POLITECNICO DI TORINO
Repository ISTITUZIONALE

Inference in conditioned dynamics through causality restoration

*Original*
Inference in conditioned dynamics through causality restoration / Braunstein, A., Catania, G., Dall'Asta, L., Mariani, M., Muntoni, A.P.. - In: SCIENTIFIC REPORTS. - ISSN 2045-2322. - 13:(2023). [10.1038/s41598-023-33770-3]

*Availability:*
This version is available at: 11583/2980214 since: 2023-07-12T09:49:06Z

*Publisher:*
Nature Publishing Group

*Published*
DOI:10.1038/s41598-023-33770-3

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

25 November 2026

scientific reports

Check for updates

OPEN

# Inference in conditioned dynamics through causality restoration

Alfredo Braunstein[1,2,3], Giovanni Catania[4], Luca Dall'Asta[1,2,3,5], Matteo Mariani[1]✉ & Anna Paola Muntoni[1]

**Estimating observables from conditioned dynamics is typically computationally hard. While obtaining independent samples efficiently from unconditioned dynamics is usually feasible, most of them do not satisfy the imposed conditions and must be discarded. On the other hand, conditioning breaks the causal properties of the dynamics, which ultimately renders the sampling of the conditioned dynamics non-trivial and inefficient. In this work, a Causal Variational Approach is proposed, as an approximate method to generate independent samples from a conditioned distribution. The procedure relies on learning the parameters of a generalized dynamical model that optimally describes the conditioned distribution in a variational sense. The outcome is an effective and unconditioned dynamical model from which one can trivially obtain independent samples, effectively restoring the causality of the conditioned dynamics. The consequences are twofold: the method allows one to efficiently compute observables from the conditioned dynamics by averaging over independent samples; moreover, it provides an effective unconditioned distribution that is easy to interpret. This approximation can be applied virtually to any dynamics. The application of the method to epidemic inference is discussed in detail. The results of direct comparison with state-of-the-art inference methods, including the soft-margin approach and mean-field methods, are promising.**

The method we will present is rather general and applies to a wide family of stochastic processes. We will thus first describe it below in complete generality, and delay its description for a specific important application (namely the risk assessment problem in epidemic spreading processes) to the following section.

Let us denote by $\mathbb{P}[\mathbf{x}] = \mathbb{P}[x(0), \ldots, x(k\Delta t)]$ the probability distribution of trajectories $\mathbf{x}$ of a (known) dynamical model. Given a (hidden) realization $\mathbf{x}^*$, consider a set of observations $\mathscr{O} = (O_1, \ldots, O_M)$ sampled from a (known) conditional distribution $\mathbb{P}[\mathscr{O}|\mathbf{x}^*]$. The scope of this work is to devise an efficient method to infer information about $\mathbf{x}^*$ given $\mathscr{O}$, in particular, to be able to estimate averages over the posterior distribution

$$\mathbb{P}[\mathbf{x}|\mathscr{O}] = \mathbb{P}[\mathbf{x}]\mathbb{P}[\mathscr{O}|\mathbf{x}]\mathbb{P}[\mathscr{O}]^{-1}. \tag{1}$$

Although it might be generally feasible to sample efficiently from the prior $\mathbb{P}[\mathbf{x}]$, sampling from $\mathbb{P}[\mathbf{x}|\mathscr{O}]$ is normally difficult. A naive approach is given by importance sampling[1,2], that consists in evaluating the average of a function $f$ by first generating $M$ independent samples $\mathbf{x}^1, \ldots, \mathbf{x}^M$ from $\mathbb{P}[\mathbf{x}]$ and then computing

$$\langle f \rangle \approx \frac{\sum_{\mu=1}^{M} f(\mathbf{x}^\mu)\mathbb{P}[\mathscr{O}|\mathbf{x}^\mu]}{\sum_{\mu=1}^{M} \mathbb{P}[\mathscr{O}|\mathbf{x}^\mu]}. \tag{2}$$

Unfortunately, this method is impractical when observations deviate significantly from the typical case, as for the case in which $\mathbb{P}[\mathscr{O}|\mathbf{x}^\mu]$ becomes very small (or even zero), rendering the convergence of $\langle f \rangle$ to the true average value inefficient.

One reason for which sampling from $\mathbb{P}[\mathbf{x}]$ is usually feasible is that the causal structure induced by the dynamical nature of the stochastic process can be exploited to efficiently generate trajectories. The *causal* property of the stochastic dynamics lies in the fact that the state of the system at a given time depends naturally (in a stochastic way) on states at previous times. When considering discrete time-steps or epochs $0, \Delta t, 2\Delta t, \ldots$ (in the following discussion, for simplicity of notation, we will assume $\Delta t = 1$), this property implies that the distribution of trajectories of the stochastic dynamics assumes the following factorized form:

[1]DISAT, Politecnico di Torino, Corso Duca Degli Abruzzi 24, 10129 Turin, Italy. [2]INFN, Sezione di Torino, Turin, Italy. [3]Italian Institute for Genomic Medicine, IRCCS Candiolo, SP-142, 10060 Candiolo, TO, Italy. [4]Departamento de Física Téorica I, Universidad Complutense, 28040 Madrid, Spain. [5]Collegio Carlo Alberto, P.za Arbarello 8, 10122 Turin, Italy. ✉email: matteo.mariani@polito.it

$$\mathbb{P}[x(0),\dots,x(T)] = \prod_{t=0}^{T} \mathbb{P}[x(t)|x(t-1),\dots,x(0)], \tag{3}$$

where in the $t = 0$ term the conditioning part is empty and thus the probability is unconditioned. In most models, it is computationally simple (or at least feasible) to sample $x(t)$ from $\mathbb{P}[x(t)|x(t-1),\dots,x(0)]$, implying that (3) can be exploited to generate trajectories by sequentially sampling $x(0)$, then $x(1)$, etc.

The intrinsic difficulty associated with sampling from the conditioned distribution $\mathbb{P}\big[\mathbf{x}|\mathscr{O}\ \big]$ is a consequence of *causality breaking*[3] induced by the addition of the extra information in $\mathscr{O}$. In general, $\mathbb{P}[x(t)|x(t-1),\dots,x(0)]$ and $\mathbb{P}\big[x(t)|x(t-1),\dots,x(0),\mathscr{O}\ \big]$ are very different objects. For example, even if the former is time and space invariant, the latter is generally not, because this symmetry is typically broken by the observations. This difference ultimately implies that we cannot sample from the posterior distribution sequentially as in the unconditioned case (3). Although we can write an exact expression similar to (3),

$$\mathbb{P}\big[\mathbf{x}|\mathscr{O}\ \big] = \prod_{t=0}^{T} \mathbb{P}\big[x(t)|x(t-1),\dots,x(0),\mathscr{O}\ \big], \tag{4}$$

sampling from (4) is unfortunately still problematic. Indeed, the expression for $\mathbb{P}\big[x(t)|x(t-1),\dots,x(0),\mathscr{O}\ \big]$ is in general extremely difficult to compute and involves a marginalization over times $t' > t$ (with an exponential number of terms):

$$\mathbb{P}\big[x(t)|x(t-1),\dots,x(0),\mathscr{O}\ \big] \propto \sum_{x(t+1),\dots,x(T)} \prod_{t'=0}^{T} \mathbb{P}\big[x(t')|x(t'-1),\dots,x(0)\big]\mathbb{P}\big[\mathscr{O}\ |x(T),\dots,x(0)\big]. \tag{5}$$

This dependence on future times is in our opinion the real source of the causality breaking phenomenon.

