Cluster Variational Approximations for Structure Learning of Continuous-Time Bayesian Networks from Incomplete Data

Dominik Linzner[*]
Department of Electrical Engineering and Information Technology
Technische Universitaet Darmstadt
dominik.linzner@bcs.tu-darmstadt.de

Heinz Koeppl[*]
Department of Electrical Engineering and Information Technology
Department of Biology
Technische Universitaet Darmstadt
heinz.koeppl@bcs.tu-darmstadt.de

Abstract

Continuous-time Bayesian networks (CTBNs) constitute a general and powerful framework for modeling continuous-time stochastic processes on networks. This makes them particularly attractive for learning the directed structures among interacting entities. However, if the available data is incomplete, one needs to simulate the prohibitively complex CTBN dynamics. Existing approximation techniques, such as sampling and low-order variational methods, either scale unfavorably in system size, or are unsatisfactory in terms of accuracy. Inspired by recent advances in statistical physics, we present a new approximation scheme based on cluster-variational methods significantly improving upon existing variational approximations. We can analytically marginalize the parameters of the approximate CTBN, as these are of secondary importance for structure learning. This recovers a scalable scheme for direct structure learning from incomplete and noisy time-series data. Our approach outperforms existing methods in terms of scalability.

1 Introduction

Learning directed structures among multiple entities from data is an important problem with broad applicability, especially in biological sciences, such as genomics [1] or neuroscience [18]. With prevalent methods of high-throughput biology, thousands of molecular components can be monitored simultaneously in abundance and time. Changes of biological processes can be modeled as transitions of a latent state, such as expression or non-expression of a gene or activation/ inactivation of protein activity. However, processes at the bio-molecular level evolve across vastly different time-scales [11]. Hence tracking every transition between states is unrealistic. Additionally, biological systems are, in general, strongly corrupted by measurement or intrinsic noise.

In previous numerical studies, continuous-time Bayesian networks (CTBNs) [12] have been shown to outperform competing methods for reconstruction of directed networks, such as ones based on Granger causality or the closely related dynamic Bayesian networks [1]. Yet, CTBNs suffer from

---

the curse of dimensionality, prevalent in multi-component systems. This becomes problematic if observations are incomplete, as then the latent state of a CTBN has to be laboriously estimated [14]. In order to tackle this problem, approximation methods through sampling, e.g., [8, 7, 17], or variational approaches [5, 6] have been investigated. These, however, either fail to treat high-dimensional spaces because of sample sparsity, are unsatisfactory in terms of accuracy, or provide good accuracy at the cost of an only locally consistent description.

In this manuscript, we present, to the best of our knowledge, the first direct structure learning method for CTBNs based on variational inference. Our method combines two key ideas. We extend the framework of variational inference for multi-component Markov chains by borrowing results from statistical physics on cluster-variational-methods [21, 20, 16]. Here the previous result in [5] is recovered as a special case. We show how to calculate parameter-free dynamics of CTBNs in form of ordinary differential equations (ODEs), depending only on the observations, prior assumptions, and the graph structure. Lastly, we derive an approximation for the structure score, which we use to implement a scalable structure learning algorithm. The notion of using marginal CTBN dynamics for network reconstruction from noisy and incomplete observations was recently explored in [19] to successfully reconstruct networks of up to 11 nodes by sampling from the exact marginal posterior of the process, albeit using large computational effort. Yet, the method is sampling-based and thus still scales unfavorably in high dimensions. In contrast, we can recover the marginal CTBN dynamics at once using a standard ODE solver.

2 Background

2.1 Continuous-time bayesian networks

We consider continuous-time Markov chains (CTMCs) $\{X(t)\}_{t \geq 0}$ taking values in a countable state-space $\mathcal{X}$. A time-homogeneous Markov chain evolves according to an intensity matrix $\mathcal{R}$: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, whose elements are denoted by $\mathcal{R}(x, y)$, where $x, y \in \mathcal{X}$.

A continuous-time Bayesian network [12] is defined as an N-component process over a factorized state-space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$ evolving jointly as a CTMC. In the following, we will make use of the shorthand $x = (x_1, \ldots, x_N)$ with $x \in \mathcal{X}$ and $x_n \in \mathcal{X}_n$. However, as mostly no ambiguity arises we write x for $x \in \mathcal{X}_n$ in these cases to lighten the notation. We impose a directed graph structure $\mathcal{G} = (V, E)$, encoding the relationship among the components $V \equiv \{V_1, \ldots, V_N\}$, which we refer to as nodes. These are connected via an edge set $E \subseteq V \times V$. This quantity – the structure – is what we will later learn. The instantaneous state of each component is denoted by $X_n(t)$ assuming values in $\mathcal{X}_n$, which depends only on the states of a subset of nodes, called the parent set $\mathrm{pa}(n) \equiv \{m \mid (m, n) \in E\}$. Conversely, we define the child set $\mathrm{ch}(n) \equiv \{m \mid (n, m) \in E\}$. The dynamics of a local state $X_n(t)$ are modeled as a Markov process, when conditioned on the current state of all its parents $U_n(t)$ taking values in $\mathcal{U}_n \equiv \{\mathcal{X}_m \mid m \in \mathrm{pa}(n)\}$. They can then be expressed by means of the conditional intensity matrices (CIMs) $\mathcal{R}_n^u : \mathcal{X}_n \times \mathcal{X}_n \to \mathbb{R}$, where $u \equiv (u_1, \ldots u_L) \in \mathcal{U}_n$ denotes the current state of the parents ($L = |\mathrm{pa}(n)|$). Specifically, we can express the probability of finding node n in state y after some small time-step δt, given that it was in state x at time t with $x, y \in \mathcal{X}_n$ as

$$P(X_n(t + \delta t) = y \mid X_n(t) = x, U_n(t) = u) = \delta(x, y) + \mathcal{R}_n^u(x, y)\delta t + o(\delta t),$$

where $\mathcal{R}_n^u(x, y)$ is the matrix element of $\mathcal{R}_n^u$ corresponding to the transition $x \to y$ given the parents' state u. Additionally, we have to enforce $\mathcal{R}_n^u(x, x) \equiv -\sum_{y \neq x} \mathcal{R}_n^u(x, y)$ for $\mathcal{R}_n^u$ to be a proper CIM. The CIMs are connected to the joint intensity matrix $\mathcal{R}$ of the CTMC via amalgamation – see, for example, [12].

2.2 Variational energy

The foundation of this work is to derive a lower bound on the evidence of the data for a CTMC in the form of a variational energy. Such variational lower bounds are of great practical significance and pave the way to a multitude of approximate inference methods, in this context called *variational inference*. We consider paths $\mathbf{X} \equiv \{X(s) \mid 0 \leq s \leq T\}$ of a CTMC with a series of noisy state observations $\mathbf{Y} \equiv (Y^0, \ldots, Y^I)$ at times $(t^0, \ldots, t^I)$, drawn according to an observation model $Y^i \sim P(Y^i \mid X(t^i))$. We consider the posterior Kullback–Leibler (KL) divergence $D_{KL}(Q(\mathbf{X}) \| P(\mathbf{X} \mid \mathbf{Y}))$ given
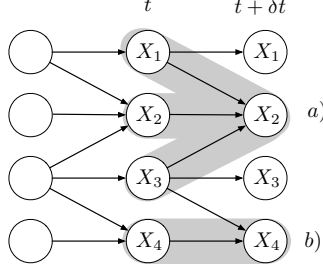
Figure 1: Sketch of different cluster choices for a CTBN in discretized time : a) star approximation b) naive mean-field.

a candidate distribution $Q(\mathbf{X})$, which can be decomposed as

$$D_{KL}(Q(\mathbf{X})||P(\mathbf{X} \mid \mathbf{Y})) = D_{KL}(Q(\mathbf{X})||P(\mathbf{X})P(\mathbf{Y} \mid \mathbf{X})) + \ln P(\mathbf{Y}).$$

As $D_{KL}(Q(\mathbf{X})||P(\mathbf{X} \mid \mathbf{Y})) \geq 0$ this recovers a lower bound on the evidence

$$\ln P(\mathbf{Y}) \geq \mathcal{F}, \tag{1}$$

where the bound $\mathcal{F} \equiv -D_{KL}(Q(\mathbf{X})||P(\mathbf{X})P(\mathbf{Y} \mid \mathbf{X}))$ is also known as the Kikuchi functional [10], or the Kikuchi variational energy. The Kikuchi functional has recently found heavy use in variational approximations for probabilistic models [21, 20, 16], because of the freedom it provides for choosing clusters in space and time. We will now make use of this feature.

