Bayesian structural inference for hidden processes

Christopher C. Strelioff[1,*] and James P. Crutchfield[1,2,†]
[1]*Complexity Sciences Center and Physics Department, University of California at Davis, One Shields Avenue, Davis, California 95616, USA*
[2]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*

(Received 10 September 2013; published 10 April 2014)

We introduce a Bayesian approach to discovering patterns in structurally complex processes. The proposed method of Bayesian structural inference (BSI) relies on a set of candidate unifilar hidden Markov model (uHMM) topologies for inference of process structure from a data series. We employ a recently developed exact enumeration of topological ϵ-machines. (A sequel then removes the topological restriction.) This subset of the uHMM topologies has the added benefit that inferred models are guaranteed to be ϵ-machines, irrespective of estimated transition probabilities. Properties of ϵ-machines and uHMMs allow for the derivation of analytic expressions for estimating transition probabilities, inferring start states, and comparing the posterior probability of candidate model topologies, despite process internal structure being only indirectly present in data. We demonstrate BSI's effectiveness in estimating a process's randomness, as reflected by the Shannon entropy rate, and its structure, as quantified by the statistical complexity. We also compare using the posterior distribution over candidate models and the single, maximum *a posteriori* model for point estimation and show that the former more accurately reflects uncertainty in estimated values. We apply BSI to in-class examples of finite- and infinite-order Markov processes, as well to an out-of-class, infinite-state hidden process.

DOI: 10.1103/PhysRevE.89.042119 PACS number(s): 02.50.Ga, 05.45.Tp, 02.50.Ey

I. INTRODUCTION

Emergent patterns are a hallmark of complex, adaptive behavior, whether exhibited by natural or designed systems. Practically, discovering and quantifying the structures making up emergent patterns from a sequence of observations lies at the heart of our ability to understand, predict, and control the world. But what are the statistical signatures of structure? A common modeling assumption is that observations are independent and identically distributed (IID). This is tantamount, though, to assuming a system is structureless. Therefore, pattern discovery depends critically on testing when the IID assumption is violated. Said more directly, successful pattern discovery extracts the (typically hidden) mechanisms that create departures from IID structurelessness. In many applications, the search for structure is made all the more challenging by limited available data. The very real consequences, when pattern discovery is done incorrectly with finite data, are that structure can be mistaken for randomness and randomness for structure.

The search for meaningful or appropriate structure to describe the mechanisms generating observed data is fundamental to all areas of science. Due to this central importance, a wide variety of approaches to finding structure or topology has resulted from focusing on different types of data, as well as from varying assumptions about the appropriate type of mathematical model for the system of interest. Before introducing a specific type of data and our model classes, it is helpful to point the interested reader to complementary efforts on structural inference. Examples include employing information-theoretic measures to infer nonlinear ordinary differential equations (ODEs) [1] and particle filtering to infer ODEs and continuous-time Markov processes [2], as well as a variety of hybrid Bayesian techniques [3,4]. Beyond specific methods of structural inference, some have discussed why modeling works and how it can be difficult [5,6]. Though incomplete as a broad overview of structural inference in different settings, the above references and citations therein provide a useful starting point.

Here, we develop an approach to pattern discovery that attempts to remove the confusions between randomness and structure, focusing on data series consisting of a sequence of symbols from a finite alphabet. That is, we wish to discover temporal patterns, as they occur in discrete-time and discrete-state time series. (The approach also applies to spatial data exhibiting one-dimensional patterns.) Inferring structure from data series of this type is integral to many fields of science ranging from bioinformatics [7,8], dynamical systems [9–12], and linguistics [13,14] to single-molecule spectroscopy [15,16], neuroscience [17,18], and crystallography [19,20]. Inferred structure assumes a meaning distinctive to each field. For example, in single-molecule dynamics structure reflects stable molecular configurations, as well as the rates and types of transition between them. In the study of coarse-grained dynamical systems and linguistics, structure often reflects forbidden words and relative frequencies of symbolic strings that make the language or dynamical system functional. Thus, the results of successful pattern discovery teach one much more about a process than models that are only highly predictive.

Our goal is to infer structure using a finite data sample from some process of interest and a set of candidate ϵ-machine model topologies. This choice of model class is made because ϵ-machines provide optimal prediction as well as being a minimal and unique representation [21]. In addition, given an ϵ-machine, structure and randomness can be quantified using the statistical complexity C_μ and Shannon entropy rate h_μ. Previous efforts to infer ϵ-machines from finite data include *subtree merging* (SM) [22], *ϵ-machine spectral reconstruction*

[*] strelioff@ucdavis.edu
[†] chaos@ucdavis.edu

(ϵMSR) [23], and *causal-state splitting reconstruction* (CSSR) [24,25]. These methods produce a single, best estimate of the appropriate ϵ-machine given the available data.

The following develops a distinctively different approach to the problem of structural inference—*Bayesian structural inference* (BSI). BSI requires a data series D and a set of candidate unifilar hidden Markov model (uHMM) topologies, which we denote $\mathcal{M}$. However, for our present goal of introducing BSI, we consider only a subset of unifilar hidden Markov models—the topological ϵ-machines—that are guaranteed to be ϵ-machines irrespective of estimated transition probabilities [26]. Unlike the inference methods cited above, BSI's output is not a single best estimate. Instead, BSI determines the posterior probability of each model topology conditioned on D and $\mathcal{M}$. One result is that many model topologies are viable candidates for a given data set. The shorter the data series, the more prominent this effect becomes. We argue, in this light, that the most careful approach to structural inference and estimation is to use the complete set of model topologies according to their posterior probability. Another consequence, familiar in a Bayesian setting, is that principled estimates of uncertainty—including uncertainty in model topology—can be straightforwardly obtained from the posterior distribution.

The methods developed here draw from several fields, ranging from computational mechanics [21] and dynamical systems [27–29] to methods of Bayesian statistical inference [30]. As a result, elements of the following will be unfamiliar to some readers. To create a bridge, we provide an informal overview of foundational concepts in Sec. II before moving to BSI's technical details in Sec. III.

II. PROCESS STRUCTURE, MODEL TOPOLOGIES, AND FINITE DATA

To start, we offer a nontechnical introduction to structural inference to be clear how we distinguish (i) a process and its inherent structure from (ii) model topology and these from (iii) sampled data series. A *process* represents all possible behaviors of a system of interest. It is the object of our focus. Saying that we infer *structure* means we want to find the process's organization—the internal mechanisms that generate its observed behavior. However, in any empirical setting we only have samples of the process's behavior in the form of finite *data series*. A data series necessarily provides an incomplete picture of the process due to the finite nature of the observation. Finally, we use a *model* or, more precisely, a *model topology* to express the process's structure. The model topology—the set of states and transitions, their connections, and observed output symbols—explicitly represents the process's structure. Typically, there are many model topologies that accurately describe the probabilistic structure of a given process. ϵ-Machines are special within the set of accurate models, however, in that they are the model topology that provides the unique and minimal representation of process structure.

To ground this further, let us graphically survey different model topologies and consider what processes they represent and how they generate finite data samples. Figure 1 shows models with one or two states that generate binary processes—observed behavior is a sequence of 0s and 1s. For example, the smallest model topology is shown in Fig. 1(a) and represents the IID binary process. This model generates data by starting in state A and outputs a 0 with probability p and a 1 with probability $1 - p$, always returning to state A.

A more complex model topology, shown in Fig. 1(g), has two states and four edges. In this case, when the model is in state A it generates a 0 with probability p and returns to state A or it generates a 1 with probability $1 - p$ and moves to state B. When in state B, a 0 is generated with probability q and 1 with probability $1 - q$, moving to state A in both cases. If $p \neq q$ this model topology represents a unique, structured process. However, if $p = q$ the probability of generating a 0 or 1 does not depend on states A and B and the resulting process is IID. Thus, this model topology with $p = q$ becomes an overly verbose representation of the IID process, which requires only a single state—the topology of Fig. 1(a). This setting of the transition probabilities is an example where a model topology describes the probabilistic behavior of a process but does not reflect the structure. In fact, the model topology in Fig. 1(g) is not an ϵ-machine when $p = q$. Rather, the process structure is properly represented by Fig. 1(a), which is.

This example and other cases where specific model topologies are not minimal and unique representations of a process's structure motivate identifying a subclass of model topologies. All model topologies in Fig. 1 are unifilar hidden Markov models (defined shortly). However, the six model topologies with two states and four edges, Fig. 1(g)–1(i) and 1(l)–1(n), are not minimal when $p = q$. As with the previous example, they all become overly complex representations of the IID process for this parameter setting. Excluding these uHMMs leaves a subset of topologies called *topological ϵ-machines*, Fig. 1(a)–1(f), 1(j), and 1(k), that are guaranteed to be minimal and unique representations of process structure for any transition probabilities setting, other than 0 or 1. Partly to emphasize the role of process structure and partly to simplify technicalities, in this first introduction to BSI we only consider topological ϵ-machines. A sequel lifts this restriction, adapting BSI to work with all ϵ-machines.

In this way, we see how a process's structure is expressed in model topology and how possible ambiguities arise. This is the *forward* problem of statistical inference. Now consider the complementary *inverse* problem: Given an observed data series, find the model topology that most effectively describes the unknown process structure. In a Bayesian setting, the first step is to identify those model topologies that can generate the observed data. As just discussed, we do this by choosing a specific model topology and start state and attempting to trace the hidden-state path through the model, using the observed symbols to determine the edges to follow. If there is a path for at least one start state, the model topology is a viable candidate. This process is repeated for each model topology in a specified set, such as that displayed in Fig. 1. The procedure that lists, and tests, model topologies in a set of candidates we call *enumeration*.

To clarify the procedure for tracing hidden-state paths let us consider a specific example of observed data consisting of the following short binary sequence:

$$11101100111101111001. \qquad (1)$$

If tested against each candidate in Fig. 1, 8 of the 14 model topologies are possible: (a), (e), (g)–(i), and (l)–(n). For

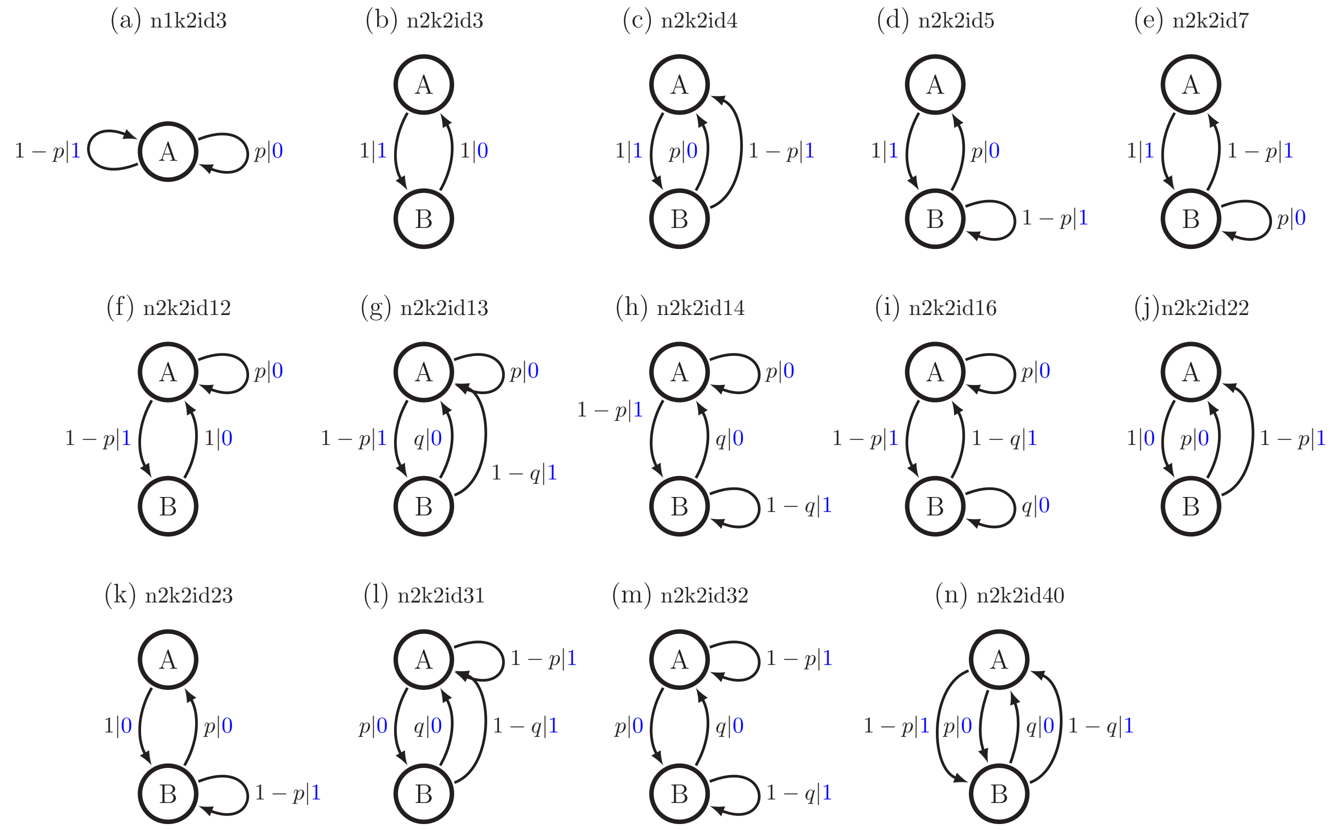

FIG. 1. (Color online) All binary, unifilar hidden Markov model topologies with one or two states. Each topology, designated (a) through (n), also has a unique label that provides the number of states $n = 1,2$, the alphabet size $k = 2$, and a unique id that comes from the algorithm used to enumerate all possible model topologies [26]. Model edges are labeled with a transition probability and output symbol using the following format: *probability* | *symbol*.

example, using Fig. 1(i) and starting in state A, the observed data are generated by the hidden-state path as follows:

$$ABABBABBBABABBABABBBA. \tag{2}$$

One way to describe this path—one that is central to statistical estimation—is to count the number of times each edge in the model was traversed. Using $n(\sigma\; x\;|\sigma_0)$ to denote the number of times that symbol x is generated using an edge from state σ given that the sequence starts in state σ_0, we obtain $n(A0|A) = 0$, $n(A1|A) = 7$, $n(B0|A) = 6$, and $n(B1|A) = 7$, again assuming $\sigma_0 = A$. Similar paths and sets of edge counts are found for the eight viable topologies cited above. These counts are the basis for estimating a topology's transition and start-state probabilities. From these, one can then calculate the probability that each model topology produced the observed data series—each candidate's posterior probability.