When dynamics are unconditioned, i.e. causality applies, information is intuitively flowing from past to future. Although it is a very intuitive concept, the study of information flow is actually rather involved and it opens to interesting insights into the collective interactions among agents in agent-based systems. We refer the interested reader to[4,5]. The approach proposed here, called *Causal Variational Approach* (CVA), aims at providing a variational approximation of the posterior distribution $\mathbb{P}\big[\mathbf{x}|\mathscr{O}\ \big]$, for which causality features are restored and, therefore, independent samples can be efficiently generated from it. In particular, we propose to approximate $\mathbb{P}\big[\mathbf{x}|\mathscr{O}\ \big] \approx Q(\mathbf{x})$, where $Q(x(0),\dots,x(T)) = \prod_{t=0}^{T} q_t(x(t)|x(t-1),\dots x(0))$. This approach is formally exact. Indeed, if we set each $q_t(x(t)|x(t-1),\dots x(0)) = \mathbb{P}\big[x(t)|x(t-1),\dots,x(0),\mathscr{O}\ \big]$, we would recover the exact posterior, due to equation (4). However, this is in practice unfeasible, because it would require $q_t$ to depend on a huge (i.e. exponential in the size of the system) number of parameters. The general idea of CVA method is to restrict the functional space of $Q$ assuming the $q_t(x(t)|x(t-1),\dots,x(0))$ to have the same broad functional form of the unconstrained prior distribution $\mathbb{P}[x(t)|x(t-1),\dots,x(0)]$, retaining then the ability to efficiently compute it and sample from it, but generalizing it by the addition of extra parameters. This generalization will naturally allow for the spatial and/or time heterogeneity that is present in the corresponding terms in the posterior, and will be explained in detail for the specific models in the next sections. In particular, we chose in the approximating distribution of CVA to maintain the following properties (if they are present) of the prior distribution:

1. *Spatial Independence*. In agent-based models[6] on $N$ agents, $x(t) = (x_1(t),\dots,x_N(t))$ and most dynamical processes satisfy a spatial conditional independence property[7], namely that:

$$\mathbb{P}[x(t)|x(t-1),\dots,x(0)] = \prod_{i=1}^{N} \mathbb{P}[x_i(t)|x(t-1),\dots,x(0)], \tag{6}$$

   As this property is often crucial for efficient sampling from $\mathbb{P}[\mathbf{x}]$, CVA maintains it on the approximating $q_t$ i.e. $q_t(x(t)|x(t-1),\dots,x(0)) = \prod_i q_t^i(x^i(t)|x(t-1),\dots,x(0))$.
2. *Local interactions*. Each variable of the prior process, moreover, might depend only on a restricted (local) set of variables on a given contact network; we choose to preserve this dependence in $q_t$. Note that this property is in general not present in the posterior.
3. *Markovianity*. If the prior distribution defines a memory-less stochastic process[8], namely

$$\mathbb{P}[x(t)|x(t-1),\dots,x(0)] = \mathbb{P}[x(t)|x(t-1)], \tag{7}$$

   CVA extends this property to the approximating factors as well, $q_t^i\big(x^i(t)|x(t-1),\dots,x(0)\big) = q_t^i\big(x^i(t)|x(t-1)\big)$. Note that if additionally the observations are time-factorized, namely $\mathbb{P}[\mathscr{O}\ |\mathbf{x}] = \prod_{t=0}^{T} \mathbb{P}[\mathscr{O}_t|x(t)]$, then it can be shown (See Section IX of the Supplementary Information) that Markovianity extends to the posterior distribution, $\mathbb{P}[x(t)|x(t-1),\dots,x(0),\mathscr{O}\ ] = \mathbb{P}[x(t)|x(t-1),\mathscr{O}\ ]$.

There are simple but instructive examples where CVA leads to the exact posterior, see for example the SI epidemic model for $N = 2$ individuals in Section I of the Supplementary Information.

CVA can be used to tackle some difficult problems emerging in the field of epidemic inference, such as epidemic risk assessment from partial and time-scattered observations of cases, or the detection of the sources of infection. These problems have been recently addressed within a Bayesian probabilistic framework using

computational methods inspired by statistical physics[9–11], and generative neural networks[12]. With respect to existing similar approaches based on variational autoencoders (e.g.[12,13]), the CVA ansatz for the posterior distribution does not employ neural networks, has comparatively a much smaller set of parameters and allows for much simpler physical interpretations. In particular, risk assessment from contact tracing data is of major importance for epidemic containment, because having access to an accurate measure of the individual risk can pave the way to effective targeted quarantine plans based on contact tracing devices[14–16]. Moreover, in epidemic problems, there are quantities of interest that are not known a priori. An example is the infection rate of the disease. Our method can be used to compute such quantities, treated as hyperparameters of the CVA distribution. Being able to find the prior parameters of a distribution gives also the possibility to simplify the inference problem by adopting a simpler model. For example, in the context of epidemic inference, CVA allows one to study inference problems related to the SEIR model (introduced later) with an effective SI model, which is simpler than SEIR. This part, which we call *model reduction* is illustrated in detail in the Results section. After presenting the CVA in a general setting, its main features will be discussed by exploiting a conditioned random walk[17–19] as a toy model. Then, an application to the important problem of epidemic inference and risk assessment on dynamic contact networks is developed and analyzed in detail. We stress that the two reference cases, i.e. the epidemic inference and the conditioned random walk, represent two very different dynamical processes, the former continuous in time while the latter advances in discrete time-steps.

## Methods

The method is based on approximating the original constrained process by introducing an effective unconstrained causal process that is naturally consistent with the observations.

Let $Q_\theta(\mathbf{x})$ be the probability distribution of a generalized dynamics, parametrized by the vector $\theta$ of parameters. The best approximation to $\mathbb{P}\left[\mathbf{x}|\mathscr{O}\right]$ (in a precise variational sense) can be obtained by observing that Eq. (1) can be interpreted as a Boltzmann distribution $Z^{-1}\exp\left[-H(\mathbf{x})\right]$ with $H(\mathbf{x}) = -\log\mathbb{P}\left[\mathbf{x},\mathscr{O}\right]$ and $Z = \mathbb{P}\left[\mathscr{O}\right]$ and by minimizing the corresponding variational free energy[20], i.e.

$$\mathscr{F}(Q_\theta) := \int \mathrm{d}\mathbf{x} Q_\theta(\mathbf{x}) \log\frac{Q_\theta(\mathbf{x})}{\mathbb{P}\left[\mathbf{x},\mathscr{O}\right]} \tag{8}$$

$$= \left\langle \log\frac{Q_\theta(\mathbf{x})}{\mathbb{P}\left[\mathbf{x},\mathscr{O}\right]} \right\rangle_{Q_\theta}. \tag{9}$$

This quantity can be estimated efficiently by sampling the distribution $Q_\theta$. Note that $\mathscr{F}(Q_\theta) = D_{KL}\left(Q_\theta||\mathbb{P}\left[\mathbf{x}|\mathscr{O}\right]\right) - \log\mathbb{P}\left[\mathscr{O}\right]$ where $D_{KL}$ is the Kullback-Leibler divergence[21]. To optimize $\mathscr{F}$, a gradient descent method can be employed (see Section V of the Supplementary Information), where the gradient can be also estimated by means of sampling, indeed

$$\nabla_\theta\mathscr{F}(Q_\theta) = \left\langle \nabla_\theta \log Q_\theta(\mathbf{x}) \log\frac{Q_\theta(\mathbf{x})}{\mathbb{P}\left[\mathbf{x},\mathscr{O}\right]} \right\rangle_{Q_\theta}. \tag{10}$$

The crucial point in this estimation is that both $Q_\theta$ and $\mathbb{P}\left[\mathbf{x},\mathscr{O}\right] = \mathbb{P}[\mathbf{x}]\mathbb{P}\left[\mathscr{O}|\mathbf{x}\right]$ have explicit expressions that, due to their causal structure, can be efficiently computed using rejection-free sampling, at difference with $\mathbb{P}\left[\mathbf{x}|\mathscr{O}\right]$ and $\mathbb{P}\left[\mathscr{O}\right]$ which do not benefit from this property. The fact that the samples are independent allows for trivial parallelism in implementation. For a detailed description of the gradient descent optimization adopted in this work, we refer the reader to Section V of the Supplementary Information.