3   Cluster variational approximations for CTBNs

The idea behind cluster variational approximations, derived subsequently, is to find a decomposition of the Kikuchi functional over M cluster functionals $\mathcal{F}_j$ of smaller sub-graphs $A_j(t)$ for a CTBN using its δt-discretization (see Figure 1):

$$\mathcal{F} \simeq \int_0^T \mathrm{d}t \sum_{j=1}^M \mathcal{F}_j(A_j(t)).$$

Examples for $A_j(t)$ are the the completely local *naive mean-field approximation* $A_j^{\mathrm{mf}}(t) = \{X_j(t + \delta t), X_j(t)\}$, or the *star approximation* $A_j^{\mathrm{s}}(t) = \{X_j(t + \delta t), U_j(t), X_j(t)\}$ on which our method is based. We notice that the formulation of CTBNs already imposes structure on the transition matrix

$$P(X(t + \delta t) \mid X(t)) = \prod_{n=1}^N P(X_n(t + \delta t) \mid X_n(t), U_n(t)), \tag{2}$$

suggesting a node-wise factorization to be a natural choice. Our goal is to find an expansion of $\mathcal{F}$ for finite δt for different cluster choices and subsequently consider the continuous-time limit $\delta t \to 0$. In order to arrive at the Kikuchi functional in the star approximation, we assume that $Q(\mathbf{X})$ describes a CTBN, i.e. satisfies (2). However, to render our approximation tractable, we further restrict the set of approximating processes by assuming them to be only weakly dependent in the transitions. Specifically, we require the existence of some expansion in orders of the *coupling strength* ε

$$Q(X_n(t + \delta t) \mid X_n(t), U_n(t)) = Q(X_n(t + \delta t) \mid X_n(t)) + \mathcal{O}(\varepsilon),$$

where the remainder $\mathcal{O}(\varepsilon)$ contains the dependency on the parents.[2] Because the derivation is quite lengthy, we have to leave the details of the calculation to Appendix B.1. The Kikuchi functional $\mathcal{F}$ can then be expanded in first order of ε and decomposes on the δt-discretized network spanned by the CTBN process, into local star-shaped terms – see, for example, Figure 1. We emphasize that the expanded variational energy in star approximation is no longer a lower bound on the evidence, but provides an approximation. We define a set of marginals, completely specifying a CTBN

$$m_n(x) \equiv Q(X_n(t) = x),$$

$$\tau_n^u(x, y) \equiv \lim_{\delta t \to 0} \frac{Q(X_n(t + \delta t) = y, X_n(t) = x, U_n(t) = u)}{\delta t} \qquad \text{for } x \neq y,$$

---

[2]An example of a function with such an expansion is a Markov random field with coupling strength ε.

the shorthand $m_n^u \equiv \prod_{l \in \mathrm{pa}(n)} m_l(u_l)$ and $\tau_n^u(x,x) \equiv -\sum_{y \neq x} \tau_n^u(x,y)$. Checking self-consistency of these quantities via marginalization of $Q(X_n(t+\delta t) = y, X_n(t) = x, U_n(t) = u)$ recovers an inhomogeneous Master equation

$$\dot{m}_n(x) = \sum_{y \neq x, u} [\tau_n^u(y,x) - \tau_n^u(x,y)]. \tag{3}$$

Because of the intrinsic asynchronous update constraint on CTBNs, only local probability flow inside the state-space $\mathcal{X}_n$ is allowed. This renders this equation equivalent to a continuity constraint on the global probability distribution. The resulting functional is only dependent on the marginal distributions. Performing the limit of $\delta t \to 0$, we arrive at a sum of node-wise functionals in continuous-time (see Appendix B.2)

$$\mathcal{F} \simeq \mathcal{F}_S, \quad \mathcal{F}_S \equiv \sum_{n=1}^{N} (H_n + E_n) + \mathcal{F}_0,$$

where we identified the entropy H_n and the energy E_n respectively

$$H_n = \int_0^T \mathrm{d}t \sum_{x,u} \sum_{y \neq x} \tau_n^u(x,y) \left[1 - \ln \tau_n^u(x,y) + \ln(m_n(x)m_n^u) \right],$$

$$E_n = \int_0^T \mathrm{d}t \left[\sum_x m_n(x)\mathbb{E}_n[\mathcal{R}_n^u(x,x)] + \sum_{x,u} \sum_{y \neq x} \tau_n^u(x,y) \ln \mathcal{R}_n^u(x,y) \right].$$

The expectation value is defined as $\mathbb{E}_n[f(u)] \equiv \sum_{u'} m_n^{u'} f(u')$ for any function $f(u)$. As expected, the functional is very similar to the one derived in [5] as both are derived from the KL divergence between true and approximate distributions from a set of marginals. Indeed, if we replace our star-shape cluster by the completely local one $A_j^{\mathrm{mf}}(t)$, we recover exactly their previous result, demonstrating the generality of our method (see Appendix B.3). In principle, higher-order clusters can be considered [20, 16]. Lastly, we enforce continuity by (3) fulfilling the constraint. We can then derive the Euler-Lagrange equations corresponding to the Lagrangian,

$$\mathcal{L} = \mathcal{F} - \int_0^T \mathrm{d}t \sum_{n,x} \lambda_n(x) \left\{ \dot{m}_n(x) - \sum_{y \neq x, u} [\tau_n^u(y,x) - \tau_n^u(x,y)] \right\},$$

with Lagrange multipliers $\lambda_n(x) \equiv \lambda_n(x,t)$.

3.1 CTBN dynamics in star approximation

The approximate dynamics of the CTBN can be recovered as stationary points of the Lagrangian, satisfying the Euler–Lagrange equation. Differentiating $\mathcal{L}$ with respect to $m_n(x)$, its time-derivative $\dot{m}_n(x)$, $\tau_n^u(x,y)$ and the Lagrange multiplier $\lambda_n(x)$ yield a closed set of coupled ODEs for the posterior process of the marginal distributions $m_n(x)$ and transformed Lagrange multipliers $\rho_n(x) \equiv \exp(\lambda_n(x))$, eliminating $\tau_n^u(x,y)$

$$\dot{\rho}_n(x) = \{\mathbb{E}_n[\mathcal{R}_n^u(x,x)] + \psi_n(x)\}\rho_n(x) - \sum_{y \neq x} \mathbb{E}_n[\mathcal{R}_n^u(x,y)]\rho_n(y), \tag{4}$$

$$\dot{m}_n(x) = \sum_{y \neq x} m_n(y)\mathbb{E}_n[\mathcal{R}_n^u(y,x)] \frac{\rho_n(x)}{\rho_n(y)} - m_n(x)\mathbb{E}_n[\mathcal{R}_n^u(x,y)] \frac{\rho_n(y)}{\rho_n(x)}, \tag{5}$$

with

$$\psi_n(x') = \sum_{j \in \mathrm{ch}(n)} \sum_x m_j(x) \left\{ \sum_{y \neq x} \frac{\rho_j(y)}{\rho_j(x)} \mathbb{E}_j[\mathcal{R}_j^u(x,y)|X_n(t) = x'] + \mathbb{E}_j[\mathcal{R}_j^u(x,x)|X_n(t) = x'] \right\}.$$

