*Author postprint of manuscript published in Pattern Recognition Letters 34 (2013) 587–594.*
*https://doi.org/10.1016/j.patrec.2012.11.013*

# Picking Up the Pieces: Causal States in Noisy Data, and How to Recover Them

Gustav Eje Henter[a,*], W. Bastiaan Kleijn[a,b]

[a] *Sound and Image Processing Laboratory, School of Electrical Engineering,*
*KTH – Royal Institute of Technology, SE-100 44 Stockholm, Sweden*
[b] *School of Engineering and Computer Science, Victoria University of Wellington,*
*PO Box 600, Wellington 6140, New Zealand*

## Abstract

Automatic structure discovery is desirable in many Markov model applications where a good topology (states and transitions) is not known a priori. CSSR is an established pattern discovery algorithm for stationary and ergodic stochastic symbol sequences that learns a predictively optimal Markov representation consisting of so-called causal states. By means of a novel algebraic criterion, we prove that the causal states of a simple process disturbed by random errors frequently are too complex to be learned fully, making CSSR diverge. In fact, the causal state representation of many hidden Markov models, representing simple but noise-disturbed data, has infinite cardinality. We also report that these problems can be solved by endowing CSSR with the ability to make approximations. The resulting algorithm, robust causal states (RCS), is able to recover the underlying causal structure from data corrupted by random substitutions, as is demonstrated both theoretically and in an experiment. The algorithm has potential applications in areas such as error correction and learning stochastic grammars.

*Keywords:* computational mechanics, causal states, CSSR, hidden Markov model, HMM, learnability

## 1. Introduction

The world is full of noisy data, for which we want to create models, recognize patterns, and make predictions. Many traditional modelling approaches introduce an unobservable state which governs process dynamics, but because the state is hidden, it is difficult to propose a proper state-space. An alternative is to model the data with states defined directly by the observations. This paper considers *causal-state splitting reconstruction* (CSSR), a technique for

[*] Corresponding author. Tel: +46 8 790 7420.
*Email addresses:* gustav.henter@ee.kth.se (Gustav Eje Henter), bastiaan.kleijn@ecs.vuw.ac.nz (W. Bastiaan Kleijn)

learning such observable-state representations—functions $\varepsilon(\cdot)$ from observation sequences to Markov states—from empirical data, automatically inferring states and their structure. We show that CSSR is very sensitive to disturbances in the observations, such as random substitutions. However, we also show that with a slight change to the algorithm, more robust output is obtained, which matches the observation-to-state mapping $\varepsilon$ of the underlying, undisturbed process.

Hidden Markov models (HMMs) constitute a simple class of traditional hidden-state models, widely used for pattern recognition and prediction in practical applications. Example uses include recognition tasks in speech and video, e.g., (Woodland et al., 1995; Starner and Pentland, 1995), and bioinformatics (Eddy, 1998). Typically, the expectation-maximization algorithm is used for parameter estimation, whereafter the forward algorithm or the Viterbi algorithm can be utilized for prediction or classification of new data (Rabiner, 1989).

Standard HMM-training techniques do not address the problem of selecting a suitable state-space topology, i.e., the number of states and the permitted transitions between them. While a natural state concept may sometimes be known in advance, many situations exist where it is difficult to suggest a good state structure a priori. Numerous proposals for learning this structure from data exist, but there is no de-facto standard.

Some recent efforts have concentrated on directly modeling the observed data, letting the observations themselves define the states, e.g., (Gavaldà et al., 2006; Holmes and Isbell, Jr., 2006). This facilitates discovering structure automatically. One such technique is CSSR (Shalizi and Shalizi, 2004; Shalizi et al., 2002). CSSR converges on a specific Markov process representation known as the *causal states*, which constitutes the unique minimal optimal predictor of the process. Shalizi and Shalizi (2004) show the procedure can produce more accurate and parsimonious models than selecting the number of states through cross-validation over EM-trained HMMs. The algorithm has been used in NLP tasks such as chunking and named entity recognition (Padró and Padró, 2007), for anomaly detection (Ray, 2004), and for assessing the complexity and dynamics of various natural systems, e.g., neuronal spike-times (Klinkner et al., 2006) and the stock market (Park et al., 2007).

In this paper, we first show that causal states, though conceptually elegant, can be impossible to learn with CSSR even for simple processes, especially if noise or other disturbances are present. In particular, we present an algebraic criterion useful for showing that many HMMs are non-learnable since they have an infinite number of causal states, and provide a version of the criterion that can be checked algorithmically. The criterion establishes that the causal states of a simple two-state Markov process become impossible to learn with CSSR when random substitutions are introduced into the observations. (The disturbed observations can be represented by an HMM which satisfies the non-learnability criterion.)

Second, this paper demonstrates how the noise-sensitivity problem can be solved by allowing CSSR to make approximations. The result is the robust causal states (RCS) algorithm, which can recover the finite underlying causal structure also in the presence of defects such as random deletions, insertions, and substitutions, and demonstrates substantially improved performance over CSSR in learning the same two-state Markov process from noisy data.

The RCS algorithm shares certain characteristics with a technique called CCSA (Schmiedekamp et al., 2006), but models returned by CCSA are typically not first-order Markovian and cannot represent causal states. We anticipate models obtained through RCS can be used for detecting and correcting errors similar to the corrupted text example in Ron et al. (1996), particularly given the use of CSSR-derived models for anomaly detection (Ray, 2004).

The remainder of the paper is organized as follows: Section 2 introduces causal states and CSSR. Section 3 provides new results on processes CSSR cannot learn. Section 4 consequently describes how problems with non-learnable noisy processes can be solved. Section 5 then presents a practical example of learning a noisy process, while Section 6 discusses related work and Section 7 concludes.

## 2. Background

### 2.1. Causal-state systems

Let $X_t$ for $t \in \mathbb{Z}$ be a bi-infinite, conditionally stationary, ergodic stochastic process, each $X_t$ being a discrete random variable over a symbol alphabet $\mathcal{A}$ of finite size $k = |\mathcal{A}|$. A specific outcome $x_{-\infty}^{t}$ of the random sequence up until and including some time $t$ is termed a *history*. Similarly, a *future* is a hypothetical outcome $x_{t+1}^{\infty}$ of the process from $t+1$ onward.

Past observations give information about the future. Our beliefs about the future are specified by the joint distribution $\mathbb{P}\left(X_{t+1}^{\infty} = x_{t+1}^{\infty} \mid X_{-\infty}^{t} = x_{-\infty}^{t}\right)$ of possible futures $\left\{x_{t+1}^{\infty}\right\}$. Two histories $u$ and $v$ are considered *equivalent for prediction* if they induce the same beliefs, that is, $\mathbb{P}\left(X_{t+1}^{\infty} = f \mid X_{-\infty}^{t} = u\right) = \mathbb{P}\left(X_{t+1}^{\infty} = f \mid X_{-\infty}^{t} = v\right) \forall f$. This partitions the set of all histories into well-defined equivalence classes, each with a distinct distribution of futures. The classes are dubbed *causal states*;[1] together, they form a *causal-state system*. We may introduce the deterministic function $\varepsilon\left(\cdot\right)$ that maps histories to equivalence classes, and say that the causal state at $t$ is $s_t = \varepsilon\left(x_{-\infty}^{t}\right)$.

The history $x_{-\infty}^{t}$ at $t$ and the outcome of the next observation $X_{t+1}$ together form a new history $x_{-\infty}^{t+1}$. This history also belongs in a causal state, thus establishing a transition. We write $s_{t+1} = T\left(s_t,\, x_{t+1}\right)$ where $T$ is a *transition function*. For causal states, $T$ is deterministic; the states are then said to be *recursively calculable*. In automata theory, the states form a *deterministic machine*. (In the *nondeterministic* alternative, the state at $t$ and the observation $x_{t+1}$ do not unambiguously determine the state at $t+1$.)

