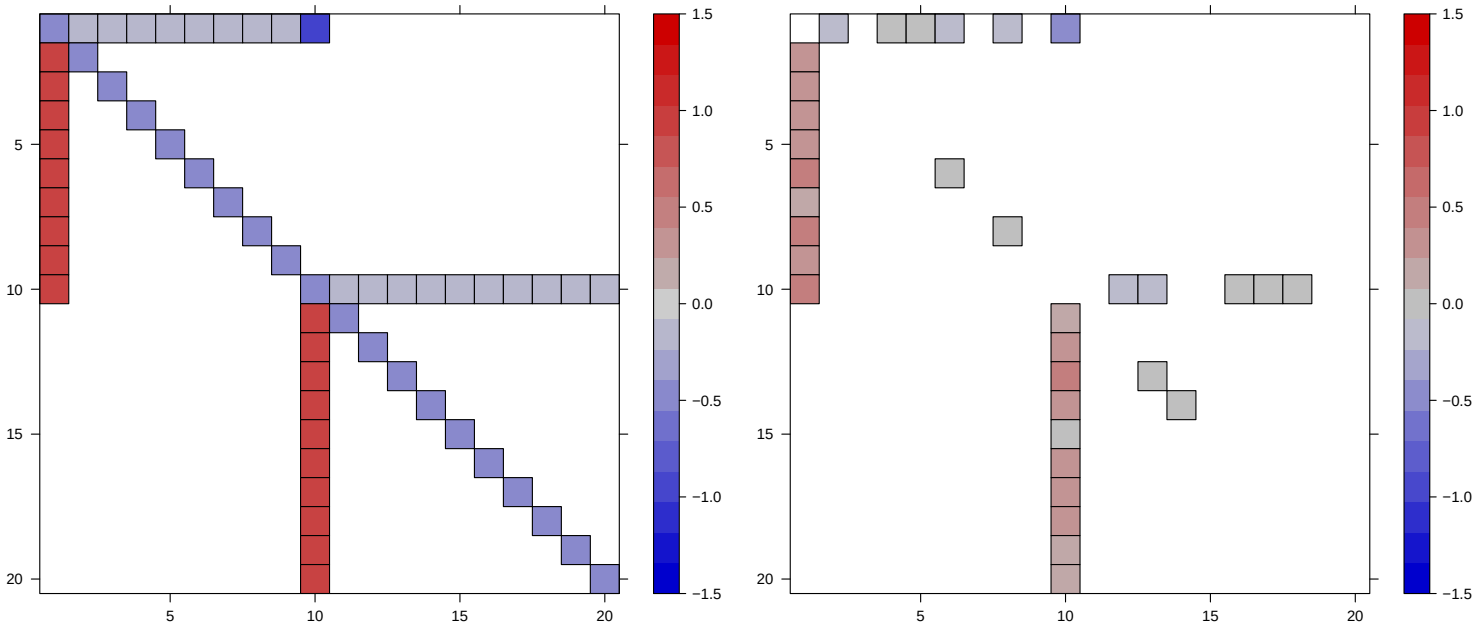
Some numerical results from simulated data:



(Left matrix: True B, right matrix: $\hat{B}_\lambda$ with $\lambda = 0.300$).

Degrees of freedom in nonlinear regression

Some numerical results from simulated data:



(Left matrix: True B, right matrix: $\hat{B}_\lambda$ with $\lambda = 0.500$).

Degrees of freedom in nonlinear regression

Conclusion. Estimation through L^1-penalized estimation with loss function R yields reasonable sparse estimates.

Degrees of freedom in nonlinear regression

Conclusion. Estimation through L^1-penalized estimation with loss function R yields reasonable sparse estimates.

Problem. How should λ be chosen? This can be thought of as a **model selection** type of problem, as we are choosing a "submodel" based on the zeroes of our estimate of B.

Degrees of freedom in nonlinear regression

Conclusion. Estimation through L^1-penalized estimation with loss function R yields reasonable sparse estimates.

Problem. How should λ be chosen? This can be thought of as a **model selection** type of problem, as we are choosing a "submodel" based on the zeroes of our estimate of B.

Sad fact. We don't really know how to choose λ.

Degrees of freedom in nonlinear regression

Conclusion. Estimation through L^1-penalized estimation with loss function R yields reasonable sparse estimates.

Problem. How should λ be chosen? This can be thought of as a **model selection** type of problem, as we are choosing a "submodel" based on the zeroes of our estimate of B.

Sad fact. We don't really know how to choose λ.