By way of outlining what is to follow, let us formalize the procedure just sketched in terms of the primary goal of estimating candidates' posterior probabilities. First, Sec. III recapitulates what is known about the space of structured processes, reviewing how they are represented as ϵ-machines and how topological ϵ-machines are exactly enumerated. Then Sec. IV adapts Bayesian inference methods to this model class, analyzing transition probability and start-state estimation for a single, known topology. Next, setting the context for comparing model topologies, it explores the organization of the prior over the set $\mathcal{M}$ of candidate models. Section IV closes with a discussion of how to estimate various process statistics from functions of model parameters. Finally, Sec. V applies BSI to a series of increasingly complex processes: (i) the Golden Mean Process, a finite-order Markov process; (ii) the Even Process, an infinite-order Markov process; and, finally, (iii) the Simple Nonunifilar Source (SNS), an infinite-memory process. Although these example processes appear simple at first glance, the range of dynamics is substantial. The Golden Mean Process is a simple (first-order) Markov chain, meaning that each state corresponds to the previous observed symbol [31]. Although the Even Process has two states when illustrated as a uHMM, a Markov chain representation would require an infinite number of states. Finally, the SNS, because it is nonunifilar, would take an infinite number of states to represent as a uHMM. This example is truly *out of class* and it is not possible to capture the structure of the SNS using the set of topological ϵ-machines. Given these varied data sources, we illustrate BSI's effectiveness by emphasizing its ability to accurately estimate a process's randomness (Shannon entropy rate h_μ) in all examples and stored information (statistical complexity C_μ) in all examples except the SNS.

III. STRUCTURED PROCESSES

We describe a system of interest in terms of its observed behavior, following the approach of computational mechanics, as reviewed in Ref. [21]. Again, a *process* is the collection of behaviors that the system produces. A process's probabilistic description is a bi-infinite chain of random variables, denoted by capital letters, $\ldots X_{t-2}\, X_{t-1}\, X_t\, X_{t+1}\, X_{t+2} \ldots$. A realization is indicated by lowercase letters, $\ldots x_{t-2}\, x_{t-1}\, x_t\, x_{t+1}\, x_{t+2} \ldots$. We assume the value x_t belongs to a discrete alphabet $\mathcal{X}$. We work with blocks $X_{t:t'} = X_t \ldots X_{t'-1}$, where the first index is inclusive and the second exclusive.

ϵ-Machines were originally defined in terms of prediction in the so-called history formulation [21,22]. Given a past realization $x_{-\infty:t} = \ldots x_{t-2}\, x_{t-1}$ and future random variables $X_{t:\infty} = X_t\, X_{t+1} \ldots$, the conditional distributions $\mathbb{P}(X_{t:\infty}|x_{-\infty:t})$ define the predictive equivalence relation over pasts as follows:

$$x_{-\infty:t} \sim x_{-\infty:t'} \Leftrightarrow \mathbb{P}(X_{t:\infty}|x_{-\infty:t}) = \mathbb{P}(X_{t':\infty}|x_{-\infty:t'}). \quad (3)$$

Within the history formulation, a process determines the ϵ-machine topology through $\sim$: The *causal states* $\boldsymbol{\mathcal{S}}$ are its equivalence classes and these, in turn, induce state-transition dynamics [21]. This way of connecting a process and its ϵ-machine influenced previous approaches to structural inference [22,25,32].

The ϵ-machine generator formulation, an alternative, was motivated by the problem of synchronization [33,34]. There an ϵ-machine topology defines the process that can be generated by it. Recently, the generator and history formulations were proven to be equivalent [35]. Although the history view is sometimes more intuitive, the generator view is useful in a variety of applications, especially the approach to structural inference developed here.

Following [33–35], we start with four definitions that delineate the model classes relevant for temporal pattern discovery.

Definition 1: A finite-state, edge-labeled hidden Markov model (HMM) consists of the following:

(1) A finite set of hidden states $\boldsymbol{\mathcal{S}} = \{\sigma_1, \ldots, \sigma_N\}$.

(2) A finite output alphabet $\mathcal{X}$.

(3) A set of $N \times N$ symbol-labeled transition matrices $T^{(x)}$, $x \in \mathcal{X}$, where $T^{(x)}_{i,j}$ is the probability of transitioning from state σ_i to state σ_j and emitting symbol x. The corresponding overall state-to-state transition matrix is denoted $T = \sum_{x \in \mathcal{X}} T^{(x)}$.

Definition 2: A finite-state, edge-labeled, unifilar HMM (uHMM) is a finite-state, edge-labeled HMM with the following property:

(1) *Unifilarity*: For each state $\sigma_i \in \boldsymbol{\mathcal{S}}$ and each symbol $x \in \mathcal{X}$ there is at most one outgoing edge from state σ_i that outputs symbol x.

Definition 3: A finite-state ϵ-machine is a uHMM with the following property:

(1) *Probabilistically distinct states*: For each pair of distinct states $\sigma_k, \sigma_j \in \boldsymbol{\mathcal{S}}$ there exists some finite word $w = x_0 x_1 \ldots x_{L-1}$ such that

$$\mathbb{P}(w|\sigma_0 = \sigma_k) \neq \mathbb{P}(w|\sigma_0 = \sigma_j).$$

Definition 4: A topological ϵ-machine is a finite-state ϵ-machine where the transition probabilities for leaving each state are equal for all outgoing edges.

These definitions provide a hierarchy in the model topologies to be considered. The most general set (Definition 1) consists of finite-state, edge-labeled HMM topologies with few restrictions. These are similar to models employed in many machine learning and bioinformatics applications; see, e.g., Ref. [7]. Using Definition 2, the class of HMMs is further restricted to be unifilar. The inference methods developed here apply to all model topologies in this class, as well as all more restricted subclasses. As a point of reference, Fig. 1 shows all binary, full-alphabet (able to generate both 0s and 1s) uHMM topologies with one or two states. If all states in the model are probabilistically distinct, following Definition 3, these model topologies are also valid generator ϵ-machines. Whether a uHMM is also a valid ϵ-machine often depends on the specific transition probabilities for the machine; see Sec. II for an example. This dependence motivates the final restriction to topological ϵ-machines (Definition 4), which are guaranteed to be minimal even if transition probabilities are equal.

Here, we employ the set of topological ϵ-machines for structural inference. Although specific settings of the transition probabilities are used to *define the set of allowed model topologies* this does not affect the actual inference procedure. For example, in Fig. 1 only Figs. 1(a)–1(f), 1(j), and 1(k) are topological ϵ-machines. However, the set of topological ϵ-machines does exclude a variety of model topologies that might be useful for general time-series inference. For example, when Definition 4 is applied, all processes with full support (all words allowed) reduce to a single-state model. However, broadening the class of topologies beyond the set considered here is straightforward and so we address extending the present methods to them in a sequel. The net result emphasizes structure arising from the distribution's support and guarantees that inferred models can be interpreted as valid ϵ-machines. And the goal is to present BSI's essential ideas for one class of structured processes—the topological ϵ-machines.

The set of topological ϵ-machines can be exactly and efficiently enumerated [26], motivating the use of this model class as our first example application of BSI. Table I lists the number $F_{n,k}$ of full-alphabet topologies with $n = 1, \ldots, 5$ states and alphabet size $k = 2$. Compare this table with the model topologies in Fig. 1, where all $n = 1$ and $n = 2$ uHMMs are shown. Only Fig. 1(a)–1(f), 1(j), and 1(k) are topological ϵ-machines, accounting for the difference between the 8 models in Table I and the 14 in Fig. 1. For comparison, the library has been enumerated up to eight states, containing approximately 2×10^9 distinct topologies. However, for the examples to follow we employ all 36 660 binary model topologies up to and including five states as the candidate basis for structural inference.

TABLE I. Size $F_{n,2}$ of the enumerated library of full-alphabet, binary topological ϵ-machines from one to five states. Reproduced with permission from Ref. [26].

States n	ϵ-Machines $F_{n,2}$
1	1
2	7
3	78
4	1388
5	35 186

IV. BAYESIAN INFERENCE

Previously, we developed methods for kth-order Markov chains to infer models of discrete stochastic processes and coarse-grained continuous chaotic dynamical systems [12,36]. There, we demonstrated that correct models for in-class data sources could be effectively and parsimoniously estimated. In addition, we showed that the hidden-state nature of out-of-class data sources could be extracted via model comparison between Markov orders as a function of data series length. Notably, we also found that the entropy rate can be accurately estimated, even when out-of-class data were considered.

The following extends the Markov chain methods to the topologically richer model class of unifilar hidden Markov models. The starting point depends on the unifilar nature of the HMM topologies considered here (Definition 2)—transitions from each state have a unique emitted symbol and destination state. As we demonstrated in Sec. II, unifilarity also means that, given an assumed start state, an observed data series corresponds to at most one path through the hidden states. The ability to directly connect observed data and hidden-state paths is not possible in the more general class of HMMs (Definition 1) because they can have many, often exponentially many, possible hidden paths for a single observed data series. In contrast, as a result of unifilarity, our analytic methods previously developed for "nonhidden" Markov chains [36] can be applied to infer uHMMs and ϵ-machines by adding a latent (hidden) variable for the unknown start state. We note in passing that for the more general class of HMMs, including nonunifilar topologies, there are two approaches to statistical inference. The first is to convert them to a uHMM (if possible), using mixed states [37]. The second is to use more conventional computational methods, such as Baum-Welch [38].

Setting aside these alternatives for now, we formalize the connection between observed data series and a candidate uHMM topology discussed in Sec. II. We assume that a data series $D_{0:T} = x_0 x_1 \ldots x_{T-2} x_{T-1}$ of length T has been obtained from the process of interest, with x_t taking values in a discrete alphabet $\mathcal{X}$. When a specific model topology and start state are assumed, a hidden-state sequence corresponding to the observed data can sometimes, but not always, be found. We denote a hidden state at time t as σ_t and a hidden-state sequence corresponding to $D_{0:T}$ as $S_{0:T+1} = \sigma_0 \sigma_1 \ldots \sigma_{T-1} \sigma_T$. Note that the state sequence is longer than the observed data series since the start and final states are included. Using this notation, an observed symbol x_t is emitted when transitioning from state σ_t to state σ_{t+1}. For example, using the observed data in Eq. (1), a hidden-state path corresponding to Eq. (2) can be obtained by assuming topology Fig. 1(i) and start state A.

We can now write out the probability of an observed data series. We assume a stationary uHMM topology M_i with a set of hidden states $\sigma_i \in \boldsymbol{\mathcal{S}}_i$. We add the subscript i to make it clear that we are analyzing a set of distinct, enumerated model topologies. As demonstrated in the example from Sec. II, edge counts $n(\sigma_i x | \sigma_{i,0})$ are obtained by tracing the hidden-state path given an assumed start state $\sigma_{i,0}$. Putting this all together, the probability of observed data $D_{0:T}$ and corresponding state-path $S_{0:T+1}$ is as follows:

$$\mathbb{P}(S_{0:T+1}, D_{0:T}) = p(\sigma_{i,0}) \prod_{\sigma_i \in \boldsymbol{\mathcal{S}}_i} \prod_{x \in \mathcal{X}} p(x|\sigma_i)^{n(\sigma_i x|\sigma_{i,0})}. \quad (4)$$

A slight manipulation of Eq. (4) lets us write the probability of observed data and hidden dynamics, given an assumed start state $\sigma_{i,0}$, as follows:

$$\mathbb{P}(S_{0:T+1}, D_{0:T} | \sigma_{i,0}) = \prod_{\sigma_i \in \boldsymbol{\mathcal{S}}_i} \prod_{x \in \mathcal{X}} p(x|\sigma_i)^{n(\sigma_i x|\sigma_{i,0})}. \quad (5)$$

The development of Eq. (5) and the simple example provided in Sec. II lay the groundwork for our application of Bayesian methods. That is, given topology M_i and start state $\sigma_{i,0}$, the probability of *observed* data $D_{0:T}$ and *hidden* dynamics $S_{0:T+1}$ can be calculated. For the purposes of inference, the combination of observed and hidden sequences is our data $\mathbf{D} = (D_{0:T}, S_{0:T+1})$.