The causal state constitutes a *sufficient statistic* for prediction: the index of the current state provides all information past observations $X_{-\infty}^{t}$ contain

[1] Shalizi and Shalizi (2004) indicate that "causal states," despite the name, do not have any obvious connection to traditional causality. However, this remains the standard term.

about the future, i.e., $I\left(X_{t+1}^{\infty};\, \varepsilon\left(X_{-\infty}^{t}\right)\right) = I\left(X_{t+1}^{\infty};\, X_{-\infty}^{t}\right)$. As $\varepsilon(\cdot)$ can be computed from any other sufficient statistic, the causal states are further the *minimal* sufficient statistic, the unique minimum-entropy representation capable of optimally predicting $X_{t+1}^{\infty}$ from $X_{-\infty}^{t}$.[2]

Unfortunately, being minimal does not imply being small; as shown in Section 3, causal-state representations frequently have infinite cardinality, particularly when observations are subject to disturbances. This prompted our investigation of robust and approximate causal states as given by RCS later on.

Since $\{S_t\} = \left\{\varepsilon\left(X_{-\infty}^{t}\right)\right\}$ is a Markov process (Shalizi and Crutchfield, 2001), the causal-state framework can describe any conditionally stationary, stochastic discrete process as a Markov process. However, a causal-state system is not an HMM, but a so-called a *probabilistic deterministic finite automaton* (PDFA), assuming the number of states is finite. An (infinite-duration) PDFA is a five-tuple

$$(\mathcal{A},\, Q,\, q_0,\, A,\, \delta)\,, \tag{1}$$

where $\mathcal{A}$ is a finite alphabet, $Q$ is a finite set of states, $q_0$ is an initial probability function $Q \mapsto [0,\,1]$, $A$ is a next-symbol probability function $Q \times \mathcal{A} \mapsto [0,\,1]$, and $\delta$ is a transition function $Q \times \mathcal{A} \mapsto Q$. The constraints

$$\sum_{q \in Q} q_0(q) = 1 \tag{2}$$

$$\sum_{\sigma \in \mathcal{A}} A(q, \sigma) = 1\, \forall q \in Q \tag{3}$$

must be satisfied. For stationarity, we require that $q_0$ corresponds to the stationary distribution of the random walk over $Q$ defined by $A$ and $\delta$.

PDFA and HMMs differ in that transitions in HMMs are random functions of the state, conditionally independent of the emissions, whereas in PDFA $s_{t+1}$ can always be identified from $s_t$ and the observation $x_{t+1}$.

For a finite causal-states set, the associated PDFA is the minimal PDFA (meaning it has the lowest entropy $H(S_t)$) generating the process $X_t$. Conversely, not all stationary PDFA represent causal-state systems. The four-state PDFA in Figure 1, for example, trivially has only two causal states, and its state index is not a minimal sufficient statistic for prediction.

A more formal and comprehensive treatment of causal states is provided in Shalizi and Crutchfield (2001).

### *2.2. CSSR*

Causal-state splitting reconstruction (CSSR) is an algorithm for identifying causal states in practical applications, where process statistics must be estimated

[2] The *minimal generative HMMs* of Löhr and Ay (2009) employ nondeterministic functions $\xi(\cdot)$ to map histories to state spaces smaller than the causal states, while retaining the correct conditional future distributions $\mathbb{E}_{\xi}\mathbb{P}\left(X_{t+1}^{\infty} \mid \xi(h)\right) = \mathbb{P}\left(X_{t+1}^{\infty} \mid X_{-\infty}^{t} = h\right)$. However, the corresponding hidden-state process $\left\{\xi\left(X_{-\infty}^{t}\right)\right\}$ is necessarily non-Markovian.

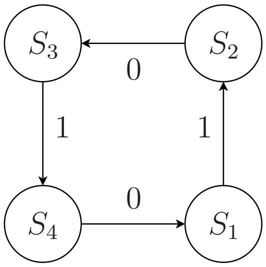

Figure 1: Four-state PDFA over $\mathcal{A} = \{0, 1\}$.

from a finite number of observations. Specifically, applying CSSR to data returns a PDFA approximating the causal states. This approximation converges in probability on the true causal states as the number of samples $N \to \infty$, if the process satisfies three conditions (Shalizi et al., 2002):

1. The process is conditionally stationary, so that $\mathbb{P}\left(X_{t+1}^{\infty}{=}f \mid X_{-\infty}^{t}{=}h\right) = \mathbb{P}\left(X_{t+1+\tau}^{\infty}{=}f \mid X_{-\infty}^{t+\tau}{=}h\right)$ for all $\tau \in \mathbb{Z}$, possible histories $h$, and futures $f$.
2. The process has a finite number of causal states.
3. Every state can be uniquely identified from observation strings of length $\Lambda$. Specifically, we require that for every causal state $s$, there exists at least one corresponding string $u$ not longer than $\Lambda$ such that $\varepsilon\left(hu\right) = s \, \forall h$;[3] thus all histories ending in $u$ belong in $s$.

Since there are no stipulations on the precise number of states or their transition structure, CSSR can perform unsupervised structure discovery.

To get statistically reliable results from finite-length data, CSSR constructs states only from short *suffixes* of the history string. A suffix is an $l$-symbol string representing the most recent observations $x_{t-l}^{t-1}$ at time $t-1$. The maximum suffix length considered by CSSR is $L$, a user-set parameter. A core CSSR output is an approximation $\widehat{\varepsilon}$ of the partitioning function $\varepsilon$, restricted to partition history-string suffixes of length $L$ or less. Despite the length limitation, CSSR can learn certain PDFA equivalent to certain processes with unbounded memory (Shalizi et al., 2002), such as the even process of Weiss (1973).

For convergence we require $L \geq \Lambda$, otherwise $\widehat{\varepsilon}$ cannot identify all causal states from the suffixes considered in the procedure. Setting $L$ too large will result in a shortage of data and unreliable results, leading to an upper bound $L \leq L\left(N\right)$ dependent on $N$, the amount of data available (Shalizi and Shalizi, 2004).

We now present the idea behind CSSR and outline algorithm operation. An in-depth description is available in Shalizi et al. (2002), while Shalizi and Shalizi (2004) provide pseudocode.

Central to CSSR is the observation that a statistic $\eta$ satisfying $I\left(X_{t+1};\, \eta\left(X_{-\infty}^{t}\right)\right) = I\left(X_{t+1};\, X_{-\infty}^{t}\right)$, meaning it can predict a single time step

[3] Strictly speaking, this need not hold for *all* $h$; a set of measure zero may be exempted.

optimally, which also is recursively calculable (has a deterministic transition function), must be a sufficient statistic for the *entire* future. To obtain a minimal sufficient statistic, CSSR first creates a minimal set of states for predicting the next symbol, and then modifies these to also have deterministic transitions. Starting from a single state containing the empty suffix, work proceeds in two stages:

1. **Homogenization**
   This step considers all suffixes of length $L$ or less, starting from the shortest ones. Suffixes are collected into preliminary *working states* based on what distribution they imply for the next symbol. As next-symbol distributions have to be estimated from the data and include random variation, distributions are compared through a statistical hypothesis test such as the $\chi^2$ test or Kolmogorov-Smirnov, at a user-specified significance level $\alpha$. Suffixes whose next-symbol distributions differ significantly from all working states form new states of their of own. The *precausal states* obtained at the end of homogenization can predict a single step into the future optimally.
   We show in Section 4 that this part of the procedure is not appropriate for corrupted data, and present a more robust homogenization criterion.
2. **Determinization**
   This step makes states recursively calculable while retaining sufficiency. This is done by splitting states that have nondeterministic transitions, i.e., states $s$ where, for some $\sigma \in \mathcal{A}$, $\widehat{\varepsilon}(u\sigma)$ does not take on the same value for all parent suffixes $u \in s$. Splitting creates one state for each possible value of $\widehat{\varepsilon}(u\sigma)$ for $u$ previously in $s$. Splits may cause other states to become nondeterministic, so the states need to be checked again, but the procedure must eventually terminate.