Workaround. Ignore the hard problem and solve a simpler problem.

Degrees of freedom in nonlinear regression

Example. Consider the linear regression model

$$Y = X\beta + \varepsilon$$

where $X \in \mathbb{M}(n, p)$, $\beta \in \mathbb{R}^p$ and ε is $\mathcal{N}(0, \sigma^2 I_n)$.

Degrees of freedom in nonlinear regression

Example. Consider the linear regression model

$$Y = X\beta + \varepsilon$$

where $X \in \mathbb{M}(n, p)$, $\beta \in \mathbb{R}^p$ and ε is $\mathcal{N}(0, \sigma^2 I_n)$.

Sparse estimation can be obtained by the LASSO estimator

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \, \| Y - X\beta \|_2^2 + \lambda \|\beta\|_1$$

Degrees of freedom in nonlinear regression

Example. Consider the linear regression model

$$Y = X\beta + \varepsilon$$

where $X \in \mathbb{M}(n, p)$, $\beta \in \mathbb{R}^p$ and ε is $\mathcal{N}(0, \sigma^2 I_n)$.

Sparse estimation can be obtained by the LASSO estimator

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \| Y - X\beta \|_2^2 + \lambda \|\beta\|_1$$

and λ can be chosen by introducing the **generalization error** of $\hat{\beta}_\lambda$

$$\mathrm{Err}_\beta(\hat{\beta}_\lambda) = E_\beta \| Y^* - X\hat{\beta}_\lambda \|_2^2,$$

where Y^* is independent of Y and has the same distribution as Y, and choosing λ as the minimizer of an estimate of $\mathrm{Err}_\beta(\hat{\beta}_\lambda)$.

Degrees of freedom in nonlinear regression

In the linear regression model, Tibshirani and Taylor (2012) showed that

$$E_\beta \| Y^* - X\hat{\beta}_\lambda \|_2^2 = E_\beta \| Y - X\hat{\beta}_\lambda \|_2^2 + 2\sigma^2 E_\beta \text{ rank } X_{\mathcal{A}_\lambda},$$

where $\mathcal{A}_\lambda = \{i \leq p \mid \hat{\beta}_\lambda \neq 0\}$.

Degrees of freedom in nonlinear regression

In the linear regression model, Tibshirani and Taylor (2012) showed that

$$E_\beta \| Y^* - X\hat{\beta}_\lambda \|_2^2 = E_\beta \| Y - X\hat{\beta}_\lambda \|_2^2 + 2\sigma^2 E_\beta \text{ rank } X_{\mathcal{A}_\lambda},$$

where $\mathcal{A}_\lambda = \{i \leq p \mid \hat{\beta}_\lambda \neq 0\}$. Therefore, we may estimate $\text{Err}_\beta(\hat{\beta}_\lambda)$ by

$$\widehat{\text{Err}}_\beta(\hat{\beta}_\lambda) = \| Y - X\hat{\beta}_\lambda \|_2^2 + 2\hat{\sigma}^2 \text{ rank } X_{\mathcal{A}_\lambda}$$

and define $\hat{\lambda} = \text{argmin}_{\lambda \geq 0} \widehat{\text{Err}}_\beta(\hat{\beta}_\lambda)$.

Degrees of freedom in nonlinear regression

In the linear regression model, Tibshirani and Taylor (2012) showed that

$$E_\beta \| Y^* - X\hat{\beta}_\lambda \|_2^2 = E_\beta \| Y - X\hat{\beta}_\lambda \|_2^2 + 2\sigma^2 E_\beta \text{ rank } X_{\mathcal{A}_\lambda},$$

where $\mathcal{A}_\lambda = \{i \leq p \mid \hat{\beta}_\lambda \neq 0\}$. Therefore, we may estimate $\text{Err}_\beta(\hat{\beta}_\lambda)$ by

$$\widehat{\text{Err}}_\beta(\hat{\beta}_\lambda) = \| Y - X\hat{\beta}_\lambda \|_2^2 + 2\hat{\sigma}^2 \text{ rank } X_{\mathcal{A}_\lambda}$$

and define $\hat{\lambda} = \text{argmin}_{\lambda \geq 0} \widehat{\text{Err}}_\beta(\hat{\beta}_\lambda)$.

This yields a methodology for **choosing λ in the linear regression case**.

