Graphical structure learning in multivariate stochastic processes

Søren Wengel Mogensen (Lund University)

joint work with
Niels Richard Hansen (University of Copenhagen) and
Daniel Malinsky (Columbia University)

January 25, 2022

INDEPENDENT RESEARCH FUND DENMARK

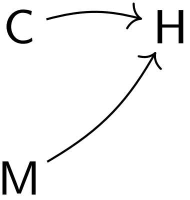
Conditional independence

Conditional independence of random variables is found in many places in the quantitative fields of research. We can even use it to distinguish data-generating structures (see, e.g., Spirtes and Zhang [2018])

Conditional independence

Conditional independence of random variables is found in many places in the quantitative fields of research. We can even use it to distinguish data-generating structures (see, e.g., Spirtes and Zhang [2018])

An example from everyday life (**C**offee, **H**eadache, **M**onday?),



M is exogenous

C is exogenous

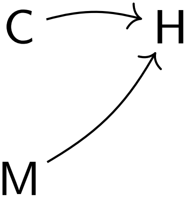
$H = f_H(M, C, \varepsilon_H)$

C and M are (marginally) independent.

Conditional independence

Conditional independence of random variables is found in many places in the quantitative fields of research. We can even use it to distinguish data-generating structures (see, e.g., Spirtes and Zhang [2018])

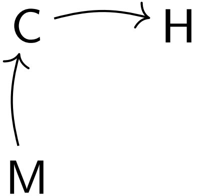
An example from everyday life (**C**offee, **H**eadache, **M**onday?),



M is exogenous

C is exogenous

$H = f_H(M, C, \varepsilon_H)$



M is exogenous

$C = f_C(M, \varepsilon_C)$

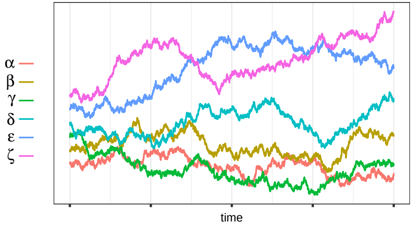
$H = f_H(C, \varepsilon_H)$

C and M are (marginally) independent.

H and M are conditionally independent given C.

Conditional independence

Let's say we sample data from a dynamical system. What is a useful independence relation to understand the data? Can we learn about the underlying structure from such a relation?

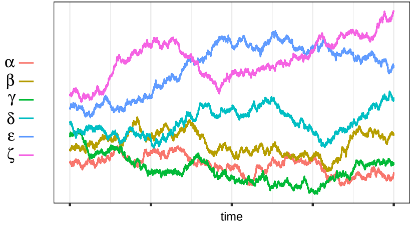


The Ornstein-Uhlenbeck process

As a particularly simple and well-behaved example model[1], we consider an Ornstein-Uhlenbeck process, $X_t = (X_t^1, \dots, X_t^n)^T$,

$$\mathrm{d}X_t = M(X_t - A)\mathrm{d}t + D\mathrm{d}W_t,$$

where M, D are $n \times n$ matrices, D is diagonal, and A is a vector of length n.



[1] It is possible to use this framework in much more general processes.

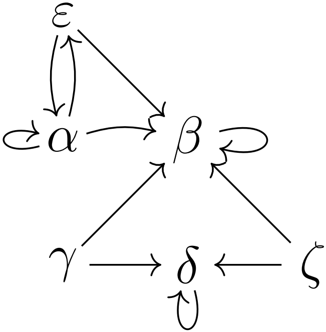
Graphical representation

A *directed graph* is a pair (V, E) where V is a set of vertices and E set of edges. The structure of the $n \times n$ matrix M can be represented by a directed graph, $\mathcal{D}$, with $V = \{1, \ldots, n\}$ and

$$\alpha \rightarrow_{\mathcal{D}} \beta \Leftrightarrow M_{\beta\alpha} \neq 0.$$

$$M = \begin{array}{c} \\ \end{array} \begin{array}{cccccc} \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \end{array}$$

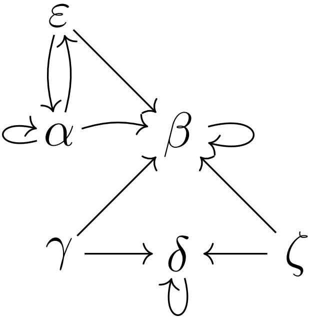
$$M = \begin{pmatrix} * & & & & * & \\ * & * & * & & * & * \\ & & & & & \\ & & * & * & & * \\ * & & & & & \\ & & & & & \end{pmatrix} \begin{array}{c} \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \\ \zeta \end{array}$$

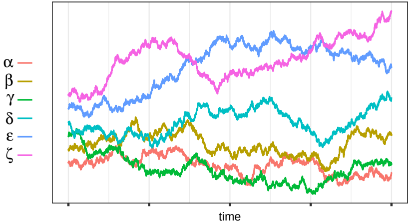


We use Greek letters to denote the vertices, $1 \sim \alpha$, $2 \sim \beta$, etc.

