Stochastic differential equations as causal models

Alexander Sokol

Joint work with Niels Richard Hansen

19 April, 2013

Agenda

1. Review of stochastic differential equations
2. SDEs as models of causality
3. Identifiability of intervention distributions
4. Comparison with the do-calculus
5. Comparison with time series models

Review of stochastic differential equations

We consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ satisfying the usual conditions.

The stochastic integral $\int_0^t H_s \,\mathrm{d}X_s$ can be defined for integrand processes $(H_t)_{t\geq 0}$ and integrators $(X_t)_{t\geq 0}$ satisfying suitable regularity properties, and behaves "like an integral" in the sense that for example

$$\lim_n \sum_{k=1}^{2^n} H_{t(k-1)/2^n}(X_{tk/2^n} - X_{t(k-1)/2^n}) = \int_0^t H_s \,\mathrm{d}X_s,$$

where the limit is in probability.

The class of "well-behaved integrators" is the space of **semimartingales**.

Review of stochastic differential equations

Because of the irregularity of the sample paths of the integrators, the stochastic integral behaves differently than the ordinary Lebesgue integral. Itô's formula states that for $f \in C^2(\mathbb{R})$ and a semimartingale $(X_t)_{t\geq 0}$,

$$
\begin{aligned}
f(X_t) = f(X_0) + \int_0^t f'(X_{s-})\,\mathrm{d}X_s + \frac{1}{2}\int_0^t f''(X_{s-})\,\mathrm{d}[X^c]_s \\
+ \sum_{0<s\leq t} f(X_s) - f(X_{s-}) - f'(X_s)\Delta X_s
\end{aligned}
$$

where $\Delta X_t = X_t - X_{t-}$, the jump of X at time t and $[X^c]$ is the quadratic variation process of the continuous local martingale part X^c of X.

Review of stochastic differential equations

Example. For a Brownian motion $(W_t)_{t\geq 0}$, it holds that

$$W_t^2 = 2\int_0^t W_s \,\mathrm{d}W_s + t,$$

which is distinct from the ordinary result

$$t^2 = 2\int_0^t s\,\mathrm{d}s.$$

This difference essentially arises from the fact that the sample paths of Brownian motion have nonzero quadratic variation, meaning that

$$[W]_t = \lim_{n\to\infty}\sum_{k=1}^{2^n}(W_{tk/2^n} - W_{t(k-1)/2^n})^2 = t.$$

Review of stochastic differential equations

Using the stochastic integral, we may consider equations in X such as

$$X_t = 1 + \int_0^t X_{s-} \, \mathrm{d}Z_s,$$

where Z is some semimartingale. This equation can be written more suggestively in "differential form" as

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

We refer to such equations as **stochastic differential equations** (SDEs). The solution to this particular equation, for example, is

$$X_t = \exp\left(Z_t - \frac{1}{2}[Z^c]_t\right) \prod_{0<s\leq t} (1 + \Delta Z_s) \exp\left(-\Delta Z_s\right).$$

Review of stochastic differential equations

In general, we are most interested in multivariate SDEs. Consider:

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

where $\mathbb{M}(p, d)$ is the space of real $p \times d$ matrices. We then consider the equation

$$X_t^i = X_0^i + \sum_{j=1}^{d} \int_0^t a_{ij}(X_{s-}) \, \mathrm{d}Z_s^j \qquad \text{for } i \leq p \tag{$\dagger$}$$

which has a unique solution whenever a is Lipschitz. (†) can be written more compactly using differential and matrix notation as

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

Review of stochastic differential equations

SDEs of this type are regularly used in fields such as for example:

- Mathematical finance, for modeling financial assets.
- Chemistry, for modeling reaction networks.
- Biology, for modeling membrane potentials of neurons.

Review of stochastic differential equations

Some benefits of SDE modeling are:

- In simple cases, tractable formulas for expressions of interest.
- Well suited for capturing time dependency.
- Well suited for high-frequency data.

Some drawbacks are:

- In advanced cases, closed-form expressions are rarely available.
- Numerical computations of interest can be quite demanding.
- The behaviour of multidimensional SDEs is not well understood.

SDEs as models of causality

As the argument t in X_t frequently denotes time, it is often natural to interpret the solution (X_t) to (†) in terms of causality: 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 $X_{t+\Delta}$.

In the following, we formalize a notion of intervention in SDEs and show that this notion essentially is equivalent to a version of the above idea.