A. Inferring transition probabilities

The first step is to infer transition probabilities for a single uHMM or topological ϵ-machine M_i. As noted above, we must assume a start state $\sigma_{i,0}$ so edge counts $n(\sigma_i, x|\sigma_{i,0})$ can be obtained from $D_{0:T}$. This requirement means that the inferred transition probabilities also depend on the assumed start state. At a later stage, when comparing model topologies, we demonstrate that the uncertainty in start state can be averaged over.

The set $\{\theta_i\}$ of parameters to estimate consists of those transition probabilities defined to be neither one nor zero by the assumed topology: $\theta_i = \{0 < p(x|\sigma_i, \sigma_{i,0}) < 1 : \sigma_i \in \boldsymbol{\mathcal{S}}_i^*, \sigma_{i,0} \in \boldsymbol{\mathcal{S}}_i\}$, where $\boldsymbol{\mathcal{S}}_i^* \subseteq \boldsymbol{\mathcal{S}}_i$ is the subset of hidden states with more than one outgoing edge. The resulting likelihood follows directly from Eq. (5):

$$\mathbb{P}(\mathbf{D}|\theta_i, \sigma_{i,0}, M_i) = \prod_{\sigma_i \in \boldsymbol{\mathcal{S}}_i} \prod_{x \in \mathcal{X}} p(x|\sigma_i, \sigma_{i,0})^{n(\sigma_i, x|\sigma_{i,0})}. \quad (6)$$

We note that the set of transition probabilities used in the above expression are unknown when doing statistical inference. However, we can still write the probability of the observed data given a setting for these unknown values, as indicated by the notation for the likelihood: $\mathbb{P}(\mathbf{D}|\theta_i, \sigma_{i,0}, M_i)$. Although not made explicit above, there is also a possibility that the likelihood vanishes for some, or all, start states if the observed data is not compatible with the topology. For example, if we attempt to use Fig. 1(d) for the observed data in Eq. (1) we find that neither $\sigma_{i,0} = A$ nor $\sigma_{i,0} = B$ leads to viable paths for the observed data, resulting in zero likelihood. For later use, we denote the number of times a hidden state is visited by $n(\sigma_i \bullet |\sigma_{i,0}) = \sum_{x \in \mathcal{X}} n(\sigma_i, x|\sigma_{i,0})$.

Equation (6) exposes the Markov nature of the dynamics on the hidden states and suggests adapting the methods we previously developed for Markov chains [36]. Said simply, states that corresponded there to histories of length k for Markov chain models are replaced by a hidden state σ_i. Mirroring the earlier approach, we employ a conjugate prior for transition probabilities. This choice means that the posterior distribution has the same form as the prior but with modified parameters. In the present case, the conjugate prior is a product of Dirichlet distributions as follows:

$$\mathbb{P}(\theta_i|\sigma_{i,0},M_i) = \prod_{\sigma_i\in\boldsymbol{\mathcal{S}}_i^*}\left\{\frac{\Gamma(\alpha(\sigma_i\bullet|\sigma_{i,0}))}{\prod_{x\in\mathcal{X}}\Gamma(\alpha(\sigma_i x|\sigma_{i,0}))}\right. \times\delta\left(1-\sum_{x\in\mathcal{X}}p(x|\sigma_i,\sigma_{i,0})\right) \left.\times\prod_{x\in\mathcal{X}}p(x|\sigma_i,\sigma_{i,0})^{\alpha(\sigma_i x|\sigma_{i,0})-1}\right\}, \qquad (7)$$

where $\alpha(\sigma_i\bullet|\sigma_{i,0}) = \sum_{x\in\mathcal{X}}\alpha(\sigma_i x|\sigma_{i,0})$. In the examples to follow we take $\alpha(\sigma_i x|\sigma_{i,0}) = 1$ for all parameters of the prior. This results in a uniform density over the simplex for all transition probabilities to be inferred, irrespective of start state [39].

The product of Dirichlet distributions includes transition probabilities only from hidden states in $\boldsymbol{\mathcal{S}}_i^*$ because these states have more than one outgoing edge. For transition probabilities from states $\sigma_i \notin \boldsymbol{\mathcal{S}}_i^*$ there is no need for an explicit prior since the transition probability must be zero or 1 by definition of the uHMM topology. As a result, the prior expectation for transition probabilities is as follows:

$$\boldsymbol{E}_{\text{prior}}[p(x|\sigma_i,\sigma_{i,0})] = \frac{\alpha(\sigma_i x|\sigma_{i,0})}{\alpha(\sigma_i\bullet|\sigma_{i,0})}, \qquad (8)$$

for states $\sigma_i \in \boldsymbol{\mathcal{S}}_i^*$.

Next, we employ Bayes's theorem to obtain the posterior distribution for the transition probabilities given data and prior assumptions. In this context, it takes the following form:

$$\mathbb{P}(\theta_i|\mathbf{D},\sigma_{i,0},M_i) = \frac{\mathbb{P}(\mathbf{D}|\theta_i,\sigma_{i,0},M_i)\mathbb{P}(\theta_i|\sigma_{i,0},M_i)}{\mathbb{P}(\mathbf{D}|\sigma_{i,0},M_i)}. \qquad (9)$$

The terms in the numerator are already specified above as the likelihood and the prior, Eqs. (6) and (7), respectively.

The normalization factor in Eq. (9) is called the *evidence* or *marginal likelihood*. This term integrates the product of the likelihood and prior with respect to the set of transition probabilities θ_i as follows:

$$\begin{aligned}\mathbb{P}(\mathbf{D}|\sigma_{i,0},M_i) &= \int d\theta_i\ \mathbb{P}(\mathbf{D}|\theta_i,\sigma_{i,0},M_i)\mathbb{P}(\theta_i|\sigma_{i,0},M_i)\\ &= \prod_{\sigma_i\in\boldsymbol{\mathcal{S}}_i^*}\left\{\frac{\Gamma(\alpha(\sigma_i\bullet|\sigma_{i,0}))}{\prod_{x\in\mathcal{X}}\Gamma(\alpha(\sigma_i x|\sigma_{i,0}))}\right.\\ &\quad\left.\times\frac{\prod_{x\in\mathcal{X}}\Gamma(\alpha(\sigma_i x|\sigma_{i,0})+n(\sigma_i x|\sigma_{i,0}))}{\Gamma(\alpha(\sigma_i\bullet|\sigma_{i,0})+n(\sigma_i\bullet|\sigma_{i,0}))}\right\},\end{aligned} \qquad (10)$$

resulting in the average of the likelihood with respect to the prior. In addition to normalizing the posterior distribution [Eq. (9)], the evidence is important in our subsequent applications of Bayes's theorem. In particular, the quantity is central to the model selection to follow and is used to (i) determine the start state given the model and (ii) compare model topologies.

As discussed above, conjugate priors result in a posterior distribution of the same form, with prior parameters modified by observed counts as follows:

$$\begin{aligned}P(\theta_i|\mathbf{D},\sigma_{i,0},M_i) &= \prod_{\sigma_i\in\boldsymbol{\mathcal{S}}_i^*}\left\{\frac{\Gamma(\alpha(\sigma_i\bullet|\sigma_{i,0})+n(\sigma_i\bullet|\sigma_{i,0}))}{\prod_{x\in\mathcal{X}}\Gamma(\alpha(\sigma_i x|\sigma_{i,0})+n(\sigma_i x|\sigma_{i,0}))}\right.\\ &\quad\times\delta\left(1-\sum_{x\in\mathcal{X}}p(x|\sigma_i,\sigma_{i,0})\right)\\ &\quad\left.\times\prod_{x\in\mathcal{X}}p(x|\sigma_i,\sigma_{i,0})^{\alpha(\sigma_i x|\sigma_{i,0})+n(\sigma_i x|\sigma_{i,0})-1}\right\}.\end{aligned} \qquad (11)$$

Comparing Eqs. (7) and (11)—prior and posterior, respectively—shows that the distributions are very similar: $\alpha(\sigma_i x|\sigma_{i,0})$ (prior only) is replaced by $\alpha(\sigma_i x|\sigma_{i,0}) + n(\sigma_i x|\sigma_{i,0})$ (prior plus data). Thus, one can immediately write down the posterior mean for the transition probabilities:

$$\boldsymbol{E}_{\text{post}}[p(x|\sigma_i,\sigma_{i,0})] = \frac{\alpha(\sigma_i x|\sigma_{i,0})+n(\sigma_i x|\sigma_{i,0})}{\alpha(\sigma_i\bullet|\sigma_{i,0})+n(\sigma_i\bullet|\sigma_{i,0})}, \qquad (12)$$

for states $\sigma_i \in \boldsymbol{\mathcal{S}}_i^*$. As with the prior, probabilities for transitions from states $\sigma_i \notin \boldsymbol{\mathcal{S}}_i^*$ are zero or 1, as defined by the model topology.

Notably, the posterior mean for the transition probabilities does not completely specify our knowledge since the uncertainty, reflected in functions of the posterior's higher moments, can be large. These moments are available elsewhere [39]. However, using methods detailed below, we employ sampling from the posterior at this level, as well as other inference levels, to capture estimation uncertainty.

B. Inferring start states

The next task is to calculate the probabilities for each start state given a proposed machine topology and observed data. Although we are not typically interested in the actual start state, introducing this latent variable is necessary to develop the previous section's analytic methods. And, in any case, another level of Bayes's theorem allows us to average over uncertainty in the start state to obtain the probability of observed data for the topology, independent of start state.

We begin with the evidence $\mathbb{P}(\mathbf{D}|\sigma_{i,0},M_i)$ derived in Eq. (10) to estimate transition probabilities. When determining the start state, the evidence (marginal likelihood) from inferring transition probabilities becomes the likelihood for start-state estimation. As before, we apply Bayes's theorem, this time with unknown start states, instead of unknown transition probabilities,

$$\mathbb{P}(\sigma_{i,0}|\mathbf{D},M_i) = \frac{\mathbb{P}(\mathbf{D}|\sigma_{i,0},M_i)\mathbb{P}(\sigma_{i,0}|M_i)}{\mathbb{P}(\mathbf{D}|M_i)}. \qquad (13)$$

This calculation requires defining a prior over start states $\mathbb{P}(\sigma_{i,0}|M_i)$. In practice, setting start states as equally probable *a priori* is a sensible choice in light of the larger goal of structural inference. The normalization $\mathbb{P}(\mathbf{D}|M_i)$, or evidence, at this level follows by averaging over the uncertainty in $\sigma_{i,0}$,

$$\mathbb{P}(\mathbf{D}|M_i) = \sum_{\sigma_{i,0}\in\boldsymbol{\mathcal{S}}_i}\mathbb{P}(\mathbf{D}|\sigma_{i,0},M_i)\mathbb{P}(\sigma_{i,0}|M_i). \qquad (14)$$

The result of this calculation no longer explicitly depends on start states or transition probabilities. The uncertainty created by these unknowns has been averaged over, producing a very useful quantity for comparing different topologies: $\mathbb{P}(\mathbf{D}|M_i)$. However, one must not forget that inferring transition and start-state probabilities underlies the structural comparisons to follow. In particular, the priors set at the levels of transition probabilities and start states can impact the structures detected due to the hierarchical nature of the inference: $\mathbb{P}(\mathbf{D}|\theta_i,\sigma_{i,0},M_i) \to \mathbb{P}(\mathbf{D}|\sigma_{i,0},M_i) \to \mathbb{P}(\mathbf{D}|M_i)$.

C. Inferring model topology

So far, we inferred transition probabilities and start states for a given model topology. Now we are ready to compare different topologies in a set $\boldsymbol{\mathcal{M}}$ of candidate models. As with inferring start states given a topology, we write down yet another version Bayes's theorem, except one for model topology,

$$\mathbb{P}(M_i|\mathbf{D},\boldsymbol{\mathcal{M}}) = \frac{\mathbb{P}(\mathbf{D}|M_i,\boldsymbol{\mathcal{M}})\mathbb{P}(M_i|\boldsymbol{\mathcal{M}})}{\mathbb{P}(\mathbf{D}|\boldsymbol{\mathcal{M}})}, \tag{15}$$

writing the likelihood as $\mathbb{P}(\mathbf{D}|M_i,\boldsymbol{\mathcal{M}})$ to make the nature of the conditional distributions clear. This is exactly the same, however, as the evidence derived above in Eq. (14): $\mathbb{P}(\mathbf{D}|M_i) = \mathbb{P}(\mathbf{D}|M_i,\boldsymbol{\mathcal{M}})$. Equality holds because nothing in calculating the previous evidence term directly depends on the *set of models* considered. The evidence $\mathbb{P}(\mathbf{D}|\boldsymbol{\mathcal{M}})$, or normalization term, in Eq. (15) has the following general form:

$$\mathbb{P}(\mathbf{D}|\boldsymbol{\mathcal{M}}) = \sum_{M_j \in \boldsymbol{\mathcal{M}}} \mathbb{P}(\mathbf{D}|M_j,\boldsymbol{\mathcal{M}})\mathbb{P}(M_j|\boldsymbol{\mathcal{M}}). \tag{16}$$

To apply Eq. (15) we must first provide an explicit prior over model topologies. One general form, tuned by single parameter β, is

$$\mathbb{P}(M_i|\boldsymbol{\mathcal{M}}) = \frac{\exp\left(-\beta\phi(M_i)\right)}{\sum_{M_j\in\boldsymbol{\mathcal{M}}} \exp\left(-\beta\phi(M_j)\right)}, \tag{17}$$

where $\phi(M_i)$ is some desired function of model topology. In the examples to follow we use the number of causal states—$\phi(M_i) = |M_i|$—thereby penalizing for model size. This is particularly important when a short data series is being investigated. However, setting $\beta = 0$ removes the penalty, making all models in $\boldsymbol{\mathcal{M}}$ *a priori* equally likely. It is important to investigate the effects of choosing a specific β for a given set of candidate topologies. Below, we first demonstrate the effect of choosing $\beta = 0$, 2, or 4. After that, however, we employ $\beta = 4$ since this value, in combination with the set of one- to five-state binary-alphabet topological ϵ-machines, produces a preference for one- and two-state machines for short data series and still allows for inferring larger machines with only a few thousand symbols. Experience with this β shows that it is structurally conservative.