For estimating the next-step distributions in stage 1, we here consider a simple maximum-likelihood scheme, taking $\widehat{\mathbb{P}}\left(X_t = \sigma \mid X_{t-l}^{t-1} = u\right)$ as the frequency with which the string $u$ has been followed by a $\sigma$ in the data, and subsequently forming $\widehat{\mathbb{P}}\left(X_t = \sigma \mid X_{t-l}^{t-1} \in s\right)$ as a weighted mean over the suffixes in $s$.

Some states found by the procedure may be so-called *transient states*: states visited only finitely often with positive probability. These cannot be causal states and are therefore removed before and after determinization. In practice, the particular scheme chosen for identifying transitions between states built from suffixes of restricted length—the *closure* discussed in Klinkner and Shalizi (2005)—is of importance to remove transient states and perform determinization such that results agree with the example in Shalizi et al. (2002).

Shalizi et al. (2002) show that the worst-case computational complexity and data requirements of CSSR increase exponentially in $L$, but that complexity is polynomial in the alphabet size $k$ for $L$ fixed. By only parsing the data once and storing requisite suffix statistics in a context tree or similar structure, complexity becomes linear in the data size $N$.

## 3. Limitations of CSSR

As stated, the causal-state representations of simple (i.e., finite state-space) processes are not always simple. A particular problem is observation noise. This section demonstrates that the causal-state representation of finite state-space HMMs may be infinite, meaning they cannot be learned by CSSR, and presents a sufficient criterion for this to occur. With the criterion, for instance using the associated pseudocode, one can verify that even small PDFA may become non-learnable if observations are corrupted by noise, as such data can equivalently be described by a non-learnable HMM.

Being unable to learn simple processes like noisy PDFA limits the usefulness of CSSR, but we introduce a method to overcome the noise problem in Section 4.

These results are not necessarily surprising. While the causal-states framework can represent any conditionally stationary process, CSSR can only learn processes with a finite number of causal states, which must be representable by PDFA. It is known that many processes generated by finite state-machines such as HMMs cannot be cast as a PDFA of any size (Dupont et al., 2005), and thus require an infinite causal-state representation. However, our criterion lets us identify specific finite-state HMMs that CSSR fails to learn; see Section 5 for an example involving PDFA output corrupted by random substitutions.

### *3.1. Practical consequences*

In applications to real-world data, the number of states identified by CSSR is typically seen to diverge rapidly with increasing $L$, as longer memory provides an ever-growing set of suffixes that the algorithm can differentiate, suggestive of an infinite number of causal states. This is evident in Padró and Padró (2005), one of few publications that consider CSSR output stability over $L$. They also found that the large representations produced by CSSR at high $L$ performed worse than models from small $L$, when used in a named entity recognition task. Methods to curb the divergent growth of CSSR output are therefore of interest.

Notably, a common application of CSSR is to assess the statistical complexity of natural processes by computing the entropy of the causal states, $C_\mu = H(\varepsilon(X_t)) \approx \widehat{C}_\mu = H(\widehat{\varepsilon}(X_t))$, $\widehat{\varepsilon}(\cdot)$ being estimated from data using CSSR (Crutchfield and Shalizi, 1999; Park et al., 2007). Our results suggest that the causal-state space often is infinite. There is then no guarantee that $C_\mu$ is finite. Hence, one may wish to interpret empirically estimated $\widehat{C}_\mu$ with caution, particularly if convergence of the estimate has not been investigated. Nerukh (2008) noticed diverging estimates $\widehat{C}_\mu$ with increasing dataset size for a model of molecular dynamics.

### *3.2. An HMM non-learnability criterion*

We now restrict ourselves to the causal states of HMMs. While some HMMs have a finite and learnable causal-state representation, including the process in Section 5 as long as noise is absent, we here present a sufficient criterion for when a general finite-state HMM $H_{AB}$ cannot be represented by a finite number of causal states.

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be the transition matrix of a stationary, ergodic Markov chain $S_{-\infty}^{\infty}$ with $n < \infty$ states $\{1, 2, \ldots, n\}$, and let $\operatorname{rank} \boldsymbol{A} = r$. Let further $X_{-\infty}^{t}$ be a random sequence of symbols from a finite alphabet $\mathcal{A}$ of cardinality $k$, say $\mathcal{A} = \{1, 2, \ldots, k\}$. The distribution of $X_t$ depends only on the Markov chain state $S_t$ at $t$, and can be represented by a matrix $\boldsymbol{B} \in \mathbb{R}^{n \times k}$ with $b_{ij} = \mathbb{P}\left(X_t = j \mid S_t = i\right)$. Together, $\boldsymbol{A}$ and $\boldsymbol{B}$ define a finite-size HMM $H_{AB}$.

Knowing only previous observations, the hidden state $S_t$ can frequently not be determined with certainty. The distribution of future HMM observations, given previous observations, is then completely determined by what we *can* know about the state, specifically the probabilities of the hidden states:

$$\begin{aligned}&\mathbb{P}\left(X_{t+1}^{\infty} = f \mid X_{-\infty}^{t} = h\right) \\ &\qquad = \sum_{i=1}^{n} \mathbb{P}\left(X_{t+1}^{\infty}{=}f \mid S_t{=}i\right) \mathbb{P}\left(S_t{=}i \mid X_{-\infty}^{t}{=}h\right). \quad (4)\end{aligned}$$

As the hidden-state probability distribution $\left(\boldsymbol{p}_t\right)_i = \mathbb{P}\left(S_t = i \mid X_{-\infty}^{t} = h\right)$ formally constitutes a deterministic function of the data $h$, it forms a sufficient statistic from which the causal state may be computed. We represent these $n$ probabilities using an $n$-vector $\boldsymbol{p}_t$. However, we need not assume that $\boldsymbol{p}_t$ is normalized; we just require that $\boldsymbol{p}_t \neq \boldsymbol{0}$ has entries proportional to $\mathbb{P}\left(S_t = i \mid X_{-\infty}^{t} = h\right)$, uniquely determining the hidden-state distribution.

Differences $\delta \boldsymbol{p} = \boldsymbol{p}_t' - \boldsymbol{p}_t$ in hidden-state probabilities at $t$ that are in the null space of $\boldsymbol{A}^T$ will not affect the state distribution at time $t+1$ (since $\boldsymbol{p}_{t+1}' = \boldsymbol{A}^T \boldsymbol{p}_t' = \boldsymbol{A}^T \left(\delta \boldsymbol{p} + \boldsymbol{p}_t\right) = \boldsymbol{p}_{t+1}$), and do not influence future behavior. Hence an $r$-dimensional vector $\boldsymbol{v}_t$, a linear function of $\boldsymbol{p}_t$, suffices to identify the causal state in the row space of $\boldsymbol{A}^T$.