Furthermore, we recover the *reset condition*

$$\lim_{t \to t^{i-}} \rho_n(x) = \lim_{t \to t^{i+}} \rho_n(x) \exp \left\{ \sum_{x' \in \mathcal{X}|x'_n = x} \ln P(Y^i \mid x') \prod_{k=1, k \neq n}^{N} m_k(x'_k) \right\}, \tag{6}$$
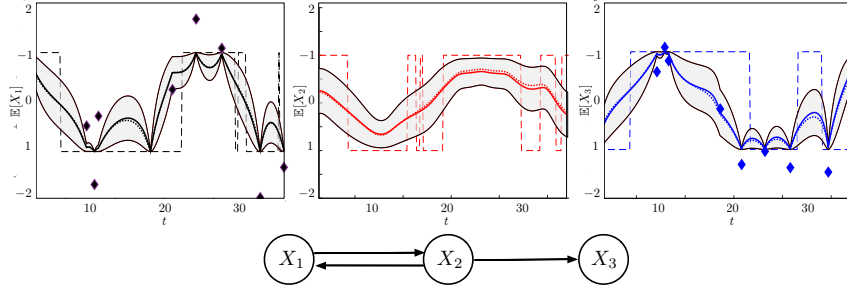
Figure 2: Dynamics in star approximation of a 3 node CTBN following Glauber dynamics at $a = 1$ and $b = 0.6$ conditioned on noisy observations (diamonds). We plotted the expected state (blue) plus variance (grey area). The observation model is the latent state plus gaussian random noise of variance $\sigma = 0.8$ and zero mean. The latent state (dashed) is well estimated for X_2, even when no data has been provided. For comparison, we plotted the exact posterior mean (dots). We did not plot the exact variance, which depends only on the mean, for better visibility.

and x_k denotes the kth component of $x \in \mathcal{X}$. This incorporates the conditioning of the dynamics on noisy observations. For the full derivation we refer the reader to Appendix B.4. We require boundary conditions for the evolution interval in order to determine a unique solution to the set of equations (4) and (5). We thus set either $m_n(x, 0) = Y_n^0$ and $\rho_n(x, T) = Y_n^I$ in the case of noiseless observations, or – if the observations have been corrupted by noise – $m_n(x, 0) = \frac{1}{2}$ and $\rho_n(x, T) = 1$ as boundaries before and after the first and the last observation, respectively. The coupled set of ODEs can then be solved iteratively as a fixed-point procedure in the same manner as in previous works [15, 5] (see Appendix A.1 for details) in a forward-backward procedure. As the Kikuchi functional is convex, this procedure is guaranteed to converge. In order to incorporate noisy observations into the CTBN dynamics, we need to assume an observation model. In the following we assume that the data likelihood factorizes $P(Y^i \mid X) = \prod_n P(Y_n^i \mid X_n)$, allowing us to condition on the data by enforcing $\lim_{t \to t^{i-}} \rho_n(x) = \lim_{t \to t^{i+}} P_n(Y^i \mid x) \rho_n(x)$. In Figure 2, we exemplify CTBN dynamics ($N = 3$) conditioned on observations corrupted by independent Gaussian noise. We find close agreement with the exact posterior dynamics. Because we only need to solve $2N$ ODEs to approximate the dynamics of an N-component system, we recover a linear complexity in the number of components, rendering our method scalable.

3.2 Parameter estimation

Maximization of the variational energy with respect to transition rates $\mathcal{R}_n^u(x, y)$ yields the expected result for the estimator of transition rates

$$\hat{\mathcal{R}}_n^u(x, y) = \frac{\mathbb{E}[M_n^u(x, y)]}{\mathbb{E}[T_n^u(x)]},$$

given the *expected sufficient statistics* [14]:

$$\mathbb{E}[T_n^u(x)] = \int_0^T \mathrm{d}t \, m_n(x) m_n^u, \quad \mathbb{E}[M_n^u(x, y)] = \int_0^T \mathrm{d}t \, \tau_n^u(x, y),$$

where $\mathbb{E}[T_n^u(x)]$ are the *expected dwelling times* and $\mathbb{E}[M_n^u(x, y)]$ are the *expected number of transitions*. Following standard *expectation–maximization* (EM) procedure, e.g. [15], we can estimate the systems' parameters given the underlying network.

3.3 Benchmark

In the following we compare the accuracy of the star approximation with the naive mean-field approximation. Throughout this section we will consider a binary local state-space (spins) $\mathcal{X}_n = \{+1, -1\}$. We consider a system obeying Glauber dynamics with the rates $\mathcal{R}_n^u(x, -x) = \frac{a}{2}\left(1 + x \tanh\left(b \sum_{l \in \mathrm{pa}(n)} u_l\right)\right)$. Here b is the inverse temperature of the system. With increasing b
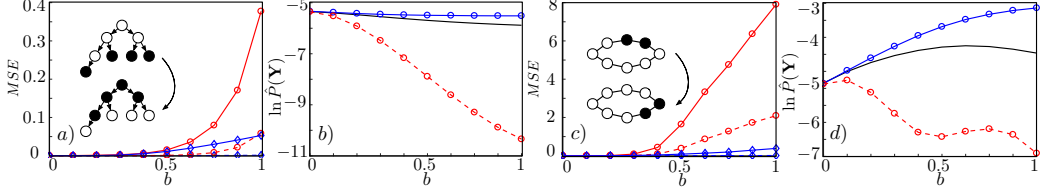
Figure 3: We perform inference on a tree network a) and b), and an undirected chain c) and (d), both consisting of 8 nodes with noiseless evidence as denoted in sketch inlet (black: $x = -1$, white: $x = 1$) in a) and c) obeying Glauber dynamics with $a = 8$. In a) we plotted mean-squared-error (MSE) for the expected dwelling times (dashed) and the the expected number of transitions for the naive mean-field (circle, red) and star approximation (diamond, blue) with respect to the predictions of the exact simulation as a function of temperature b. In b) and d) we plot the approximation of logarithmic evidence as a function of temperature. We find that for both approximations (star approximation in blue, naive mean field in red dashed and exact result in black) better performance on the tree network, while the star approximation clearly improves upon naive mean-field in both scenarios.

the dynamics of each node depend more strongly on the dynamics of its neighbors and thus harder to describe using mean-field dynamics. The pre-factor a scales the overall rate of the process. This system can be an appropriate toy-example for biological networks as it encodes additive threshold behavior. In Figure 3 a) and c), we show how closely the expected sufficient statistics match the true ones for a tree network and an undirected chain with periodic boundaries of 8 nodes, so that we can still compare to the exact result. In this application, we restrict ourselves to noiseless observations to better connect to previous results as in [5]. We compare the estimation of the evidence using the variational energy in Figure 3 b) and d). We find that while our estimate using the star approximation is a much closer approximation, it does not provide a lower bound.

4   Cluster variational structure learning for CTBNs

For structure learning tasks, knowing the exact parameters of a model is in general unnecessary. For this reason we will derive an analogous but parameter-free formulation of the variational approximation for evidence and the latent state dynamics, analogous to the ones in the previous section.

4.1   Variational structure score for CTBNs

In the following we derive an approximate CTBN structure score, for which we need to marginalize over the parameters of the variational energy. To this end, we assume that the parameters of the CTBN are random variables distributed according to a product of local and independent *Gamma distributions* $P(\mathcal{R} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{G}) = \prod_n \prod_{x,u} \prod_{y \neq x} Gam\left[\mathcal{R}_n^u(x,y) \mid \alpha_n^u(x,y), \beta_n^u(x)\right]$ given a graph structure $\mathcal{G}$. With our cluster approximation, the evidence is approximately given by $P(\mathbf{Y} \mid \mathcal{R}, \mathcal{G}) \approx \exp(\mathcal{F}_S)$. By a simple analytical integration we recover an approximation to the CTBN structure score

$$P(\mathcal{G} \mid \mathbf{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \approx P(\mathcal{G}) \int_0^\infty d\mathcal{R} \; e^{\mathcal{F}_S} P(\mathcal{R} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{G})$$

$$\propto e^H \prod_n \prod_{x,u} \prod_{y \neq x} \left(\frac{\beta_n^u(x)}{(\mathbb{E}[T_n^u(x)] + \beta_n^u(x))^{M_n^u(x,y)}} \right)^{\alpha_n^u(x,y)} \frac{\Gamma\left(\mathbb{E}[M_n^u(x,y)] + \alpha_n^n(x,y)\right)}{\Gamma\left(\alpha_n^u(x,y)\right)}, \quad (7)$$

with $\Gamma(\bullet)$ being the Gamma-function. The approximated CTBN structure score still satisfies structural modularity, if not broken by the structure prior $P(\mathcal{G})$. However, we can not implement a *k-learn* structure learning strategy as originally proposed in [13], as in the latent state estimation nodes become coupled and depend on each others' estimate, which in turn depend on the chosen parent set. For a detailed derivation, – see Appendix B.5. Finally we note, that in contrast to the evidence in Figure 3, we have no analytical expression for the structure score (the integral is intractable) so that we can not compare with the exact result after integration.

