LICORS: Light Cone Reconstruction of States for Non-parametric Forecasting of Spatio-Temporal Systems

Georg M. Goerg Cosma Rohilla Shalizi*

August 7, 2012

Abstract

We present a new, non-parametric forecasting method for data where continuous values are observed discretely in space and time. Our method, *light-cone reconstruction of states* (LICORS), uses physical principles to identify predictive states which are local properties of the system, both in space and

*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 USA; { gmg, cshalizi } @ stat.cmu.edu. This work was partially supported by grants from INET, from the NIH (# 2 R01 NS047493), and from the NSF (DMS1207759). The authors thank Stacey Ackerman-Alexee , Dave Albers, Chris Genovese, Rob Haslinger, Martin Nilsson Jacobi, Heike Janicke, Kristina Klinkner, Cristopher Moore, Jean-Baptiste Rouquier, Chad Schafer, Rafael Stern, and Chris Wiggins for valuable discussion, and Larry Wasserman for detailed suggestions that have improved all aspects of this work.

time. LICORS discovers the number of predictive states and their predictive distributions automatically, and consistently, under mild assumptions on the data source. We provide an algorithm to implement our method, along with a cross-validation scheme to pick control settings. Simulations show that CV-tuned LICORS outperforms standard methods in forecasting challenging spatio-temporal dynamics. Our work provides applied researchers with a new, highly automatic method to analyze and forecast spatio-temporal data.

Keywords: non-parametric prediction, dynamical system, forecasting, predictive state reconstruction, spatio-temporal data.

1 Introduction

Many important scientific and data-analytic problems involve fields which vary over both space and time, e.g., data from functional magnetic resonance imaging, meteorological observations, or experimental studies in physics and chemistry. An outstanding objective in studying such data is prediction, where we want to describe the field in the future.

Spatio-temporal data being increasingly easy to acquire, manipulate and visualize, statisticians have developed corresponding methods for statistical inference, reviewed in works like Finkenstädt et al. (2007); Cressie and Wikle (2011). The usual tools are a combination of ways of describing the distribution of the random field (e.g., various dependency measures), and stochastic modeling, focusing primarily on parametric inference, and secondarily on parameter-conditional predictions.

While these approaches are valuable, there is a complementary role for direct,

non-parametric prediction of spatio-temporal data, just as with time series (Bosq, 1998; Fan and Yao, 2003). Our aim here is to blend modern methods of non-parametric prediction with insights from nonlinear physics on the organization of spatial dynamics, yielding predictors of spatio-temporal evolution that are computationally efficient and make minimal assumptions on the data source, but are still accurate and even interpretable.

The idea behind our approach is simply that it takes time for influences to propagate across space, so we can constrain the search for predictors to a spatio-temporally local neighborhood at each point. We combine this with a novel form of non-parametric smoothing, which infers the prediction (regression or conditional probability) function by averaging together similar observations, where "similarity" is defined in terms of predictive consequences, effectively replacing the original geometry of the predictor variables with a new one, optimized for forecasting. The combination of these two tools lets us discover underlying structures, as well make fast and accurate predictions.

Section 2 formally defines our prediction problem and introduces our non-parametric localized approach. Section 3 gives the statistical methods to estimate these optimal predictors from discretely-observed continuous-valued fields. Section 4 shows, under weak conditions on the data-generating process, that our method consistently estimates the predictive distributions. Section 5 proposes a cross-validation scheme to choose our control settings, and compares our predictive accuracy to standard time series techniques. Finally, Section 6 summarizes this new methodology and discusses future work. Proofs and implementation details can be found in the Supplementary

Material.

2 Local Prediction of Spatio-temporal Fields

2.1 Setting and Notation; Light Cones

We observe a random field $(X(\mathbf{r},t))_{\mathbf{r}\in\mathbf{S},t\in\mathbb{T}}$ in discrete space and time. The field takes values in a set $\mathcal{X}$, which may be discrete or continuous. Space $\mathbf{S}$ is a regular lattice, equipped with norm $\|\mathbf{r}\|$. Time $\mathbb{T}$ is taken to be the positive integers up to T.

Suppose that disturbances or influences in the system have a maximum speed of propagation, c. Then the only events which could affect what happens at a given $(\mathbf{r},t)$ are those where $s \leq t$ and $\|\mathbf{r}-\mathbf{u}\| \leq c(t-s)$. Since this set grows as s recedes into the past, we call this the **past light cone** (PLC) of $(\mathbf{r},t)$. The **future light cone** (FLC) are all events which could be affected by the present moment $(\mathbf{r},t)$; it thus consists of all those $(\mathbf{u},s)$, where $s > t$ and $\|\mathbf{r}-\mathbf{u}\| \leq c(s-t)$. Light cones look like triangles in $(1+1)D$ fields, and in $(2+1)D$, pyramids (Fig. 1). Denote the configuration in the past cone of $(\mathbf{r},t)$ by $L^-(\mathbf{r},t)$:

$$L^-(\mathbf{r},t) = \{X(\mathbf{u},s) \mid s \leq t, \|\mathbf{r}-\mathbf{u}\| \leq c(t-s)\} \tag{1}$$

$L^+(\mathbf{r},t)$ is, similarly, the configuration in the future cone.

The spatio-temporal prediction problem is thus: use the configuration of the past cone, $L^-(\mathbf{r},t)$, to forecast the configuration of the future cone, $L^+(\mathbf{r},t)$. Light-cone prediction compromises between capturing global patterns and needing only local

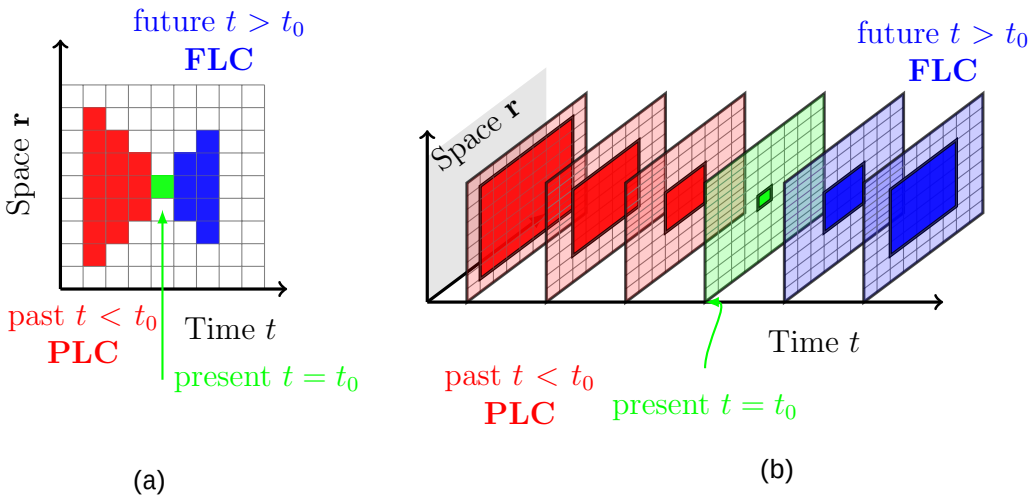

Figure 1: Past (red) and future (blue) light cones in a $(1+1)D$ (a) and $(2+1)D$ (b) system. Here c, the velocity of signal propagation, is set to 1. The past cone is truncated at a horizon of $h_p = 3$ steps, while the future cone's horizon is only $h_f = 2$. Whether the present (green) is included in the past or the future cone is a matter of convention; see Section 5.

information. We will construct optimal predictors for light cones presently. Light cones can be defined for spatial extended patches of points. (When the "patch" becomes the whole spatial lattice, we are back to global prediction.) This leads to a parallel theory of prediction, but it turns out that the predictive state of a patch is determined by the predictive states of its points (Shalizi, 2003, §3.3, Lemma 2 and Theorem 3), so we lose no information, and gain tractability, by not considering cones for patches.

Computationally, we need to truncate the cones at a finite number of time steps — we will call these the *past horizon* h_p of L^-, and likewise the *future horizon* h_f of L^+. Doing this reduces L^+ and L^- to finite-dimensional random vectors. (For instance, in Fig. 1, with $h_p = 3$ and $c = 1$, $\ell^-(\mathbf{r}, t)$ has 15 degrees of freedom.)

The horizons are control settings, and may be tuned through (for example) cross-validation (§5.2). Similarly, when the maximum speed of propagation c is not given from background knowledge, it is also a control setting.

2.2 Predictive States

To predict the future $L^+(\mathbf{r}, t)$ from a particular past configuration, say ℓ^-, requires knowing the conditional distribution

$$\mathbb{P}\left(L^+(\mathbf{r}, t) \mid L^-(\mathbf{r}, t) = \ell^-\right) \tag{2}$$

for all ℓ^-. (Subsequently $(\mathbf{r}, t)$ may be omitted for readability.) Since treating this conditional distribution as an arbitrary function of ℓ^- is not feasible statistically or computationally, we try to find a *sufficient statistic* η of past configurations that keeps the predictive information:

$$\mathbb{P}\left(L^+(\mathbf{r}, t) \mid H(\mathbf{r}, t) = \eta(\ell^-)\right) = \mathbb{P}\left(L^+(\mathbf{r}, t) \mid L^-(\mathbf{r}, t) = \ell^-\right) \ . \tag{3}$$

There are usually many sufficient statistics $\eta, \eta', \ldots$. When η and η' are both sufficient, but $\eta(\ell^-) = f(\eta'(\ell^-))$ for some f, then η is a smaller, more compressed, summary of the data than η', and so the former is preferred by Occam's Razor. The minimal sufficient statistic ϵ compresses the data as much as can be done without losing any predictive power, retaining only what is needed for optimal predictions.