To define $\boldsymbol{v}_t$, let $\boldsymbol{G}_r \in \mathbb{C}^{n \times r}$ and $\boldsymbol{G}_0 \in \mathbb{C}^{n \times n-r}$ be bases for the row and null spaces of $\boldsymbol{A}^T$, and let $\boldsymbol{H}_r \in \mathbb{C}^{r \times n}$ and $\boldsymbol{H}_0 \in \mathbb{C}^{n-r \times n}$ denote the corresponding matrices that transform distribution vectors to their row and null space coordinates, so that $\left[\boldsymbol{H}_r^T \, \boldsymbol{H}_0^T\right]^T = \left[\boldsymbol{G}_r \, \boldsymbol{G}_0\right]^{-1}$. We can then write $\boldsymbol{p}_t = \boldsymbol{G}_r \boldsymbol{v}_t$.

Our non-learnability criterion considers how $\boldsymbol{v}_t$ evolves as additional observations become available. For hidden-state distribution vectors $\boldsymbol{p}_t$, the belief state, upon observing a symbol $\sigma$, is updated following the forward algorithm (Rabiner, 1989);

$$\begin{aligned}&\mathbb{P}\left(S_{t+1} = j \mid X_{-\infty}^{t+1} = h\sigma\right) \\ &\qquad \propto b_{j\sigma} \sum_{i=1}^{n} a_{ij} \mathbb{P}\left(S_t = i \mid X_{-\infty}^{t} = h\right). \quad (5)\end{aligned}$$

For $\boldsymbol{v}_t$ we can formulate a similar update $\boldsymbol{v}_{t+1} = \boldsymbol{C}_\sigma \boldsymbol{v}_t$, defining the *forward matrices* $\boldsymbol{C}_\sigma \in \mathbb{C}^{r \times r}$ by

$$\boldsymbol{C}_\sigma = \boldsymbol{H}_r \operatorname{diag}\left(\boldsymbol{b}_{\cdot\sigma}\right) \boldsymbol{A}^T \boldsymbol{G}_r \quad (6)$$

for all $\sigma$ in $\mathcal{A}$, $\boldsymbol{b}_{\cdot\sigma}$ being the $\sigma$th column of $\boldsymbol{B}$.

The forward matrices describe how our beliefs about future HMM behavior, represented by $\boldsymbol{v}_{t+1}$, are updated from previous beliefs $\boldsymbol{v}_t$ upon observing the symbol $\sigma$. The use of linear operators to represent the influence of observations on future behavior is similar to the observable operator models of Jaeger (2000).

With these definitions we can state our first main result:

**Theorem 1.** *Let* $r = \operatorname{rank} \boldsymbol{A}$*. The causal-state representation of* $H_{AB}$ *must have infinite cardinality if* $\operatorname{rank}\big(\boldsymbol{B}^T \boldsymbol{A}^T \boldsymbol{G}_r\big) = r$ *and a nonempty* $\mathcal{A}_{\text{sub}} \subseteq \mathcal{A}$ *exists that satisfies:*

1. $\boldsymbol{b}_{\cdot\sigma} > \boldsymbol{0}$ *for all* $\sigma \in \mathcal{A}_{\text{sub}}$*.*
2. *The forward matrices are nonsingular for all* $\sigma \in \mathcal{A}_{\text{sub}}$*. They can then be eigendecomposed as* $\boldsymbol{C}_\sigma = \boldsymbol{Q}_\sigma \boldsymbol{\Lambda}_\sigma \boldsymbol{Q}_\sigma^{-1}$*.*
3. *For any nonzero* $\boldsymbol{v} \in \mathbb{C}^r$ *there exists a* $\sigma \in \mathcal{A}_{\text{sub}}$ *such that* $\boldsymbol{q} = \boldsymbol{Q}_\sigma^{-1} \boldsymbol{v}$ *has more than one nonzero element, and the eigenvalues* $(\boldsymbol{\Lambda}_\sigma)_{ii}$ *corresponding to these nonzero elements* $q_i$ *of* $\boldsymbol{q}$ *do not all have the same absolute value.*

The theorem is useful to show that CSSR has trouble learning many processes, including simple PDFA disturbed by random substitutions as exemplified in Section 5. The proof is sketched below; for technical details please consult appendix A.

*Proof.* We assume that the number of causal states, $M$, is finite, and derive a contradiction. Because $\boldsymbol{G}_r \boldsymbol{v}_t$ for $\boldsymbol{v}_t \in \mathbb{C}^r \setminus \boldsymbol{0}$ spans the row space of $\boldsymbol{A}^T$, knowing $\boldsymbol{v}_t$ suffices to determine the next-symbol distribution at $t+1$ (through $\boldsymbol{p}_{t+1} = \boldsymbol{A}^T \boldsymbol{G}_r \boldsymbol{v}_t$). Moreover, because $\operatorname{rank}\big(\boldsymbol{B}^T \boldsymbol{A}^T \boldsymbol{G}_r\big) = r$, any change in direction for $\boldsymbol{v}_t$ alters the next-symbol probability distribution, so the correspondence between distinct expectations for future HMM behavior and different $\boldsymbol{v}_t$-vector directions is one to one. Not all these expectations need to be causal states, however.

Take any $\boldsymbol{v} \in \mathbb{C}^r \setminus \boldsymbol{0}$ and assume it represents beliefs that correspond to a causal state. Choose an $\sigma \in \mathcal{A}_{\text{sub}}$ according to 3, which hinges on 2. Upon observing a string of $M' \geq M$ contiguous $\sigma$-symbols—which has nonzero probability by 1—associated causal-state directions are generated by repeatedly multiplying $\boldsymbol{v}_t$ by $\boldsymbol{C}_\sigma$; $\boldsymbol{v}_{t+m} = (\boldsymbol{C}_\sigma)^m \boldsymbol{v}_t$ for $m \in \{0, 1, \ldots, M'\}$. If we consider the ratios between the absolute values of the nonzero $\boldsymbol{q}_{t+m}$-vector elements from 3 that arise, these are either strictly increasing or decreasing with $m$. Hence none of the generated $\boldsymbol{v}_{t+m}$-vectors (causal-state directions) are collinear. Since this is true for any $\boldsymbol{v} \in \mathbb{C}^r$, there exist at least $M' + 1 > M$ distinct causal states, contradicting the original assumption of a finite causal-state machine with only $M$ states. □

### *3.3. Checking non-learnability*

Item 3 in the criterion may appear complex, but is for instance satisfied if the $\boldsymbol{C}_\sigma$-matrices do not share any eigenvectors and the eigenvalues of each matrix have distinct absolute values. This slightly stronger requirement is easier