Table 1: Experimental results with datasets generated from random CTBNs ($N = 5$) with families of up to k_{max} parents. To demonstrate that our score prevents over-fitting we search for families of up to $k = 2$ parents. When changing one parameter the other default values are fixed to $D = 10$, $b = 0.6$ and $\sigma = 0.2$.

k_{max}	EXPERIMENT	VARIABLE	AUROC	AUPR
1	NUMBER OF TRAJECTORIES	$D = 5$	0.78 ± 0.03	0.64 ± 0.01
		$D = 10$	0.87 ± 0.03	0.76 ± 0.00
		$D = 20$	0.96 ± 0.02	0.92 ± 0.00
	MEASUREMENT NOISE	$\sigma = 0.6$	0.81 ± 0.10	0.71 ± 0.00
		$\sigma = 1.0$	0.69 ± 0.07	0.49 ± 0.01
2	NUMBER OF TRAJECTORIES	$D = 5$	0.64 ± 0.09	0.50 ± 0.17
		$D = 10$	0.68 ± 0.12	0.54 ± 0.14
		$D = 20$	0.75 ± 0.11	0.68 ± 0.16
	MEASUREMENT NOISE	$\sigma = 0.6$	0.71 ± 0.13	0.58 ± 0.20
		$\sigma = 1.0$	0.64 ± 0.11	0.53 ± 0.15

4.2 Marginal dynamics of CTBNs

The evaluation of the approximate CTBN structure score requires the calculation of the latent state dynamics of the marginal CTBN. For this, we approximate the Gamma function in (7) via *Stirling's approximation*. As Stirling's approximation becomes accurate asymptotically, we imply that sufficiently many transitions have been recorded across samples or have been introduced via a sufficiently strong prior assumption. By extremization of the marginal variational energy, we recover a set of integro-differential equations describing the marginal self-exciting dynamics of the CTBN (see Appendix B.6). Surprisingly, the only difference of this parameter-free version compared to (4) and (5) is that the conditional intensity matrix has been replaced by its posterior estimate

$$\bar{\mathcal{R}}_n^u(x, y) \equiv \frac{\mathbb{E}[M_n^u(x, y)] + \alpha_n^u(x, y)}{\mathbb{E}[T_n^u(x)] + \beta_n^u(x)}. \tag{8}$$

The rate $\bar{\mathcal{R}}_n^u(x, y)$ is thus determined recursively by the dynamics generated by itself conditioned on the observations and prior information. We notice the similarity of our result to the one recovered in [19], where, however, the expected sufficient statistics had to be computed self-consistently during each sample path. We employ a fixed-point iteration scheme to solve the integro-differential equation for the marginal dynamics in a manner similar to EM (for the detailed algorithm, see Appendix A.2).

5 Results and discussion

For the purpose of learning, we employ a *greedy hill-climbing* strategy where we exhaustively score all possible families for each node with up to k parents and set the highest scoring family as the current one. We do this repeatedly until our network estimate converges, which usually takes only 2 of such *sweeps*. We can transform the scores to probabilities and generate *Reciever-Operator-Characteristics* (ROCs) and *Precision-Recall* (PR) curves by thresholding the averaged graphs. As a measure of performance, we calculate the averaged *Area-Under-Curve* (AUC) for both. We evaluate our method using both synthetic and real-world data from molecular biology. In order to stabilize our method in the presence of sparse data, we augment our algorithm with a prior $\alpha = 5$ and $\beta = 10$, which is uninformative of the structure, for both experiments. We want to stress that, while we removed the bottleneck of exponential scaling of latent state estimation of CTBNs, Bayesian structure learning via scoring still scales super-exponentially in the number of components [9]. Our method can thus not be compared to shrinkage based network inference methods such as graphical lasso.

The synthetic experiments are performed on CTBNs encoded with Glauber dynamics. For each of the D trajectories, we recorded 10 observations Y^i at random time-points t^i and corrupted them with Gaussian noise with variance $\sigma = 0.6$ and zero mean. In Table 1, we apply our method to random graphs consisting of $N = 5$ nodes, up to k_{max} parents. We note that fixing k_{max} does not fix the possible degree of the node (which can go up to $N - 1$). For random graphs with $k_{max} = 1$ our method performs best, as expected, and we are able to reliably recover the correct graph if enough
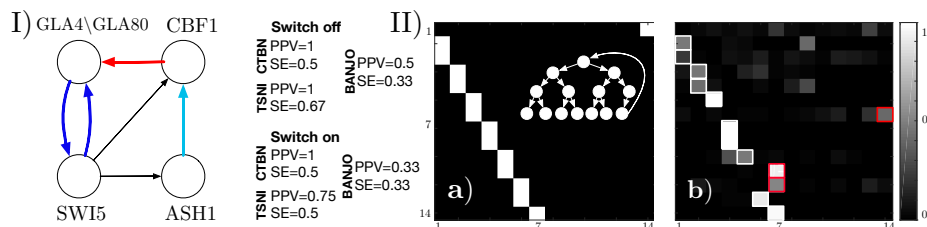
Figure 4: I) Reconstruction of a gene regulatory network (IRMA) from real-world data. To the left we show the inferred network for the "switch off" and "switch on" dataset. The ground truth network is displayed by black thin edges, the correctly inferred edges are thick (all inferred edges were correct). The the red edge was identified only in "switch on", the teal edge only in "switch off". On the right we show a small table summarizing the reconstruction capabilities of our method, TSNI and BANJO (PPV of random guess is 0.5). II) Reconstruction of large graphs. We tested our method on a ground truth graph with 14 nodes, as displayed in $a)$ with node-relations sketched in the inlet, encoded with Glauber dynamics and searched for a maximum of $k = 1$ parents. Although we used relatively few observations that have been strongly corrupted, the averaged learned graph $b)$ is visibly close to the ground truth. We framed the prediction of the highest scoring graph, where correctly learned edges are framed white and the incorrect ones are framed red.

data are provided. To demonstrate that our score penalizes over-fitting we search for families of up to $k = 2$ parents. For the more challenging scenario of $k_{max} = 2$ we find a drop in performance. This can be explained by the presence of strong correlations in more connected graphs and the increased model dimension with larger k_{max}. In order to prove that our method outperforms existing methods in terms of scalability, we successfully learn a tree-network, with a leaf-to-root feedback, of 14 nodes with $a = 1$, $b = 0.6$, see Figure 4 II). This is the largest inferred CTBN from incomplete data reported (in [19] a CTBN of 11 nodes is learned, albeit with incomparably larger computational effort).

Finally, we apply our algorithm to the *In vivo Reverse-engineering and Modeling Assessment* (IRMA) network [4], a synthetic gene regulatory network that has been implemented on cultures of yeast, as a benchmark for network reconstruction algorithms, see Figure 4 I). It is, to best of our knowledge, the only molecular biological network with a ground truth. The authors of [4] provide time course data from two perturbation experiments, referred to as "switch on" and "switch off", and attempted reconstruction using different methods. In order to compare to their results we adopt their metrics *Positive Predicted Value* (PPV) and the *Sensitivity score* (SE) [2]. The best performing method is ODE-based (TSNI [3]) and required additional information on the perturbed genes in each experiment, which may not always be available. As can be seen in Figure 4 I) our method performs accurately on the "switch off" and the "switch on" data set regarding the PPV. The SE is slightly worse than for TSNI on "switch off". In both cases, we perform better than the other method based on Bayesian networks (BANJO [22]). Lastly, we note that in [1] more correct edges could be inferred using CTBNs, however with parameters tuned with respect to the ground truth to reproduce the IRMA network. For details on our processing of the IRMA data, see Appendix C.