We now construct the minimal sufficient statistic, following Shalizi (2003), to which we refer for some mathematical details.

Definition 2.1 (Equivalent configurations)**.** *The past configurations* ℓ_i^- *at* $(\mathbf{r}, t)$ *and* ℓ_j^- *at* $(\mathbf{u}, s)$ *are* predictively equivalent*,* $(\ell_i^-, (\mathbf{r}, t)) \sim (\ell_j^-, u(\mathbf{s},))$*, if they predict the same future with equal probabilities, i.e. if*

$$\mathbb{P}\left(L^+(\mathbf{r}, t) \mid L^-(\mathbf{r}, t) = \ell_i^-\right) = \mathbb{P}\left(L^+(\mathbf{u}, s) \mid L^-(\mathbf{u}, s) = \ell_j^-\right) \tag{4}$$

Let $[(\ell^-, (\mathbf{r}, t))]$ be the equivalence class of $(\ell^-, (\mathbf{r}, t))$, i.e., the set of all past configurations and coordinates that predict the same future as ℓ^- does at $(\mathbf{r}, t)$. Let

$$\epsilon(\ell^-, (\mathbf{r}, t)) \equiv \left[\ell^-\right] \tag{5}$$

be the function mapping each $(\ell^-, (\mathbf{r}, t))$ to its predictive equivalence class. The values ϵ can take are the *predictive states*; they are the minimal statistics which are sufficient for predicting L^+ from L^- (Shalizi, 2003).

Since each predictive state has a unique predictive distribution and vice versa. We will thus slightly abuse notation to denote by $\mathcal{E}$ both the set of equivalence classes and the set of predictive distributions, whose elements we will write ϵ_j. We will further abuse notation by writing the mapping from past cone configurations to predictive distributions as $\epsilon(\cdot)$, leading to the measure-valued random field

$$S(\mathbf{r}, t) := \epsilon\left(L^-(\mathbf{r}, t)\right) \ . \tag{6}$$

One can show (Shalizi, 2003) that $S(\mathbf{r}, t)$ is Markov even if $X(\mathbf{r}, t)$ is not. However, X is not an ordinary hidden Markov random field, since there is an unusual *deterministic*

dependence between transitions in S and the realization of X, analogous to that of a chain with complete connections (Fernández and Maillard, 2005).

To be able to draw useful inferences from a single realization of the process, we must assume some form of homogeneity or invariance of the conditional distributions.

Assumption 2.2 (Conditional invariance)**.** *The predictive distribution of a PLC configuration ℓ^- does not change over time or space. That is, for all $\mathbf{r}, t$, all $\mathbf{u}, s$, and all past light-cone configurations ℓ^-,*

$$(\ell^-, (\mathbf{r}, t)) \sim (\ell^-, (\mathbf{u}, s)) \tag{7}$$

We may thus regard $\sim$ as an equivalence relation among PLC configurations, and ϵ as a function over ℓ^- alone.

This is just *conditional* invariance, like the conditional stationarity for time series used in Caires and Ferreira (2005). It would be implied by the field being a Markov random field with homogeneous transitions, or of course by full stationarity and spatial invariance, but it is weaker. Assumption 2.2 lets us talk about *the* predictive distribution of a PLC configuration, regardless of when or where it was observed, and to draw inferences by pooling such observations. If this assumption fails, we could in principle still learn a different set of predictive states for each moment of time and/or each point of space (as in Shalizi (2003)), but this would need data from multiple realizations of the same process.

3 Estimating Predictive States

We extend the work of Shalizi (2003); Shalizi et al. (2004) to continuous-valued fields, introducing statistical methods to estimate and predict non-linear dynamics accurately and efficiently, while still obtaining insight into the spatio-temporal structure. Algorithmic details are given in the Supplementary Material.

Assume we have T consecutive measurements of the field $X(\mathbf{r}, t)$, observed over the lattice $\mathbf{S}$, with $N = |\mathbf{S}| \cdot T$ space-time coordinates $(\mathbf{r}, t)$ in all. Each one of these N point-instants has a past and a future light-cone configuration, $\ell^-(\mathbf{r}, t)$ and $\ell^+(\mathbf{r}, t)$, represented as, respectively, n_p and n_f dimensional vectors. Since predictive states are sets of PLC configurations with the same predictive distribution, we need to test this sameness, based on conditional samples $\{\ell^+ \mid \ell_i^-\}_{i=1}^N$ from the observed field. We will apply non-parametric two-sample tests for $H_0 : \mathbb{P}\left(L^+ \mid L^- = \ell_i^j\right) = \mathbb{P}\left(L^+ \mid L^- = \ell_i^-\right)$ pairwise for all i and j. Because there are typically a great many past light cones (one for each point-instant), and light-cone configurations are themselves high-dimensional objects, we generally must do this step-wise.

3.1 Partitioning PLC Configurations: Similar Pasts Have Similar Futures

It is often reasonable to assume that the mapping from the past to predictive distributions is regular, so that if two historical configurations are close (in some suitable metric), then their predictive distributions are also close. This lets us avoid having to do some pairwise tests, as their results can be deduced from others.

Assumption 3.1 (Continuous histories)**.** *For every $\rho > 0$, there exists a $\delta > 0$ such that*

$$\|\ell_i^- - \ell_j^-\| < \delta \Rightarrow \mathcal{D}_{KL}\left(\mathbb{P}\left(L^+ \mid \ell_i^-\right) \,||\, \mathbb{P}\left(L^+ \mid \ell_j^-\right)\right) < \rho, \tag{8}$$

where $\mathcal{D}_{KL}\left(p \,||\, q\right)$ is the Kullback-Leibler divergence between distributions p and q (Kullback, 1968).

Assumption 3.1 requires that sufficiently small changes ($< \delta$) in the local past make only negligible ($< \rho$) changes to the distribution of local future outcomes. Statistically, such smoothness-in-distribution lets us pool observations from highly similar PLC configurations, enhancing efficiency; physically, it reflects the smoothness of reasonable dynamical mechanisms. Chaotic systems, where the exact trajectory depends sensitively on initial conditions, do not present difficulties, since Assumption 3.1 is about the conditional *distribution* of the future given partial information on the past, and chaos has long been recognized as a way to stabilize such distributions, forming the basis for prediction and control of chaos (Kantz and Schreiber, 2004).

We use Assumption 3.1 to justify an initial "pre-clustering" of the PLC configuration space, greatly reducing computational cost with little damage to predictions. We first divide the PLC configuration space using fast clustering algorithms into $K \ll N$ clusters, and then test equality of distributions between clusters ($\mathcal{O}(K^2)$), rather than light cones ($\mathcal{O}(N^2)$).

When N is small enough, we can skip this initial pre-clustering. To simplify exposition, we treat this as assigning each distinct past cone to its own cluster.

3.2 Partitioning Clusters into Predictive States

Each cluster P_k contains a set of similar PLC configurations, and also defines a sample of conditional FLCs, $\mathbf{F}_k(\delta) = \{\ell_j^+ \mid \ell_j^- \in P_k\} \in \mathbb{R}^{N_k \times n_f}$, $k = 1, \ldots, K$. Since all $\ell_j^- \in P_k$ have very similar distribution, $\mathbf{F}_k \sim Q$ is an approximate sample from the predictive distribution $\mathbb{P}\left(L^+\left(\mathbf{r}, t\right) \mid \ell^- \in P_k\right)$. Lemma 4.7, below, shows that for sufficiently small δ, $\mathbf{F}_{k_i}(\delta)$ is an exact sample of $p(\epsilon(P_{k_i}))$. Thus, to simplify the exposition, we ignore the ρ difference in this section.

Thus, finding equivalent clusters reduces to testing hypotheses of the form $H_0 : p_{k_i} = p_{k_j}$ based on the two samples $\mathbf{F}_{k_i}(\delta)$ and $\mathbf{F}_{k_j}(\delta)$. For $h_f = 0$ and $c = 1$, FLCs are one-dimensional and we can use a Kolmogorov-Smirnov test (or any other two-sample univariate test). In general, however, $\mathbf{F}_k$ are samples from a very high-dimensional distribution, and we use non-parametric, multivariate, two-sample tests (see e.g. Rosenbaum, 2005; Rizzo and Székely, 2010; Gretton et al., 2007). Any test satisfying Assumption 4.11 could be used.