SDEs as models of causality

Definition. Consider some $m \leq p$ and $c \in \mathbb{R}$. The postintervention SDE arising from making the intervention $X^m := c$ in the SDE (†) is the $p-1$ dimensional SDE

$$Y_t^i = Y_0^i + \sum_{j=1}^{d} \int_0^t b_{ij}(Y_{s-}) \, \mathrm{d}Z_s^j \qquad \text{for } i \neq m \tag{††}$$

where $b : \mathbb{R}^{p-1} \to \mathbb{M}(p-1, d)$, $b(y_1, \ldots, y_p) = a(y_1, \ldots, c, \ldots, y_p)$ and the c is on the m'th coordinate, and $Y_t^m = c$. In differential form, the intervention takes us from the p dimensional SDE

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

to the $p-1$ dimensional SDE

$$\mathrm{d}Y_t = b(Y_t) \, \mathrm{d}Z_t.$$

SDEs as models of causality

Our first objective is to give an interpretation of what this definition means. We will do this using the notion of Euler schemes.

Definition. Fix $T > 0$ and consider $\Delta > 0$ such that $N = T/\Delta$ is natural. The Euler scheme $(X^{\Delta}_{k\Delta})_{k\leq N}$ for $(\dagger)$ over $[0, T]$ with step size Δ is the set of p-dimensional variables defined by putting $X^{\Delta}_0 = X_0$ and recursively defining

$$X^{\Delta}_{k\Delta} = X^{\Delta}_{(k-1)\Delta} + a(X^{\Delta}_{(k-1)\Delta})(Z_{k\Delta} - Z_{(k-1)\Delta}).$$

Theorem (Protter, Émery). Assume that a is Lipschitz. Connecting the variables $(X^{\Delta}_{k\Delta})$ into a function $(X^{\Delta}_t)_{t\leq T}$ in a suitable way and letting Δ tend to zero, it holds that $\sup_{t\leq T} \mid X_t - X^{\Delta}_t \mid$ tends to zero in probability.

SDEs as models of causality

A structural equation model is implicit in the recursive definition of the Euler scheme: $(X^{\Delta}_{k\Delta})_{k\leq N}$ constitutes a SEM of $p(N+1)$ one-dimensional variables with:

- Functional relationships:

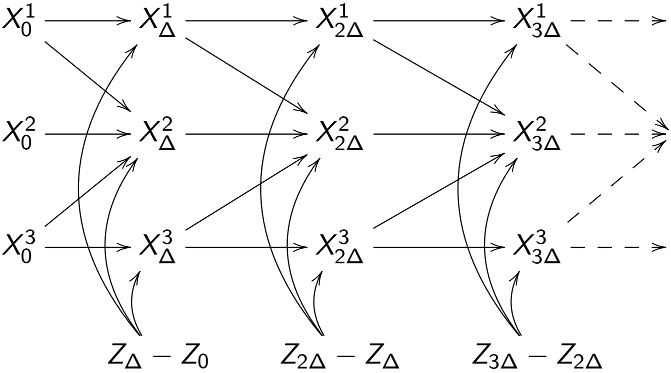
$$(X^{\Delta}_{k\Delta})^i = (X^{\Delta}_{(k-1)\Delta})^i + \sum_{j=1}^{d} a_{ij}(X^{\Delta}_{(k-1)\Delta})(Z^j_{k\Delta} - Z^j_{(k-1)\Delta}).$$

- Noise variables $(Z_{k\Delta} - Z_{(k-1)\Delta})_{k\leq N}$.
- DAG G defined by letting $((i_1, j_1), (i_2, j_2))$ be an edge if and only if $i_2 = i_1 + 1$ and $j_1 = j_2$ or $j_1 \neq j_2$ and $a_{j_1 \cdot}$ is not independent of the j_2'th coordinate.

Note that given the mapping a, this DAG is not learned: Rather, for fixed a, the above constitutes a rule for which DAG we think of as the “true” DAG for the SEM, and in practice it is a which is to be estimated.

SDEs as models of causality

Below, an illustration of the DAG:



SDEs as models of causality

The following result is then immediate:

Proposition. Fix $T > 0$ and $\Delta > 0$ such that $N = T/\Delta$ is natural. The Euler scheme SEM for the SDE (††) is equal to the postintervention SEM obtained by making the do-calculus intervention $X^{\Delta}_{k\Delta} := c$ for all $k \leq N$ in the Euler scheme SEM for the SDE (†).

Essentially, this means that the following diagram commutes:

$$\begin{tikzcd}
\boxed{\text{Observational Euler SEM}} \arrow[r] \arrow[d] & \boxed{\text{Observational SDE}} \arrow[d] \\
\boxed{\text{Postintervention Euler SEM}} \arrow[r] & \boxed{\text{Postintervention SDE}}
\end{tikzcd}$$

SDEs as models of causality

What is necessary to justify the use of this DAG?