**Algorithm 1** Automatic check if $H_{AB}$ is non-learnable.

```
function canBeProvedUnlearnable($\boldsymbol{A}, \boldsymbol{B}, \mathcal{A}_{\text{sub}}$)

  if $|\mathcal{A}_{\text{sub}}| < 2$
    return false # Two symbols or more are necessary

  eigendecompose $\boldsymbol{A}^T$ to obtain $\{\lambda_i\}_{i=1}^n$ and $\{\boldsymbol{g}_i\}_{i=1}^n$
  let $r \leftarrow |\{i\}_{i=1:\lambda_i \neq 0}^n|$ # Count the nonzero eigenvalues
  let $\boldsymbol{G}_r \leftarrow [\boldsymbol{g}_i]_{i=1:\,\lambda_i \neq 0}^n$ # Concatenate nonsingular eigenvectors
  let $\boldsymbol{H}_r \leftarrow (\boldsymbol{G}_r^* \boldsymbol{G}_r)^{-1} \boldsymbol{G}_r^*$ # Moore–Penrose pseudoinverse
  if $\operatorname{rank}\big(\boldsymbol{B}^T \boldsymbol{A}^T \boldsymbol{G}_r\big) < r$
    return false # Rank requirement not satisfied

  let $\mathcal{V} \leftarrow \emptyset$
  for each $\sigma \in \mathcal{A}_{\text{sub}}$
    if $\min_i b_{i\sigma} = 0$
      return false # Point 1 not satisfied

    let $\boldsymbol{C}_\sigma \leftarrow \boldsymbol{H}_r \operatorname{diag}(\boldsymbol{b}_{\cdot\sigma})\, \boldsymbol{A}^T \boldsymbol{G}_r$
    eigendecompose $\boldsymbol{C}_\sigma$ to obtain $\{\lambda_i^{(\sigma)}\}_{i=1}^r$ and $\{\boldsymbol{\nu}^{(\sigma,\,i)}\}_{i=1}^r$
    if $0 \in \{\lambda_i^{(\sigma)}\}_{i=1}^r$
      return false # Point 2 not satisfied

    let $\Lambda^\sigma \leftarrow \emptyset$
    for each $i \in \{1,\, 2,\, \ldots,\, r\}$

      # $\boldsymbol{C}_\sigma$ must have distinct absolute eigenvalues
      if $|\lambda_i^{(\sigma)}| \in \Lambda^\sigma$
        return false # Point 3 not surely satisfied
      let $\Lambda^\sigma \leftarrow \Lambda^\sigma \cup |\lambda_i^{(\sigma)}|$

      # $\boldsymbol{C}_\sigma$-matrices should not share any eigenvectors
      let $\boldsymbol{\nu}^{(\sigma,\,i)} \leftarrow \big\|\boldsymbol{\nu}^{(\sigma,\,i)}\big\|^{-1} \boldsymbol{\nu}^{(\sigma,\,i)}$ # Normalize eigenvector
      if $\big(\boldsymbol{\nu}^{(\sigma,\,i)} \in \mathcal{V}\big) \vee \big(-\boldsymbol{\nu}^{(\sigma,\,i)} \in \mathcal{V}\big)$
        return false # Point 3 not surely satisfied
      let $\mathcal{V} \leftarrow \mathcal{V} \cup \boldsymbol{\nu}^{(\sigma,\,i)}$

    end for
  end for

  return true # Entire criterion satisfied
```

to verify in practice, and can be used to check the non-learnability criterion algorithmically. Pseudocode for such a procedure is provided in Algorithm 1.

When applying Algorithm 1, care must be taken to account for numerical precision, so that all comparisons are numerically robust. By storing eigenvalues and eigenvectors in hash tables, the computational complexity of set membership checks can be made independent of set cardinality, speeding up the procedure.

The only user choice in the algorithm is to select $\mathcal{A}_{\mathrm{sub}}$ for non-binary alphabets. One possibility is to take $\mathcal{A}_{\mathrm{sub}} = \mathcal{A}$, but additional (or even all) subsets may be investigated; finding any $\mathcal{A}_{\mathrm{sub}} \subseteq \mathcal{A}$ for which the theorem applies suffices to establish non-learnability.

## 4. Robust causal states

For disturbed PDFA data, the causal-states representation can frequently be infinite, as shown. CSSR then runs afoul of criterion 2, Section 2.2, and does not learn a stable representation. We here introduce a method capable of learning the underlying causal structure from observations corrupted by random errors like deletions, insertions, and substitutions at a limited rate.

The key improvement is to endow the homogenization step with a maximum resolution, so that fragments of the underlying causal states, which appear in applications to disturbed data, can be brought back together.

In standard CSSR, the test that clusters suffixes according to next-step behavior is intentionally very sensitive to distribution differences, and can, with enough data, eventually separate any possible set of precausal states for a given $L$, no matter how closely spaced their next-symbol distributions are.

Disturbances in the data may affect both the observed suffix frequencies and the associated empirical next-step symbol distributions. The latter effect typically causes suffixes previously in the same precausal state to separate. CSSR then puts these suffixes in different states, producing complex, fractured output after determinization. However, if disturbances are small, suffixes from the same underlying precausal state remain tightly clustered in next-step distribution space. A homogenization that does not resolve the intra-cluster differences can then recover the underlying topology.

### *4.1. Robust homogenization*

We now describe a modified learning algorithm with resolution as described above, which provably learns the underlying causal-state topology.

In this section, we let $\boldsymbol{p}$ and $\boldsymbol{q}$ be vectors on the unit $(k-1)$-simplex

$$\triangle^{k-1} = \{\boldsymbol{p} \in [0,\, 1]^k : \mathbf{1}^T\boldsymbol{p} = 1\}, \tag{7}$$

representing probability distributions over $\mathcal{A}$. The vector element $(\boldsymbol{p})_i$ gives the probability of the $i$th symbol $\sigma_i \in \mathcal{A}$.

In practice, conclusions must be based on distributions estimated from data. Let $\widehat{\boldsymbol{p}}_a$ and $\widehat{\boldsymbol{p}}_b$ be vectors on $\triangle^{k-1}$ representing estimates of the next-symbol probability distributions $\boldsymbol{p}_a$ and $\boldsymbol{p}_b$, with $n_a$ and $n_b$ being the sample sizes upon which the estimates are based. We consider comparing the estimated distributions through a statistical test at significance level $\alpha$, where the criterion for rejecting the null hypothesis $H_0\colon \boldsymbol{p}_a = \boldsymbol{p}_b$ can be written

$$d\left(\widehat{\boldsymbol{p}}_a,\, n_a,\, \widehat{\boldsymbol{p}}_b,\, n_b\right) > F_{\text{sig}}\left(\alpha\right) \tag{8}$$

for some functions $d$ and $F_{\text{sig}}$, $F_{\text{sig}}$ being nondecreasing in $\alpha$.

To limit the maximum resolution of the test we introduce a parameter $n_{\max}$ and modify the criterion into

$$d\left(\widehat{\boldsymbol{p}}_a, \min\left(n_a, n_{\max}\right), \widehat{\boldsymbol{p}}_b, \min\left(n_b, n_{\max}\right)\right) > F_{\text{sig}}\left(\alpha\right), \tag{9}$$

consequently defining a dissimilarity measure $d_m$ between distributions as

$$d_m\left(\widehat{\boldsymbol{p}}_a,\, \widehat{\boldsymbol{p}}_b\right) = d\left(\widehat{\boldsymbol{p}}_a,\, n_{\max},\, \widehat{\boldsymbol{p}}_b,\, n_{\max}\right). \tag{10}$$

We require that

1. $d_m$ is continuous and a metric on $\Delta^{k-1} \times \Delta^{k-1}$, convex in each argument, and
2. $F_{\text{sig}}$ is monotonic and continuous on $\alpha \in [0,\, 1]$ with a range that includes the range of $d_m$.

When these criteria are satisfied, we refer to homogenization using (9) as *robust homogenization.*[4]

While the distribution estimates in test (9) are based on all available samples and converge on their expected values as data becomes plentiful, the "safety margin" of the test will not shrink below a certain point. Together with determinization and transient removal as in ordinary CSSR, this forms the *robust causal states algorithm* (RCS). For $n_{\max} > (L+1)\,N$, RCS coincides with ordinary CSSR.

The two-sample Kolmogorov-Smirnov test satisfies the criteria for robust homogenization, but the $\chi^2$ statistic does not, as it does not obey the triangle inequality and thus cannot be a metric. (Nevertheless, it has still shown itself capable of structure recovery from noisy data in preliminary experiments.)

### *4.2. Recovering causal structure*

We now define a corrupted process and show that RCS can recover the underlying causal states.

[4] Alternatively, one might consider testing a revised null hypothesis $H_0\colon \left\|\boldsymbol{p}_a - \boldsymbol{p}_b\right\|_p \leq \epsilon_{\max}$ on the distance between distributions. Although this could potentially be more data-efficient, robust homogenization as proposed is straightforward to implement as it makes use of standard tests.