In the examples we explore two approaches to using the results of structural inference. The first takes into account all model topologies in the set considered, weighted according to the posterior distribution given in Eq. (15). The second selects a single model M_{map} that is the *maximum a posteriori* (MAP) topology,

$$M_{\text{map}} = \underset{M_i\in\boldsymbol{\mathcal{M}}}{\text{argmax}}\, \mathbb{P}(M_i|\mathbf{D},\boldsymbol{\mathcal{M}}). \tag{18}$$

The difference between these methods is most dramatic for short data series. Also, using the MAP topology often underestimates the uncertainty in functions of the model parameters, which we discuss shortly. Of course, since one throws away any number of comparable models, estimating uncertainty in any quantity that explicitly depends on the model topology cannot be done properly if MAP selection is employed. However, we expect some will want or need to use a single model topology, so we consider both methods.

D. Estimating functions of model parameters

A primary goal in inference is estimating functions that depend on an inferred model's parameters. We denote this $f(\theta_i)$ to indicate the dependence on transition probabilities. Unfortunately, substituting the posterior mean for the transition probabilities into some function of interest does not provide the desired expectation. In general, obtaining analytic expressions for the posterior mean of desired functions is quite difficult; see, for example, Refs. [40,41]. Deriving expressions for the uncertainty in the resulting estimates is equally involved and typically not done; although see Ref. [40].

Above, the inference method required inferring transition probabilities, start state, and topology. Function estimation, as a result, should take into account all these sources of uncertainty. Instead of deriving analytic expressions for posterior means (if possible), we turn to numerical methods to estimate function means and uncertainties in great detail. We do this by repeatedly sampling from the posterior distribution at each level to obtain a sample ϵ-machine and evaluating the function of interest for the sampled parameter values. The algorithms in Fig. 2 detail the process of sampling $f(\theta_i)$ using all candidate models $\boldsymbol{\mathcal{M}}$ (Algorithm 1) or the single M_{MAP} model (Algorithm 2). Given a set of samples of the function of interest, any summary statistic can be employed. In the examples, we generate $N_s = 50\,000$ samples from which we estimate a variety of properties. More specifically, these samples are employed to estimate the posterior mean and the

ALGORITHM 1: Sample using all topologies in $\boldsymbol{\mathcal{M}}$

```
for $n$ in $(1, N_s)$ do:
    $M_i^* \sim \mathbb{P}(M_i|\mathbf{D},\boldsymbol{\mathcal{M}})$    # sample topology
    $\sigma_{i,0}^* \sim \mathbb{P}(\sigma_{i,0}|\mathbf{D}, M_i^*)$    # sample start state
    $\theta_i^* \sim \mathbb{P}(\theta_i|\mathbf{D}, \sigma_{i,0}^*, M_i^*)$    # sample parameters
    $f_n = f(\theta_i^*)$    # store sample
```

ALGORITHM 2: Sample using MAP topology

```
$M_{\text{map}} = \text{argmax}_{M_i\in\boldsymbol{\mathcal{M}}} P(M_i|\mathbf{D},\boldsymbol{\mathcal{M}})$    # find MAP topology
for $n$ in $(1, N_s)$ do:
    $\sigma_{i,0}^* \sim \mathbb{P}(\sigma_{i,0}|\mathbf{D}, M_{\text{map}})$    # sample start state
    $\theta_i^* \sim \mathbb{P}(\theta_i|\mathbf{D}, \sigma_{i,0}^*, M_{\text{map}})$    # sample parameters
    $f_n = f(\theta_i^*)$    # store sample
```

FIG. 2. Pseudocode for generating N_s samples of a function $f(\theta_i)$ of model parameters $\{\theta_i\}$. Algorithm 1 samples a topology each time through the loop, whereas Algorithm 2 uses the MAP topology for all iterations. The sampling at each stage allows for the creation of a set of samples $\{f_n\}$ that accurately reflect the many sources of uncertainty in the posterior distribution.

95%, equal-tailed, credible interval (CI) [30]. This means there is a 5% probability of samples being outside the specified interval, with equal probability of being above or below the interval. Finally, a Gaussian kernel density estimation (Gkde) is used to visualize the posterior density for the functions of interest.

The examples demonstrate estimating process randomness and structure from data series using the two algorithms introduced above. For a known ϵ-machine topology M_i, with specified transition probabilities $\{p(x|\sigma_i)\}$, these properties are quantified using the entropy rate h_μ and statistical complexity C_μ, respectively. The entropy rate is

$$h_\mu = -\sum_{\sigma_i \in \boldsymbol{\mathcal{S}}_i} p(\sigma_i) \sum_{x \in \mathcal{X}} p(x|\sigma_i) \log_2 p(x|\sigma_i) \tag{19}$$

and the statistical complexity is

$$C_\mu = -\sum_{\sigma_i \in \boldsymbol{\mathcal{S}}_i} p(\sigma_i) \log_2 p(\sigma_i). \tag{20}$$

In these expressions, the $p(\sigma_i)$ are the asymptotic state probabilities determined by the left eigenvector (normalized in probability) of the internal Markov chain transition matrix $T = \sum_{x \in \mathcal{X}} T^{(x)}$. Of course, h_μ and C_μ are also functions of the model topology and transition probabilities, so these quantities provide good examples of how to estimate functions of model parameters in general.

V. EXAMPLES

We divide the examples into two parts. First, we demonstrate inferring transition probabilities and start states for a known topology. Second, we focus on inferring ϵ-machine topology using the set of all binary, one- to five-state topological ϵ-machines, consisting of 36 660 candidates; see Table I. We use the convergence of estimates for the information-theoretic values h_μ and C_μ to monitor structure discovery. However, estimating model parameters is at the core of the later examples and so we start with this procedure.

For each example we generate a single data series $D_{0:T}$ of length $T = 2^{17}$. When analyzing convergence, we consider subsamples $D_{0:L}$ of lengths $L = 2^i$, using $i = 0,1,2,\ldots,17$. For example, a four-symbol sequence starting at the first data point is designated $D_{0:4} = x_0x_1x_2x_3$. The overlapping analysis of a single data series gives insight into convergence for the inferred models and for the statistics estimated.

A. Estimating parameters

1. Even Process

We first explore a single example of inferring properties of a known data source using Eqs. (6)–(11). We generate a data series from the Even Process and then, using the correct topology (Fig. 3), we infer start states and transition probabilities and estimate the entropy rate and statistical complexity. We do not concentrate on this level of inference in subsequent examples, preferring to focus instead on model topology and its representation of the unknown process structure. Nonetheless, the procedure detailed here underlies all of the examples.

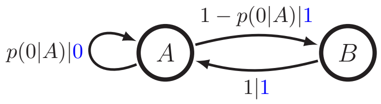

FIG. 3. (Color online) State-transition diagram for the Even Process's ϵ-machine topology. The "true" value of the unspecified transition probability is $p(0|A) = 1/2$. For this topology, $\boldsymbol{\mathcal{S}}_{\text{even}} = \{A,B\}$ and $\boldsymbol{\mathcal{S}}^*_{\text{even}} = \{A\}$ because state B has only one outgoing transition.

The Even Process is notable because it has infinite Markov order. This means no finite-order Markov chain can reproduce its word distribution [36]. It can be finitely modeled, though, with a finite-state unifilar HMM—the ϵ-machine of Fig. 3. A single data series was generated using the Even Process ϵ-machine with $p(0|A) = 1/2$. The start state was randomized before generating sequence data of length $T = 2^{17}$. As it turned out, the initial segment was $D_{0:T} = 100\ldots$, indicating that the unknown start state was B on that realization. This is so because the first symbol is a 1, which can be generated starting in either state A or B, but the sequence 10 is only possible by starting at node B.

Next, we estimate the transition probabilities from the generated data series using length-L subsamples $D_{0:L} = x_0x_1\ldots x_{L-1}$ to track convergence. Although the mean and other moments of the Dirichlet posterior can be calculated analytically [39], we sample values using Algorithm 2 in Fig. 2. However, in this example we employ M_{even} instead of M_{map} because we are focused on the model parameters for a known topology. The posterior density for each subsample $D_{0:L}$ is plotted in Fig. 4 using Gaussian kernel density estimation (Gkde). The true value of $p(0|A)$ is shown as a black, dashed line and the posterior mean as a solid, gray line. (Both lines connect values evaluated at each length $L = 2^0, 2^1, \ldots 2^{17}$.) The convergence of the posterior density to the correct value of

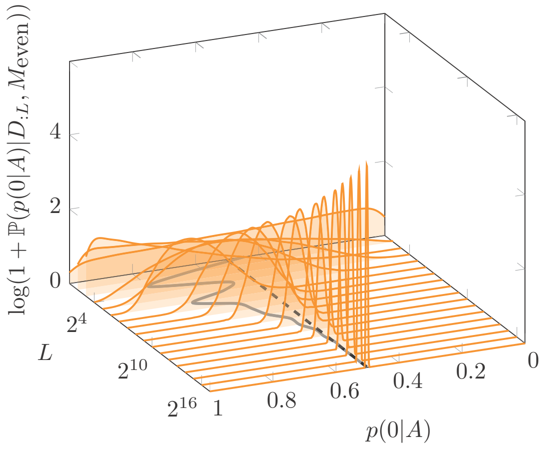

FIG. 4. (Color online) Convergence of posterior density $\mathbb{P}(p(0|A)|D_{0:L}, M_{\text{even}})$ as a function of subsample length $L = 2^i$, $i = 0,1,2,\ldots,17$. Each posterior density plot uses a Gaussian kernel density estimator with 50 000 samples from the posterior. The true value of $p(0|A) = 1/2$ appears as a black, dashed line and the posterior mean as a gray, solid line. A natural logarithm is used in plotting probability densities.