When a fixed point is reached, the corresponding distribution $Q_\theta(\mathbf{x})$ is the argument that (locally) minimizes the free energy in (9) therefore providing an approximation of the posterior distribution $\mathbb{P}\left[\mathbf{x}|\mathscr{O}\right]$. Finally, the result can be used to generate samples satisfying the constraints given by $\mathscr{O}$ and to compute interesting observables from them by efficiently computing sample averages.

**A toy model application: conditioned random walk.** Before introducing the main scenario where CVA is employed—i.e. on epidemic spreading models -, we first discuss a simple but instructive application of the method, that consists in generating an approximate probabilistic description for a conditioned random walk. A simple realization of the latter is a one-dimensional random walk, starting at site $x(0) = 0$. If the generating process is spatially homogeneous, the probability of every feasible trajectory $\mathbf{x}$ of length $T$ is $\mathbb{P}[\mathbf{x}] = 2^{-T}$. Note that every possible trajectory can be directly sampled by means of a causal generative process, namely a discrete-time Markov chain in which the conditional probability of a jump is $\mathbb{P}[x(t+1) = x(t) \pm 1\,|x(t)] = 1/2$. Fig. 1(a) displays a space-time representation for a set of realizations of such an unbiased random walk (black paths). For this process, let us imagine a procedure that, given a time instant $t^\mu$ and position $x^\mu$ in space, can test if the trajectory was at time $t^\mu$ to the left or right of position $x^\mu$, and denote the corresponding half-line as $W^\mu \subseteq \mathbb{Z}$. Assume that for a given unknown trajectory, we have $M$ observations of this kind $\mathscr{O} = (t^\mu, W^\mu)_\mu$ with $\mu = 1,\ldots,M$. The posterior probability of a trajectory $\mathbf{x}$ can be written as

$$\mathbb{P}\left[\mathbf{x}|\mathscr{O}\right] = \frac{\prod_{\mu=1}^{M}\mathbb{I}[x(t^\mu)\in W^\mu]}{\sum_{\mathbf{y}}\prod_{\mu=1}^{M}\mathbb{I}[y(t^\mu)\in W^\mu]}, \tag{11}$$

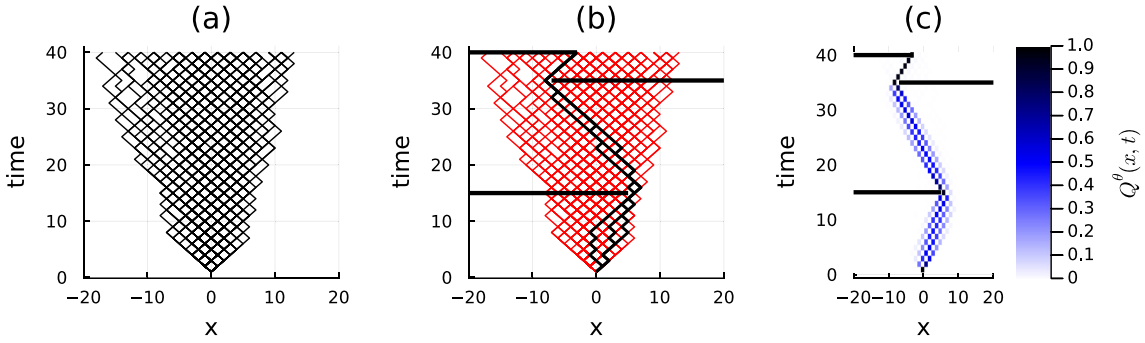


**Figure 1.** Panel (**a**) Unconditioned homogeneous random walk on a one-dimensional lattice. Time is reported on the vertical axis (up to $T = 40$) and the spatial coordinate $x$ is on the horizontal axis. Panel (**b**) Some trajectories are sampled from the unconditioned homogeneous distribution. The black (red) ones (do not) satisfy the constraints, i.e. they (do not) avoid the black horizontal barriers. The fraction of feasible trajectories among a given pool can be numerically estimated, and it approaches $10^{-6}$. In other words, only one of a million trajectories sampled from the unconditioned distribution satisfies the constraint. Panel (**c**) The distribution of the trajectories sampled from the CVA distribution. The color of each pixel indicates the probability for a trajectory to visit the corresponding state at a specific time.

where the numerator is 1 only if the trajectory $\mathbf{x}$ satisfies all observations, and zero otherwise. The denominator is the sum over all trajectories (so the variable $\mathbf{y}$ runs over the space of all the possible trajectories) of the numerator and plays the role of a normalization term for the posterior. The effect of $\mathbb{O}$ is to select (or constrain to) a subset of the trajectories of a free random walk, i.e. those compatible with the observations. One could naively sample trajectories from the free dynamics and then select only those compatible with $\mathbb{O}$. However, as depicted in Fig. 1, the fraction of trajectories compatible with the constraints might be very small to allow for a feasible computation: in the example of Fig. 1(b), where it is assumed that three regions at specific time steps (black horizontal barriers) cannot be crossed, several realizations of the unconstrained dynamics are discarded (red paths), while only a small fraction is kept (black paths). In this regard, the CVA provides an efficient way of generating trajectories compatible with the constraints by building up an effective probability distribution that is - by construction - compatible with the former. Within the framework provided by CVA, the following causal *ansatz* can be introduced for the conditioned random walk problem:

$$Q_\theta(\mathbf{x}) = \delta_{x(0),0} \prod_{t=0}^{T-1} \left[ r^t_{x(t)} \delta_{x(t+1),x(t)+1} + l^t_{x(t)} \delta_{x(t+1),x(t)-1} \right], \tag{12}$$

with $\theta = \{r^t_x\}^{t=1,\dots,T}_{x=-T,\dots,T}$ being the set of site-dependent and time-dependent rates to jump to the right, and $l^t_{x(t)} = 1 - r^t_{x(t)}$ the associated probabilities to jump to the left. We remark that Eq. (12) has the same functional form as the unconstrained distribution, i.e. it still represents the probability distribution of a random walk, but with heterogeneous (in general, both in space and time) jump rates. The distribution $Q_\theta$ requires $T(2T+1)$ parameters, where $2T+1$ is the total number of sites that can be visited by the realizations of the random walk. These parameters are sought by minimizing the KL distance between $Q_\theta$ and the posterior distribution Eq. (11). The resulting probability $Q_\theta(\mathbf{x})$ obtained using the CVA is characterized by heterogeneous rates $r^t_x$ whose dependence in time and space perfectly mirrors the constraints introduced by the barriers. The marginal distributions of the trajectories sampled from $Q_\theta(\mathbf{x})$ are represented in Fig. 1(c), where the color gradient is associated with the marginal probability of occurrence of each step.