Let $X_t$ be a stationary, ergodic CSSR-learnable process, meaning it satisfies the criteria in Section 2.2. Fix an $L \geq \Lambda$ and let $\Sigma_X^L$ be the set of nonzero-probability strings (suffixes) of length $L$ or shorter generated by $X_t$, and similarly for $\Sigma_X^{L+1}$. We define the *distinguishability* $d_{\min}$ of $X_t$ under $d_m$ as the minimum next-step distribution dissimilarity between precausal states,

$$d_{\min} = \min_{u,\, v \in \Sigma_X^L :\, \boldsymbol{p}_u \neq \boldsymbol{p}_v} d_m\left(\boldsymbol{p}_u,\, \boldsymbol{p}_v\right). \tag{11}$$

Here $\boldsymbol{p}_u$ denotes the next-symbol probability vector of history suffix $u$, and similarly for $v$. This definition is similar to the PDFA distinguishability in Ron et al. (1998); Clark and Thollard (2004); Gavaldà et al. (2006), except that these consider multi-step distributions (distributions over finite strings), and restrict themselves to the $\ell_\infty$-norm as distance measure, whereas we allow an arbitrary metric.

We say that the process $Y_t$ is a *distinguishable corruption* of $X_t$ given $L$ if

1. it is stationary and ergodic,
2. it generates the same nonzero-probability strings of length $L+1$ or shorter as $X_t$, namely $\Sigma_X^{L+1}$, and
3. there is a bound $\widetilde{d}$ on the effect of the disturbances in $Y_t$ such that

$$d_m\left(\widetilde{\boldsymbol{p}}_u,\, \boldsymbol{p}_u\right) \leq \widetilde{d} < \frac{1}{4} d_{\min} \tag{12}$$

   holds for all $u \in \Sigma_X^L$, where $\widetilde{\boldsymbol{p}}_u$ denotes the next-symbol probability vector for $Y_t$ given history suffix $u$.

As a practically important special case, random symbol substitutions, deletions, and insertions create distinguishable corruptions, provided $\Sigma_X^{L+1}$ remains unchanged and the proportion of disturbed symbols is not too large. The concept of distinguishable corruptions thus covers a broad range of natural degradations. We also note that with greater distinguishability, more noise can be tolerated in (12).

Given the above definitions, the following theorem provides sufficient but not necessary conditions for RCS to recover the causal structure of $X_t$ using data from $Y_t$. A proof is outlined in Appendix B.

**Theorem 2.** *Consider RCS with a given hypothesis test and $n_{\max}$, thus defining $d_m$. Let $X_t$ be a stationary, ergodic CSSR-learnable process. Choose an $L \geq \Lambda$, and let $Y_t$ be a distinguishable corruption of $X_t$. Then there exists a nonempty interval $I_{\text{sig}}$ such that RCS with significance parameter $\alpha \in I_{\text{sig}}$, applied to data from $Y_t$, converges in probability on the causal-states suffix clustering of $X_t$.*

In other words, the causal structure of $X_t$ can be recovered by applying robust homogenization to data from $Y_t$, if disturbances are small compared to the causal-state separation, stabilizing the output structure in a situation where the causal-state representation can be highly sensitive to noise. The proof is provided below, except for technical details described in Appendix B.

*Proof.* Following the law of large numbers, all strings in $\Sigma_X^L$ will, with probability one, appear $n_{\max}$ times or more in the $Y_t$-data already after some finite but stochastic time, by condition 2. Homogenization applied after this time considers the same set of strings as homogenization of $X_t$ and is based on $d_m$-distances from (10) only.

Consider a test during homogenization, where $u \in \Sigma_X^L$ is compared against the collection of suffixes (working state) $V \subset \Sigma_X^L$. Assuming no previous test has made an error, all suffixes in $V$ belong to the same precausal state of $X_t$. Let $\widehat{\boldsymbol{p}}_u$ denote the empirical next-symbol probability distribution given history suffix $u$, estimated from the noisy data—as opposed to the actual next-symbol distribution $\widetilde{\boldsymbol{p}}_u$—and similarly for $V$. Since $d_m$ is a metric, we use the triangle inequality to find

$$\begin{aligned} d_m\left(\widehat{\boldsymbol{p}}_u, \widehat{\boldsymbol{q}}_V\right) \leq{}& d_m\left(\widehat{\boldsymbol{p}}_u, \widetilde{\boldsymbol{p}}_u\right) + d_m\left(\widetilde{\boldsymbol{p}}_u, \boldsymbol{p}_u\right) \\ &+ d_m\left(\boldsymbol{q}_V, \widetilde{\boldsymbol{q}}_V\right) + d_m\left(\widetilde{\boldsymbol{q}}_V, \widehat{\boldsymbol{q}}_V\right) \end{aligned} \tag{13}$$

if $u$ belongs in the same precausal state of $X_t$ as $V$, and

$$\begin{aligned} d_m\left(\widehat{\boldsymbol{p}}_u, \widehat{\boldsymbol{q}}_V\right) \geq{}& d_m\left(\boldsymbol{p}_u, \boldsymbol{q}_V\right) - d_m\left(\widehat{\boldsymbol{p}}_u, \widetilde{\boldsymbol{p}}_u\right) - d_m\left(\widetilde{\boldsymbol{p}}_u, \boldsymbol{p}_u\right) \\ &- d_m\left(\boldsymbol{q}_V, \widetilde{\boldsymbol{q}}_V\right) - d_m\left(\widetilde{\boldsymbol{q}}_V, \widehat{\boldsymbol{q}}_V\right), \end{aligned} \tag{14}$$

otherwise. Invoking convexity, distinguishability (11), and the disturbance bound (12) yields

$$d_m\left(\widehat{\boldsymbol{p}}_u, \widehat{\boldsymbol{q}}_V\right) \leq 2\widetilde{d} + d_m\left(\widehat{\boldsymbol{p}}_u, \widetilde{\boldsymbol{p}}_u\right) + d_m\left(\widetilde{\boldsymbol{q}}_V, \widehat{\boldsymbol{q}}_V\right), \tag{15}$$

if $u$ belongs to $V$, and

$$\begin{aligned} d_m\left(\widehat{\boldsymbol{p}}_u, \widehat{\boldsymbol{q}}_V\right) \geq{}& d_{\min} - 2\widetilde{d} \\ &- d_m\left(\widehat{\boldsymbol{p}}_u, \widetilde{\boldsymbol{p}}_u\right) - d_m\left(\widetilde{\boldsymbol{q}}_V, \widehat{\boldsymbol{q}}_V\right), \end{aligned} \tag{16}$$

otherwise. We see that there exists a nonempty $I_{\text{sig}}$ such that $\alpha \in I_{\text{sig}} \Rightarrow F_{\text{sig}}(\alpha) \in (2\widetilde{d}, d_{\min} - 2\widetilde{d})$. Ignoring terms due to estimation from finite data, these $\alpha$ group together suffixes belonging to the same unperturbed precausal state, but not those from different states.

Lastly, the maximum deviation between the disturbed distributions (tildes) and the empirical estimates thereof (hats) with probability one becomes arbitrarily small as $N \to \infty$, by the law of large numbers and the continuity of $d_m$. Therefore robust homogenization, and subsequent determinization, will in the limit return the same state-to-suffix mapping as $X_t$ when $\alpha \in I_{\text{sig}}$. □

Since by assumption the same set of strings is considered by CSSR and RCS, the upper bound on computational complexity from Section 2.2 applies equally to both algorithms. For noisy data, RCS may be the faster method, as it creates fewer precausal states, leading to fewer comparisons during homogenization, and fewer states to check during determinization.