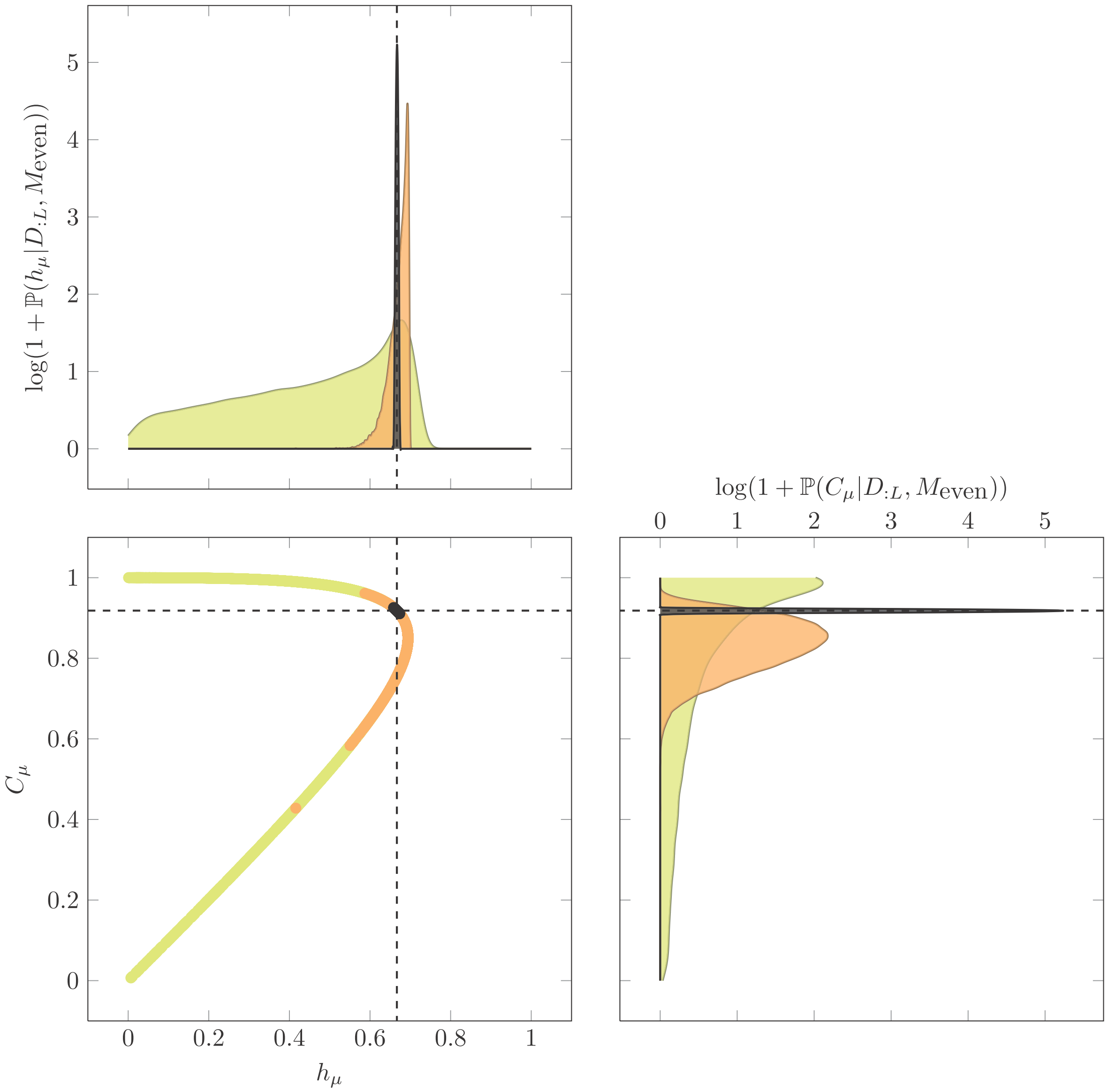

FIG. 5. (Color online) Convergence of randomness (h_μ) and structure (C_μ) calculated with transition probabilities and start states estimated from Even Process data, assuming the correct topology. Fifty thousand samples were taken from the joint posterior $\mathbb{P}(h_\mu, C_\mu | D_{0:L}, M_{\text{even}})$. (Lower left) A subsample of size 5000 for data sizes $L = 1$ [(light) beige], $L = 64$ [(medium) orange], and $L = 16\,384$ [(dark) black]. Gaussian kernel density estimates (using all 50 000 samples) of the marginal distributions $\mathbb{P}(h_\mu | D_{0:L}, M_{\text{even}})$ (top) and $\mathbb{P}(C_\mu | D_{0:L}, M_{\text{even}})$ (right) for the same values of L. Dashed lines indicate the true values of h_μ and C_μ for the Even Process. The distinctive arc traced out by the (light) beige samples in the h_μ-C_μ plane reflect the constrained nature of the structure-randomness relationship that the topology in Fig. 3 creates as $p(0|A)$ is varied between zero and 1. A natural logarithm is used in plotting probability densities.

$p(0|A) = 1/2$ with increasing data size is clear and, moreover, the true value is always in a region of positive probability.

For our final example using a known topology we estimate h_μ and C_μ from the Even Process data. This illustrates estimating these functions of model parameters when the ϵ-machine topology is known but there is uncertainty in start state and transition probabilities. As above, we use Algorithm 2 in Fig. 2 and employ the known machine structure. We sample start states and transition probabilities, followed by calculating h_μ and C_μ—via Eqs. (19) and (20), respectively—to build a posterior density for these quantities.

Figure 5 presents the joint distribution for C_μ and h_μ along with the Gkde estimation of their marginal densities. Samples from the joint posterior distribution are plotted in the lower left panel for subsample lengths $L = 1{,}64$, and 16 384. Only 5000 of the available samples are displayed in this panel to minimize the graphic's size. The marginal densities for h_μ (top panel) and C_μ (right panel) are plotted using a Gkde with all 50 000 samples. Small data size [$L = 1$, indicated by (light) beige points] samples allow a wide range of structure and randomness. The arc in the h_μ-C_μ plane reflects the flat priors set for start states and transition probabilities. These priors allow for almost uniform samples (modified by $L = 1$ data point) of the $p(0|A)$ transition probability. The resulting values of C_μ and h_μ create an arc that is only constrained by the model topology. We note that a uniform prior distribution over transition probabilities and start states does not produce a uniform distribution over h_μ or C_μ. Increasing the size of the data subsample to $L = 64$ [(medium) orange points] results in a considerable reduction in the uncertainty for both functions. For this amount of data, the possible values of entropy rate and statistical complexity curve around the true value in the h_μ-C_μ plane and result in a shifted peak for the marginal density for h_μ. For subsample length $L = 16\,384$ [(dark) black points] the estimates of both functions of model parameters converge to the true values, indicated by the black, dashed lines.

B. Inferring process structure

We are now ready to demonstrate BSI's efficacy for structural inference via a series of increasingly complex processes, monitoring convergence using data subsamples up to a length of $L = 2^{17}$. In this, we determine the number of hidden states, number of edges connecting them, and symbols output on each transition. As discussed above, we use the set of topological ϵ-machines as candidates because an efficient and exhaustive enumeration is available.

For comparison, we first explore the organization of the prior over the set of candidate ϵ-machines using intrinsic informational coordinates—the process entropy rate h_μ and statistical complexity C_μ. We focus on their joint distribution, as induced by various settings of the prior parameter β. The results lead us to use $\beta = 4$ for the subsequent examples. This value creates a preference for small models when few data are available but allows for a larger number of states when reasonable amounts of data support it.

We establish the BSI's effectiveness by inferring the structure of a finite-order Markov (Golden Mean) Process, an infinite-order Markov (Even) Process, and an infinite memory process (SNS). Again, the proxy for convergence is estimating structure and randomness as a function of the data subsample length L. Comparing these quantities' posterior distributions with their prior ones illustrates uncertainty reduction as more data are analyzed.

1. *Priors for structured processes*

Here, we use a prior over all binary-alphabet, topological ϵ-machines with one to five states (recall Table I). We denote the set of topological ϵ-machines detailed in Table I as $\boldsymbol{\mathcal{M}}$. Equation (17) allows specifying a preference for smaller ϵ-machines by setting $\beta > 0$ and defining the function of model structure to be the number of states: $\phi(M_i) = |M_i|$. Beyond setting this explicitly, there is an inherent bias to smaller models inversely proportional to the parameter space dimension. The parameter space is that of the estimated transition probabilities. Its dimension is the number of states with more than one out-going transition. However, candidate ϵ-machine topologies with many states and few transitions result in a small parameter space and so may be assigned high probability for short data series. In addition, the prior over topologies must take into account the increasing number of candidates as the number of states increases. Setting β sufficiently high so large models are not given high probability under these conditions is reasonable, as we would like to approach structure estimates (C_μ) monotonically from below, as data size increases.

Figure 6 plots samples from the resulting joint prior for (h_μ, C_μ) as well as the corresponding Gkde for marginal densities of both quantities. The data are generated by using the method of Sec. IV D and replacing the posterior density with the prior density. Specifically, rather than sampling a topology M_i from $\mathbb{P}(M_i|D,\boldsymbol{\mathcal{M}})$, we sample from $\mathbb{P}(M_i|\boldsymbol{\mathcal{M}})$. Similar substitutions are made at each level, using the distributions that do not depend on observed data, resulting in samples from the prior. Each color in the figure reflects samples using all ϵ-machines in $\boldsymbol{\mathcal{M}}$ with different values for the prior parameter: $\beta = 0$ [(light) beige], $\beta = 2$ [(medium) orange], and $\beta = 4$ [(dark) black]. We note that there is substantial overlap in the $\beta = 0$ and $\beta = 2$, resulting in distributions that are difficult to distinguish in many ways. While $\beta = 0$ has many samples at high C_μ, reflecting the large number of five-state ϵ-machines, increasing to $\beta = 2$ results in noticeable bands in the h_μ-C_μ plane and peaks at $C_\mu = \log_2 1, C_\mu = \log_2 2, C_\mu = \log_2 3$ bits, and so on. This reflects the fact that larger β makes smaller machines more likely. As a consequence, the emergence of patterns due to one-, two-, and three-state topologies is seen. Setting $\beta = 4$ shows a stronger *a priori* preference for one- and two-state machines, reflected by the strong peaks at $C_\mu = 0$ bits and $C_\mu = 1$ bit. Again, the dark lines and curves in the h_μ-C_μ are created by repeated sampling of the single-state IID process, resulting in the $C_\mu = 0$ line, and various two-state machines (compare with Fig. 5). Interestingly, the prior distribution over h_μ and C_μ is quite similar for $\beta = 0$ and 2, with more distributional structure due to smaller ϵ-machines at $\beta = 2$. However, the prior distribution for h_μ and C_μ differs markedly for $\beta = 4$, creating a strong preference for one- and two-state topologies. This results in an *a priori* preference for low C_μ and high h_μ that, as we demonstrate shortly, is modified for moderate amounts of data. We employ $\beta = 4$ as a reasonable value in all subsequent examples. In practice, sensitivity to this choice should be tested in each application to verify that the resulting behavior is appropriate. We suggest small, nonzero values as reasonable starting points. As always, sufficient data make the choice relatively unimportant for the resulting inference.

2. *Markov example: The Golden Mean Process*

The first example of structural inference explores the Golden Mean Process, pictured in Fig. 7. Although it is illustrated as an HMM in the figure, it is effectively a Markov chain with no hidden states: observing a 1 corresponds to state A, whereas observing 0 means the process is in state B. Previously, we showed that this data source can be inferred using the model class of kth order Markov chains, as expected [36]. However, the Golden Mean Process is also a member of the class of binary-alphabet, topological ϵ-machines considered here. As a result, structural inference from golden mean data is an example of in-class modeling.

We proceed using the approach laid out above for the Even Process transition probabilities and start states. We generated a single data series by randomizing the start state and creating a symbol sequence of length $T = 2^{17}$ using the Golden Mean Process ϵ-machine. As above, we monitor the convergence using subsamples $D_{0:L} = x_0 x_1 \ldots x_{L-1}$ for lengths $L = 2^i$, $i = 0,1,\ldots 17$. The candidate machines $\boldsymbol{\mathcal{M}}$ consist of all

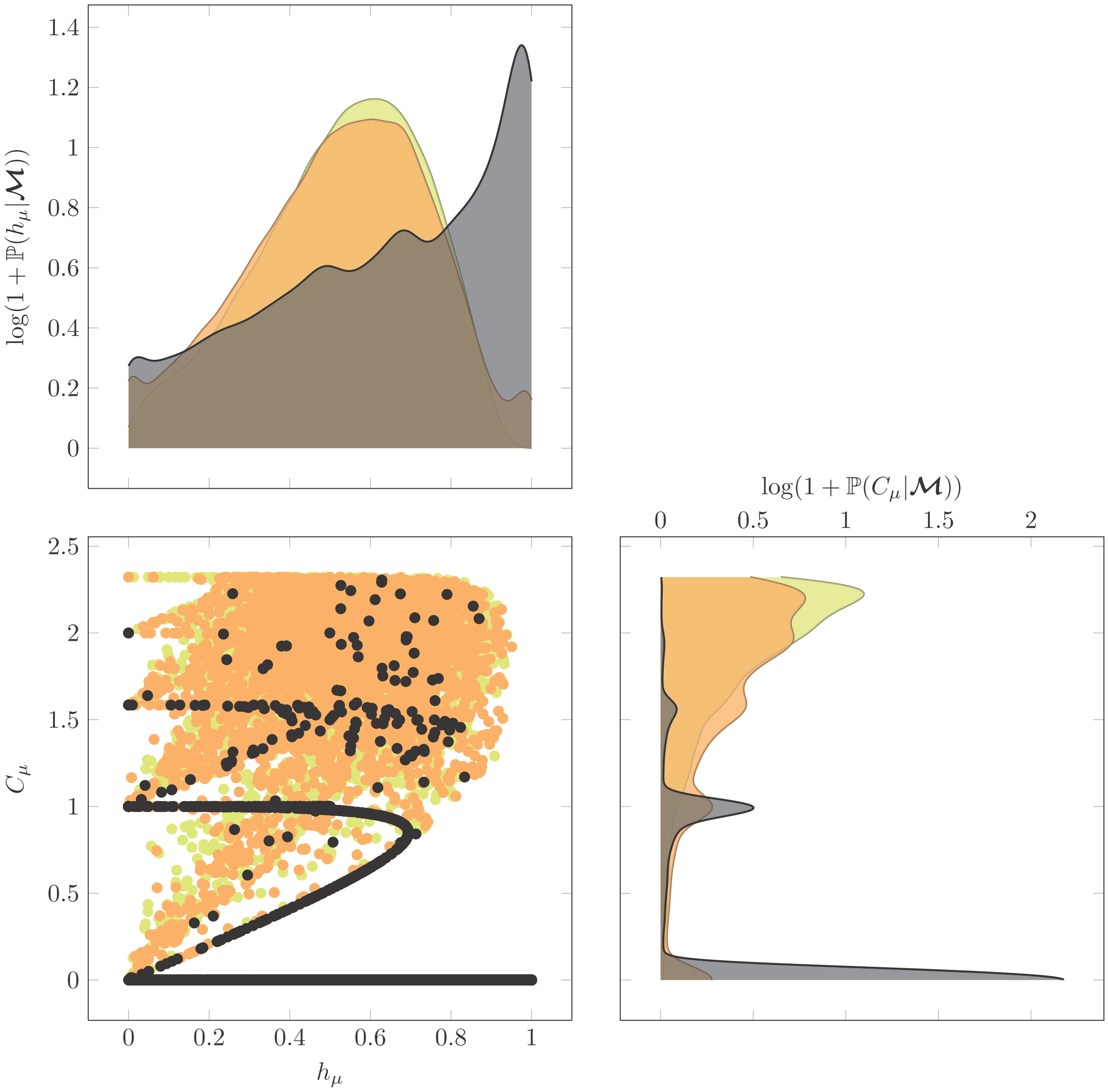

FIG. 6. (Color online) Model prior dependence on penalty parameter β: Fifty thousand samples were taken from the joint prior $\mathbb{P}(h_\mu, C_\mu|\mathcal{M})$ using all binary-alphabet, topological ϵ-machines with one to five states and parameters: $\beta = 0$ [(light) beige], $\beta = 2$ [(medium) orange], and $\beta = 4$ [(dark) black]. (Lower left) A subsample of size 5000 from the joint distribution is shown for each value of β. A Gaussian kernel density estimation, using all 50 000 samples for each value of β, of the marginal distributions $\mathbb{P}(h_\mu|\mathcal{M})$ (top) and $\mathbb{P}(C_\mu|\mathcal{M})$ (right) are also provided. Visible arcs and lines in the h_μ-C_μ plane for $\beta = 4$ are created by high prior probability of the single-state topology, creating the $C_\mu = 0$ line, and two-state models, creating the parabolic arc. For values $\beta = 0$ and 2, many model topologies are sampled creating overlapping and diffuse clouds of samples at higher C_μ and moderate h_μ. In these cases, no single model topology is sufficiently resampled to create the distinctive curves visible for $\beta = 4$. A natural logarithm is used in plotting probability densities.