6 Conclusion

We develop a novel method for learning directed graphs from incomplete and noisy data based on a continuous-time Bayesian network. To this end, we approximate the exact but intractable latent process by a simpler one using cluster variational methods. We recover a closed set of ordinary differential equations that are simple to implement using standard solvers and retain a consistent and accurate approximation of the original process. Additionally, we provide a close approximation to the evidence in the form of a variational energy that can be used for learning tasks. Lastly, we demonstrate how marginal dynamics of continuous-time Bayesian networks, which only depend on data, prior assumptions, and the underlying graph structure, can be derived by the marginalization of the variational free energy. Marginalization of the variational energy provides an approximate structure score. We use this to detect the best scoring graph using a greedy hill-climbing procedure. It would be beneficial to identify higher-order approximations of the variational energy in the future.

We test our method on synthetic as well as real data and show that our method produces meaningful results while outperforming existing methods in terms of scalability.

References

[1] Enzo Acerbi, Teresa Zelante, Vipin Narang, and Fabio Stella. Gene network inference using continuous time Bayesian networks: a comparative study and application to Th17 cell differentiation. *BMC Bioinformatics*, 15, 2014.

[2] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego di Bernardo. How to infer gene networks from expression profiles. *Molecular systems biology*, 3:78, 2007.

[3] Mukesh Bansal, Giusy Della Gatta, and Diego di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, apr 2006.

[4] Irene Cantone, Lucia Marucci, Francesco Iorio, Maria Aurelia Ricci, Vincenzo Belcastro, Mukesh Bansal, Stefania Santini, Mario Di Bernardo, Diego di Bernardo, and Maria Pia Cosma. A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell*, 137(1):172–181, apr 2009.

[5] Ido Cohn, Tal El-Hay, Nir Friedman, and Raz Kupferman. Mean field variational approximation for continuous-time Bayesian networks. *Journal Of Machine Learning Research*, 11:2745–2783, 2010.

[6] Tal El-Hay, Ido Cohn, Nir Friedman, and Raz Kupferman. Continuous-Time Belief Propagation. *Proceedings of the 27th International Conference on Machine Learning*, pages 343–350, 2010.

[7] Tal El-Hay, R Kupferman, and N Friedman. Gibbs sampling in factorized continuous-time Markov processes. *Proceedings of the 22th Conference on Uncertainty in Artificial Intelligence*, 2011.

[8] Yu Fan and CR Shelton. Sampling for approximate inference in continuous time Bayesian networks. *AI and Math*, 2008.

[9] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning Probabilistic Relational Models. *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*, August 1999.

[10] Ryoichi Kikuchi. A theory of cooperative phenomena. *Physical Review*, 81(6):988–1003, mar 1951.

[11] Michael Klann and Heinz Koeppl. Spatial Simulations in Systems Biology: From Molecules to Cells. *International Journal of Molecular Sciences*, 13(6):7798–7827, 2012.

[12] Uri Nodelman, Christian R Shelton, and Daphne Koller. Continuous Time Bayesian Networks. *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 378–387, 1995.

[13] Uri Nodelman, Christian R. Shelton, and Daphne Koller. Learning continuous time Bayesian networks. *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pages 451–458, 2003.

[14] Uri Nodelman, Christian R Shelton, and Daphne Koller. Expectation Maximization and Complex Duration Distributions for Continuous Time Bayesian Networks. *Proc. Twenty-first Conference on Uncertainty in Artificial Intelligence*, pages pages 421–430, 2005.

[15] Manfred Opper and Guido Sanguinetti. Variational inference for Markov jump processes. *Advances in Neural Information Processing Systems 20*, pages 1105–1112, 2008.

[16] Alessandro Pelizzola and Marco Pretti. Variational approximations for stochastic dynamics on graphs. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):1–28, 2017.

[17] Vinayak Rao and Yee Whye Teh. Fast MCMC sampling for Markov jump processes and extensions. *Journal of Machine Learning Research*, 14:3295–3320, 2012.

[18] Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj Guha Thakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, Pek Yee Lum, Amy Leonardson, Rolf Thieringer, Joseph M Metzger, Liming Yang, John Castle, Haoyuan Zhu, Shera F Kash, Thomas A Drake, Alan Sachs, and Aldons J Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, jul 2005.

[19] Lukas Studer, Christoph Zechner, and Matthias Reumann. Marginalized Continuous Time Bayesian Networks for Network Reconstruction from Incomplete Observations. *Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016)*, pages 2051–2057, 2016.

[20] Eduardo Domínguez Vázquez, Gino Del Ferraro, and Federico Ricci-Tersenghi. A simple analytical description of the non-stationary dynamics in Ising spin systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(3):033303, 2017.

[21] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. *Advances in neural information*, 13:657–663, 2000.

[22] Jing Yu, V. Anne Smith, Paul P. Wang, Alexander J. Hartemink, and Erich D. Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603, dec 2004.

Supplementary Material

A   Algorithms

In this section we give the detailed algorithms described in the main text. All equation references point to the main text.

A.1   Stationary points of Euler–Lagrange equation

Algorithm 1 Stationary points of Euler–Lagrange equation

1: **Input:** Legal set of initial trajectories $m_n(x)$, boundary conditions $m(x, 0)$ and $\rho(x, T)$,
   Observations $\mathbf{Y}$.
2: **repeat**
3:     **for all** $n \in \{1, \ldots, N\}$ **do**
4:         **for all** $Y^i \in \mathbf{Y}$ **do**
5:             Update $\rho_n(x)$ by backward propagation from t_i to t_{i-1} using (5) fulfilling reset conditions
               (6).
6:         **end for**
7:         Update $m_n(x)$ by forward propagation using (4) given $\rho_n(x)$.
8:     **end for**
9: **until** Convergence
10: **Output:** Set of $m_n(x)$ and $\rho_n(x)$.

A.2   Marginal CTBN dynamics

Algorithm 2 Marginal CTBN dynamics

1: **Input:** Propose set of initial trajectories $m_n(x)$,
   observations $\mathbf{Y}$, prior assumption on sufficient statistics $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, initial guess for $\bar{\mathcal{R}}_n^u(x, y)$.
2: **repeat**
3:     Set current $\bar{\mathcal{R}}_n^u(x, y)$ as current CIM.
4:     Solve marginal dynamic equation with $\bar{\mathcal{R}}_n^u(x, y)$ using Algorithm1.
5:     Use expected sufficient statistics to update $\bar{\mathcal{R}}_n^u(x, y)$ via (8).
6: **until** Convergence
7: **Output:** Set of $m_n(x)$ and $\rho_n(x)$.

B   Derivations

B.1   Variational energy in star approximation

In the following we derive the star approximation of a factorized stochastic process. In order to lighten the notation we omit the corresponding process to each variable from now on, with $X(t + \delta t) = y$, $X(t) = x$. The exact expression of the variational energy $\mathcal{F}[Q]$ for a continuous-time Markov process decomposes into time-wise energies $\mathcal{F}[Q] = \sum_t f(t)$ with

$$f(t) \equiv \underbrace{\int_{\mathcal{X}} \mathrm{d}x \int_{\mathcal{X}} \mathrm{d}y\, Q(y, x) \ln P(y \mid x)}_{\equiv \langle H(t) \rangle_Q} - \underbrace{\int_{\mathcal{X}} \mathrm{d}x \int_{\mathcal{X}} \mathrm{d}y\, Q(y, x) \ln Q(y, x) + \int_{\mathcal{X}} \mathrm{d}x\, Q(x) \ln Q(x)}_{\equiv S[Q](t)},$$

where we identified the time-dependent energy function $H(x, y, t) \equiv \ln P(y \mid x)$ and the entropy $S[Q](t)$. In the following, we explicitly use $x = (x_1, \ldots, x_N)$ and $y = (y_1, \ldots, y_N)$ for $x, y \in \mathcal{X}$ with $x_n \in \mathcal{X}_n$ and $y_n \in \mathcal{X}_n$. We assume $Q(\mathbf{X})$ to describe a factorized stochastic process, i.e. $Q(y, x) = \frac{\prod_n Q(y_n, x_n, u_n)}{\prod_n Q(x_n, u_n)} Q(x)$, where we introduced the process of neighbours $U_n(t) = u_n$. It