Degrees of freedom in nonlinear regression

In the linear regression model, Tibshirani and Taylor (2012) showed that

$$E_\beta \| Y^* - X\hat{\beta}_\lambda \|_2^2 = E_\beta \| Y - X\hat{\beta}_\lambda \|_2^2 + 2\sigma^2 E_\beta \text{ rank } X_{\mathcal{A}_\lambda},$$

where $\mathcal{A}_\lambda = \{i \leq p \mid \hat{\beta}_\lambda \neq 0\}$. Therefore, we may estimate $\text{Err}_\beta(\hat{\beta}_\lambda)$ by

$$\widehat{\text{Err}}_\beta(\hat{\beta}_\lambda) = \| Y - X\hat{\beta}_\lambda \|_2^2 + 2\hat{\sigma}^2 \text{ rank } X_{\mathcal{A}_\lambda}$$

and define $\hat{\lambda} = \text{argmin}_{\lambda \geq 0} \widehat{\text{Err}}_\beta(\hat{\beta}_\lambda)$.

This yields a methodology for **choosing λ in the linear regression case**.

Let us try this for a more complicated model (though still less complicated than the Ornstein-Uhlenbeck model).

Degrees of freedom in nonlinear regression

We consider the nonlinear regression model

$$Y = \varphi(\beta) + \varepsilon$$

where $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ is continuous and ε is $\mathcal{N}(0, \sigma^2 I_n)$.

Degrees of freedom in nonlinear regression

We consider the nonlinear regression model

$$Y = \varphi(\beta) + \varepsilon$$

where $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ is continuous and ε is $\mathcal{N}(0, \sigma^2 I_n)$.

Consider an estimator $\hat{\beta}$ and introduce

Degrees of freedom in nonlinear regression

We consider the nonlinear regression model

$$Y = \varphi(\beta) + \varepsilon$$

where $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ is continuous and ε is $\mathcal{N}(0, \sigma^2 I_n)$.

Consider an estimator $\hat{\beta}$ and introduce

$$\mathrm{df}(\hat{\beta}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}_\beta(Y_i, \varphi(\hat{\beta})_i)$$

Degrees of freedom in nonlinear regression

We consider the nonlinear regression model

$$Y = \varphi(\beta) + \varepsilon$$

where $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ is continuous and ε is $\mathcal{N}(0, \sigma^2 I_n)$.

Consider an estimator $\hat{\beta}$ and introduce

$$\mathrm{df}(\hat{\beta}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}_\beta(Y_i, \varphi(\hat{\beta})_i)$$

$$\mathrm{df}_S(\hat{\beta}) = E_\beta \operatorname{div} \varphi(\hat{\beta})$$

Degrees of freedom in nonlinear regression

We consider the nonlinear regression model

$$Y = \varphi(\beta) + \varepsilon$$

where $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ is continuous and ε is $\mathcal{N}(0, \sigma^2 I_n)$.

Consider an estimator $\hat{\beta}$ and introduce

$$\mathrm{df}(\hat{\beta}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}_\beta(Y_i, \varphi(\hat{\beta})_i)$$
$$\mathrm{df}_S(\hat{\beta}) = E_\beta \operatorname{div} \varphi(\hat{\beta})$$

which we call the **degrees of freedom** and the **Steinian degrees of freedom**, respectively, when defined. Here, the divergence can be a weak type of divergence.

Degrees of freedom in nonlinear regression

From Efron (2004), it holds that

$$\mathsf{Err}(\hat{\beta}) = \mathsf{MSPE}(\hat{\beta}) + 2\sigma^2 \mathsf{df}(\hat{\beta})$$

Degrees of freedom in nonlinear regression

From Efron (2004), it holds that

$$\mathsf{Err}(\hat{\beta}) = \mathsf{MSPE}(\hat{\beta}) + 2\sigma^2 \mathsf{df}(\hat{\beta})$$

while Stein (1981) showed that under certain differentiability conditions (almost differentiability) on $\hat{\beta}$, we have

$$\mathsf{df}(\hat{\beta}) = \mathsf{df}_S(\hat{\beta}).$$

Degrees of freedom in nonlinear regression

From Efron (2004), it holds that

$$\mathsf{Err}(\hat{\beta}) = \mathsf{MSPE}(\hat{\beta}) + 2\sigma^2 \mathsf{df}(\hat{\beta})$$

while Stein (1981) showed that under certain differentiability conditions (almost differentiability) on $\hat{\beta}$, we have

$$\mathsf{df}(\hat{\beta}) = \mathsf{df}_S(\hat{\beta}).$$

This is important because while $\mathsf{df}(\hat{\beta})$ is **hard to estimate**, we can always obtain an unbiased estimate of $\mathsf{df}_S(\hat{\beta})$ as