To estimate the predictive states from an initial partitioning of PLC configurations, we iterate through the list of configurations, recursively testing equality of distributions. To initialize the algorithm, create the first predictive state ϵ_1, containing the first configuration ℓ_1^-. Then take ℓ_2^- and test if its distribution is equal to that of ϵ_1. If it is (at the level α), then put ℓ_2^- in ϵ_1; otherwise generate a new predictive state ϵ_2 with ℓ_2^-. Then test the next configuration against all previously established predictive states and proceed as before. This continues until all configurations have been assigned to a predictive state.

The predictive distribution of each predictive state can be found by applying any

consistent non-parametric density estimator to the future cone samples belonging to that state. If we only want point forecasts, we can skip estimating the whole predictive distribution and just get (e.g.) the mean of the samples.

4 Consistency

LICORS consistently recovers the correct assignment of past cone configurations to predictive states, and the predictive distributions, under weak assumptions on the data-generating process. These allow for the number of predictive states to grow slowly with the sample size, so that we have non-parametric consistency. We give all assumptions and lemmas in the main text; proofs are in the Supplementary Material.

4.1 Assumptions

Let $N = |\mathbf{S} \times \mathbb{T}|$ be the total number of space-time points at which we observe both the past and future light cone. We presume that $N \to \infty$, without caring whether $|\mathbf{S}| \to \infty$, $|\mathbb{T}| \to \infty$, or both.

Assumption 4.1 (Slowly growing number of predictive states)**.** *The number of predictive states,* $|\mathcal{E}| = m(N) = \mathrm{o}\big(N\big)$*, and always* $\leq N$*.*

Assumption 4.1 only guarantees that *at least* one of the predictive states grows in size. To bound testing error probabilities, the number of light cones seen in *every* state must grow as N grows.

Assumption 4.2 (Increasing number of light cones in each state)**.** *The number of light cones in each state,* $N_j := |\epsilon_j|$*, grows with* N*: for all* $\epsilon_j \in \mathcal{E}$*,*

$$\lim_{N\to\infty} N_j(N) = \infty \tag{9}$$

Let $N_{\min} = \min_j N_j$ be the number of samples in the smallest predictive state; thus also $N_{\min} \to \infty$ for $N \to \infty$. Assumption 4.2 means that the system re-visits each predictive state as it evolves, i.e., all states are recurrent. This lets us learn the predictive distribution of each state, from a growing sample of its behavior.

Assumption 4.3 (Bounded conditional distributions)**.** *All predictive distributions* $\epsilon_j \in \mathcal{E}$ *have densities with respect to a common reference measure* ν*, and* $0 < \iota < d\epsilon_j/d\nu < \kappa < \infty$*, for some constants* ι *and* κ*.*

This merely technical assumption guarantees bounded likelihood ratios.

Assumption 4.4 (Distinguishable predictive states)**.** *The KL divergence between states is bounded from below:* $\forall i \neq j$*,*

$$0 < d_{\min} \leq \mathcal{D}_{KL}\left(\epsilon_i \,||\, \epsilon_j\right) =: d_{i,j} \tag{10}$$

We do not need $d_{i,j} < \infty$. (In fact, $\mathcal{D}_{KL}\left(\epsilon_i \,||\, \epsilon_j\right) = \infty$ is helpful.) (10) is automatically satisfied for any fixed number of states. For an increasing state space, $m = m(N)$, assume

$$\inf_{i,j\in m(N)} d_{i,j} = d_{\min} > 0 \text{ for } N \to \infty. \tag{11}$$

Lemma 4.5 (Conditionally independent FLCs)**.** *If the cones* $L^+(\mathbf{r},t)$ *and* $L^+(\mathbf{u},s)$ *do not overlap, then*

$$L^+(\mathbf{r},t) \perp\!\!\!\perp L^+(\mathbf{u},s) \mid S(\mathbf{r},t), S(\mathbf{u},s). \tag{12}$$

In particular,

$$\mathbb{P}\left(L^+(\mathbf{r},t), L^+(\mathbf{u},s) \mid S(\mathbf{r},t), S(\mathbf{u},s)\right) = \mathbb{P}\left(L^+(\mathbf{r},t) \mid S(\mathbf{r},t)\right)\mathbb{P}\left(L^+(\mathbf{u},s) \mid S(\mathbf{u},s)\right). \tag{13}$$

Corollary 4.6. *If* $h_f = 0$*, then FLCs are conditionally independent given their predictive state.*

4.1.1 Getting samples from ϵ_i

We get a sample of FLCs from the predictive distribution of ℓ_i by first taking all PLCs in a δ-neighborhood around ℓ_i,

$$I_i(\delta) = \{j \mid \|\ell_i^- - \ell_j^-\| < \delta\}. \tag{14}$$

For later use, we denote by $S_i(N,\delta) = |I_i(\delta)|$ the number of such light cones. By Assumption 3.1, we get our sample from ϵ_i by collecting the corresponding future cone configurations:

$$\mathbf{F}_i(\delta) = \{\ell_j^+ \mid j \in I_i(\delta)\}, \tag{15}$$

Lemma 4.7. *For sufficiently small* $\delta > 0$*, all past configurations in* $I_i(\delta)$ *are predic-*

tively equivalent: $\forall j, k \in I_i(\delta), \ell_j^- \sim \ell_k^-$ *. Consequently, all* ℓ_j^+*,* $j \in I_i(\delta)$*, are drawn from the same distribution* $\epsilon(\ell_i^-)$*.*

For finite N, it may not be possible in practice to find and use a sufficiently small δ. With pre-clustering, for instance, some of the clusters may have diameters greater than the δ which guarantees equality of distribution. Then the samples $\mathbf{F}_i(\delta)$ are actually from multiple states. One could circumvent this by using more clusters, which generally shrinks cluster diameters, but this would also reduce the number of samples per neighborhood, increasing the error rate of our two-sample tests. In practice, then, one must trade off decreasing δ to discover all predictive states and keeping a low testing error.

Corollary 4.8. *For sufficiently small* $\delta > 0$*, and non-overlapping FLCs, all the future configurations in* $\mathbf{F}_i(\delta)$ *are IID samples from* $\epsilon(\ell_i^-)$*.*

In general, for $h_f > 0$ the FLCs in $\mathbf{F}_i(\delta)$ can be overlapping and the conditional likelihood does not factorize. Yet, without loss of generality, we can consider only non-overlapping FLCs. This is because we can explicitly exclude overlapping FLCs from $\mathbf{F}_i(\delta)$, at the cost of reducing the sample size to $\tilde{S}_i(N, \delta) \leq S_i(N, \delta)$. For each ℓ_i, the maximum number of FLCs which we must thereby exclude, say w, is fixed geometrically, by c, h_f and the dimension of the space $\mathbf{S}$, and does not grow with N. The exclusion thus is asymptotically irrelevant, since $\frac{S_i(N,\delta)}{w} \leq \tilde{S}_i(N, \delta) \leq S_i(N, \delta)$.

Further, note that, at least formally, it's enough to analyze the univariate, zero-horizon FLC distributions, which rules out overlaps. This is because longer-horizon FLC distributions must be consistent with the one-step ahead distributions and

the transition relations of the underlying predictive states. Thus we could get the n_f-dimensional FLC distribution by iteratively combining the univariate FLC distributions and the predictive state transitions, i.e., by chaining together one-step-ahead predictions, as in Shalizi and Crutchfield (2001, Corollary 2).

Assumption 4.9 (Number of samples from each cone)**.** *For each fixed $\delta > 0$, and each past light cone ℓ_i, $S_i(N, \delta) \xrightarrow[N\to\infty]{} \infty$.*

For each δ, $S_i(N, \delta)$ is a random variable, and to establish consistency we need some regularity conditions on how S_i grows with N. Let $S_{\min}(N, \delta) = \min_j S_j(N, \delta)$ be the smallest number of samples per δ-neighborhood for each N and δ.

Assumption 4.10. *For some $\tilde{c} > 0$,*

$$N \cdot m(N) \cdot \mathbb{E}e^{-\tilde{c}d_{\min}^2 S_{\min}(N,\delta)} \xrightarrow[N\to\infty]{} 0 \ . \tag{16}$$

Since $\mathbb{E}e^{tS_{\min}(N,\delta)}$ is the moment generating function of $S_{\min}$, this amounts to asserting that the number of samples concentrates around its mean while growing, ruling out pathological cases where $S_i(N, \delta)$ grows to infinity, but concentrates around small values.

4.2 Unknown Predictive States: Two-sample Problem

With a finite number N of observations, recovering the states is the same as determining which past cone configurations are predictively equivalent. We represent this with an $N \times N$ binary matrix $\mathbf{A}$, where $A_{ij} = 1$ if and only if $\ell_i \sim \ell_j$. LICORS gives

us an estimate of this matrix, $\widehat{\mathbf{A}}$, and we will say that the predictive states can be recovered consistently when

$$\mathbb{P}\left(\widehat{\mathbf{A}} \neq \mathbf{A}\right) \xrightarrow[N\to\infty]{} 0 \ . \tag{17}$$

Since the predictive distributions are unknown, we use non-parametric two-sample tests to determine whether two past cone configurations are predictively equivalent. While simulations can always be used to approximate the power of particular tests against particular alternatives, there do not (yet) seem to be any general expressions for the power of such tests, analogous to the bounds on likelihood tests in terms of KL divergence (Kullback, 1968). Nonetheless, we expect that for $N \to \infty$, the probability of error approaches zero, as long as the true distributions are far enough apart. We thus make the following assumption.