**Epidemic models and observations.** From now on, we will consider a class of individual-based epidemic models describing a spreading process in a community of $N$ individuals, interacting through a (possibly dynamic) contact network. The overall state of the system at time $t$ (consisting of the state of each individual) is described by a vector $\mathbf{x}(t) \in \mathcal{X}^N$, where $\mathcal{X}$ is a finite set of possible health conditions (called compartments). The simplest, but already non-trivial, model of epidemic spreading is the discrete-time Susceptible-Infected (SI) model[22], in which $\mathcal{X} = \{S, I\}$ (corresponding to an individual being "susceptible" and "infected", respectively) where the only allowed transition occurs from state $S$ to state $I$. More precisely, each time $t$, if an infected individual $j$ is in contact with a susceptible individual $i$, the former can infect the latter (which moves into state $I$) with a transmission probability $\tilde{\lambda}_{ji}(t)$, sometimes called *transmissivity*. Since transmissions are independent, the individual transition probabilities are

$$\mathbb{P}[x_i(t+\Delta t) = S | x(t)] = \delta_{x_i(t),S} \prod_{j \neq i} \left(1 - \tilde{\lambda}_{ji}(t) \delta_{x_j(t),I}\right) \tag{13}$$

and $\mathbb{P}[x_i(t+\Delta t) = I | x(t)] = 1 - \mathbb{P}[x_i(t+\Delta t) = S | x(t)]$. A simple assumption is that time dependence only enters to describe the dynamic nature of the contact network (with $\tilde{\lambda}_{ji}(t) = 0$ if there is no contact between $j$

and $i$ at epoch $t$). More realistically, the transmission probability $\tilde{\lambda}_{ji}(t)$ should also depend on the current stage of infection of the infector $j$ (e.g. presence/absence of symptoms) and thus mainly on the time elapsed since her own infection, making the epidemic dynamics non-Markovian.

Assuming transmission probabilities proportional to $\Delta t$ in Eq. (13) and defining transmission rates as $\lambda_{ij}(t) = \lim_{\Delta t \to 0^+} \tilde{\lambda}_{ij}(t)/\Delta t$, a continuous-time version of the SI model is obtained. Both for the discrete-time and continuous-time SI models, the history of the epidemic process can be fully specified by the "infection times" $\{t_i\}_{i=1}^N$ of all individuals; we conventionally set $t_i = 0$ if individual $i$ is already infected at the initial time and $t_i = +\infty$ if $i$ is never infected during the whole epidemic process. In terms of the trajectory vector $\mathbf{t} := (t_1, \ldots, t_N)$, the probability pseudo-density of an epidemic history can be generally written as

$$\mathbb{P}[\mathbf{t}] = \prod_i \left\{ \gamma \delta(t_i) + (1-\gamma) \Lambda \left( \sum_{j \neq i} \mathbb{I}\left[t_j \leq t\right] \lambda_{ji}(t), t_i \right) \right\}, \tag{14}$$

where $\gamma$ denotes the probability of each individual to be a patient zero, and $\Lambda\left(f(t), b\right) = f(b) e^{-\int_{-\infty}^{b} f(t) dt}$ is the first-success distribution density of an event with rate $f(t)$. Hence, the quantity $\Lambda\left(\sum_{j \neq i} \mathbb{I}\left[t_j \leq t\right] \lambda_{ji}(t), t_i\right)$ is the probability density associated with the infection, at time $t_i$, of individual $i$ by one of its infectious contacts at previous times. Notice that when $\int_{-\infty}^{b} \Lambda\left(f, t\right) dt < 1$, it means that there is a non-zero probability of the individual remaining susceptible. In that case, we will formally assign the defect mass $1 - \int_{-\infty}^{b} \Lambda\left(f, t\right) dt$ to $t = \infty$. In addition to the epidemic model, a set of observations has to be defined. In real epidemics, observations mirror the outcomes of medical tests, namely the state of an individual $i$ at time $t$. For the sake of simplicity, an auxiliary variable $r \in \{+.-\}$ representing positive or negative tests, respectively, is defined, such that each observation can be encoded as a triplet $(i, t, r)$. Given the stochastic nature of clinical tests, it is assumed that the outcome $r$ of a test performed on individual $i$ at time $t$ obeys a known conditional distribution law $\mathbb{P}[r|t_i]$ where $t_i$ represents the infection time of individual $i$. When medical tests are affected by uncertainty, i.e. there exist non-zero false positive and false negative rates of the diagnostic tests, the conditional probability states

$$\mathbb{P}[r = +|t_i] = \left(1 - p_{\mathrm{FNR}}\right) \mathbb{I}[t_i \leq t] + p_{\mathrm{FPR}}\, \mathbb{I}[t_i > t] \tag{15a}$$

$$\mathbb{P}[r = -|t_i] = p_{\mathrm{FNR}}\, \mathbb{I}[t_i \leq t] + \left(1 - p_{\mathrm{FPR}}\right) \mathbb{I}[t_i > t] \tag{15b}$$

For a population undergoing $M$ individual test events, the set $\mathcal{O}$ of observations is then identified with the set of triplets $(i^\mu, t^\mu, r^\mu)$ for $\mu = 1, \ldots, M$. As in the random walk example, each observation constrains the dynamics: in the noise-less case (i.e. $p_{\mathrm{FNR}} = p_{\mathrm{FPR}} = 0$) the posterior distribution gives zero measure to all the epidemic trajectories that violate the observations. Given a realization of an epidemic model defined on a contact network, and the (possibly noisy) observation of the states of a subset of the individuals (at possibly different times), the epidemic risk assessment problem consists of estimating the epidemic risk, i.e. the risk of being infected, of the unobserved individuals at some specific time. In practice, it amounts to computing marginal probabilities from the posterior distribution

$$\mathbb{P}\left[x_i(t) = I | \mathcal{O}\right] = \int d\mathbf{t} \mathbb{I}[t_i \leq t] \mathbb{P}\left[\mathbf{t} | \mathcal{O}\right] \tag{16}$$

where $\int d\mathbf{t}$ denotes the integral over all infections times $t_1, \ldots, t_N$.

A richer epidemic model, that is often used as a testing ground for more realistic scenarios (see e.g. Refs.[23,24]), is the SEIR model[25], which includes also the Exposed (E) and Recovered (R) states, i.e. $\mathcal{X} = \{S, E, I, R\}$. The only allowed transitions are $S \to E$ (representing the contagion event), $E \to I$, and $I \to R$; the latter ones occur for each individual independently of the others, with latency and recovery rates $\nu_i$ and $\mu_i$, respectively. The previous representation in terms of transmission times can be straightforwardly generalized to the SEIR model (by introducing individual-wise infective and recovery times) as well as the definition of observations from clinical tests and the measure of the individual risk (see Supplementary Information online for additional details).

The choice of the parameters $\theta$ of the CVA *ansatz* reflects and somehow generalizes the features of the generative model $\mathbb{P}[\mathbf{t}]$: in the SI case for instance, for each individual $i$, heterogeneous infection rates $\lambda_i(t)$, zero-patient probabilities $\gamma_i$, and self-infection rates $\omega_i(t)$ are defined. Since $\lambda_i(t)$ and $\omega_i(t)$ are time-dependent quantities, an additional parametrization is introduced for computational purposes. Then, the trial distribution $Q_\theta(\mathbf{t})$ is optimized with respect to the full set of parameters. The total number of parameters for the inference in the SI model in a population of $N$ individuals is $7N$, while for the SEIR model is $13N$. We refer to Section III of the Supplementary Information for a more detailed discussion of the parameter choice and of the implementation of the gradient descent.