$$\widehat{\mathsf{df}}_S(\hat{\beta}) = \mathsf{div}\, \varphi(\hat{\beta}).$$

Degrees of freedom in nonlinear regression

From Efron (2004), it holds that

$$\mathsf{Err}(\hat{\beta}) = \mathsf{MSPE}(\hat{\beta}) + 2\sigma^2 \mathsf{df}(\hat{\beta})$$

while Stein (1981) showed that under certain differentiability conditions (almost differentiability) on $\hat{\beta}$, we have

$$\mathsf{df}(\hat{\beta}) = \mathsf{df}_S(\hat{\beta}).$$

This is important because while $\mathsf{df}(\hat{\beta})$ is **hard to estimate**, we can always obtain an unbiased estimate of $\mathsf{df}_S(\hat{\beta})$ as

$$\widehat{\mathsf{df}}_S(\hat{\beta}) = \mathsf{div}\, \varphi(\hat{\beta}).$$

This allows estimation of the generalization error and therefore allows **model selection**, in our particular case of interest **sparse model selection** through choice of λ.

Degrees of freedom in nonlinear regression

The almost differentiability conditions are problematic. We showed the following (under different regularity conditions):

Degrees of freedom in nonlinear regression

The almost differentiability conditions are problematic. We showed the following (under different regularity conditions):

Theorem I. For estimators of the type $\hat{\beta} = \text{argmin}_{\beta \in K} \|Y - \varphi(\beta)\|_2^2$, with $K \subseteq \mathbb{R}^p$ compact, it holds that $\text{df}(\hat{\beta}) \geq \text{df}_S(\hat{\beta})$.

Degrees of freedom in nonlinear regression

The almost differentiability conditions are problematic. We showed the following (under different regularity conditions):

Theorem I. For estimators of the type $\hat{\beta} = \text{argmin}_{\beta \in K} \|Y - \varphi(\beta)\|_2^2$, with $K \subseteq \mathbb{R}^p$ compact, it holds that $\text{df}(\hat{\beta}) \geq \text{df}_S(\hat{\beta})$.

Theorem II. With $\hat{\beta}$ as above and K the centered L^1-ball of radius s,

$$\widehat{\text{df}}_S(\hat{\beta}) = \text{tr}\, J^{-1}_{(\mathcal{A},\mathcal{A})} G_{(\mathcal{A},\mathcal{A})} - \frac{\gamma_{\mathcal{A}}^t J^{-1}_{(\mathcal{A},\mathcal{A})} G_{(\mathcal{A},\mathcal{A})} J^{-1}_{(\mathcal{A},\mathcal{A})} \gamma_{\mathcal{A}}}{\gamma_{\mathcal{A}}^t J^{-1}_{(\mathcal{A},\mathcal{A})} \gamma_{\mathcal{A}}}.$$

Degrees of freedom in nonlinear regression

The almost differentiability conditions are problematic. We showed the following (under different regularity conditions):

Theorem I. For estimators of the type $\hat{\beta} = \text{argmin}_{\beta \in K} \|Y - \varphi(\beta)\|_2^2$, with $K \subseteq \mathbb{R}^p$ compact, it holds that $\text{df}(\hat{\beta}) \geq \text{df}_S(\hat{\beta})$.

Theorem II. With $\hat{\beta}$ as above and K the centered L^1-ball of radius s,

$$\widehat{\text{df}}_S(\hat{\beta}) = \text{tr}\, J^{-1}_{(\mathcal{A},\mathcal{A})} G_{(\mathcal{A},\mathcal{A})} - \frac{\gamma_{\mathcal{A}}^t J^{-1}_{(\mathcal{A},\mathcal{A})} G_{(\mathcal{A},\mathcal{A})} J^{-1}_{(\mathcal{A},\mathcal{A})} \gamma_{\mathcal{A}}}{\gamma_{\mathcal{A}}^t J^{-1}_{(\mathcal{A},\mathcal{A})} \gamma_{\mathcal{A}}}.$$

Theorem III. For an estimator $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \|Y - \varphi(\beta)\|_2^2 + \lambda \|\beta\|_1$,

$$\widehat{\text{df}}_S(\hat{\beta}) = \text{tr}\, J^{-1}_{(\mathcal{A},\mathcal{A})} G_{(\cdot,\mathcal{A})}.$$

Thanks to:

My advisor

Niels R. Hansen

My non-advisor co-authors

Marloes H. Maathuis

Benjamin Falkeborg

Particular long-suffering teachers

Martin Jacobsen

Ernst Hansen