Assumption 4.11. *Suppose we have n samples from distribution p, and n' samples from distribution q, all IID. Then there exist a positive constants $d_{n,n'}$ tending to 0 as $n, n' \to \infty$, and a sequence of tests $T_{n,n'}$ of $H_0 : p = q$ vs. $H_1 : p \neq q$ with size $\alpha = \mathrm{o}\big(\min(n,n')^{-2}\big)$, and type II error rate $\beta(\alpha, n, n') = \mathrm{o}\big(\min(n,n')^{-2}\big)$ so long as p and q are mutually absolutely continuous and $\mathcal{D}_{KL}\left(p \,||\, q\right) \geq d_{n,n'}$.*

Note that if the number of predictive states is constant in N, we can weaken the assumption to just a sequence of tests whose type I and type II error probabilities both go to zero supra-quadratically when $\mathcal{D}_{KL}\left(p \,||\, q\right) \geq d_{\min}$.

Theorem 4.12 (Consistent predictive state estimation)**.** *Under Assumptions 2.2,*

3.1, 4.1, 4.2, 4.3, 4.4, 4.9, 4.10, and 4.11,

$$\mathbb{P}\left(\widehat{\mathbf{A}} \neq \mathbf{A}\right) \xrightarrow[N\to\infty]{} 0. \tag{18}$$

5 Simulations

To evaluate the non-asymptotic predictive ability of LICORS, and to compare it to more conventional methods, we use the following simulation, designed to be challenging, but not impossible. $X(\mathbf{r}, t)$ is a continuous-valued field in $(1+1)D$, with a discrete latent state $d(\mathbf{r}, t)$. We use "wrap-around" boundary conditions, so sites 0 and $|S|-1$ are adjacent, and the one spatial dimension is a torus. The observable field $X(\mathbf{r}, t)$ is conditionally Gaussian,

$$X(\mathbf{r}, t) \mid d(\mathbf{r}, t) \sim \begin{cases} \mathcal{N}(d(\mathbf{r}, t), 1), & \text{if } |d(\mathbf{r}, t)| < 4, \\ \mathcal{N}(0, 1), & \text{otherwise,} \end{cases} \tag{19}$$

with initial conditions $X(\cdot, 1) = X(\cdot, 2) = \mathbf{0}$. The state space $d(\mathbf{r}, t)$ evolves with the observable field,

$$d(\mathbf{r}, t) = \left[\frac{\sum_{i=-2}^{2} X(\mathbf{r} + \mathbf{i} \bmod |\mathbf{S}|, t-2)}{5} - \frac{\sum_{i=-1}^{i} X(\mathbf{r} + \mathbf{i} \bmod |\mathbf{S}|, t-1)}{3} \right], \tag{20}$$

where $[x]$ is the closest integer to x. In words, Eq. (20) says that the latent state $d(\mathbf{r}, t)$ is the rounded difference between the sample average of the 5 nearest sites at $t-2$ and the sample average of the 3 nearest sites at $t-1$. Thus $h_p = 2$ and $c = 1$.

If we include the present in the FLC, (19) gives $h_f = 0$, making FLC distributions one-dimensional and letting us use the Kolmogorov-Smirnov test. As

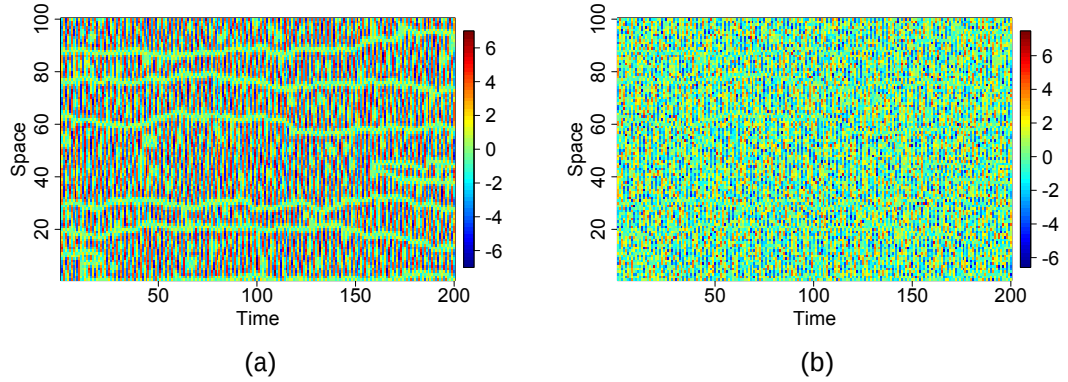

Figure 2: Simulation of (19)–(20): (a) state-space $d(\mathbf{r}, t)$, (b) observed field $X(\mathbf{r}, t)$. Space (100 cells) runs vertically, time (200 steps, first 100 discarded for burn-in) runs from left to right.

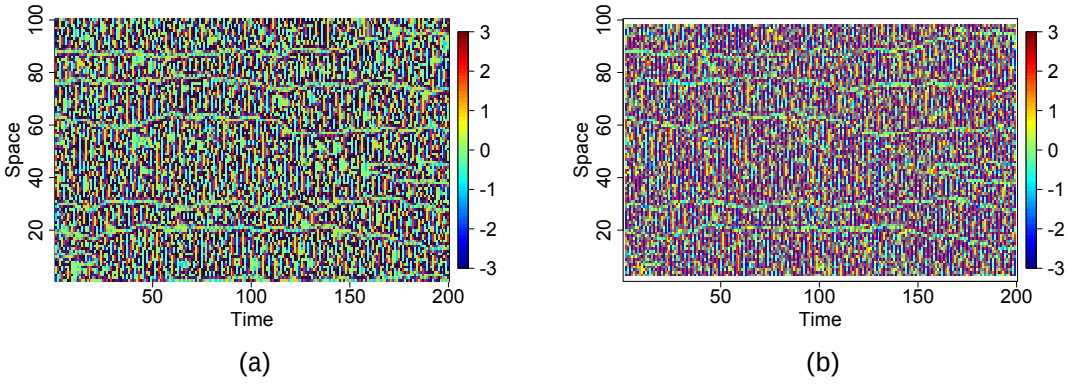

Figure 3: Comparison of true and estimated predictive distributions. (a) true predictive state $S(\mathbf{r}, t)$, with points colored by conditional expectations; (b) LICORS predictions, with states and distributions reconstructed using $k = 50$ nearest neighbors (fixed) and $h_p = 2, \alpha = 0.2$ (chosen by cross-validation).

$d(\mathbf{r}, t)$ is integer-valued, a little calculation shows there are 7 predictive states, which we may label with their conditional means as $\{\epsilon_{-3}, \epsilon_{-2}, \ldots, \epsilon_2, \epsilon_3\}$. Thus $X(\mathbf{r}, t) \mid \epsilon_k \sim \mathcal{N}(k, 1)$.

Figure 2 shows one realization of (19)–(20). The latent states have clear spatial

structures, which is obscured in the observed field. Figure 3a shows the true predictive state space $S(\mathbf{r}, t)$ (expected value at each at each $(\mathbf{r}, t)$); the LICORS estimate $\widehat{S}(\mathbf{r}, t)$ is shown in Fig. 3b. LICORS not only accurately estimates $S(\mathbf{r}, t)$, but also learns the prediction rule (19) from the observed field $X(\mathbf{r}, t)$.

5.1 Forecasting Competition: AR, VAR, and LICORS

A brute-force approach to spatio-temporal prediction would treat the whole spatial configuration at any one time as a single high-dimensional vector, and then use ordinary, parametric time-series methods such as vector auto-regressions (VAR) (Lütkepohl, 2007), or non- or semi- non-parametric models (Bosq, 1998; Fan and Yao, 2003). Such global approaches suffer under the curse of dimensionality: real data sets may contain millions of space-time points, so fitting global models becomes impractical, even with strong regularization (Bosq and Blanke, 2007). Moreover, such global models will not be good representations of complex spatial dynamics.

On the other hand, space can be broken up into small patches (in the limit, single points), and then one can fit standard time series models to each patch's low-dimensional time series. Such local strategies (partially) lift the curse of dimensionality, and thus make VAR or non-parametric time-series prediction practical, but creates the problem of selecting good sizes and shapes for these patches, and ignores spatial dependence across patches.