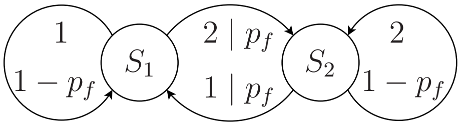

Figure 2: The flip process. Transitions are drawn as arrows labeled by emitted symbols and transition probabilities.

In practical applications the generating process is typically unknown, and one may not know if there exists a simple, CSSR-learnable underlying causal-state representation that well describes the data. Lacking other information for choosing $n_{\max}$, one may explore the material at different parameter values looking for parsimonious models, similar to the suggestions for deciding $L$ and $\alpha$ given by Klinkner and Shalizi (2005).

For some tests—including Kolmogorov-Smirnov—the function $d$ in the criterion can be factored as

$$d\left(\widehat{\boldsymbol{p}}_a,\, n_a,\, \widehat{\boldsymbol{p}}_b,\, n_b\right) = d_m\left(\widehat{\boldsymbol{p}}_a,\, \widehat{\boldsymbol{p}}_b\right) d_n\left(n_a,\, n_b\right). \tag{17}$$

Given enough data, only $L$ and the *critical distance* $d_{\text{crit}} = F_{\text{sig}}\left(\alpha\right)/d_n\left(n_{\max},\, n_{\max}\right)$ then determine the result of applying RCS. This reduces the parameter-space dimensionality since $\alpha$ and $n_{\max}$ need not be considered independently. Equation (17) also indicates that an alternative route to robust state clustering, not investigated further here, is not to limit $n_a$ or $n_b$, but instead choose $\alpha$ for each test such that small differences are never resolved.

## 5. A practical example

We now present a simple experiment showing how RCS can reliably return the underlying causal structure of a noise-disturbed process. CSSR, in contrast, fails to do so, consistent with the fact that the noisy data $Y_t$ satisfies the non-learnability criterion in Theorem 1.

### *5.1. The flip process*

Like the original CSSR paper (Shalizi and Shalizi, 2004) we consider learning a two-state toy process, specifically the "flip process" in Figure 2. This will be our underlying data source $X_t$. The process has two symbols, $\{1,\, 2\}$, and two causal states. The current state is revealed by the latest symbol, wherefore the states are labeled $\{S_1,\, S_2\}$.

We shall explore the effect of introducing random symbol substitutions in the data. Specifically, each observation, independently of all others, will with probability $\epsilon \in [0,\, 1]$ be replaced by a (uniformly) random character from the alphabet. The resulting noisy flip process, in Figure 3, can equivalently be represented by an HMM. As discussed in Dupont et al. (2005), transforming finite automata to HMMs generally requires one state per transition, since automata

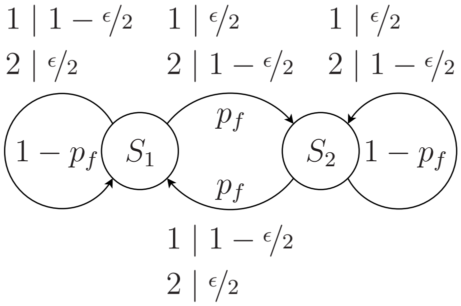

Figure 3: Nondeterministic automaton representing the noisy flip process. Transition probabilities are specified inside arcs, with conditional emission probabilities given outside.

associate symbol emissions with transitions, whereas HMMs associate emissions with states. This leads to a four-state HMM representation of the noisy flip process, defined by the matrices

$$\boldsymbol{A}=\begin{bmatrix} 1-p_f & p_f & 0 & 0 \\ 0 & 0 & p_f & 1-p_f \\ 1-p_f & p_f & 0 & 0 \\ 0 & 0 & p_f & 1-p_f \end{bmatrix}, \boldsymbol{B}=\begin{bmatrix} 1-\epsilon/2 & \epsilon/2 \\ \epsilon/2 & 1-\epsilon/2 \\ 1-\epsilon/2 & \epsilon/2 \\ \epsilon/2 & 1-\epsilon/2 \end{bmatrix}.$$

This HMM provides the observed $Y_t$ in the experiments.

Trivially, the noisy flip process has a single causal state for $p_f = 1/2$ or $\epsilon = 1$, and two for $\epsilon = 0$. For other parameter values Theorem 1 applies and the number of causal states must be infinite, meaning that the causal-state description of the observations $Y_t$ is sensitive to small data impurities. This non-learnability of $Y_t$ can be verified numerically using Algorithm 1. Alternatively, a general, algebraic proof, following the same steps as the numerical procedure, is provided in Appendix C. The algebraic calculations are relatively simple since all relevant characteristic polynomials here have degree two.

### *5.2. Recovering the flip process*

We applied RCS and CSSR to flip-process data with $p_f = 1/3$ and substitution probability $\epsilon = 0.2$. Like Shalizi and Shalizi (2004), the experiments used the Kolmogorov-Smirnov test with the significance $\alpha$ fixed at $10^{-3}$. $n_{\max}$ was correspondingly set to 1095, placing the boundary of the critical region near $\frac{1}{2}d_{\min}$, half the $d_m$-distance between the precausal states. Several different dataset sizes between $N = 10^3$ and $10^8$ were investigated, with 200 trials performed for each $N$. In each trial, CSSR and RCS were applied to a new string of $N$ symbols generated by the flip automaton with substitution noise.

The results in Figure 4 show that, once $N$ reached $10^6$, RCS with the given parameters in all trials returned a set of states and symbol-labeled nonzero-probability transitions isomorphic (in a graph-theoretic sense) to the causal states of $X_t$. The RCS-estimated suffix partitioning $\widehat{\varepsilon}$ is then equivalent to the

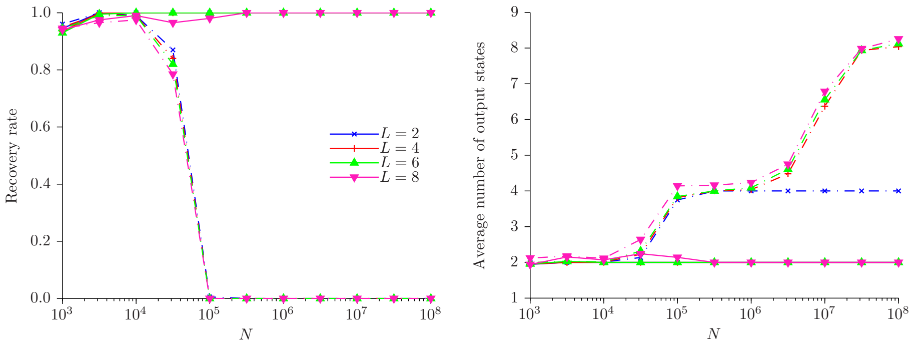

Figure 4: Underlying structure reliably recovered by RCS at large $N$. Results for RCS are drawn with solid lines; original CSSR is dash-dotted.

underlying $\varepsilon$. CSSR output, in contrast, never matched the underlying causal-state partitioning $\varepsilon$ for sample sizes exceeding $10^5$, typically giving larger models instead. RCS is thus robust to the added corruptions, whereas the original CSSR algorithm is not.

The graphs illustrate that the error probability is greater and convergence slower for large $L$. This is because each increase in $L$ includes longer and more uncommon suffixes for consideration, which require more data to converge to their expected next-step distributions. For memory lengths below 7, $10^4$ samples sufficed for RCS to return the underlying structure in virtually every trial. The observed convergence is assured by Theorem 2.

## 6. Related approaches

RCS and CSSR exist within a greater context of learning schemes with states defined directly over the observations, e.g., Ron et al. (1996); Gavaldà et al. (2006); Holmes and Isbell, Jr. (2006). Among these, the algorithm of Ron et al. (1996) for learning approximate variable-length Markov models (VLMMs, a kind of Markov chain) has polynomial complexity given a maximal memory length $L$ and a desired accuracy, but also has lower expressive power than causal states. For instance, the even process of Weiss (1973) has finite and learnable causal states, but cannot be represented fully by any finite Markov chain or finite VLMM.