## Results on epidemic inference

**Performances on synthetic networks.** The performance of the CVA in reconstructing epidemic trajectories can be tested by measuring its ability in classifying the state of the unobserved individuals based on their predicted risk. The results are directly compared with those obtained using other inference techniques previously proposed in the literature, such as Belief Propagation[26] as implemented in[9] (sib), a Monte Carlo method (MC), a Soft Margin method (soft) adapted from[27], and simple heuristic methods based on Mean Field equations

(MF)[9] or on sampling (heu). A description of the implementation of the several methods used for comparison is provided in the Supplementary Information (Section VII) online.

For the sake of simplicity, we considered SI epidemic processes on proximity graphs[28], i.e. random graphs generated by proximity relationships between $N = 50$ individuals randomly drawn from a uniform distribution on a two-dimensional square region (a definition of proximity graphs is given in Section VIII of the Supplementary Information). Each instance corresponds to a different realization of both the dynamical network and the forward epidemic propagation. Observations $\mathcal{O}$ are noiseless, i.e. $p_{\mathrm{FNR}} = 0$, and performed on a randomly chosen fraction of the population at a fixed time $t_{\mathrm{obs}} = T$. The comparison among the different inference methods is performed by ranking the individual marginal probabilities of being in state $I$ at a chosen time $t^*$ and building the corresponding *receiving operating characteristic* (ROC) curve[29]. The AUC (area under the ROC curve) at a time $t^*$ is an indicator of the accuracy of the method in reconstructing the state of the individuals at that time. The AUC at initial time ($t^* = 0$, patient-zero problem) and at the final time ($t^* = T$, risk assessment) are shown in Fig. 2 as function of the number of observations available.

As one may expect, in both cases the average performances of all methods improve when the number of observations increases. In particular, the Soft Margin method is expected to converge to the exact results for this type of experiment when the number $N$ of individuals is small. The results obtained with CVA are very close (and closer than any other technique) to those obtained by means of Soft Margin (soft), even in the interesting and challenging regime with only a few observations.

To further investigate the performances of CVA against other state-of-the-art techniques, the AUC associated with the prediction of individual risk is quantified as a function of time, with two different observation protocols: (i) the states of a fraction of individuals are observed at observation times scattered over the duration of the simulation or (ii) observations are performed at the last time $t_{\mathrm{obs}} = T$. More precisely, in the first protocol observation times are randomly drawn a priori with uniform distribution in the interval $[1, T]$, and observations are biased towards tested-positive outcomes to mimic a realistic scenario where symptomatic, i.e. infected individuals, are more likely tested than susceptible ones. For these experiments, two realistic dynamic contact network instances are considered, one generated using the Spatio-temporal Epidemic Model (StEM) in continuous-time in Ref.[30] and the other using the discrete-time OpenABM model in Ref.[14] (see Section VIII of the Supplementary Information for a brief description of StEM and OpenABM models). For sake of simplicity, instead of adopting the complex epidemic dynamics described in Ref.[14] and Ref.[30], epidemic realizations are generated using a continuous-time SI model on these contact graphs.

A measure of the individual risk is computed according to all different methods (CVA, sib, soft, and MC), and the corresponding AUCs are shown as functions of time (in days), in Fig. 3(a, c) and (b, d) for OpenABM and the StEM respectively. For the latter only, we also consider different MC parameterizations, in particular when using $\delta \in \{12, 24, 48, 96\}$ hours; a further increase of $\delta$ does not carry any improvement of the results. The quantity $\delta$ is associated with the MC move's proposal (see Section VII of the Supplementary Information for a detailed description). Panels (a) and (b) are associated with the observations scattered in time, while panels (c) and (d) use observations at the last time only. In panel (a) simulations are run for $N = 2000$, while in panel (c) we set $N = 1000$. It is easy to see that, in panels (a) and (c), CVA (blue dots) is the best-performing method in terms of AUC; only MC (pink triangles) reaches comparable AUC for $t \sim T$. The results achieved by Belief Propagation are similar to those produced by CVA when the size of the graph is $N = 1000$ (panel (c)), and significantly

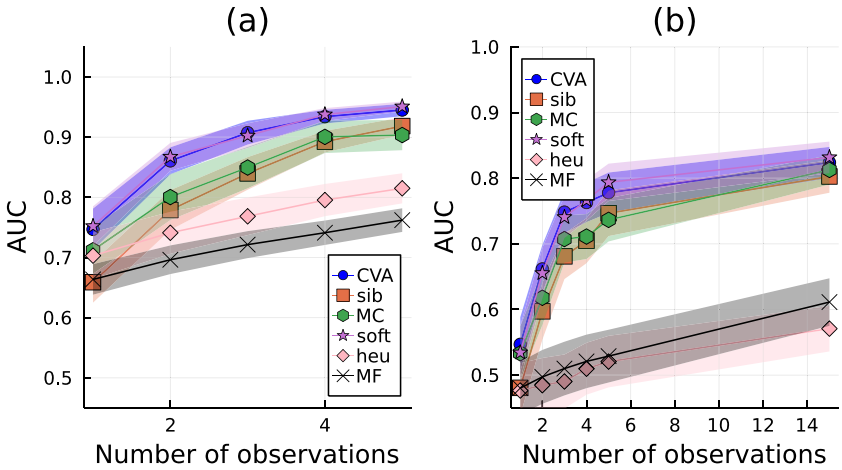


**Figure 2.** Area under the ROC (AUC) as a function of the number of observations for the risk assessment problem, i.e. $t^\star = T$, in panel (**a**), and for the patient-zero problem, $t^\star = 0$, panel (**b**). The simulated contact graph is a proximity network with average connectivity $2.2/N$. For both simulations in panels (**a**) and (**b**), the total number of individuals is $N = 50$, the probability of being the zero patient is set to $\gamma = 1/N$, and the infection rate is $\lambda = 0.1$. For each epidemic realization, the inference is performed for an increasing number of noiseless observations (here $p_{\mathrm{FNR}} = 0$) at time $t_{\mathrm{obs}} = T$. Thick lines and shaded areas indicate the averages and the standard errors computed over 40 different instances.