To show how LICORS escapes this dilemma, we compare it to other forecasting techniques in a simulation. Using 100 replications of (19) – (20), with $n = 100$ points in space, and $T = 200$ steps in time, we compared LICORS, with and without pre-

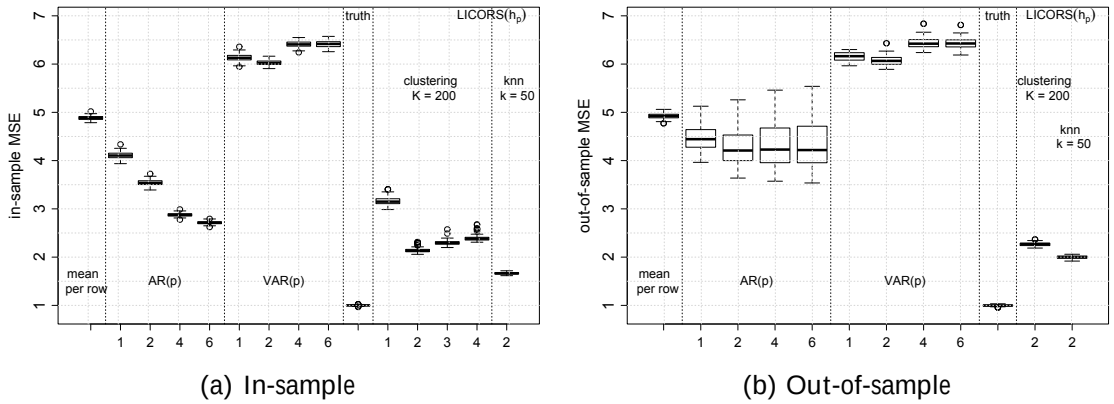

Figure 4: MSEs for LICORS and parametric competitors on (19)–(20). LICORS with pre-clustering used $K = 200$ clusters and varying past horizons; LICORS without pre-clustering use $k = 50$ neighbors and $h_p = 2$; both variants fixed $\alpha = 0.05$.

clustering, to (a) the empirical time-average of each spatial point; (b) a separate, univariate $AR(p)$ model for each point; a (c) separate $VAR(p)$ for each non-overlapping spatial patch of 5 points; and the true conditional expectation function. (See §B.1 in the Supplemental Information for details of the competing methods.)

Figure 4 shows for each predictor the estimated mean squared error (MSE) for the in-sample (Fig. 4a) as well as out-of-sample (Fig. 4b) one-step ahead prediction error. Splitting up space while using standard methods appears not to help and may even hurt. LICORS performs best among all methods, once $h_p \geq 2$. While pre-clustering performs worse than direct estimation, it still predicts much better than the other methods.

Overall, LICORS with $h_p = 2$ gives the best forecasts, where $\alpha = 0.05$ was set in advance. At no point did we make an assumption about the number of predictive states or the shape of the conditional distribution. Even though the true system

is conditionally Gaussian, LICORS out-performed the parametric Gaussian models. Thus we expect to do even better on non-Gaussian fields.

Even though we know the true light cone size in simulations, the "true" α can not be obtained directly. It controls the number of estimated predictive states: larger α implies less merging of clusters, and thus more number of predictive states; smaller α leads to more merging and hence less states.

In practice, one does not know the true light cone size nor the true number of states; they are rather control settings which affect the predictive performance. As we can accurately measure predictive performance by out-of-sample MSE, we propose a cross-validation (CV) procedure to tune h_p and α.

5.2 Cross-validation to Choose Optimal Control Settings

A good method should learn the invariant predictive structures of the system, avoiding over-fitting to the accidents of the observed sample. Ideally, the method should estimate nearly the same predictive states from (almost) any two realizations of the same system, while still being sensitive to differences between distinct systems.

Cross-validation is the classic way to handle this sensitivity-stability trade-off, and we use a data-set splitting version of it here. We simply divide the data set at its mid-point in time, use its earlier half to find predictive states, and evaluate the states' performance on the data's later half; see Supplemental Figure 7. (Assumption 2.2, of conditional stationarity, is important here.) While quite basic, simulations show that it does indeed find good control settings.

Using the same realizations of the model system as in the forecasting competition,

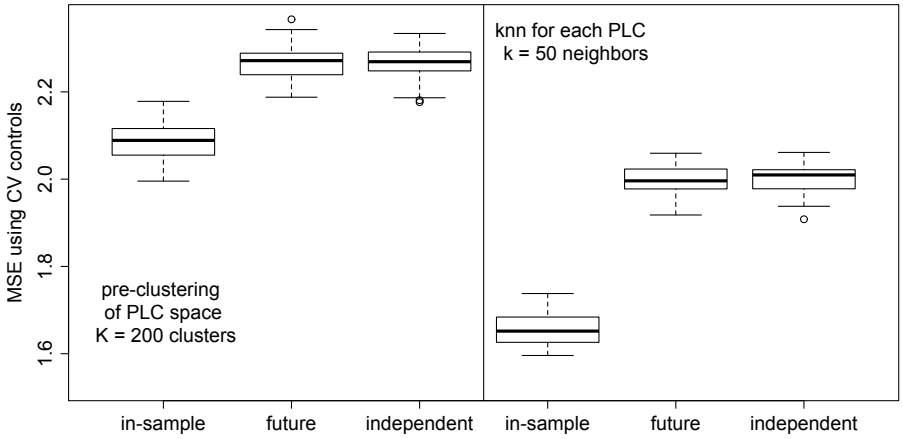

Figure 5: Cross-validation for LICORS: MSE, using the CV-picked control settings, on the first half of each realization ("in-sample"), on the second half ("future"), and on all of an independent realization ("independent").

we tried all combinations of $h_p \in \{1, 2, 3\}$ and $\alpha \in \{0.3, 0.2, 0.15, 0.1, 0.05, 0.01, 0.001\}$. We picked the control settings to do well on the continuation of the sample realization, but since this is a simulation, we can also check that these settings perform well on an *independent* realization of the same process. Figure 5 compares, for the selected control settings, the in-sample MSE on the first half of each realization, the MSE on the second half, and the MSE on all of a completely independent realization, for both the direct and the pre-clustered versions of LICORS. (As before, direct estimation does a bit better than pre-clustering.) There is little difference between the MSEs on the continuation of the training data and on independent data, indicating little over-fitting to accidents of particular sample paths. (See §B.2 in the supplemental information for further details.) Notably, CV picked the optimal h_p, namely 2, on all 100 trials.

As expected, the smaller the value of α picked by CV, the more merging between

clusters, and the smaller the number of states (see Supplemental Figure 8). Here, the true number of states $m = 7$, but both pre-clustering and direct estimation give much higher $\widehat{m}$ (10–30 with pre-clustering, 30–90 without). The gap appears to be due to cross-validating pushing (in this context) for lower approximation error and more states, rather than fewer states and lower estimation error (§B.2 in the supplemental information). Having $\widehat{m}$ be substantially larger than m thus does not degrade out-of-sample predictions.

6 Discussion

6.1 Related Work

Predictive state reconstruction estimates the prediction processes introduced by Knight (1975). Knight's construction is for stochastic processes X with a single, continuous time index; but since X_t can take values in infinite-dimensional spaces, most useful spatial models can implicitly be handled in this way, and by considering discrete time we avoid many measure-theoretic complications. After Knight, the same basic construction of the prediction process was independently rediscovered in nonlinear dynamics and physics (Crutchfield and Young, 1989; Shalizi and Crutchfield, 2001), in machine learning (Jaeger, 2000; Littman et al., 2002; Langford et al., 2009), and in the philosophy of science (Salmon, 1984).

Spatio-temporally local prediction processes were introduced in Shalizi (2003); Shalizi et al. (2004) to study self-organization and system complexity, along lines suggested by Grassberger (1986); Crutchfield and Young (1989). A related proposal

was made by was made by Parlitz and Merkwirth (2000), and light cones have been used in stochastic models of crystallization (Capasso and Micheletti, 2002), going back to Kolmogorov (1937).

While the prediction-process formalism allows for continuous-valued observable fields, the prior work by Shalizi *et al.* only gave procedures for discrete-valued fields. Jänicke *et al.* used those procedures on continuous-valued data by discretizing them (Jänicke, 2009; Jänicke and Scheuermann, 2010; Jänicke et al., 2007). We avoid discretization by using methods to estimate and compare continuous, high-dimensional distributions.

6.2 Conclusion

We introduce a new non-parametric method, LICORS, for spatio-temporal prediction. LICORS learns the predictive geometry in the state space underlying the system, by clustering observations according to the similarity of their local predictive distributions. Together with our cross-validation scheme, LICORS is a fully data-driven, non-parametric method to learn and use the non-linear, high-dimensional dynamics of a large class of spatio-temporal systems. The good performance of the CV procedure (Fig. 7) suggests that using it to find control settings in applications will avoid over-fitting.

Under weak assumptions, LICORS consistently estimates predictive distributions. Simulations show that it largely outperforms standard prediction methods. We have motivated presented results for $(1 + 1)D$ fields, but both the theory and practice extend without modification to higher-dimensional fields. While it will be

good to extend LICORS to handle continuous predictive states, and to derive theoretical guarantees about its behavior under cross-validation, it can already be applied to experimental data. It provides a powerful, principled tool for forecasting complex spatio-temporal systems.

References

Denis Bosq. *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer-Verlag, Berlin, second edition, 1998.

Denis Bosq and Delphine Blanke. *Inference and Prediction in Large Dimensions*. Wiley, New York, 2007.

S. Caires and J. A. Ferreira. On the non-parametric prediction of conditionally stationary sequences. *Statistical Inference for Stochastic Processes*, 8:151–184, 2005. doi: 10.1007/s11203-004-0383-2. Correction, vol. 9 (2006), 109–110.