11

holds that $Q(x) = \prod_n Q(x_n \mid u_n)$. This yields

$$Q(y, x) = \prod_n Q(y_n, x_n \mid u_n),$$

Consider now

$$\langle H(t) \rangle_Q = \int \int_{\mathcal{X}} \mathrm{d}x \, \mathrm{d}y \prod_n Q(y_n, x_n \mid u_n) \ln \prod_n P(y_n \mid x_n, u_n)$$

Assuming temporal correlations with neighboring nodes scale with ε

$$P(y_n \mid x_n, u_n) \equiv P(y_n \mid x_n) + \mathcal{O}(\varepsilon), \tag{9}$$

where the remainder $\mathcal{O}(\varepsilon)$ contains dependency on the parents. Naturally, we assume that the approximating distribution Q has a similar expansion in the same parameter ε. We get the approximate entropy using Appendix B.1.1 in first order of ε

$$\langle H(t) \rangle_Q = \int \int_{\mathcal{X}} \mathrm{d}x \, \mathrm{d}y \left[\sum_n Q(y_n, x_n \mid u_n) \prod_{m \neq n} Q(y_m, x_m) - (N-1) \prod_m Q(y_m, x_m) \right] \sum_k \ln P(y_n \mid x_n, u_n).$$

If $n \neq k$ we can use Appendix B.1.1 and Appendix B.1.2:

$$\langle H(t) \rangle_Q = \varepsilon \int \int_{\mathcal{X}} \mathrm{d}x \, \mathrm{d}y \sum_n f(y_n, x_n, u_n) \prod_{m \neq n} Q(y_m, x_m) \ln P(y_k, x_k) + \mathcal{O}(\varepsilon^2),$$

where we introduced the explicit correction $Q(y_n, x_n \mid u_n) = Q(y_n, x_n) + \varepsilon f(y_n, x_n, u_n)$. By normalization constraint it follows that $\int \int_{\mathcal{X}_n} \mathrm{d}x_n \, \mathrm{d}y_n \, f(y_n, x_n, u_n) = 0$. Thus for $n \neq k$, $\langle H(t) \rangle_Q$ is of order $\mathcal{O}(\varepsilon^2)$. This leaves us with

$$\langle H(t) \rangle_Q = \int \int_{\mathcal{X}} \mathrm{d}x \, \mathrm{d}y \sum_n Q(y_n, x_n \mid u_n) \prod_{m \in \mathrm{par}(n)} Q(x_m) \ln P(y_n \mid x_n, u_n) + \mathcal{O}(\varepsilon^2).$$

The exact same treatment can be done for the entropy term and we arrive at variational free energy in *star-approximation*

$$\mathcal{F}[Q] \approx \sum_{t,n} \int \int_{\mathcal{X}_n} \mathrm{d}x_n \, \mathrm{d}y_n \, Q(y_n, x_n \mid u_n) \prod_{m \in \mathrm{par}(n)} Q(x_m) \times [\ln P(y_n \mid x_n, u_n) - \ln Q(y_n, x_n \mid u_n)]$$

$$- \sum_{t,n} \int_{\mathcal{X}_n} \mathrm{d}x \, Q(x_n \mid u_n) \ln Q(x_n \mid u_n).$$

B.1.1 Expansion formula I

Note that for $\prod_{n=1}^N Q_n$ with $Q_n = a_n + \varepsilon b_n$ for any $a_n, b_n \in \mathbb{R}$ holds

$$\prod_{n=1}^N Q_n = \sum_{m=1}^N Q_m \prod_{n \neq m, n=1}^N a_n - (N-1) \prod_{n=1}^N a_n + \mathcal{O}(\varepsilon^2)$$

Proof:

$$\prod_{n=1}^N Q_n = \prod_{n=1}^N a_n + \epsilon \sum_{m=1}^N b_m \prod_{n \neq m, n=1}^N a_n + \sum_{m=1}^N a_m \prod_{n \neq m, n=1}^N a_n - \sum_{m=1}^N a_m \prod_{n \neq m, n=1}^N a_n + \mathcal{O}(\varepsilon^2)$$

$$= \sum_{m=1}^N [a_m + \epsilon b_m] \prod_{n \neq m, n=1}^N a_n + \prod_{n=1}^N a_n - \sum_{m=1}^N \prod_{n=1}^N a_n + \mathcal{O}(\varepsilon^2)$$

$$= \sum_{m=1}^N Q_m \prod_{n \neq m, n=1}^N a_n - (N-1) \prod_{n=1}^N a_n + \mathcal{O}(\varepsilon^2).$$

B.1.2 Expansion formula II

For $\prod_{n=1}^{N} Q_n \ln Q_k$ with $Q_n = a_n + \varepsilon b_n$ for any $a_n, b_n \in \mathbb{R}$ holds

$$\prod_{n=1}^{N} Q_n \ln Q_k = \varepsilon \sum_{m=1}^{N} b_m \prod_{n \neq m, n=1}^{N} a_n \ln a_k + \mathcal{O}(\varepsilon^2)$$

Proof: In a first step we use B.1.1:

$$\left[\sum_{m=1, m \neq k}^{N} Q_m \prod_{n \neq m, n=1}^{N} a_n - (N-1) \prod_{n=1}^{N} a_n \right] \ln Q_k$$

$$= \left[\sum_{m=1, m \neq k}^{N} (a_m + \varepsilon b_m) \prod_{n \neq m, n=1}^{N} a_n - (N-1) \prod_{n=1}^{N} a_n \right] \ln Q_k$$

$$= \varepsilon \sum_{m=1}^{N} b_m \prod_{n \neq m, n=1}^{N} a_n \ln Q_k$$

$$= \varepsilon \sum_{m=1}^{N} b_m \prod_{n \neq m, n=1}^{N} a_n \ln a_k + \mathcal{O}(\varepsilon^2),$$

where in the last line we used $\ln(x + \varepsilon) \approx x$.

B.2 Continuous-time variational energy in star approximation

Throughout we will consider $\delta t \ll 1$. By inserting $Q's$ representation in marginals into $\mathcal{F}$

$$Q(y_n, x_n \mid u_n) = \delta(x, y) m_n(x) + \frac{\tau_n^u(x, y)}{m_n^u} \delta t + o(\delta t),$$

and $\tau_n^u(x, x) = -\sum_{y \neq x} \tau_n^u(y, x)$ we get

$$\mathcal{F} = -\sum_n \sum_t \sum_{x, y \neq x, u} \delta t \tau_n^u(x, y) \left[\ln \delta t \frac{\tau_n^u(x, y)}{m_n^u} - \delta t \ln \mathcal{R}_n^u(x, y) \right]$$

$$- \sum_n \sum_t \sum_{x, u} m_n^u \left[m_n(x) - \frac{\delta t \sum_{y \neq x} \tau_n^u(y, x)}{m_n^u} \right] \times \left[\ln[m_n(x) - \frac{\delta t \sum_{y \neq x} \tau_n^u(y, x)}{m_n^u}] - \ln[1 + \mathcal{R}_n^u(x, x) \delta t] \right]$$

$$+ \sum_n \sum_k \sum_x m_n^u \left[m_n(x) - \delta t \frac{\sum_{y \neq x} [\tau_n^u(x, y) - \tau_n^u(y, x)]}{m_n^u} \right] \ln \left[m_n(x) - \delta t \frac{\sum_{y \neq x} [\tau_n^u(x, y) - \tau_n^u(y, x)]}{m_n^u} \right]$$

where we also inserted $P(X_n(t) = x_n = y \mid X_n(t) = x_n = x, U_n(t) = u = u) = \delta_{x,y} + \mathcal{R}_n^u(x, y) \delta t$. With the asymptotic identity $\ln(1 + \delta t x) = \delta t x$ we can simplify

$$\mathcal{F} = -\sum_n \sum_t \sum_{x, y \neq x, u} \delta t \tau_n^u(x, y) \left[\ln \frac{\tau_n^u(x, y)}{m_n^u} - \ln \mathcal{R}_n^u(x, y) \right]$$

$$- \sum_n \sum_t \sum_{x, u} m_n^u \left[m_n(x) - \frac{\delta t \sum_{y \neq x} \tau_n^u(x, y)}{m_n^u} \right] \times \left[\ln m_n(x) - \frac{\delta t \sum_{y \neq x} \tau_n^u(x, y)}{m_n^u m_n(x)} - \mathcal{R}_n^u(x, x) \delta t \right]$$

$$+ \sum_n \sum_k \sum_x m_n^u m_n(x) \ln(m_n(x)) - \delta t \sum_{y \neq x} [\tau_n^u(x, y) - \tau_n^u(y, x)] - \delta t \sum_{y \neq x} [\tau_n^u(x, y) - \tau_n^u(y, x)] \ln(m_n(x)),$$

which becomes in the continuous-time limit $\delta t \to 0$

$$\mathcal{F} = \sum_n \int dt \sum_{x, y \neq x, u} \tau_n^u(x, y)[1 - \ln \tau_n^u(x, y) + \ln(m_n^u m_n(x))]$$

$$+ \sum_n \int dt \left[\sum_{x, u} m_n(x) m_n^u \mathcal{R}_n^u(x, x) + \sum_{x, y \neq x, u} \tau_n^u(x, y) \ln \mathcal{R}_n^u(x, y) \right].$$