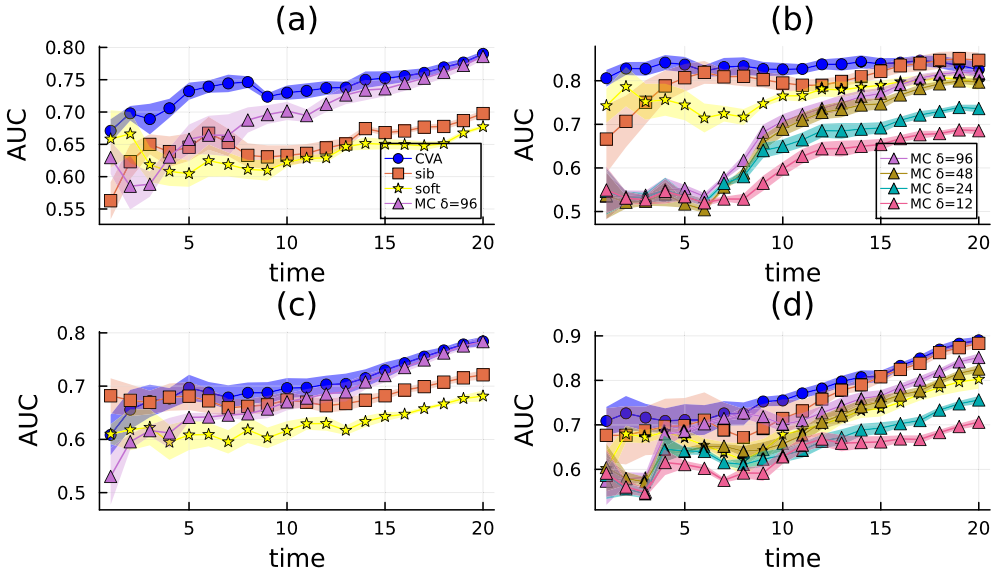


**Figure 3.** AUC associated with the prediction of the infected individuals, for the Causal Variational Approach (CVA), Belief Propagation (sib) and SoftMargin (soft), and MCMC (MC) as a function of time during the epidemic propagation of a SI model on several instances of dynamic contact network generated using the OpenABM model[14] (in panel (**a**) $N = 2000$, in (**c**) $N = 1000$) and the StEM in Ref.[30] (panels (**b**) and (**d**)) for $N = 904$. The infection rate is set to $\lambda = 0.15$ for the latter and $\lambda = 0.02$ for the former; observations are noiseless in both cases. For panels (**c**) and (**d**), observations are performed at the last time of the dynamics, i.e. $t_{\text{obs}} = T$. For the results in panels (**a**) and (**b**) observation times are extracted uniformly in the range $[1, T]$; at each observation time $t_{\text{obs}}$, infected nodes are observed with a biased probability equal to $1.1 \times N_I(t_{\text{obs}})/N$ where $N_I(t_{\text{obs}})$ is the number of infected individuals at time $t_{\text{obs}}$ and $N$ is the total number of individuals. The total number of observations is $n_{\text{obs}} = N \cdot 0.1$ for OpenABM and $n_{\text{obs}} = 100$ for the StEM.

deteriorate for $N = 2000$ (panel (a)). For the instances generated according to StEM in panels (b) and (d), the comparison reveals that CVA achieves the largest values of the AUC at all times and only Belief Propagation (sib, orange squares) performs comparably to CVA for the risk assessment problem, i.e. the inference at the last time of the dynamics. MC for $\delta \in \{48, 96\}$, approaches CVA performances in the last days while is not able to predict the zero patient. Indeed, the AUC associated with MC predictions for all parametrizations is slightly larger than 0.5 for $t < 5$ when observations are performed at the last time of the dynamics.

**Hyperparameters inference.** In the previous numerical experiments, the parameters of the generative SI model, i.e. the (homogeneous) infection rate $\lambda$ and the probability of being the zero patient $\gamma$, are assumed to be known. These quantities enter the CVA formalism as hyperparameters of the prior distribution which are often inaccessible in realistic applications, but can be estimated as those realizing the minimum of the free-energy $F = -\log \mathbb{P}[\mathcal{O}\ ]$. This can be achieved by gradient descent if the number $n_{\text{obs}}$ of available observations is sufficiently large (see Section VI of the Supplementary Information for details). An example of the quality of the parameter inference is provided by the following experiment. For a SI model with $n_{\text{obs}}$ and true parameters $(\gamma^\star, \lambda^\star) = (1/N, 0.1)$ (see the caption of Fig. 4 for further details), Fig. 4(a) shows a heatmap of $F$ computed at the convergence of CVA as function of the pair of values $(\gamma, \lambda)$ used as the hyperparameters of the corresponding prior distribution. The region attaining the lowest values also contains the true values $(\gamma^*, \lambda^*)$. The oriented paths (white arrows) in Fig. 4(a) represent the sequences of intermediate values of $\lambda$ and $\gamma$ obtained during the convergence process of CVA, starting from three different initial conditions. These traces show that trajectories end up in the same region, very close to where the true values are located (green star). Similar experiments, where the value of the zero patient probability is set to $\gamma = 1/N$ and the infection probability varies in the range $[0.05, 0.20]$ are performed. Similarly to the previous set-up, CVA is applied to infer the parameter $\lambda$. Figure 4(b) displays a scatter plot of the inferred values against the true ones. Results suggest a good agreement between the result of the inference and the generative process.

**Model reduction.** For viral diseases with sufficiently known transmission mechanisms, agent-based modeling using discrete-state stochastic processes has proven useful to build large-scale simulators of epidemic outbreaks and design containment strategies[24,30]. Such mathematical models are much more complex than the SI model analyzed previously, as they need to include additional specific features of real-world diseases. In particular, models may assume different infected states, characterized by a different capability of transmitting the virus and diverse sensitivity to diagnostic tests. Another important feature that can emerge from realistic transmissions is that individuals can stop being infectious, even before recovering from the infected state, because of the decay of their viral load. These ingredients may be effectively included in the SI and SEIR models by introduc-

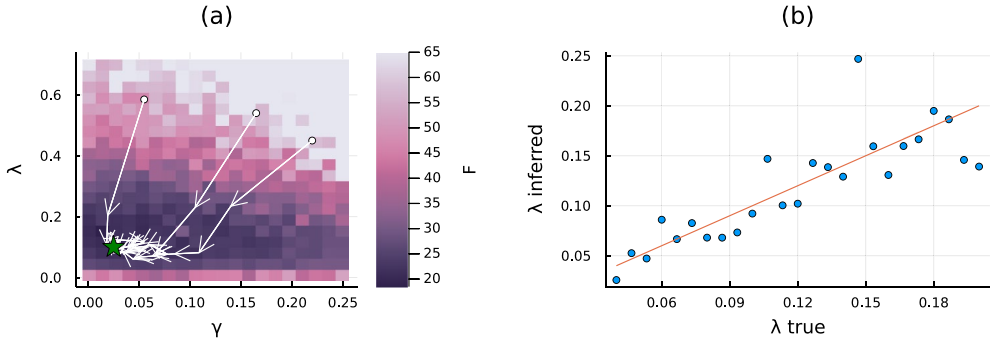


**Figure 4.** Panel (**a**) Heat map of the free energy ($F := -\log \mathbb{P}(\mathcal{O})$) computed at the convergence of CVA as a function of the assumed hyperparameters of the generative SI model. The experiment is performed on a proximity graph with $N = 50$ individuals and density $\rho = 2/N$; the epidemic model is characterized by the zero-patient probability $\gamma^* = 1/N$ and the infection rate $\lambda^* = 0.1$, shown here as a green star. We perform a large number of observations ($n_{\text{obs}} = 2N$) at uniformly randomly distributed times. As expected, the lowest values of this free energy are concentrated around the exact value ($\gamma^*$, $\lambda^*$). The oriented paths (white arrows) represent the convergence towards the minimum of $-\log \mathbb{P}[\mathcal{O}]$ obtained by performing a gradient descent algorithm over the hyperparameters starting from three different initial points in the plane ($\gamma$, $\lambda$). Panel (**b**) Scatter plot of inferred values for the infection probability against the ground truth. In these experiments, we fix and assume to know the zero patient probability $\gamma = 1/N$ while the infection parameter $\lambda$ is varied. For each $\lambda$ an epidemic simulation is performed and $n_{\text{obs}} = 10N$ observations are taken at uniformly randomly distributed times.