Graphical representation

What is a general way to define a graph which describes some sort of structure in the stochastic process, X?





Stochastic differential equations

We consider a filtered probability space, $(\Omega, \mathcal{F}, P, \mathcal{F}_t)$. We will assume that $X_t = (X_t^1, \ldots, X_t^n)^T$ solves

$$X_t = X_0 + \int_0^t \lambda(X_s)\ \mathrm{d}s + \int_0^t \sigma(X_s)\ \mathrm{d}W_s,\ \ t \in [0, T],$$

where W_t is an n-dimensional standard Brownian motion. The process $\lambda_t = (\lambda_t^1, \ldots, \lambda_t^n)$ is called the *drift*. In differential notation,

$$\mathrm{d}X_t = \lambda(X_t)\ \mathrm{d}t + \sigma(X_t)\ \mathrm{d}W_t,\ \ t \in [0, T].$$

In this talk, $\sigma(X_t) = D$ for a fixed diagonal matrix with positive entries on the diagonal. One can also allow a non-diagonal D in which case a more general class of graphs can be used [Mogensen and Hansen, 2022].

Local independence

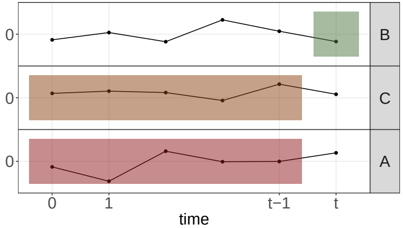
Let λ_t^β be the drift of the β-process. For $C \subseteq V$, let $\mathcal{F}_t^C$ be the completed and right-continuous version of $\sigma(\{X_s^\gamma, s \leq t, \gamma \in C\})$.

Definition (Local independence, Schweder [1970])

We say that B is locally independent of A given C, and write $A \not\to B \mid C$, if for all $\beta \in B$, there exists an $\mathcal{F}_t^C$-adapted version of the process

$$t \mapsto E(\lambda_t^\beta \mid \mathcal{F}_t^{A \cup C}).$$

Local independence is analogous to Granger causality in discrete-time processes.



Local independence, example

Let's say $n = 3$ so $V = \{1, 2, 3\}$,

$$\mathrm{d}X_t = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix} X_t \, \mathrm{d}t + \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix} \mathrm{d}W_t$$

where $W_t = (W_t^1, W_t^2, W_t^3)^T$. Is $\{1\}$ (the B-set) locally independent of $\{3\}$ (the A-set) given $\{1, 2\}$ (the C-set)? The drift of 1 is

$$-X_t^1 + X_t^2,$$

so we look at $E\left(-X_t^1 + X_t^2 \mid \mathcal{F}_t^{\{1,2,3\}}\right)$.

Note that local independence is not symmetric, i.e., $A \not\rightarrow B \mid C$ does not imply $B \not\rightarrow A \mid C$.

Local independence

Why is local independence a natural thing to consider?

Theorem (Kailath and Poor [1998], Rogers and Williams [2000])

(Some regularity conditions omitted). Let

$$X_t = X_0 + \int_0^t \lambda_s \, \mathrm{d}s + W_t$$

such that W_t is an n-dimensional standard Brownian motion. Let $(\mathcal{G}_t)$ be a subfiltration. Then

$$X_t = X_0 + \int_0^t \mathsf{E}(\lambda_s \mid \mathcal{G}_s) \, \mathrm{d}s + \bar{W}_t$$

and W and $\bar{W}$ have the same distribution.

However, if $\mathcal{G}_t = \sigma(X_s : s \leq t)$, W is a martingale with respect to $\mathcal{F}_t$, $\bar{W}$ with respect to $\mathcal{G}_t$.