The contribution of the likelihood term can be derived to be

$$\mathcal{F}_0 = \sum_t \delta t \sum_i \mathbb{E}_N[\ln P(Y^i \mid x)] \frac{\delta(t, t^i)}{\delta t} \underset{\delta t \to 0}{=} \int_0^T dt \sum_i \mathbb{E}_N[\ln P(Y^i \mid x)]\delta(t - t^i),$$

$$\mathbb{E}_N[f(x)] = \sum_{x \in \mathcal{X}} f(x) \prod_{k=1}^N m_k(x_k), \quad x_k \in \mathcal{X}_k.$$

B.3 Naive mean-field approximation

We recover the variational energy in naive mean-field approximation by only consider the zeroth order expansion in the correlations ε, meaning

$$Q(y_n \mid x_n, u_n) = Q(y_n \mid x_n)$$

Then for the entropy $S(t)$ from Appendix B.1 holds

$$S(t) \equiv \sum_{y,x} Q(y, x) \ln Q(y, x) = \sum_{y,x} \prod_n Q(y_n, x_n) \ln\left[\prod_m Q(y_m, x_m)\right]$$

Thus for the variational energy we arrive at the naive mean-field approximation

$$\mathcal{F} = \sum_n \sum_t \sum_{y_n, x_n} Q(y_n, x_n) \sum_u \prod_{k \in \mathrm{pa}(n)} Q(u_k) \ln P(y_n \mid x_n, u_n) - \sum_n \sum_t \sum_{y_n, x_n} Q(y_n, x_n) \ln Q(y_n, x_n)$$
$$+ \sum_n \sum_t \sum_{x_n} Q(x_n) \ln Q(x_n)$$

Finally considering the marginals of the transitions

$$Q(y_n, x_n) = \delta_{x,y} m_n(x) + \tau_n(x, y) + o(\delta t),$$

we recover the result from (Cohn, El-Hay, Friedmann and Kupfermann 2010) using an identical derivation as given in Appendix B.2.

B.4 CTBN dynamics in star approximation

We are now going to derive the dynamics of CTBNs in star approximation, defined by fulfilling the Euler–Lagrange equations

$$\partial_x \mathcal{L}[t, x, \dot{x}] - \partial_t[\partial_{\dot{x}}\mathcal{L}[t, x, \dot{x}]] = 0.$$

First lets consider the derivative with respect to $m_n(x)$:

$$\partial_{m_n(x)} H_n = \sum_u \sum_{y \neq x} \frac{\tau_n^u(x, y)}{m_n(x)}, \quad \partial_{m_n(x)} E_j = \mathbb{E}_n[\mathcal{R}_n^u(x, x)],$$

Further if node n has a child j

$$\partial_{m_n(x)} H_j = \sum_{x, u \mid X_n(t) = x_n = x} \sum_{y \neq x} \frac{\tau_j^u(x, y)}{m_n(x)}, \quad \partial_{m_n(x)} E_j = \sum_x m_j(x) \mathbb{E}_n[\mathcal{R}_n^u(x, x) \mid X_n(t) = x].$$

With respect to the derivative $\dot{m}_n(x)$ we get

$$\partial_{\dot{m}_n(x)} \mathcal{L} = -\lambda_n(x).$$

We derive with respect to the transitions

$$\partial_{\tau_n^u(x,y)} H_n = \ln[m_n(x) m_n^u] - \ln \tau_n^u(x, y), \quad \partial_{\tau_n^u(x,y)} E_n = \ln \mathcal{R}_n^u(x, y).$$

thus

$$\partial_{\tau_n^u(x,y)} \mathcal{L} = \ln[m_n(x) m_n^u] - \ln \tau_n^u(x, y) + \ln \mathcal{R}_n^u(x, y) - \lambda_n(x) + \lambda_n(y).$$

The derivative with respect to the Lagrange-multipliers yields:

$$\partial_{\lambda_n(x)}\mathcal{L} = -\left\{\dot{m}_n(x) - \left[\sum_{y\neq x,u} \tau_n^u(y,x) - \tau_n^u(x,y)\right]\right\}$$

And lastly assuming a factorized noise model $P(Y^i|X(t)=x) = \prod_n P_n(Y^i|X_n(t)=x_n)$ we have for the derivative of $\mathcal{F}_0$

$$\partial_{m_n(x)}\mathcal{F}_0 = \sum_i \ln P_n(Y^i|x)\delta(t-t^i)$$

These can then be combined as the following Euler-Lagrange equations:

(I) $0 = \sum_u \sum_{y\neq x} \dfrac{\tau_n^u(x,y)}{m_n(x)} + \mathbb{E}_n[\mathcal{R}_n^u(x,y)] + \dot{\lambda}_n(x) + \sum_i \ln P_n(Y^i|x)\delta(t-t^i)$

$\qquad + \sum_{j\in\text{child}(n)} \sum_{x,u|X_n(t)=x} \sum_{y\neq x} \dfrac{\tau_n^u(x,y)}{m_n(x)} + \sum_x m_j(x)\mathbb{E}_j[r_j^u(x,x) \mid X_n(t)=x]$

(II) $0 = \ln[m_n(x)m_n^u] - \ln\tau_n^u(x,y) + \ln\mathcal{R}_n^u(x,y) - \lambda_n(x) + \lambda_n(y)$

(III) $\dot{m}_n(x) = \sum_{y\neq x,u} \tau_n^u(y,x) - \tau_n^u(x,y).$

Exponentiating (II) gives

$$(\text{II}^*) \quad \tau_n^u(x,y) = m_n(x)m_n^u\mathcal{R}_n^u(x,y)\rho_n(y)/\rho_n(x),$$

where $\rho_n(x) \equiv \exp(\lambda_n(x))$. Assuming that $\mathcal{R}$ is irreducible, $\rho_n(x)$ and $m_n(x)$ are non-zero in $(0,T)$ and we can thus eliminate $\tau_n^u(x,y)$ in (I) and (II). Thus

$(\text{I}^*) \quad \dot{\rho}_n(x) = \sum_{y\neq x} \mathbb{E}_n[\mathcal{R}_n^u(x,y)]\rho_n(y) + \{\mathbb{E}_n[\mathcal{R}_n^u(x,x)] + \psi_n\}\rho_n(x)$

$(\text{III}^*) \quad \dot{m}_n(x) = \sum_{y\neq x} \{m_n(y)\mathbb{E}_n[\mathcal{R}_n^u(y,x)]\rho_n(x)/\rho_n(y) - m_n(x)\mathbb{E}_n[\mathcal{R}_n^u(x,y)]\rho_n(y)/\rho_n(x)\},$

where we used $\dot{\lambda}_n(x) = \frac{1}{\rho_n(x)}\dot{\rho}_n(x)$. We further summarized

$$\psi_n = \sum_{j\in\text{child}(n)} \sum_x m_j(x)\left\{\sum_{y\neq x} \frac{\rho_j(y)}{\rho_j(x)}\mathbb{E}_n[r_j^u(x,y) \mid X_n(t)=x] + \mathbb{E}_n[r_j^u(x,x) \mid X_n(t)=x]\right\}$$
$$+ \sum_i \ln P_n(Y^i|x)\delta(t-t^i).$$

The driving term $\ln P_n(Y^i|x)\delta(t-t^i)$ then conditions the dynamics on the observations by $\lim_{t\to t^{i-}} \rho_n(x) = \lim_{t\to t^{i+}} P_n(Y^i|x)\rho_n(x)$.