Vincenzo Capasso and Alessandra Micheletti. Stochastic geometry of spatially structured birth and growth processes: Application to crystallization processes. In Ely Merzbach, editor, *Topics in Spatial Stochastic Processes*, pages 1–39, Berlin, 2002. Springer-Verlag.

Noel A. C. Cressie and Christopher K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, New York, 2011.

James P. Crutchfield and Karl Young. Inferring statistical complexity. *Physical Re-*

view Letters, 63:105–108, 1989. URL http://www.santafe.edu/~cmg/compmech/pubs/ISCTitlePage.htm.

Jianqing Fan and Qiwei Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, Berlin, 2003.

Roberto Fernández and Grégory Maillard. Chains with complete connections: General theory, uniqueness, loss of memory and mixing properties. *Journal of Statistical Physics*, 118:555–588, 2005. doi: 10.1007/s10955-004-8821-5. URL http://arxiv.org/abs/math/0305026.

Bärbel Finkenstädt, Leonhard Held, and Valerie Isham, editors. *Statistical Methods for Spatio-Temporal Systems*. Chapman and Hall/CRC, Boca Raton, Florida, 2007.

Peter Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25:907–938, 1986.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel method for the two sample problem. In *Advances in neural information processing systems 19*, pages 513–520. MIT Press, 2007.

Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12:1371–1398, 2000. URL http://www.faculty.iu-bremen.de/hjaeger/pubs/oom_neco00.pdf.

Heike Jänicke. *Information Theoretic Methods for the Visual Analysis of Climate and Flow Data.* PhD thesis, Universität Leipzig, 2009.

Heike Jänicke and Gerik Scheuermann. Towards automatic feature-based visualization. In Hans Hagen, editor, *Scientific Visualization: Advanced Concepts*, volume 1 of *Dagstuhl Follow-Ups*, pages 62–77. Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2010. doi: 10.4230/DFU.SciViz.2010.62. URL `http://drops.dagstuhl.de/opus/volltexte/2010/2697`.

Heike Jänicke, Alexander Wiebel, Gerik Scheuermann, and Wolfgang Kollmann. Multifield visualization using local statistical complexity. *IEEE Transactions on Visualization and Computer Graphics*, 13:1384–1391, 2007. doi: 10.1109/TVCG.2007.70615. URL `http://www.informatik.uni-leipzig.de/bsv/Jaenicke/Papers/vis07.pdf`.

Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, England, second edition, 2004.

Frank B. Knight. A predictive view of continuous time processes. *Annals of Probability*, 3:573–596, 1975. URL `http://projecteuclid.org/euclid.aop/1176996302`.

Andrei N. Kolmogorov. A statistical theory for the recrystallization of metals. *Bulletin of the Academy of Sciences, USSR, Physical Series*, 1:355–359, 1937. In Russian.

Solomon Kullback. *Information Theory and Statistics*. Dover Books, New York, 2nd edition, 1968.

John Langford, Ruslan Salakhutdinov, and Tong Zhang. Learning nonlinear dynamic models. Electronic preprint, 2009. URL http://arxiv.org/abs/0905.3369.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pages 1555–1561, Cambridge, MA, 2002. MIT Press. URL http://www.eecs.umich.edu/~baveja/Papers/psr.pdf.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2007.

Ulrich Parlitz and Christian Merkwirth. Prediction of spatiotemporal time series based on reconstructed local states. *Physical Review Letters*, 84:1890–1893, 2000.

Maria L. Rizzo and Gábor J. Székely. DISCO analysis: A nonparametric extension of analysis of variance. *Annals of Applied Statistics*, 4:1034–1055, 2010. doi: 10.1214/09-AOAS245. URL http://arxiv.org/abs/1011.2288.

Paul R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society Series B*, 67:515–530, 2005.

Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, 1984.

Cosma Rohilla Shalizi. Optimal nonlinear prediction of random fields on networks.

Discrete Mathematics and Theoretical Computer Science, AB(DMCS):11–30, 2003. URL http://arxiv.org/abs/math.PR/0305160.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:817–879, 2001. URL http://arxiv.org/abs/cond-mat/9907176.

Cosma Rohilla Shalizi, Kristina Lisa Klinkner, and Robert Haslinger. Quantifying self-organization with optimal predictors. *Physical Review Letters*, 93:118701, 2004. URL http://arxiv.org/abs/nlin.AO/0409024.

Supplementary Material for "LICORS: Light Cone Reconstruction of States for Non-parametric Forecasting of Spatio-Temporal Systems"

A Predictive States: Details on Methodology, Implementation, and Algorithms

We partition the observed PLCs $\{\ell_i^-\}_{i=1}^N \subset \mathbb{R}^{n_p}$ into $K = K(\delta)$ disjoint groups $\{P_k\}_{k=1}^K$, choosing the number of groups so that all have diameters less than δ. This choice of $K(\delta)$ guarantees (Assumption 3.1) that all $\ell^- \in P_k$ have predictive distributions that are at most ρ apart. Thus all PLCs within a group P_k are (nearly) equivalent by Definition 2.1. This in turn means we only need to compare predictive distributions between clusters.

A.1 Lebesgue Smoothing

In a standard kernel regression approach one would compute a similarity measure on PLCs ℓ_i^- and then use a weighted mean of FLCs ℓ_i^+ to get a point prediction of the future cone, i.e.,

$$\widehat{L^+}(\mathbf{r}, t) = \sum_{(\mathbf{q},\tau)} w^-(\mathbf{q}, \tau; (\mathbf{r}, t))\, \ell^+(\mathbf{q}, \tau) \text{ for all } \tau < t, \tag{21}$$

1. Collect the PLC and FLC configurations, $\ell^-(\mathbf{r},t)$ and $\ell^+(\mathbf{r},t)$, for each $(\mathbf{r},t)$ in the observed field $X(\mathbf{r},1),\ldots,X(\mathbf{r},T)$.
2. To cluster or not to cluster:
   (a) Assign each point to its own cluster. Only for small N this is computationally feasible.
   (b) Perform an initial clustering (e.g. (e.g., K-means++ (Arthur and Vassilvitskii, 2007)) in the PLC configuration space (Section 3.1).
3. For each pair of clusters, test whether the estimated conditional FLC distributions are significantly different, at some fixed level α (Section 3.2). If not, merge them and go on. Stop when no more merges are possible.
4. Treat the remaining clusters as predictive states, and estimate the conditional distributions over FLC configurations.
5. Return the partition of PLC configurations into predictive states, and the associated predictive distributions.

Figure 6: Estimating predictive states from continuous-valued data: in 2a conditional distributions are tested for each ℓ_i^-, $i=1,\ldots,N$, using a δ-neighborhood (or k nearest neighbors) of ℓ_i^- (see Section 4.1.1 for details); 2b uses an initial clustering to reduce complexity of the testing problem from $\mathcal{O}(N^2)$ to $\mathcal{O}(K^2)$ (see also Section 3.1).

where $w^-(\mathbf{q},\tau;(\mathbf{r},t)) \propto K^-(\|\ell^-(\mathbf{r},t)-\ell^-(\mathbf{q},\tau)\|)$ are normalized weights determined by a kernel $K^-(\cdot)$ in the PLC configuration space. For example, a Gaussian kernel $K_h^-(\ell^-(\mathbf{r},t),\ell^-(\mathbf{q},\tau)) = \exp(-\frac{1}{h}\|\ell^-(\mathbf{r},t)-\ell^-(\mathbf{q},\tau)\|_2^2)$ with squared Euclidean distance and bandwidth h.

Since $(\mathbf{q},\tau)$ ranges over the entire space-time, $\mathbf{q}\in\mathbf{S}$, $\tau=1,\ldots,t-1$, computing this many similarities $\{s_{(\mathbf{q},\tau),(\mathbf{r},t)}\}$ becomes very time consuming. A typical 10-second

Algorithm 1 Test equality of conditional predictive FLC distributions $\mathbb{P}\left(L^+ \mid \text{clusterID} = k\right)$

Require:

Data:		
$\mathbf{F} = \{\ell_i^+\}_{i=1}^N \in \mathbb{R}^{N \times n_f}$	...	array with FLCs
clusterID	...	labels of the PLC partitioning (step 2a or 2b in Fig. 6)
Parameters:		
$\alpha \in [0, 1]$	...	significance level α for testing $H_0 : \mathbb{P}\left(L^+ \mid \ell_i^-\right) = \mathbb{P}\left(L^+ \mid \ell_j^-\right)$

Ensure:

$k_{\max} = \max$ clusterID
for $k = 1, \ldots, k_{\max}$ **do**
  fetch FLC samples given partition P_k: $\mathbf{F}_k = \{\ell_i^+\}_{\{i|clusterID[i]==k\}}$
  j = k
  lasttested = 0
  pvalue = 1
  while pvalue $> \alpha$ or $j \leq k_{\max}$ **do**
    j = j+ 1
    lasttested = j
    fetch FLC samples given partition P_j: $\mathbf{F}_j = \{\ell_i^+\}_{\{i|clusterID[i]==j\}}$
    pvalue $\leftarrow test(\mathbb{P}\left(L^+ \mid P_k\right) = \mathbb{P}\left(L^+ \mid P_j\right) \mid \mathbf{F}_k, \mathbf{F}_j)$
    if pvalue $< \alpha$ **then**
      merge cluster j with cluster k: clusterID[clusterID == j] = k
After no merging is possible clusterID contains the labels of the predictive states.
return clusterID

video might have $N = 3 \cdot 10^7$ space-time points.[1] To evaluate (21) needs $3 \cdot 10^7$ similarities in n_p-dimensional space — and this just to predict one FLC. If N is large, then predictive state estimation is a necessary pre-step before making predictions.