Local independence graphs

The *local independence graph* [Didelez, 2008] is the *directed graph* (DG), $\mathcal{D}$, such that for $\alpha, \beta \in V$

$$\alpha \not\rightarrow_{\mathcal{D}} \beta \Leftrightarrow \alpha \not\rightarrow \beta \mid V \setminus \{\alpha\}.$$

It is possible to construct a purely graphical algorithm to deduce (general) local independences from a local independence graph. This is called μ-separation (analogous to d/m-separation). We use the notation $A \perp_\mu B \mid C$ $[\mathcal{D}]$ to denote that B is μ-separated from A given C in $\mathcal{D}$.

We let $\mathcal{I}$ denote a set of local independences, $\mathcal{I} = \{A \not\rightarrow B \mid C\}$, and we let $\mathcal{I}(\mathcal{D})$ denote the set of μ-separations implied by a graph, $\mathcal{D}$.

Faithfulness

Definition (Faithfulness)

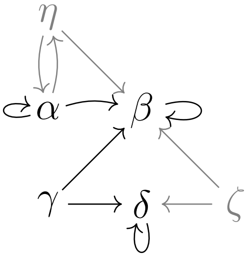
We say that $\mathcal{I}$ and $\mathcal{D}$ are *faithful* if $\mathcal{I} = \mathcal{I}(\mathcal{D})$, i.e. if

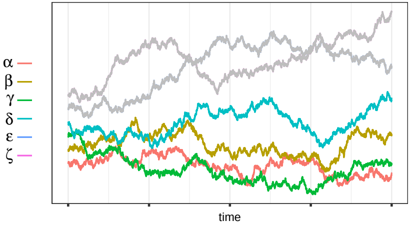
$$A \perp_\mu B \mid C \; [\mathcal{D}] \Leftrightarrow A \not\rightarrow B \mid C.$$

$\Rightarrow$ holds with respect to the underlying local independence graph, $\mathcal{D}$, under quite general conditions. If we also have *faithfulness*, then the graph encodes precisely the local independences that hold in the distribution.

Marginalization

Often it is only reasonable to assume that we observe a subset of the processes in the stochastic system, $O \subseteq V$.





No assumptions are made about the number of unobserved processes or their connections.

Graphical marginalization

We would like a graph on nodes $O \subseteq V$ that expresses the separations in $\mathcal{D} = (V, E)$, i.e., a graph $\mathcal{G} = (O, F)$ such that for all $A, B, C \subseteq O$

$$A \perp_\mu B \mid C \ [\mathcal{D}] \Leftrightarrow A \perp_\mu B \mid C \ [\mathcal{G}],$$

and a procedure to construct $\mathcal{G}$ from $\mathcal{D}$. We can use

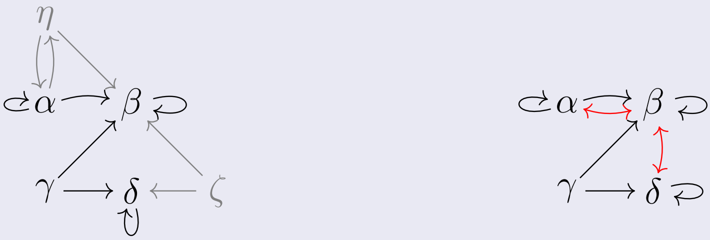
- the class of *directed mixed graphs* (DMGs), and
- *latent projection* [Verma and Pearl, 1991, Richardson et al., 2017, Mogensen and Hansen, 2020].

In a DMG two nodes α and β can be joined by any subset of edges $\{\alpha \to \beta, \beta \to \alpha, \alpha \leftrightarrow \beta\}$.

Graphical marginalization

Example

Let $O = \{\alpha, \beta, \gamma, \delta\}$ below. The marginal (over O) independence models are equal for the two DMGs.



Markov equivalence

Let $\mathcal{G} = (V, E)$ and $\bar{\mathcal{G}} = (V, \bar{E})$ be DMGs. We say that $\mathcal{G}$ and $\bar{\mathcal{G}}$ are *Markov equivalent* if for all $A, B, C \subseteq V$,

$$A \perp_\mu B \mid C \; [\mathcal{G}] \Leftrightarrow A \perp_\mu B \mid C \; [\bar{\mathcal{G}}].$$

We use $[\mathcal{G}]$ to denote the Markov equivalence class of $\mathcal{G}$. What can we say about the Markov equivalence classes?