B.5 Variational marginal score

Using $\mathcal{R}_n^u(x,x) = -\sum_{y\neq x} \mathcal{R}_n^u(x,y)$ we can write

$$E_n = \int dt \sum_{x,u} \sum_{y\neq x} [\tau_n^u(x,y)\ln\mathcal{R}_n^u(x,y) - m_n(x)m_n^u\mathcal{R}_n^u(x,y)].$$

For the approximated evidence

$$P(\mathbf{Y} \mid \mathcal{R}) \approx \prod_n \exp[E_n + H_n].$$

we get

$$P(\mathbf{Y} \mid \mathcal{R}) \approx e^H \prod_n \prod_{x,u} \prod_{y\neq x} \mathcal{R}_n^u(x,y)^{\mathbb{E}[M_n^u(x,y)]} e^{-\mathbb{E}[T_n^u(x)]\mathcal{R}_n^u(x,y)}.$$

Assuming an independent *Gamma* prior

$$P(\mathcal{R} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{G}) = \prod_n \prod_{x,u} \prod_{y \neq x} \gamma \left[r_n^u(x,y) \mid \alpha_n^u(x,y), \beta_n^u(x) \right]$$

$$= \prod_n \prod_{x,u} \prod_{y \neq x} \frac{\beta_n^u(x)^{\alpha_n^u(x,y)}}{\Gamma(\alpha_n^u(x,y))} \mathcal{R}_n^u(x,y)^{\alpha_n^u(x,y)-1} e^{-\beta_n^u(x)\mathcal{R}_n^u(x,y)}.$$

Thus we can express the graph posterior

$$P(\mathcal{G}|\mathbf{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto P(\mathcal{G}) \int_0^\infty P(\mathbf{Y}|\mathcal{R}) P(\mathcal{R}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{G}) \, \mathrm{d}\mathcal{R}$$

$$= e^H \prod_n \prod_{x,u} \prod_{y \neq x} \frac{\beta_n^u(x)^{\alpha_n^u(x,y)}}{\Gamma(\alpha_n^u(x,y))}$$

$$\times \int_0^\infty \mathcal{R}_n^u(x,y)^{\mathbb{E}[M_n^u(x,y)]+\alpha_n^u(x,y)-1} e^{-[\mathbb{E}[T_n^u(x)]+\beta_n^u(x)]\mathcal{R}_n^u(x,y)} \mathrm{d}\mathcal{R}_n^u(x,y),$$

which has an analytic solution

$$\int_0^\infty x^a e^{-bx} \mathrm{d}x = b^{-a} \Gamma(a).$$

Thus we get

$$P(\mathcal{G}|\mathbf{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto e^H \prod_n \prod_{x,u} \prod_{y \neq x} \left(\frac{\beta_n^u(x)}{(\mathbb{E}[T_n^u(x)]+\beta_n^u(x))^{\mathbb{E}[M_n^u(x,y)]}} \right)^{\alpha_n^u(x,y)} \frac{\Gamma(\mathbb{E}[M_n^u(x,y)]+\alpha_n^u(x,y))}{\Gamma(\alpha_n^u(x,y))}.$$

B.6 Marginal dynamics for CTBNs

In the following we are going to derive the dynamic equations of the marginal process for which we have to expand the Gamma-function. Assuming the sum of recorded transitions and prior transition number to be sufficiently large we can approximate the Gamma-function using Stirling's approximation $\Gamma(z) \approx \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z + \mathcal{O}\left(\frac{1}{z}\right)$ we get the approximate marginal score function

$$\ln P(\mathcal{G}|\mathbf{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \sum_n H_n + \mathcal{E}_n,$$

with

$$\mathcal{E}_n = \sum_{x,u} \sum_{y \neq x} \left(\mathbb{E}[M_n^u(x,y)] + \alpha_n^u(x,y) - \frac{1}{2} \right) \ln \left(\mathbb{E}[M_n^u(x,y)] + \alpha_n^u(x,y) \right)$$

$$- \left(\alpha_n^u(x,y) - \frac{1}{2} \right) \ln \left(\alpha_n^u(x,y) \right) + \alpha_n^u(x,y) \ln \left(\beta_n^u(x) \right)$$

$$- \left(\mathbb{E}[M_n^u(x,y)] + \alpha_n^u(x,y) \right) \ln \left(\mathbb{E}[T_n^u(x)] + \beta_n^u(x) \right) - \mathbb{E}[M_n^u(x,y)],$$

$$\partial_{m_n(x)} \mathcal{E}_n = -\sum_{y \neq x} \sum_u \mathbb{E}_n \left[\frac{\mathbb{E}[M_n^u(x,y)] + \alpha_n^u(x,y)}{\mathbb{E}[T_n^u(x)] + \beta_n^u(x)} \right],$$

$$\partial_{m_n(x)} \mathcal{E}_j = -\sum_x \sum_{y \neq x} m_j(x) \mathbb{E}_n \left[\frac{\mathbb{E}[M_j^u(x,y)] + \alpha_j^u(x,y)}{\mathbb{E}[T_j^u(x)] + \beta_j^u(x)} \mid X_n(t) = x \right] \mathbb{1}[j \in \text{child}(n)],$$

$$\partial_{\tau_n^u(x,y)} \mathcal{E}_n \approx \ln \left(\frac{\mathbb{E}[M_n^u(x,y)] + \alpha_n^u(x,y)}{\mathbb{E}[T_n^u(x)] + \beta_n^u(x)} \right),$$

where we approximated $\frac{\mathbb{E}[M_n^u(x,y)]+\alpha_n^u(x,y)-\frac{1}{2}}{\mathbb{E}[M_n^u(x,y)]+\alpha_n^u(x,y)} - 1 \approx 0$. The derivatives with respect to H_n and the constraint remain unchanged, see Appendix B.4. Finally defining the posterior-rate

$$\bar{\mathcal{R}}_n^u(x,y) \equiv \frac{\mathbb{E}[M_n^u(x,y)] + \alpha_n^u(x,y)}{\mathbb{E}[T_n^u(x)] + \beta_n^u(x)},$$

we arrive at the same set of equations as in Appendix B.4.

C   Processing IRMA data

In this section we present our approach of processing IRMA data. The IRMA dataset consists of expression data of genes, measured in concentrations, which are continuous. We can not capture continuous data using CTBNs, but need to map this data to a set of latent states. We identify two states *over-expressed* ($X = 1$) and *under-expressed* ($X = 0$) with respect to the *basal* (equilibrium) concentration c_B. This motivates the following observation model given the basal concentration

$$P(Y \mid X = 1, c_B) = \begin{cases} 1/|Y_0| & , Y \geq c_B \text{ and } Y \leq Y_0 \\ 0 & , else \end{cases}, \tag{10}$$

$$P(Y \mid X = 0, c_B) = \begin{cases} 1/|Y_0| & , Y < c_B \text{ and } Y \geq -Y_0 \\ 0 & , else \end{cases}, \tag{11}$$

where we have to choose some Y_0, so that the likelihood is normalized. We set Y_0 to some large value $Y_0 \geq \mathrm{argmax}_{|Y| \in \mathrm{DATA}}$ as our method remains invariant under each choice.

We model the basal concentration itself is a random variable, which we assume is gaussian distributed. We can estimate the parameters of the gaussian distribution μ_B and σ_B from the data. The marginal observation model is then acquired by integration

$$P(Y \mid X) = \begin{cases} 1 - \mathrm{erf}((Y - \mu_B)/\sigma_B) & , X = 1 \\ \mathrm{erf}((Y - \mu_B)/\sigma_B) & , X = 0 \end{cases}. \tag{12}$$

Given this observation model we can assign each measurement a likelihood and can process the data using our method. We note that other models for IRMA data can be thought of that may return better (or worse) results using our method.