Our approach differs in two important ways. First, we assume a discrete predictive state space which is sufficient to predict the future. Thus once we have estimated the predictive states $\epsilon_1, \ldots, \epsilon_m$, we can predict the field at any $(\mathbf{r}, t)$ using the average

[1] 25 frames per second and 300×400 pixels.

(or mode) of the estimated predictive state at $(\mathbf{r}, t)$,

$$\widehat{L^+}(\mathbf{r}, t) = \mathbb{E}_{\widehat{c}(\ell^-(\mathbf{r},t))}\left(L^+(\mathbf{r}, t)\right). \tag{22}$$

Second, we learn a new geometry on the PLC space by defining closeness in the FLC distribution space, rather than in the PLC configuration space. Thus a natural continuous state space extension of (22) is a Kernel regression with weights that depend on the similarity in the output rather than the input space, i.e.

$$\widehat{L^+}(\mathbf{r}, t) = \sum_{(\mathbf{q},\tau)} w^+(\mathbf{q}, \tau; (\mathbf{r}, t))\, \ell^+(\mathbf{q}, \tau) \text{ for all } \tau < t, \tag{23}$$

where the normalized weights $w^+(\mathbf{q}, \tau; (\mathbf{r}, t)) \propto K^+(\|\mathbb{P}(\ell^+(\mathbf{r}, t)) - \mathbb{P}(\ell^+(\mathbf{q}, \tau))\|)$ are based on a Kernel $K^+(\cdot)$ in the FLC distribution space.

One can generalize (23) to the classic non-parametric regression setting $y = m(x) + u$ and define a new Kernel regression estimator as

$$\widehat{m}^{(L)}(x) = \sum_{i=1}^{n} \frac{K_y\left(\widehat{m}^{(R)}(x_i) - m(x)\right)}{h_y} y_i\ , \tag{24}$$

where $\widehat{m}^{(R)}(\cdot)$ serves as a pilot estimate; for example the classic kernel regression smoother

$$\widehat{m}^{(R)}(x) = \sum_{i=1}^{n} \frac{K_x(x_i - x)}{h_x} y_i\ . \tag{25}$$

As we average over nearby predictions rather than nearby inputs, we may call (24) "Lebesgue smoothing", in contrast to the "Riemann" smoothing of (25). If N is

small, then we can forecast with (23) forecast without estimating predictive states. However, here we focus on predictive-state recovery, and leave Lebesgue smoothed LICORS to future work.

Further performance enhancements for testing While it is better to do $\mathcal{O}(K^2)$ high-dimensional tests than $\mathcal{O}(N^2)$, it would be better still to speed up each test. Since two distributions are the same only if their moments are, we can start by testing simply for equality of means, which is fast and powerful, and do a full distributional test only if we cannot reject on that basis. For multivariate mean tests we can use the Hotelling test (Abello et al., 1998) and its randomized generalization (Lopes et al., 2011). Yet another strategy to reduce the number of costly high-dimensional, non-parametric tests is to test various functions $f(\cdot)$ of the samples. If the distributions of $\mathbf{F}_{k_i}(\delta)$ and $\mathbf{F}_{k_j}(\delta)$ are the same, then also $\mathbb{P}\left(f\left(\mathbf{F}_{k_i}(\delta)\right)\right) = \mathbb{P}\left(f\left(\mathbf{F}_{k_j}(\delta)\right)\right)$ for any measurable f. Particularly, we can apply random projections (Lopes et al., 2011) to $\mathbf{F}_{k_i}$ to go from the high-dimensional $\mathbb{R}^{n_f}$ down to the one-dimensional $\mathbb{R}$, followed by a Kolmogorov-Smirnov test. Only if these tests can not reject equality for several projections, one uses fully non-parametric tests.

1. Split dataset at its middle in time: $\mathcal{D}_1 = \{X(\mathbf{r}, t)\}_{t=1}^{T/2}$ and $\mathcal{D}_2 = \{X(\mathbf{r}, t)\}_{t=T/2+1}^{T}$
2. For each combination of control settings, do:
   (a) Training: estimate predictive states from $\mathcal{D}_1$
   (b) Test-set prediction: find predictive state of each PLC $\in \mathcal{D}_2$ and predict its FLC $\in \mathcal{D}_2$.
   (c) Error: compare to the observed FLCs $\in \mathcal{D}_2$ and compute the loss.
3. Choose the control settings with the smallest test-set loss.

Figure 7: Cross-validation to choose control settings given data $\{X(\mathbf{r}, t)\}_{t=1}^{T}$.

B The Simulation and the Forecasting Competition

B.1 Details of Competing Methods

The local VAR models were fit with Lasso regularization (Song and Bickel, 2011), as implemented in the `fastVAR` package (Wong, 2012). We also tried un-regularized VAR models, but they performed even worse.

B.2 Excess Risk, Test Size, and Number of Estimated States

Figs. 8a and 8b show the expected relationship between α and the number of predictive states recovered $\widehat{m}$: smaller α leads to more merging, and fewer states. Here the true number of states $m = 7$, but both pre-clustering and direct estimation give much higher $\widehat{m}$. Thus for LICORS, optimal forecasting pushes for more states and

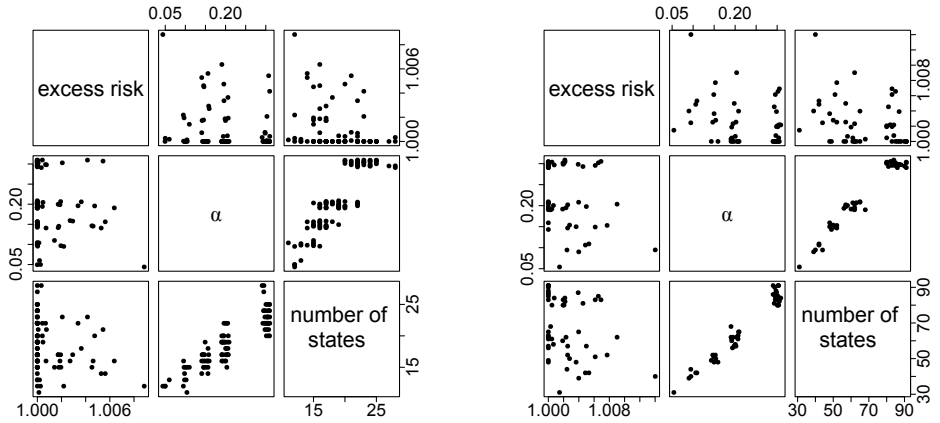

Figure 8: Relations between excess risk, test size, and the number of reconstructed states for LICORS: (a) selected α, number of estimated states, and excess risk (Eq. (26)) for pre-clustered LICORS; (b) the same for direct-estimation LICORS. α values in are jittered.

lower approximation error, rather than fewer states and lower estimation error. We can check this explanation by considering the ratio

$$\text{excess risk} := \frac{\text{MSE(sample } i+1\text{) using } (h_p, \alpha)_{i,CV_i}}{\text{MSE(sample } i+1\text{) using } (h_p, \alpha)_{i+1,\min}} \geq 1. \tag{26}$$

Recall that $(h_p, \alpha)_{i,CV_i}$ is chosen using only sample i, while $(h_p, \alpha)_{i+1,\min}$ is the minimizing pair after having evaluated the MSE on sample $i+1$. The best that any data-driven procedure could do would be to guess $(h_p, \alpha)_{i+1,\min}$ from sample i, so the excess risk is ≥ 1, with equality only if CV picked the optimal control settings. The scatter-plots show that our CV procedure has an excess risk on the order of 10^{-2} compared to the oracle pair. Hence, even though $\widehat{m}$ is substantially larger than m, the difference is practically irrelevant for predictions.

B.3 Discussion of the Simulations

The simulations showed that LICORS outperforms standard forecasting techniques by a large margin, even though it presumes very little about the data source. Especially note that the out-of-sample MSE in Fig. 5 is still much lower than the best parametric in-sample MSE in Fig. 4a — even though it uses only half the sample size. The good performance of the CV procedure (Fig. 7) suggests that using it to find control settings in applications will avoid over-fitting.

In real applications N would typically on the order of millions (rather than merely 2×10^4), making pre-clustering essential computationally — at least until $\mathcal{O}\left(N^2\right)$ comparisons for millions of data points become tractable. Pre-clustering usually leads to a performance loss as it hides fine structures in the predictive distribution space (see also the remark below Lemma 4.7). However, the in-sample and out-of-sample MSE comparison showed that this performance loss is small compared to the gain over standard parametric methods, and further attenuated with CV.