36 600 ϵ-machine topologies in Table I. Estimating h_μ and C_μ aids in monitoring convergence of inferred topology and related properties to the correct values. In addition, we provide supplementary tables and figures, using both $\mathcal{M}$ and the *maximum a posteriori* model M_{MAP} at each data length L to give a detailed view of structural inference (See Supplemental Material [42]).

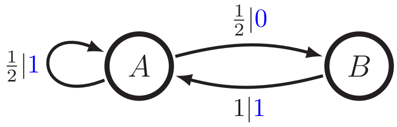

FIG. 7. (Color online) Golden Mean Process's ϵ-machine.

Figure 8 plots samples from the joint posterior over (h_μ, C_μ), as well as their marginal distributions, for three subsample lengths. As in Fig. 5, we consider $L = 1$ [(light) beige], $L = 64$ [(medium) orange], and $L = 16\,384$ [(dark) black]. However, this example employs the full set $\mathcal{M}$ of candidate topologies. For small data size ($L = 1$) the distribution closely approximates the prior distribution for $\beta = 4$, as it should. At data size $L = 64$, the samples of both

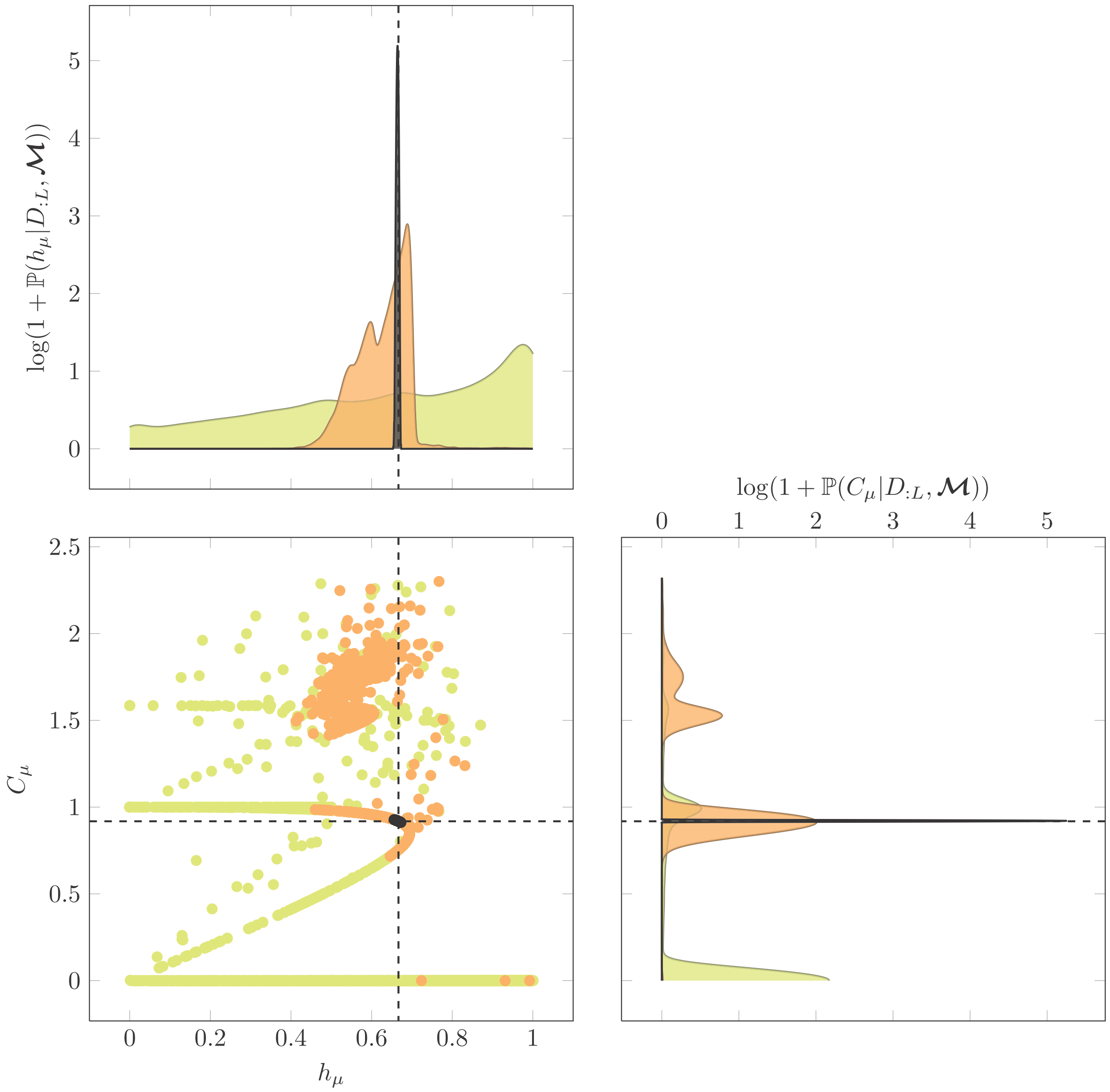

FIG. 8. (Color online) Convergence of randomness (h_μ) and structure (C_μ) calculated with model topologies, transition probabilities, and start states estimated from Golden Mean Process data, using all one- to five-state topological ϵ-machines. Fifty thousand samples were taken from the joint posterior distribution $\mathbb{P}(h_\mu, C_\mu | D_{0:L}, \boldsymbol{\mathcal{M}})$. (Lower left) Subsample of size 5000 for data sizes $L = 1$ [(light) beige], $L = 64$ [(medium) orange], and $L = 16\,384$ [(dark) black]. Gaussian kernel density estimates of the marginal distributions (using all 50 000 samples) $\mathbb{P}(h_\mu | D_{0:L}, \boldsymbol{\mathcal{M}})$ (top) and $\mathbb{P}(C_\mu | D_{:L}, \boldsymbol{\mathcal{M}})$ (right) are also provided for the same values of L. Dashed lines indicate the true values of h_μ and C_μ for the Golden Mean Process. As discussed in Figs. 5 and 6, the distinctive arcs visible for $L = 1$ are due to repeated samples of single-state and two-state model topologies at small L. A natural logarithm is used in plotting probability densities.

the h_μ and C_μ are still broad, resulting in multimodal behavior with considerable weight given to both two- and three-state topologies. Consulting Table S2 in the Supplemental Material, we see that this is the shortest length that selects the correct topology for the Golden Mean Process (denoted n2k2id5 in Table S2). For smaller L, the single-state, two-edge topology is preferred (denoted n1k2id3). However, the probability of the correct model is only 78.7%, leaving a substantial probability for alternative candidates. The uncertainty is further reflected in the large credible interval for C_μ provided by the complete set of models $\boldsymbol{\mathcal{M}}$ (see Table S1), ranging from 0.8235 bits as the lower bound to 1.797 bits as the upper bound. However, by subsample length $L = 16\,384$ the probability of the correct topology is 99.998%, given the set of candidate machines $\boldsymbol{\mathcal{M}}$, and estimates of both h_μ and C_μ have converged to accurately reflect the correct values.

In addition to Tables S1 and S2, the Supplemental Material provides Fig. S1, showing the Gkde estimates of both h_μ and C_μ using $\boldsymbol{\mathcal{M}}$ and M_{MAP} as a function of subsample length. The four panels clearly show the convergence of estimates to the correct values as L increases. For long data series, there is little difference between the inference made using the *maximum a posteriori* (MAP) model and the posterior over the entire candidate set. However, this is not true for short time series, where using the full set more accurately captures the uncertainty in estimation of the information-theoretic quantities of interest. We note that the C_μ estimates approach the true value from below, preferring small topologies when there is little data and selecting the correct, larger topology only as available data increases. This desired behavior results from setting $\beta = 4$ for the prior over $\boldsymbol{\mathcal{M}}$. Setting $\beta = 2$, shown in Fig. S2, does not have this effect. This value of β is insufficient to overcome the large number of three-, four-, and five-state ϵ-machines. Finally, Fig. S3 plots samples from the joint posterior of h_μ and C_μ using only the MAP model for subsample lengths $L = 1,64$, and 16 384. This should be compared with Fig. 8 where the complete set $\boldsymbol{\mathcal{M}}$ is used. Again, there is a substantial difference for short data series and much in common for larger L.

Before moving to the next example, let us briefly return to consider start-state inference. The data series generated to test inferring the Golden Mean Process started with the sequence $D_{0:T} = 1110\ldots$. We note that the correct start state, which happens to be state A in that realization, cannot be inferred and has lower probability than state B due to the process's structure, $\mathbb{P}(\sigma_{\mathrm{gm},0} = A|D_{0:T} = 1110\ldots,M_{\mathrm{gm}}) \approx 0.3328$, using Eq. (13). The reason for the inability to discern the start state is straightforward. Consulting Fig. 7, we can see that the string 1110 can be produced beginning in both states A and B. On the one hand, assuming $\sigma_{\mathrm{gm},0} = A$, the state path would be $AAAAB$ with probability $p(1|A)^3 p(0|A) = (1/2)^4$. On the other hand, assuming $\sigma_{\mathrm{gm},0} = B$, the state path is $BAAAB$ with probability $p(1|B)p(1|A)^2 p(0|A) = 1 \times (1/2)^3$. The only difference in the probabilities is a factor of $p(1|A) = 1/2$ versus $p(1|B) = 1$ resulting in the following:

$$\mathbb{P}(\sigma_{i,0} = A|D = 1110,M_{\mathrm{gm}}) = \frac{(1/2)^4}{(1/2)^4 + (1/2)^3}$$
$$= 1/3.$$

This calculation agrees nicely with the result stated above, using finite data and the inference calculations from Eq. (13).

It turns out that any observed data series from the Golden Mean Process that begins with a 1 will have this ambiguity in start state. However, observed sequences that begin with a 0 uniquely identify A as the start state since a 0 is not allowed leaving state B. Despite this, the correct topology is inferred and accurate estimates of h_μ and C_μ are obtained.

3. Infinite-order Markov example: The Even Process

Next we consider inferring the structure of the Even Process using the same set of binary-alphabet, one- to five-state, topological ϵ-machines. To be clear, this example differs from Sec. V A 1, where the correct topology was assumed. Now we explore Even Process structure using $\boldsymbol{\mathcal{M}}$. As noted above, the Even Process is an infinite-order Markov process and inference requires the set of topological ϵ-machines considered here. (However, see out-of-class inference of the Even Process using kth-order Markov chains in Ref. [36].) As a result, this is an example of in-class inference since the Even Process topology is contained within the set $\boldsymbol{\mathcal{M}}$. As with the previous example, a single data series was generated from the Even Process.

Figure 9 shows samples from the posterior distribution over (h_μ, C_μ) using three subsample lengths $L = 1,64$, and 16 384 as before. An equivalent plot using only the MAP model is provided in the Supplemental Material for comparison; see Fig. S6. Again, for short data series the samples mirror the prior distribution as they should. [See (light) beige points for $L = 1$.] At subsample length $L = 64$ the values of h_μ and C_μ are much more tightly delineated. Comparing samples for the Golden Mean Process in Fig. 8 shows that there is much less uncertainty in structure for the Even Process at this data size. Consulting Table S4, the MAP topology for this value of L already identifies the correct topology (denoted n2k2id7) and assigns a probability of 99.41%. This high probability is reflected by the smaller spread, when compared with the golden mean example, of the samples of h_μ and C_μ. At subsample length $L = 16\,384$ the probability of the correct topology has grown to 99.998%. Estimates of both h_μ and C_μ are also very accurate, with small uncertainties, at this L; see Table S3.