The methods of Ron et al. (1998); Clark and Thollard (2004); Gavaldà et al. (2006) are similar to RCS in that they learn PDFA and include a notion of distinguishability as the minimum distance (in $\ell_1$-norm) between probability distributions induced by states in a PDFA. A lower bound $\mu > 0$ on this distinguishability is a key ingredient in making PDFA PAC-learnable, but the publications also require additional constraints, for example a finite upper bound $L'$ on the expected output length from any state. Their results thus only apply to finite-duration automata, and not to infinite-duration, stationary and ergodic processes as considered here. Still, it seems these algorithms could in principle be made noise tolerant similar to RCS: when the effects of the noise are small compared to the distinguishability, they will not alter the decisions to separate or group together different strings in the algorithms. We are not aware of any work in this direction, however.

Holmes and Isbell, Jr. (2006) introduce the so-called looping prediction suffix tree representation for certain POMDPs, along with an algorithm for learning such representations from histories. The results are compelling, but the models are not probabilistic; transitions and observations are assumed deterministic, given a state and an action, and actions are not associated with any probability distribution.

The CSSR algorithm itself has also received several extensions and modifications. Among these, the *clustered causal state algorithm* (CCSA) of Schmiedekamp et al. (2006) is relevant in the current context. In CCSA, CSSR homogenization is replaced by an agglomerative clustering of suffixes according to empirical next-step distribution similarity. This is similar to changing $\alpha$ for every test so that the critical distance remains constant. Importantly, CCSA removes the determinization stage, so nondeterministic output automata are possible. Since such automata never can be causal states, one may rather think of CCSA as a clustered *pre*causal-state algorithm.

RCS homogenization and CCSA clustering both group together suffixes that need not have identical next-step distributions. Unlike the tests used in RCS and CSSR, CCSA clustering does not account for the number of samples on which the distribution estimates are based. Neither is there any bias to place long suffixes with poor statistics together with their parent suffix (the shorter suffix that results if the oldest symbol is deleted) by first comparing suffixes against the precausal state of the parent. The suffix-to-precausal-state assignment of CCSA will therefore be more erratic than in CSSR or RCS.

CCSA may be able to recover the precausal states of an underlying process under moderate disturbances, although this possibility has not been addressed in publications. However, recovering the underlying causal states generally requires determinization, which is not part of CCSA. CCSA output is therefore typically nondeterministic, and then cannot be the causal structure of *any* process. Such clustered states are not optimal predictors for multiple time steps, and a random walk over CCSA states need not generate statistics that are anything like $X_t$, even if noise is absent.

## 7. Conclusions

We conclude that the causal state description of stochastic processes has several appealing properties, including predictive optimality, but is sensitive to disturbances such as random substitutions. Even the slightest data impurities can cause CSSR, the canonical causal-state learning algorithm, to diverge and return a large and overfitted representation.

However, we additionally conclude that problems can be overcome, and the compact underlying causal structure can be reliably recovered also under moderate corruptions of the data, by modifying the algorithm to ignore small differences in history-string suffix behavior. Our conclusions are supported by both theory and experiment. We call the modified CSSR algorithm robust causal states, RCS. To our knowledge, no prior method has the same general ability of recovering causal structure from disturbed data.

We anticipate RCS models can be adapted for tasks such as error correction or learning stochastic grammars. Apart from exploring applications, an interesting future topic would be to consider additional mechanisms for controlling output complexity.

## Acknowledgments

This research was funded partially by the European Commission, under contract number FP6-034362, in the ACORNS project, www.acorns-project.org.

## References

Clark, A., Thollard, F., 2004. PAC-learnability of probabilistic deterministic finite state automata. Journal of Machine Learning Research 5, 473–497.

Crutchfield, J., Shalizi, C., 1999. Thermodynamic depth of causal states: Objective complexity via minimal representation. Phys. Rev. E 59, 275–283.

Dupont, P., Denis, F., Esposito, Y., 2005. Links between probabilistic automata and hidden Markov models: probability distributions, learning models and induction algorithms. Pattern Recogn. 38, 1349–1371.

Eddy, S., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763.

Gavaldà, R., Keller, P., Pineau, J., Precup, D., 2006. PAC-learning of Markov models with hidden state, in: Lect. Notes Comput. Sc. (ECML 2006), pp. 150–161.

Holmes, M., Isbell, Jr., C., 2006. Looping suffix tree-based inference of partially observable hidden state, in: Proc. 23rd ICML, pp. 409–416.

Jaeger, H., 2000. Observable operator models for discrete stochastic time series. Neural Comput. 12, 1371–1398.

Klinkner, K., Shalizi, C., 2005. "Read me" for CSSR. http://bactra.org/CSSR/CSSR-v0.1.1.tar.gz.

Klinkner, K., Shalizi, C., Camperi, M., 2006. Measuring shared information and coordinated activity in neuronal networks, in: Adv. Neu. In. (NIPS 18), pp. 667–674.

Löhr, W., Ay, N., 2009. Non-sufficient memories that are sufficient for prediction, in: Proc. Complex'2009, pp. 265–276.

Nerukh, D., 2008. Computational mechanics reveals nanosecond time correlations in molecular dynamics of liquid systems. Chem. Phys. Lett. 457, 439–443.

Padró, M., Padró, L., 2005. A named entity recognition system based on a finite automata acquisition algorithm. Proces. Leng. Nat. 35, 319–326.

Padró, M., Padró, L., 2007. Studying CSSR algorithm applicability on NLP tasks. Proces. Leng. Nat. 39, 89–96.

Park, J., Lee, J., Yang, J., Jo, H., Moon, H., 2007. Complexity analysis of the stock market. Physica A 379, 179–187.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. P. IEEE 77, 257–286.

Ray, A., 2004. Symbolic dynamic analysis of complex systems for anomaly detection. Signal Process. 84, 1115–1130.

Ron, D., Singer, Y., Tishby, N., 1996. The power of amnesia: Learning probabilistic automata with variable memory length. Mach. Learn. 25, 117–149.

Ron, D., Singer, Y., Tishby, N., 1998. On the learnability and usage of acyclic probabilistic finite automata. J. Comput. Syst. Sci. 56, 133–152.

Schmiedekamp, M., Subbu, A., Phoha, S., 2006. The clustered causal state algorithm: Efficient pattern discovery for lossy data-compression applications. Comput. Sci. Eng. 8, 59–67.

Shalizi, C., Crutchfield, J., 2001. Computational mechanics: Pattern and prediction, structure and simplicity. J. Stat. Phys. 104, 817–879.

Shalizi, C., Shalizi, K., 2004. Blind construction of optimal nonlinear recursive predictors for discrete sequences, in: Proc. 20th UAI, pp. 504–511.

Shalizi, C., Shalizi, K., Crutchfield, J., 2002. An Algorithm for Pattern Discovery in Time Series. Technical Report 02-10-060. Santa Fe Institute. http://arxiv.org/abs/cs.LG/0210025. cs/0210025.

Starner, T., Pentland, A., 1995. Real-time american sign language recognition from video using hidden Markov models, in: Proc. Int. Symp. Comput. Vis., pp. 265–270.

Weiss, B., 1973. Subshifts of finite type and sofic systems. Monatsh. Math. 77, 462–474.

Woodland, P., Leggetter, C., Odell, J., Valtchev, V., Young, S., 1995. The 1994 HTK large vocabulary speech recognition system, in: Int. Conf. Acoust. Spee., pp. 73–76.