- That for small Δ, the relationship between the variables can be approximated well by functional relationships.
- That causality propagates forward in time.
- That there are no “instantaneous” dependencies.
- That the driving semimartingales Z are unaffected by interventions.

SDEs as models of causality

A summary:

- We defined a notion of intervention of the type $X^m := c$ for SDEs.
- This notion of intervention is consistent with the limiting result of do-calculus interventions in the Euler schemes for the SDE using a particular DAG derived from the function a.
- If we believe that this notion of intervention represents the true data-generating mechanism, we will be able to understand the effect of interventions if we can identify the true SDE, in particular we need to estimate the function a.

Identifiability of intervention distributions

In practice, we only observe the distribution of the solution to an SDE. This motivates the following:

Question. If the solutions of the two SDEs

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

and

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

have the same distribution, can the distributions of the solutions to the postintervention SDEs nonetheless be different?

We will argue that in most commonly occurring cases, **the answer is no.**

Identifiability of intervention distributions

To prove our result, we will use the theory of continuous-time Markov processes.

Definition. A family of transition probabilities on $\mathbb{R}^p$ is a family of probability measures $P_t(x, \cdot)$ on $(\mathbb{R}^p, \mathcal{B}_p)$ for $t \geq 0$ and $x \in \mathbb{R}^p$ such that

- $P_0(x, \cdot)$ is the Dirac measure in x.
- $(t, x) \mapsto P_t(x, B)$ is measurable for all Borel sets B.
- It holds that $P_{t+s}(x, B) = \int P_s(y, B) P_t(x, \mathrm{d}y)$.

A process $(X_t)_{t \geq 0}$ is a Markov process with transition probabilities $P_t(x, \cdot)$ if $P(X_{t+s} \in B \mid \mathcal{F}_t) = P_s(X_t, B)$ for all $t, s \geq 0$.

Intuition. $P_t(x, B)$ is the probability of moving from x to B in a time period of t.

Identifiability of intervention distributions

For f bounded and measurable, defining $(P_t f)(x) = \int f(y) P_t(x, \mathrm{d}y)$, we can think of P_t as a linear operator on the space $\mathbf{b}\mathcal{B}_p$ of bounded Borel measurable mappings $f : \mathbb{R}^p \to \mathbb{R}^p$. (P_t) then becomes a contraction semigroup of operators $P_t : \mathbf{b}\mathcal{B}_p \to \mathbf{b}\mathcal{B}_p$.

We say that (P_t) is a Feller semigroup if it holds that:

- For all $t \geq 0$, P_t maps $C_0(\mathbb{R}^p)$ into itself.
- For all $f \in C_0(\mathbb{R}^p)$, $t \mapsto P_t f$ is continuous at zero.

Being a Feller semigroup ensures sufficient analytic properties to yield a rich semigroup and Markov process theory. In the following, we will only be considering Feller semigroups.

Identifiability of intervention distributions

For a Feller semigroup, we define

$$Af = \lim_{t \to 0^+} \frac{P_t - P_0}{t},$$

when the limit exists, and say that A is the generator of $(P_t)_{t \geq 0}$. The limit always exists on a dense subset $\mathcal{D}(A)$ of $C_0(\mathbb{R}^p)$, the **domain** of the generator. For Feller semigroups, the generator uniquely identifies the semigroup.

Identifiability of intervention distributions

Intuition. A semigroup is “like” a set of exponential operators. For example, if $\mathcal{D}(A) = C_0(\mathbb{R}^p)$ and A is bounded, then $P_t = \exp(tA)$, where the exponential is defined by its Taylor series.

Example. p-dimensional Brownian motion is a Markov process with Feller semigroup, the domain of the generator A is $C_0^2(\mathbb{R}^p)$ and for $f \in C_0^2(\mathbb{R}^p)$, it holds that

$$Af(x) = \sum_{i=1}^{p} \frac{\partial^2 f}{\partial x_i^2}(x).$$

Identifiability of intervention distributions

Next, recall that a Lévy process is a stochastic process $(X_t)_{t\geq 0}$ with initial value zero and stationary and independent increments. The following proposition is our main identifiability result.

Proposition. 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. If the solutions have the same distribution, then the postintervention distributions are the same as well.

Identifiability of intervention distributions

Sketch of proof.

1. Because Z and $\tilde{Z}$ are Lévy processes, the solutions to the SDEs are Markov process with Feller semigroups. Identify expressions for the generators on $C_c^2(\mathbb{R}^p)$.
2. Using our assumptions and the generators, identify the ways in which a and $\tilde{a}$ and the parametes of the Lévy processes are equal.
3. Note that the postintervention SDEs of both SDEs also have Lévy processes as driving semimartingales. Identify expressions for the generators on $C_c^2(\mathbb{R}^{p-1})$.
4. Use this to obtain equality of the generators of the postintervention processes on $C_c^2(\mathbb{R}^{p-1})$.