Markov equivalence

Theorem (Mogensen and Hansen [2020])

Let $\mathcal{G} = (V, E)$ be a DMG. The equivalence class of $\mathcal{G}$ has a greatest element $\mathcal{N} = (V, F)$. That is, if $\tilde{\mathcal{G}} = (V, \tilde{E})$ is Markov equivalent with $\mathcal{G}$, then $\tilde{E} \subseteq F$.

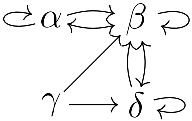
We use the theorem to define the following graph which represents the entire Markov equivalence class.

Definition

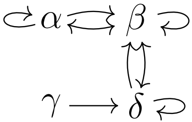
Let $\mathcal{N} = (V, F)$ be a maximal DMG. The corresponding *directed mixed equivalence graph* (DMEG) is the triple $(V, \bar{F}, \tilde{F})$ such that $F = \bar{F} \dot{\cup} \tilde{F}$ where for each $e \in F$ it holds that $e \in \bar{F}$ if and only if $e \in \mathcal{G}$ for every $\mathcal{G} \in [\mathcal{N}]$.

An equivalence class and its DMEG

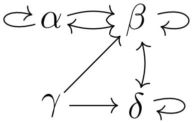
①



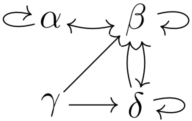
②



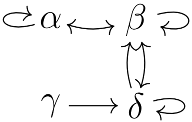
③



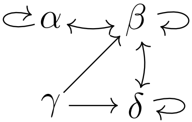
④



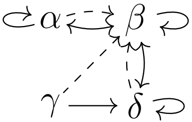
⑤



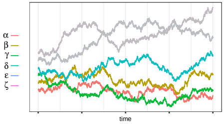
⑥



⑦



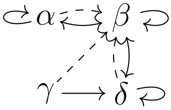
This talk in a single diagram

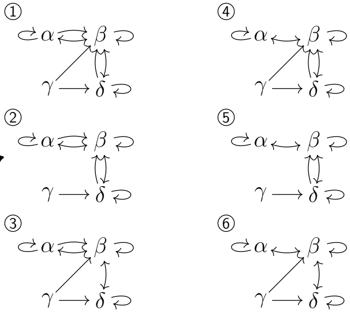


local independence test

$$\mathcal{I}^O = \{\alpha \not\to \gamma \mid \emptyset, \\ \alpha \not\to \gamma \mid \gamma, \\ \alpha \not\to \gamma \mid \delta, \ldots\}$$

learning algorithm





The components of the diagram

- Tests of local independence: Thams and Hansen [2021], Petersen and Hansen [2021]
- Algorithms: Meek [2014], Mogensen et al. [2018], Mogensen [2020]
- Markov equivalence: Mogensen and Hansen [2020, 2022]

Acknowledgments

Research sponsored by VILLUM FONDEN (research grant 13358) and Independent Research Fund Denmark (research grant 0164-00023B).

Thanks for your attention!

References I

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

Thomas Kailath and H Vincent Poor. Detection of stochastic processes. *IEEE Transactions on Information Theory*, 44(6):2230–2231, 1998.

Christopher Meek. Toward learning graphical and causal process models. In *CI at UAI*, pages 43–48, 2014.

Søren Wengel Mogensen. Causal screening in dynamical systems. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pages 310–319. PMLR, 2020.

Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1): 539–559, 2020.

Søren Wengel Mogensen and Niels Richard Hansen. Graphical modeling of stochastic processes driven by correlated noise. *Bernoulli*, 2022. to appear.

References II

Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Lasse Petersen and Niels Richard Hansen. Nonparametric conditional local independence testing. 2021.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. Available at https://arxiv.org/abs/1701.06686, 2017. URL `https://arxiv.org/pdf/1701.06686.pdf`.

L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales: Vol. 2, Itô calculus*. Cambridge University Press, 2000.

Tore Schweder. Composable markov processes. *Journal of Applied Probability*, 7 (2):400–410, 1970.

Peter Spirtes and Kun Zhang. Search for causal models. In *Handbook of Graphical Models*, pages 439–470. CRC Press, 2018.

References III

Nikolaj Thams and Niels Richard Hansen. Local independence testing for point processes. *arXiv preprint arXiv:2110.12709*, 2021.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, University of California, Los Angeles, 1991.