ing time-dependent infection rates, which is a natural assumption in the framework of the Causal Variational Approach (see Section II-III of the Supplementary Information). This property makes the latter a very suitable inference method to approximate unknown and possibly complex generative epidemic processes using classes of simpler probabilistic models. A simple test of such a potentiality is provided by the following example. Several epidemic realizations are generated with an SEIR prior model and the quality of the inference obtained by the Causal Variational Approach is evaluated when (i) the SEIR model is also used as an *ansatz* for the posterior distribution and (ii) when the posterior distribution is approximated with a simpler probabilistic model, such as the SI model. If the parameters of the generative SEIR model are known, the hyperparameters of the SEIR posterior are also known. The corresponding results (green diamonds) for the AUC as a function of time on a proximity

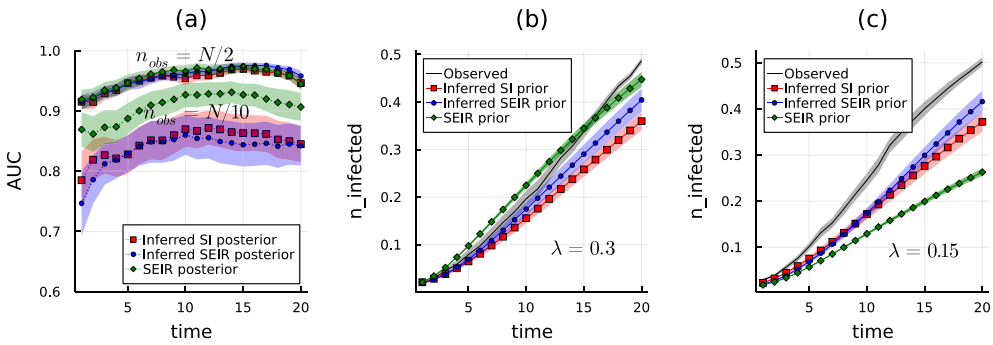


**Figure 5.** Effects of model reduction on inferential performances and generative capabilities. The numerical experiments are performed on a proximity graph with $N = 100$ individuals and density $2.2/N$. The observed epidemic realizations are generated using an SEIR model with $\gamma = 1/N$, $\lambda = 0.3$ (panels (**a**) and (**b**)) and 0.15 (panel (c)), latency delay $\nu = 0.5$ and recovery delay $\mu = 0.1$. Panel (**a**) Values of the AUC as a function of time obtained using the CVA in two observation regimes (when the number of observations is $n_{obs} = N/10$ and $n_{obs} = N/2$), with the three different inferred posterior distributions: an SEIR model with known hyperparameters (green diamonds), an SEIR model with unknown hyperparameters (blue circles), and a SI model with unknown hyperparameters (red squares). Shaded areas represent the error around the average value, computed using 22 instances. Panels (**b**) and (**c**) The average fraction of infected individuals as a function of time estimated using the correct SEIR prior model (green diamonds), an SEIR prior with the inferred hyperparameters (blue circles), and a SI prior model with the inferred hyperparameters (red squares). The regimes shown correspond to unbiased observations (center, for $\lambda = 0.3$), and to observations preferentially sampled from large outbreaks (right, for $\lambda = 0.15$). The black curves represent the same quantity computed from the observed epidemic realizations. Shaded areas represent the standard error computed from 40 realizations of the dynamics.

graph are displayed in Fig. 5(a). Otherwise, the hyperparameters of the SEIR posterior can be inferred by means of the CVA (blue circles). Finally, the *ansatz* for the posterior distribution can be simplified to a SI model, and the corresponding hyperparameters can be inferred as well within the CVA (red squares). The overall quality of the inference depends on the possibly different regimes of information contained in the observations. Strikingly, when the generative model is not known, the results from SEIR-based and SI-based inference are always very close to each other. For a sufficiently large number of observations, such results are also close to those obtained with the SEIR posterior and known hyperparameters.

From a generative perspective, the inferred hyperparameters can be also interpreted as the epidemic parameters of some prior model, from which epidemic realizations can be sampled. It is natural to ask what are the statistical properties of such generative processes compared to the original one, from which the observations were sampled. Figure 5 also shows, in two different regimes, as a function of time the average number of infected individuals estimated from the original SEIR prior model (green diamond), the SEIR prior model with inferred hyperparameters (blue circles) and the SI prior model with inferred hyperparameters (red squares). The average number of infected individuals computed over the realizations from which the observations are sampled is also displayed (black line). The regimes shown in Fig. 5 correspond to unbiased observations (panel (b), for $\lambda = 0.3$), and to observations preferentially sampled from large outbreaks (panel (c), for $\lambda = 0.15$). Although the discrepancy between the different curves is significant, the moderate difference between predictions obtained using SEIR and SI prior models with inferred hyperparameters suggests that model reduction is only a minor source of information bias.

## Conclusions

Sampling from the posterior distribution of a conditioned dynamical process can be computationally hard. In this work, a novel computational method to accomplish this task, called the Causal Variational Approach, was put forward. The Causal Variational Approach is based on the idea of inferring the posterior with an effective, unconditioned, dynamical model, whose parameters can be learned by minimizing a corresponding free energy functional. An insight into the potential of the method is obtained by analyzing a one-dimensional conditioned random walk in which some regions of space are forbidden. The CVA produces a generalized random walk process, with space-dependent and time-dependent jump rates, whose unconditioned realizations satisfy the imposed constraints. An application of greater practical interest concerns epidemic inference, in particular the risk assessment from partial and time-scattered observations. For simple stochastic epidemic models, such as SI and SEIR, taking place on contact networks of moderately small size, the CVA performs better or as well as the best methods currently available. Moreover, the variational nature of the method allows one to estimate the parameters of the original epidemic model that generated the observations, which enter the CVA in the form of hyperparameters. Since the CVA approximates the posterior distribution of the epidemic process by learning a set of generalized individual-based, time-dependent parameters, even with a rather simple *ansatz* for the epidemic model, such as an SI model, inference from observations coming from more complex epidemic processes can be performed. In fact, a generalized SI model with time-dependent infection rates and self-infection rates allows one to accommodate many features of real-world epidemic diseases, such as time-varying viral load and transmissivity, incubation, and recovery. The performances of the Causal Variational Approach do not seem to suffer from model reduction from SEIR to SI, suggesting that simplified epidemic models could be effective for inference also in real-world cases. The Causal Variational Approach is very flexible and, employing sampling to perform estimates, it can be applied virtually to any dynamics for which the latter can be carried out efficiently. In particular, the method can thus be applied to inference problems involving recurrent epidemic processes, such as the SIS model[31] or other models (e.g.[32]) where immunity decays over time. There are, however, some limitations. The CVA relies on the fact that the functional form of the posterior should be similar to the one of the prior. This is not true in general. For example, let us take an epidemic SI model in which the zero-patient probability $\gamma$ is infinitesimally small. If one individual is tested positive at a certain time, then the posterior distribution is substantially different from the prior. In particular, in the prior process each individual is the patient zero independently with probability $\gamma$, while in the posterior there is a strong (anti)correlation: indeed the measure will concentrate on trajectories with exactly one infected individual (and this is impossible to reproduce with independent patient zero probabilities). Of course, this example is extremely contrived, as the probability that infection occurs at all in this system, and thus such a test result can be obtained, is infinitesimally small as well. Moreover, in this case the problem can be simply solved by adopting a more natural distribution for the initial state (either by using a non-infinitesimal initial infection probability in the prior, or by adopting a single initial infection in the test distribution $q$, see also Supplementary Information). Nevertheless, it is a simple example in which the prior functional form is substantially different from the one of the posterior.