C Proofs

Proof of Lemma 4.5.

$$\begin{aligned}
&\mathbb{P}\left(L^+\left(\mathbf{r},t\right), L^+\left(\mathbf{u},s\right) \mid S\left(\mathbf{r},t\right), S\left(\mathbf{u},s\right)\right) && (27)\\
&\quad = \mathbb{P}\left(L^+\left(\mathbf{r},t\right) \mid L^+\left(\mathbf{u},s\right), S\left(\mathbf{r},t\right), S\left(\mathbf{u},s\right)\right) \mathbb{P}\left(L^+\left(\mathbf{u},s\right) \mid S\left(\mathbf{r},t\right), S\left(\mathbf{u},s\right)\right) && (28)\\
&\quad = \mathbb{P}\left(L^+\left(\mathbf{r},t\right) \mid L^+\left(\mathbf{u},s\right), S\left(\mathbf{r},t\right), S\left(\mathbf{u},s\right)\right) \mathbb{P}\left(L^+\left(\mathbf{u},s\right) \mid S\left(\mathbf{u},s\right)\right) && (29)\\
&\quad = \mathbb{P}\left(L^+\left(\mathbf{r},t\right) \mid S\left(\mathbf{r},t\right)\right) \mathbb{P}\left(L^+\left(\mathbf{u},s\right) \mid S\left(\mathbf{u},s\right)\right) \ , && (30)
\end{aligned}$$

The first equality is simple conditioning, the second equality holds since given the predictive state at $(\mathbf{u}, s)$ the distribution of L^+ is independent of the predictive state at another $(\mathbf{r}, t)$, and the last equality holds for the same reason as the second plus the non-overlap of the FLCs at $(\mathbf{r}, t)$ and $(\mathbf{u}, s)$. □

Proof of Corollary 4.6. The FLC of $(\mathbf{r}, t)$ with $h_f = 0$ is just the single point $X(\mathbf{r}, t)$. Since two univariate FLCs cannot overlap unless they are equal, the result follows immediately from Lemma 4.5. □

Proof of Lemma 4.7. By contradiction. Assume that ℓ_j^- and ℓ_k^-, with $j, k \in I_i(\delta)$, have different predictive states, without loss of generality ϵ_1 and ϵ_2. By Assumption 4.4, then, $\mathcal{D}_{KL}(\epsilon_1 \,||\, \epsilon_2)$ and $\mathcal{D}_{KL}(\epsilon_2 \,||\, \epsilon_1)$ are both at least $d_{\min}$. By the definition of $I_i(\delta)$, $\|\ell_j^- - \ell_k^-\| < 2\delta$. By Assumption 3.1, then, $\mathcal{D}_{KL}(\epsilon_1 \,||\, \epsilon_2)$ and $\mathcal{D}_{KL}(\epsilon_2 \,||\, \epsilon_1)$ are both at most $\rho(2\delta)$. But by making δ sufficiently small, $\rho(2\delta)$ can be made as small as desired, and in particular can be made less than $d_{\min}$. This is a contradiction, so all the past cone configurations in $I_i(\delta)$ must be predictively equivalent. □

Proof of Corollary 4.8. Immediate from combining Lemmas 4.7 and 4.5. □

Proof of Theorem 4.12. Before going into the formal proof, we make an observation regarding non-parametric two-sample tests. Most of these, to have good operating characteristics, require independent samples. Since we will be applying the tests to $\mathbf{F}_i(\delta)$ and $\mathbf{F}_j(\delta)$,

Properties C.1 (Pairwise independent samples)**.** *If*

$$I_i(\delta) \cap I_j(\delta) = \emptyset. \tag{31}$$

then the samples $\mathbf{F}_i(\delta)$ *are independent of* $\mathbf{F}_j(\delta)$, $j \neq i$ *(see* (14)*).*

Let $\Delta_{ij} := \|\ell_i^- - \ell_j^-\|$. If $\Delta_{ij} > 2\delta$, then (31) is satisfied. If $\Delta_{ij} < 2\delta$, then a sample in $\mathbf{F}_i(\delta)$ might also appear in $\mathbf{F}_j(\delta)$ and therefore violate the independence assumption for two sample tests.

For these rare cases redefine the index set $I_i(\delta)$ and $I_j(\delta)$ such that (31) holds. We can achieve this by excluding the intersection, split it in half (± 1 sample), and then re-assign these halves to each index set. For all pairs $i \neq j$ determine $I_i(\delta) \cap I_j(\delta) =: I_{i \cap j}(\delta)$. Then let

$$I_i := I_i \setminus I_{i \cap j} \cup \{i_1, \ldots, i_{|I_{i \cap j}|/2} \mid i_k \in I_{i \cap j}\} \tag{32}$$

$$\text{and } I_j := I_j \setminus I_{i \cap j} \cup \{i_{|I_{i \cap j}|/2}, \ldots, i_{|I_{i \cap j}|} \mid i_k \in I_{i \cap j}\}. \tag{33}$$

If $I_{i \cap j} = \emptyset$, (32)–(33) does not change the index set; if $I_{i \cap j} \neq \emptyset$, then (32)–(33) guarantees an empty intersection.

The proof of consistency relies crucially on a growing index set I_i. The re-definition in (32)–(33) does not change the rate at which $S_i(N, \delta)$ grows, because in the worst case (for very close PLCs) it just divides $s_i(N, \delta)$ and $s_j(N, \delta)$ in half.

Proof: We first bound the error for each row $\widehat{\mathbf{A}}_i$, and then use a union bound for the probability of error for $\widehat{\mathbf{A}}$.

Bound error per row For each row $T_{n,m}$ tests $H_0 : \ell_i \sim \ell_j$, $j > i$ (due to symmetry the cases $j < i$ have already been tested before) based on the sample $\mathbf{F}_i(\delta) \sim \epsilon_i$ and $\mathbf{F}_j(\delta) \sim \epsilon_j$. The worst-case distance d for the non-parametric test in Assumption

4.11 is $d = d_{\min}$. For simplicity consider the first row: here we have to make $N - 1$ tests, of which $N_1 - 1$ should correctly accept, and $N - N_1$ should correctly reject equality of distributions.

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mathbf{A}}_j \neq \mathbf{A}_j\right) &\leq (N_j - 1)\mathbb{P}\left((\text{type I}) + (N - N_j)\right)\mathbb{P}\left(\text{type II}\right) && (34)\\
&\leq (N_j - 1)\alpha + (N - N_j)\beta\left(\alpha, S_{\min}(N,\delta), S_{\min}(N,\delta)\right) && (35)\\
&\leq N_j\alpha + (N - N_j)\beta\left(\alpha, S_{\min}(N,\delta), S_{\min}(N,\delta)\right) && (36)
\end{aligned}$$

since the worst case, for type II error, is that both samples are as small as possible.

Bound error for entire matrix The probability of error for the entire predictive state clustering can again be bounded using the union bound:

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mathbf{A}} \neq \mathbf{A}\right) &= \mathbb{P}\left(\bigcup_{j=1}^{N}\{\widehat{\mathbf{A}}_j \neq \mathbf{A}_j\}\right) && (37)\\
&\leq \sum_{j=1}^{N}\mathbb{P}\left(\widehat{\mathbf{A}}_j \neq \mathbf{A}_j\right) && (38)\\
&\leq N\left(N_{\max}\alpha + (N - N_{\min})\beta\left(\alpha, S_{\min}(N,\delta), S_{\min}(N,\delta)\right)\right) && (39)\\
&= NN_{\max}\alpha + (N^2 - NN_{\min})\beta\left(\alpha, S_{\min}(N,\delta), S_{\min}(N,\delta)\right), && (40)
\end{aligned}$$

where $N_{\max} = \max_j N_j$ is the number of light cones in the largest predictive state.

Under Assumption 4.11, α and β are both $\mathrm{o}\big(NN_{\max}\big)$, so the over-all arrow probability tends to zero. □

Supplementary References

J. Abello, A. Buchsbaum, and J. Westbrook. A functional approach to external graph algorithms. In *Proceedings of the 6th European Symposium on Algorithms*, Berlin, 1998. Springer.

David Arthur and Sergei Vassilvitskii. `k-means++`: The advantages of careful seeding. In Harold Gabow, editor, *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms [SODA07]*, pages 1027–1035, Philadelphia, 2007. Society for Industrial and Applied Mathematics. URL `http://www.stanford.edu/~darthur/kMeansPlusPlus.pdf`.

Miles E. Lopes, Laurent J. Jacob, and Martin J. Wainwright. A more powerful two-sample test in high dimensions using random projection. Electronic pre-print, arxiv.org, 2011. URL `http://arxiv.org/abs/1108.2401`.

Song Song and Peter J. Bickel. Large vector auto regressions, 2011. URL `http://arxiv.org/abs/1106.3915`.

Jeffrey Wong. *fastVAR*, 2012. URL `http://CRAN.R-project.org/package=fastVAR`. R package version 1.2.1.