The Supplemental Material provides Figs. S4 and S5 to show the convergence of the posterior densities for h_μ and C_μ as a function of subsample length. Figure S4 shows estimates using both $\boldsymbol{\mathcal{M}}$ and M_{MAP} for $\beta = 4$. Whereas Fig. S5 demonstrates the effects of using a small penalty ($\beta = 2$) for model size. As seen with the Golden Mean Process, the difference is most apparent at small data sizes. At large L, the difference between using the complete set $\boldsymbol{\mathcal{M}}$ of models versus the MAP model is minor, as is the effect of choosing $\beta = 4$ or $\beta = 2$. However, at small data sizes the choices affect the resulting inference. In particular, the choice of $\beta = 4$ allows the inference machinery to approach the correct C_μ from below, whereas the choice of $\beta = 2$ approaches C_μ from above; see Figs. S4 and S5. This behavior, which we believe is desirable, is similar to the inference dynamics observed for the Golden Mean Process, further strengthening the apparent suitability of using $\beta = 4$.

Unlike the previous example, the start state for the correct structure is inferred with little data. In this example, the data series begins with the symbols $D_{0:T} = 10\ldots$, which can be generated only from state B. So, at $L = 2$ the start state for the correct topology is determined, but it takes more data—32 symbols in this case—for this structure to become the most probable in the set considered.

4. Out-of-class structural inference: The Simple Nonunifilar Source

The SNS is our final and most challenging example of structural inference due its being out-of-class. The SNS is not only infinite-order Markov; any unifilar presentation requires an infinite number of states. In particular, its ϵ-machine, the minimal unifilar presentation, has a countable infinity of causal states [43]. We can see the difference between the SNS and previous processes by inspecting state A, where both out-going edges emit a symbol "1." (See Fig. 10 for a hidden Markov model presentation that is not an ϵ-machine.) This makes the SNS a nonunifilar topology, as the name suggests. Importantly, even if we assume a start state, there is no longer a single, unique path through the hidden states for an observed output data series. This completely differs from the unifilar examples previously considered, where an assumed start state and observed data series either determined a unique path through hidden states or was disallowed. As a result, the inference tools developed here cannot use the HMM topology

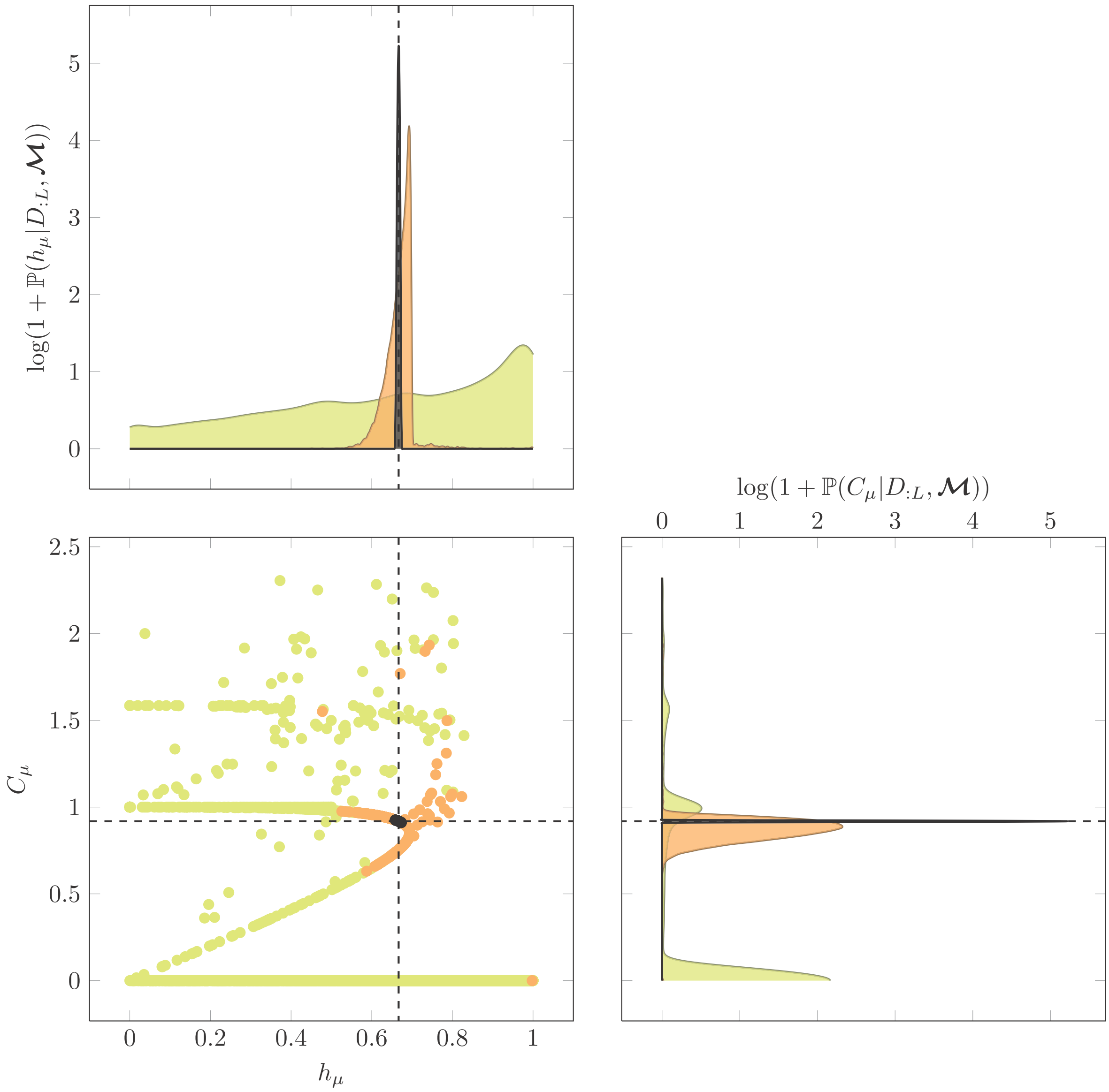

FIG. 9. (Color online) Convergence of randomness (h_μ) and structure (C_μ) calculated with model topologies, transition probabilities, and start states estimated from Even Process data, using all one- to five-state topological ϵ-machines. Fifty thousand samples from the joint posterior $\mathbb{P}(h_\mu, C_\mu|D_{:L}, \boldsymbol{\mathcal{M}})$. (Lower left) A subsample of 5000 for data sizes $L = 1$ [(light) beige], $L = 64$ [(medium) orange], and $L = 16\,384$ [(dark) black]. Gaussian kernel density estimates of the marginal distributions (using all 50 000 samples) $\mathbb{P}(h_\mu|D_{:L}, \boldsymbol{\mathcal{M}})$ (top) and $\mathbb{P}(C_\mu|D_{0:L}, \boldsymbol{\mathcal{M}})$ (right) are shown for the same values of L. Dashed lines indicate the true values of h_μ and C_μ for the Even Process. Consult captions of previous figures for a discussion of visible arcs of samples in the h_μ-C_μ plane. A natural logarithm is used in plotting probability densities.

of Fig. 10. Concretely, this class of representation breaks our method for counting transitions.

Our goal, though, is to use the set of unifilar, topological ϵ-machines at our disposal to infer properties of the Simple Nonunifilar Source. (One reason to do this is that unifilar models are *required* to calculate h_μ.) Typical data series generated by the SNS model are accepted by many of the unifilar topologies in $\boldsymbol{\mathcal{M}}$ and a posterior distribution over these models can be calculated. As with previous examples, we demonstrate estimating h_μ and C_μ for the data source. Due to the nonunifilar nature of the source, we expect C_μ estimates to increase with the size of the available data series. However, the ability to estimate h_μ accurately is unclear *a priori*. Of course, in this example we cannot find the correct model topology because infinite structures are not contained in $\boldsymbol{\mathcal{M}}$.

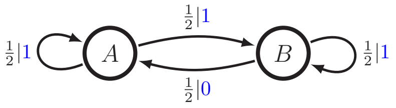

FIG. 10. (Color online) The Simple Nonunifilar Source.

Figure 11 presents the joint posterior for (h_μ, C_μ) for three subsample lengths. As previously, a single data series of length 2^{17} is generated using the SNS and analysis of subsamples $D_{0:L}$ are employed to demonstrate convergence. The short subsample ($L = 1$, light/beige points) is predictably uninteresting, reflecting the the prior distribution over models. For

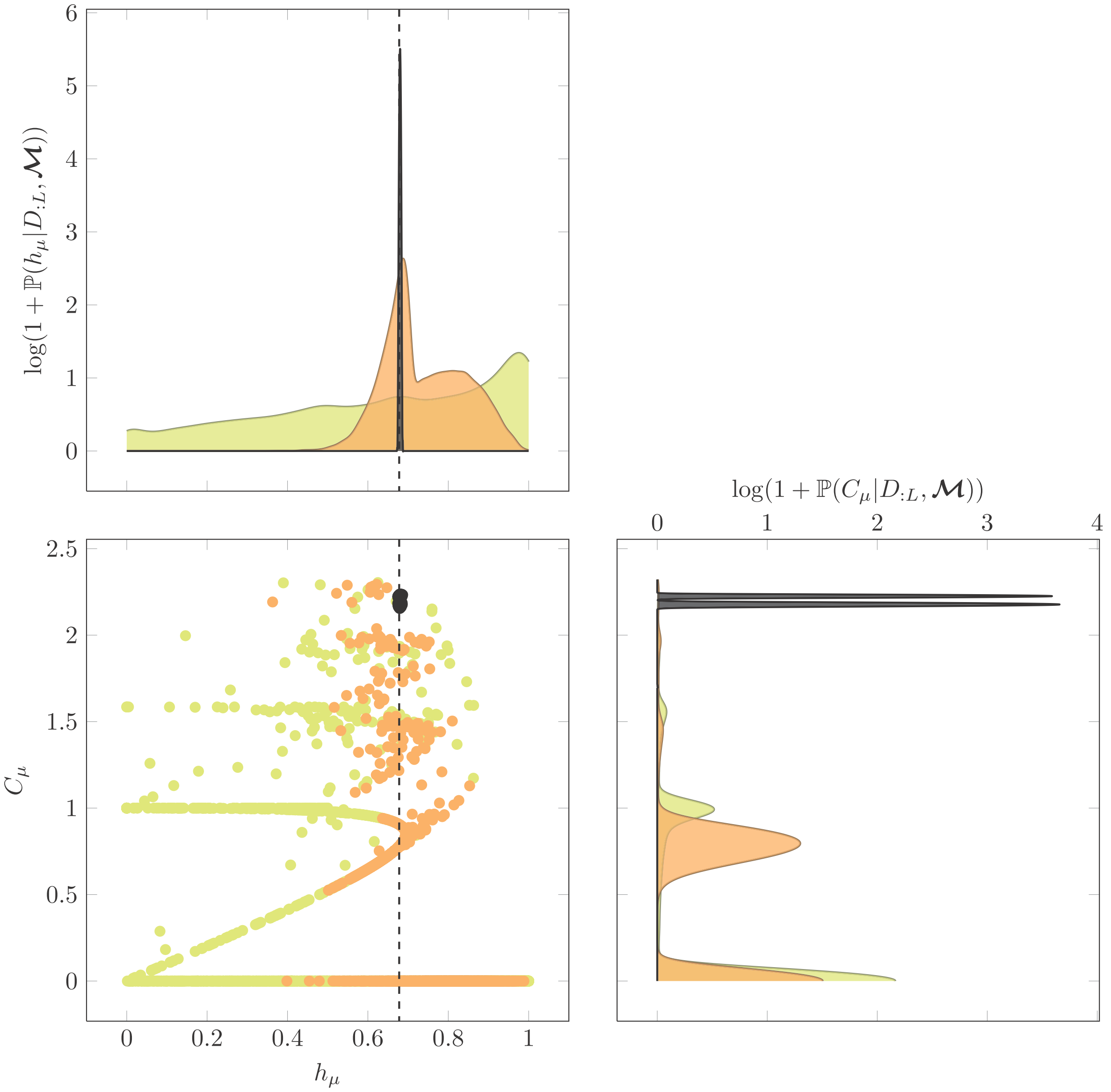

FIG. 11. (Color online) Convergence of randomness (h_μ) and structure (C_μ) calculated with model topologies, transition probabilities, and start states estimated from Simple Nonunifilar Source data, using all one- to five-state topological ϵ-machines. Sample sizes, colors and line types mirror those in previous figures. See previous figure captions for a discussion of visible lines and arcs in samples from the h_μ-C_μ plane. A natural logarithm is used in plotting probability densities.

subsamples shorter than $L = 64$ the MAP model is the single-state, two-edge topology. (Denoted n1k2id3 in Table S6.) At $L = 64$ the Golden Mean Process topology becomes most probable with a posterior probability of 53.01%. The probability of the single-state topology is still 43.98%, though, resulting in C_μ's strongly bimodal marginal posterior observed for $L = 64$. [See Fig. 11, (medium) orange points, right panel.] Bimodality also appears in the marginal posterior for h_μ, with the largest peak coming from the two-state topology and the high entropy rates being contributed by the single-state model. At large data size [$L = 16\,384$ (dark) black points] h_μ has converged on the true value, while C_μ has sharp, bimodal peaks due to many nearly equally probable five-state topologies. Consulting Table S6, we see that the MAP structure for this value of L has five states (denoted n5k2id22979) and a low posterior probability of only 8.63%. Further investigation reveals that there are four additional ϵ-machine topologies (making a total of five) with similar posterior probability. These general details persist for longer subsamples sequences including the complete data series at length 2^{17}. Although estimating h_μ converges smoothly, the inference of structure as reflected by C_μ does not show signs of graceful convergence.