Identifiability of intervention distributions

Conclusions. Assume that:

- The data-generating mechanism is an SDE with a driving semimartingale which is a Lévy process.
- The data-generating mechanism satisfies the criteria of no instantaneous dependencies and driving semimartingales unaffected by interventions.

Then it holds that:

- The notion of intervention as defined is appropriate.
- Postintervention distributions can be identified from the observational distribution.

Comparison with the do-calculus

We now give an example comparing our notion of intervention in SDEs with do-calculus type interventions. This will be used to motivate the use of SDE models. Assume given

- A real $p \times p$ matrix B.
- A p-dimensional vector A.
- A real invertible $p \times p$ matrix σ.

Consider the Ornstein-Uhlenbeck SDE

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

where W is a p-dimensional Brownian motion. This SDE has a unique solution which can be given in closed form.

Comparison with the do-calculus

If the data-generating process is such that our notion of intervention is applicable, then B determines which coordinates depend causally on others: If $B_{ij} = 0$, this means that X_i is locally unaffected by an intervention in X_j.

Now consider the case where B is lower triangular with negative values in the diagonal (corresponding to negative feedback loops ensuring stability of the system). Modulo self-loops, this means that the causality of the system exhibits no cyclic behaviour.

Comparison with the do-calculus

In this case, the SDE has a stationary distribution, which is a Gaussian distribution with mean A and variance Σ solving the Lyapounov equation

$$B\Sigma + \Sigma B^t = -\sigma\sigma^t.$$

If we observe the system in equilibrium at discrete timepoints, we obtain a stationary time series $(X^{\Delta}_{k\Delta})_{k\geq 0}$ of Gaussian distributions with parameters (A, Σ).

However, the zeroes of Σ^{-1} in general **do not** correspond to zeroes of B. Therefore, the conditional independence structure of the equilibrium distribution does not reflect the causal structure of the system. This is a reason for using an SDE model instead of a model with no time component.

Comparison with time series models

Still considering the case of the Ornstein-Uhlenbeck SDE, the time series $(X^{\Delta}_{k\Delta})_{k\geq 0}$ obtained by observing the process discretely has the distribution of an autoregressive process: $(X^{\Delta}_{k\Delta})_{k\geq 0}$ has the same distribution as $(Y^{\Delta}_{k\Delta})_{k\geq 0}$, where

$$Y_{k\Delta} = (I - \exp(\Delta B))A + \exp(\Delta B) Y_{(k-1)\Delta} + \varepsilon_k,$$

and (ε_k) are i.i.d. Gaussian with mean zero and variance

$$\int_0^{\Delta} \exp(sB)\sigma\sigma^t \exp(sB^t)\,\mathrm{d}s.$$

Question. If observations are in discrete time anyway, why not just use a discrete-time AR(1) model instead of the OU model?

Comparison with time series models

Answer. To make the question more precise, assume that the true underlying model is the OU model. We seek to understand what we should do if we "forget" about this and only model our observations as a discrete-time AR(1) model.

In an AR(1) model, we would assume that

$$X_{k\Delta} = \alpha^{\Delta} + C^{\Delta} X_{(k-1)\Delta} + \varepsilon_k^{\Delta}$$

for some Gaussian i.i.d. noise variables (ε_k^{Δ}) with mean zero and variance Γ^{Δ}.

Comparison with time series models

As the underlying true model is the OU model, we have

$$\begin{aligned}
\alpha^{\Delta} &= (I - \exp(\Delta B))A \\
C^{\Delta} &= \exp(\Delta B) \\
\Gamma^{\Delta} &= \int_0^{\Delta} \exp(sB)\sigma\sigma^t \exp(sB^t)\,\mathrm{d}s.
\end{aligned}$$

The true causal structure is reflected in the zeroes of B. This is in general **not** the same as the zeroes of C^{Δ}.

Comparison with time series models

Noting that we have the relationships

$$B = \frac{1}{\Delta} \log C^{\Delta}$$
$$B = \lim_{\Delta \to 0} \frac{1}{\Delta}(C^{\Delta} - I),$$

we can obtain the zeroes of B from C^{Δ} by using the right-hand sides in the above (expressions which are, however, motivated by the OU process).

Conclusion. Understanding the true causal structure through "direct" discrete-time modeling is possible, but it is necessary to keep in mind that the causal structure may be determined by zeroes in the **generator** of the process and not in the **transition probabilities**. This is also important when considering penalization.

Thank you for your attention!