## Data availability

All data is generated using simulations and can be reproduced by following the prescriptions provided in the main text and in Supplementary Information online. A public GitHub repository containing a Julia implementation of the algorithm and notebooks to reproduce the results of this work is available at https://github.com/abraunst/Causality.git.

Received: 25 January 2023; Accepted: 18 April 2023
Published online: 05 May 2023

## References

1. Newman, M. E. J. & Barkema, G. T. *Monte Carlo Methods in Statistical Physics* (Clarendon Press, Oxford, 1999).

2. MacKay, D. J. *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, 2003).
3. Biroli, G. & Kurchan, J. Metastable states in glassy systems. *Phys. Rev. E* **64**, 016101. https://doi.org/10.1103/PhysRevE.64.016101 (2001).
4. James, R. G., Ayala, B. D. M., Zakirov, B. & Crutchfield, J. P. *Modes of information flow* 1808.06723 (2018).
5. Sattari, S. *et al.* Modes of information flow in collective cohesion. *Sci. Adv.* **8**, eabj1720. https://doi.org/10.1126/sciadv.abj1720 (2022).
6. Macal, C. M. & North, M. J. Agent-based modeling and simulation. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 86–98. https://doi.org/10.1109/WSC.2009.5429318 (2009).
7. Dawid, A. P. Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B (Methodological)* **41**, 1–15. https://doi.org/10.1111/j.2517-6161.1979.tb01052.x (1979).
8. Norris, J. R. *Markov Chains* (Cambridge University Press, Cambridge, 1998).
9. Baker, A. *et al.* Epidemic mitigation by statistical inference from contact tracing data. *Proc. Natl. Acad. Sci.* **118**, e2106548118. https://doi.org/10.1073/pnas.2106548118 (2021).
10. Herbrich, R., Rastogi, R. & Vollgraf, R. CRISP: A Probabilistic Model for Individual-Level COVID-19 Infection Risk Estimation Based on Contact Data, https://doi.org/10.48550/arXiv.2006.04942 (2022). arXiv:2006.04942 [cs, stat].
11. O'Neill, P. D. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Math. Biosci.* **180**, 103–114. https://doi.org/10.1016/S0025-5564(02)00109-8 (2002).
12. Biazzo, I., Braunstein, A., Dall'Asta, L. & Mazza, F. A Bayesian generative neural network framework for epidemic inference problems. *Sci. Rep.* **12**, 19673. https://doi.org/10.1038/s41598-022-20898-x (2022).
13. Wu, D., Wang, L. & Zhang, P. Solving statistical mechanics using variational autoregressive networks. *Phys. Rev. Lett.* **122**, 080602. https://doi.org/10.1103/PhysRevLett.122.080602 (2019).
14. Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936. https://doi.org/10.1126/science.abb6936 (2020).
15. Cencetti, G. *et al.* Digital proximity tracing on empirical contact networks for pandemic control. *Nat. Commun.* **12**, 1655. https://doi.org/10.1038/s41467-021-21809-w (2021).
16. Eames, K. T. D. & Keeling, M. J. Contact tracing and disease control. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **270**, 2565–2571. https://doi.org/10.1098/rspb.2003.2554 (2003).
17. Foss, S. & Sakhanenko, A. Structural properties of conditioned random walks on integer lattices with random local constraints. In *In and Out of Equilibrium 3: Celebrating Vladas Sidoravicius. Progress in Probability* (eds Vares, M. E. *et al.*) 407–438 (Springer, Cham, 2021). https://doi.org/10.1007/978-3-030-60754-8_19.
18. Gantert, N., Popov, S. & Vachkovskaia, M. On the range of a two-dimensional conditioned simple random walk. *Ann. Henri Lebesgue* **2**, 349–368. https://doi.org/10.5802/ahl.20 (2019).
19. Ding, J., Fukushima, R., Sun, R. & Xu, C. Geometry of the random walk range conditioned on survival among Bernoulli obstacles. *Probab. Theory Relat. Fields* **177**, 91–145. https://doi.org/10.1007/s00440-019-00943-z (2020).
20. Parisi, G. *Statistical Field Theory* (Avalon Publishing, London, 1998).
21. Joyce, J. M. Kullback–Leibler Divergence. In *International Encyclopedia of Statistical Science* (ed. Lovric, M.) 720–722 (Springer, Berlin, 2011). https://doi.org/10.1007/978-3-642-04898-2_327.
22. Allen, L. J. S. Some discrete-time SI, SIR, and SIS epidemic models. *Math. Biosci.* **124**, 83–105. https://doi.org/10.1016/0025-5564(94)90025-6 (1994).
23. Kerr, C. C. *et al.* Covasim: An agent-based model of Covid-19 dynamics and interventions. *PLOS Comput. Biol.* **17**, e1009149 (2021).
24. Hinch, R. *et al.* OpenABM-covid19-an agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLOS Comput. Biol.* **17**, e1009146. https://doi.org/10.1371/journal.pcbi.1009146 (2021).
25. Biswas, M. H. A., Paiva, L. T. & Pinho, M. D. A SEIR model for control of infectious diseases with constraints. *Math. Biosci. Eng.* **11**, 761–784. https://doi.org/10.3934/mbe.2014.11.761 (2014).
26. Mézard, M. & Montanari, A. *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).
27. Antulov-Fantulin, N., Lančić, A., Šmuc, T., Štefančić, H. & Šikić, M. Identification of patient zero in static and temporal networks: Robustness and limitations. Phys. Rev. Lett. **114**, 248701. https://doi.org/10.1103/PhysRevLett.114.248701 (2015).
28. Mathieson, L. & Moscato, P. An introduction to proximity graphs. In *Business and Consumer Analytics: New Ideas* (eds Moscato, P. & de Vries, N. J.) 213–233 (Springer, Cham, 2019). https://doi.org/10.1007/978-3-030-06222-4_4.
29. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010 (2006).
30. Lorch, L. *et al.* Quantifying the effects of contact tracing, testing, and containment measures in the presence of infection hotspots. *ACM Transactions on Spatial Algorithms and Systems*https://doi.org/10.1145/3530774 *(2022).*
31. Ortega, E., Machado, D. & Lage-Castellanos, A. Dynamics of epidemics from cavity master equations: Susceptible-infectious-susceptible models. *Phys. Rev. E* **105**, 024308. https://doi.org/10.1103/PhysRevE.105.024308 (2022).
32. Wonham, M. J., de Camino-Beck, T. & Lewis, M. A. An epidemiological model for West Nile virus: invasion analysis and control applications. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **271**, 501–507. https://doi.org/10.1098/rspb.2003.2608 (2004).


## Acknowledgements

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (Piano Nazionale Di Ripresa e Resilienza (PNRR)-Missione 4 Componente 2, Investimento 1.3- D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors-views and opinions, neither the European Union nor the European Commission can be considered responsible for them. LDA acknowledges financial support from ICSC (Centro Nazionale di Ricerca in High-Performance Computing, Big Data, and Quantum Computing) founded by European Union – NextGenerationEU.

## Author contributions

A.B. and M.M. devised research and developed the CVA code. A.B. and L.D.A. supervised research. A.P.M. developed the two epidemic simulator interfaces. G.C. developed the MCMC code. M.M. and A.P.M. performed simulations. All authors analyzed results and wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-33770-3.

**Correspondence** and requests for materials should be addressed to M.M.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023