We provide supplementary plots in Figs. S7 and S8 that show the convergence of h_μ and C_μ using $\boldsymbol{\mathcal{M}}$ and M_{MAP} for prior parameters $\beta = 4$ and $\beta = 2$, respectively. Again, the choice of β matters most at small data sizes. While the C_μ estimate increases as function of L for $\beta = 4$, the use of $\beta = 2$ results in posterior means for C_μ that first decrease as function of L and then increase. Again, this supports the use of $\beta = 4$ for this set of binary-alphabet, topological ϵ-machines.

The need to employ the complete model set $\mathcal{M}$ versus the MAP topology is most evident at small data sizes; as was also seen in previous examples. However, the C_μ inference in this example is more complicated due to the large number of five-state topologies with roughly equal probability. The MAP method selects just one model, of course, and so cannot represent the posterior distribution's bimodal behavior. Given that the data source is out-of-class, this trouble is perhaps not surprising. Figure S9 shows samples from the joint posterior of (h_μ, C_μ) using only the MAP topology. Using the latter also suffers from requiring one to select a single exemplar topology for a posterior distribution that is simply not well represented by a single ϵ-machine.

VI. DISCUSSION

The examples demonstrated structural inference using the set of one- to five-state, binary-alphabet, topological ϵ-machines. We found that in-class examples, including the Golden Mean and Even Processes, were effectively and efficiently discovered. That is, the correct topology was accorded the largest posterior probability and estimates of information coordinates h_μ and C_μ were accurate. However, we found that a sufficiently large value of β, providing the model size penalty, was key to a conservative structural inference. Conservative means that C_μ estimates approach the true value from below, effectively counteracting the increasing number of topologies with larger state sets. For the out-of-class example, given by the Simple Nonunifilar Source, these broader patterns held true. However, structure could not be captured as reflected in the increasing number of states inferred as a function of data length. Also, many topologies had relevant posterior probability for the SNS data, reflecting a lack of consensus and a large degeneracy with regard to structure. This resulted in a multimodal posterior distribution for C_μ and a MAP model with very low posterior probability.

One of the surprises was the number of *accepting* topologies for a given data set. By this we mean the number of candidate structures for which the data series of interest has a valid path through hidden states, resulting in nonzero posterior probability. In many ways, this aspect of structural inference mirrors grammatical inference for deterministic finite automaton [44,45]. As one might expect given this comparison, the key property that determines if a data series is accepted by a specific model topology is the support: the set of strings or words that have nonzero probability. If the support of a candidate model is the same or larger than the support of the data source, the data will be "accepted" and the model will have nonzero posterior probability. For topological ϵ-machines this also means that there will be at least one model with nonzero posterior probability: the IID process, which has full support and accepts any sequence of 0's and 1's. Expanding the set of candidate model topologies to include all uHMMs, as we will do in a sequel, considerably increases the number of model topologies with full support. This will guarantee a substantial set of viable candidate models with nonzero posterior probability for any data series.

In the Supplemental Material we provide plots for the three processes considered above showing the number of accepting topologies in the set of one- to five-state ϵ-machines used for $\mathcal{M}$. (See Fig. S10 in the Supplemental Material.) For all of these topologies, a rapid decline in the number of accepting topologies occurs for the first 2^6 to 2^7 symbols, followed by a plateau at a set of accepting topologies. For smaller topologies, which come from the model class under consideration, this pattern makes sense. Often, the smaller topology is embedded within a larger set of states, some of which are never used. For out-of-class examples like the SNS this behavior is less transparent. The rejection of a data series by a given topology provides a first level of filtering by assigning zero posterior probability to the structure due to vanishing likelihood of the data given the model. For the examples given above, of the 36 660 possible topologies, 6225 accepted golden mean data, 3813 topologies accepted Even Process data, and 6225 accepted SNS data when the full data series was considered.

In all of the examples the data sources were stationary, so statistics did not change over the course of the data series. This is important because stationarity is built into the model class definition employed: the model topology and transition probabilities did not depend on time. However, given a general data series with unknown properties, it is unwise to assume stationarity holds. How can this be probed? One method is to subdivide the data into overlapping segments of equal length. Given these, inference using $\mathcal{M}$ or M_{MAP} should return similar results for each segment. For in-class data sources like the even and Golden Mean Processes, the true model should be returned for each data subsegment. For out-of-class, but stationary, models like the Simple Nonunifilar Source, the true topology cannot be returned, but a consistent model within $\mathcal{M}$ should be returned for each data segment.

However, one form of relatively simple nonstationarity—a structural change-point problem such as switching between the Golden Mean and Even Processes—can be detected by BSI applied to subsegments. The inferred topology for early segments returns the golden mean topology and later segments return the even topology. Notably, the inferred topology using all of the data or a subsegment overlapping the switch returns a more complicated model topology reflecting both structures. Of course, detection of this behavior requires sufficient data and slow switching between data sources.

In a sequel we compare BSI to alternative structural inference methods. The range of and differences with these is large and so a comparison demands its own venue. Also, the sequel addresses expanding the model candidates beyond the set of topological ϵ-machines to the full set of unifilar hidden Markov models. This change is a necessary step before useful comparisons between methods can be explored.

VII. CONCLUSION

We demonstrated effective and efficient inference of topological ϵ-machines using a library of candidate structures and the tools of Bayesian inference. Several avenues for further development are immediately obvious. First, as just noted, using full unrestricted ϵ-machines—allowing models outside the set of topological ϵ-machines—is straightforward. This will provide a broad array of candidates within the more general class of unifilar hidden Markov models. In the present setting, by way of contrast, processes with full support (all words allowed) can map only to the single-state topology.

Second, refining the eminently parallelizable Bayesian structural inference algorithms will allow them to take advantage of large compute clusters and cloud computing to dramatically expand the number of candidate topologies considered. For comparison, the current implementation uses nonoptimized PYTHON on a single thread. This configuration (running on contemporary Linux compute node) takes between 0.6 and 1.6 h, depending on the number of accepting topologies, to calculate the posterior distribution over the 36 660 candidates for a data series of length 2^{17}. An additional 10 to 20 min is needed to generate the 50 000 samples from the posterior to estimate functions of model parameters, like h_μ and C_μ.

We note that the methods of Bayesian structural inference can be applied to any set of unifilar hidden Markov models and, moreover, they do not have to employ a large, enumerated library. For example, a small set of candidate 50-state topologies could be compared for a given data series. This ability opens the door to automated methods for generating candidate structures. Of course, as always, one must keep in mind that all inferences are then conditioned on the, possibly limited or inappropriate, set of model topologies chosen.

Finally, let us return to the scientific and engineering problem areas cited in the introduction that motivated structural inference in the first place. Generally, Bayesian structural inference will find application in fields, such as those mentioned, that rely on finite-order Markov chains or the broader class of (nonunifilar) hidden Markov models. It will also find application in areas requiring accurate estimates of various system statistics. The model class considered here (ϵ-machines) consists of a novel set of topologies and usefully allows one to estimate both randomness and structure using two of the most basic informational measures, namely h_μ and C_μ. As a result, we expect Bayesian structural inference to find an array of applications in bioinformatics, linguistics, and dynamical systems.

ACKNOWLEDGMENTS

The authors thank Ryan James and Chris Ellison for helpful comments and implementation advice. Partial support was provided by ARO Grants No. W911NF-12-1-0234 and No. W911NF-12-1-0288.

[1] J. P. Crutchfield and B. S. McNamara, Complex Syst. **1**, 417 (1987).

[2] E. L. Ionides, C. Bretó, and A. A. King, Proc. Natl. Acad. Sci. USA **103**, 18438 (2006).

[3] T. Toni and M. P. H. Stumpf, Bioinformatics **26**, 104 (2009).

[4] M. Sunnåker, E. Zamora-Sillero, A. L. García de Lomana, F. Rudroff, U. Sauer, J. Stelling, and A. Wagner, Bioinformatics **30**, 221 (2013).

[5] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, PLoS Comput Biol **3**, e189 (2007).

[6] M. K. Transtrum, B. B. Machta, and J. P. Sethna, Phys. Rev. Lett. **104**, 060201 (2010).

[7] B.-J. Yoon, Curr. Genomics **10**, 402 (2009).

[8] L. Narlikar, N. Mehta, S. Galande, and M. Arjunwadkar, Nucl. Acids Res. **41**, 1416 (2012).

[9] R. L. Davidchack, Y.-C. Lai, E. M. Bollt, and M. Dhamala, Phys. Rev. E **61**, 1353 (2000).

[10] C. S. Daw, C. E. A. Finney, and E. R. Tracy, Rev. Sci. Instrum. **74**, 915 (2003).

[11] M. B. Kennel and M. Buhl, Phys. Rev. Lett. **91**, 084102 (2003).

[12] C. C. Strelioff and J. P. Crutchfield, CHAOS **17**, 043127 (2007).

[13] R. P. N. Rao, N. Yadav, M. N. Vahia, H. Joglekar, R. Adhikari, and I. Mahadevan, Proc. Natl. Acad. Sci. USA **106**, 13685 (2009).

[14] R. Lee, P. Jonathan, and P. Ziman, Proc. R. Soc. A **466**, 2545 (2010).

[15] D. Kelly, M. Dillingham, A. Hudson, and K. Wiesner, PLoS ONE **7**, e29703 (2012).

[16] C.-B. Li, H. Yang, and T. Komatsuzaki, Proc. Natl. Acad. Sci. USA **105**, 536 (2008).

[17] P. Graben, J. D. Saddy, M. Schlesewsky, and J. Kurths, Phys. Rev. E **62**, 5518 (2000).

[18] R. Haslinger, K. L. Klinkner, and C. R. Shalizi, Neural Comput. **22**, 121 (2010).

[19] D. P. Varn, G. S. Canright, and J. P. Crutchfield, Acta. Cryst. Sec. B **63**, 169 (2007).

[20] D. P. Varn, G. S. Canright, and J. P. Crutchfield, Acta. Cryst. Sec. A **69**, 197 (2013).

[21] J. P. Crutchfield, Nat. Phys. **8**, 17 (2011).

[22] J. P. Crutchfield and K. Young, Phys. Rev. Let. **63**, 105 (1989).

[23] D. P. Varn, G. S. Canright, and J. P. Crutchfield, Phys. Rev. B **66**, 174110 (2002).

[24] C. R. Shalizi, K. L. Shalizi, and J. P. Crutchfield, Santa Fe Institute Working Paper 02-10-060, 2002.

[25] C. R. Shalizi, K. L. Shalizi, and R. Haslinger, Phys. Rev. Lett. **93**, 118701 (2004).

[26] B. D. Johnson, J. P. Crutchfield, C. J. Ellison, and C. S. McTague, SFI Working Paper 10-11-027, 2012.

[27] E. Ott, *Chaos in Dynamical Systems* (Cambridge University Press, New York, 1993).

[28] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (Addison-Wesley, Reading, MA, 1994).

[29] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding* (Cambridge University Press, New York, 1995).

[30] A. B. Gelman, J. S. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis* (Chapman & Hall, CRC, 1995).

[31] J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains*, 2nd ed. (Springer-Verlag, New York, 1976).

[32] C. R. Shalizi and J. P. Crutchfield, J. Stat. Phys. **104**, 817 (2001).

[33] N. Travers and J. P. Crutchfield, J. Stat. Phys. **145**, 1181 (2011).

[34] N. Travers and J. P. Crutchfield, J. Stat. Phys. **145**, 1202 (2011).

[35] N. Travers and J. P. Crutchfield, SFI Working Paper 11-11-051, 2011.

[36] C. C. Strelioff, J. P. Crutchfield, and A. W. Hübler, Phys. Rev. E **76**, 011106 (2007).

[37] C. J. Ellison, J. R. Mahoney, and J. P. Crutchfield, J. Stat. Phys. **136**, 1005 (2009).

[38] L. Rabiner, Proc. IEEE **77**, 257 (1989).

[39] S. S. Wilks, *Mathematical Statistics* (John Wiley & Sons, Inc., New York, 1962).

[40] D. H. Wolpert and D. R. Wolf, Phys. Rev. E **52**, 6841 (1995).

[41] L. Yuan and H. K. Kesavan, Commun. Stat. Theory Methods **26**, 139 (1997).

[42] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.89.042119 for additional plots, tables, and figures related to inference examples in the main manuscript.

[43] J. P. Crutchfield, Physica D **75**, 11 (1994).

[44] K. J. Lang, B. A. Pearlmutter, and R. A. Price, in *Grammatical Inference*, Volume 1433 of Lecture Notes Comp. Sci., edited by V. Honavar and G. Slutzki (Springer, Berlin, 1998), pp. 1–12.

[45] C. de la Higuera, Patt. Recog. **38**, 1332 (2005).