Mixed LICORS: A Nonparametric Algorithm for Predictive State Reconstruction

Georg M. Goerg **Cosma Rohilla Shalizi**
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213
{ gmg, cshalizi } @stat.cmu.edu

Abstract

We introduce *mixed LICORS*, an algorithm for learning nonlinear, high-dimensional dynamics from spatio-temporal data, suitable for both prediction and simulation. Mixed LICORS extends the recent LICORS algorithm (Goerg and Shalizi, 2012) from hard clustering of predictive distributions to a non-parametric, EM-like soft clustering. This retains the asymptotic predictive optimality of LICORS, but, as we show in simulations, greatly improves out-of-sample forecasts with limited data. The new method is implemented in the publicly-available R package LICORS.

1 Introduction

Recently Goerg and Shalizi (2012) introduced light cone reconstruction of states (LICORS), a nonparametric procedure for recovering predictive states from spatio-temporal data. Every spatio-temporal process has an associated, latent prediction process, whose measure-valued states are the optimal local predictions of the future at each point in space and time (Shalizi, 2003). LICORS consistently estimates this prediction process from a single realization of the manifest space-time process, through agglomerative clustering in the space of predictive distributions; estimated states are clusters. This converges on the minimal set of states capable of optimal prediction of the original process.

Experience with other clustering problems shows that soft-threshold techniques often predict much better than hard-threshold methods. Famously, while k-means (Lloyd, 1982) is very fast and robust, the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) in Gaussian mixture models gives better clustering results. Moreover, mixture models allow clusters to be interpreted probabilistically, and new data to be assigned to old clusters.

With this inspiration, we introduce *mixed LICORS*, a soft-thresholding version of (hard) LICORS. This embeds the prediction process framework of Shalizi (2003) in a mixture-model setting, where the predictive states correspond to the optimal mixing weights on extremal distributions, which are themselves optimized for forecasting. Our proposed nonparametric, EM-like algorithm then follows naturally.

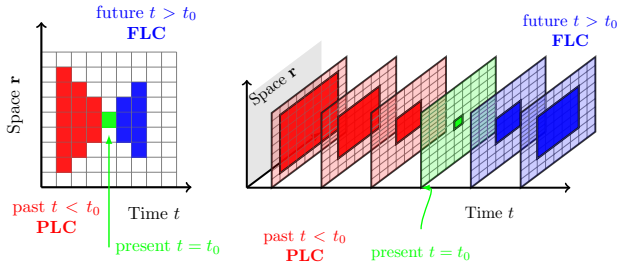
After introducing the prediction problem and fixing notation, we explain the mixture-model and hidden-state-space interpretations of predictive states (§2). We then present our nonparametric EM algorithm for estimation, with automatic selection of the number of predictive states (§3), and the corresponding prediction algorithm (§4). After demonstrating that mixed LICORS predicts better out of sample than hard-clustering procedures (§5), we review the proposed method and discuss future work (§6).

2 A Predictive States Model for Spatio-temporal Processes

We fix notation and the general set-up for predicting spatio-temporal processes, following Shalizi (2003). We observe a random field $X(\mathbf{r}, t)$, discrete- or continuous-valued, at each point $\mathbf{r}$ on a regular spatial lattice $\mathbf{S}$, at each moment t of discrete time $\mathbb{T} = 1 : T$, or $N = T|\mathbf{S}|$ observational in all. The field is $(d+1)D$ if space $\mathbf{S}$ is d dimensional (plus 1 for time); video is $(2+1)D$. $\|\mathbf{r} - \mathbf{u}\|$ is a norm (e.g., Euclidean) on the spatial coordinates $\mathbf{r}, \mathbf{u} \in \mathbf{S}$.

To optimally predict an unknown (future) $X(\mathbf{r}, t)$ given (past) observed data $\mathcal{D} = \{X(\mathbf{s}, u)\}_{\mathbf{s}\in\mathbf{S}, u\in\mathbb{T}}$, we need to know $\mathbb{P}(X(\mathbf{r}, t) \mid \mathcal{D})$. Estimating this conditional distribution is quite difficult if $X(\mathbf{r}, t)$ can de-

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.



(a) $(1+1)D$ field (b) $(2+1)D$ field

Figure 1: Geometry of past ($\ell^-(\mathbf{r},t)$) and future ($\ell^+(\mathbf{r},t)$) light cones, with $h_p = 3$, $h_f = 2$, $c = 1$.

pend on all variables in $\mathcal{D}$. Since information in a physical system can typically only propagate at a finite speed c, $X(\mathbf{r},t)$ can only depend on a subset of $\mathcal{D}$. We therefore only have to consider local, spatio-temporal neighborhoods of $(\mathbf{r},t)$ as possibly relevant for prediction.

2.1 Light Cones

The past light cone (PLC) of $(\mathbf{r},t)$ is all past events that could have influenced $(\mathbf{r},t)$,

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

Analogously the future light cone (FLC) $L^+(\mathbf{r},t)$ is all future events which could be influenced by $(\mathbf{r},t)$. In practice, we limit PLCs and FLCs to finite horizons h_p and h_f; together with c, these fix the dimensionality of light-cones; see Figure 1 for an illustration.

For physically-plausible processes, then, we only have to find the conditional distribution of $X(\mathbf{r},t)$ given its PLC $(\ell^-,(\mathbf{r},t))$. Doing this for every $(\mathbf{r},t)$ in space-time, however, is still excessive. Not only do spatio-temporal patterns recur, but different histories can have similar predictive consequences. Thus we need not find $\mathbb{P}\left(X(\mathbf{r},t) \mid (\ell^-,(\mathbf{r},t))\right)$ for each $(\mathbf{r},t)$ separately, but can first summarize PLCs by predictively sufficient statistics, $\epsilon\left(\ell^-,(\mathbf{r},t)\right)$, and then only find $\mathbb{P}\left(X(\mathbf{r},t) \mid \epsilon\left(\ell^-,(\mathbf{r},t)\right)\right)$.

We now construct the minimal sufficient statistic, $\epsilon(\ell^-(\mathbf{r},t))$, following Shalizi (2003), to which we refer for some mathematical details, and references on predictive sufficiency.

Definition 2.1 (Equivalent configurations)**.** *The past configurations ℓ_i^- at $(\mathbf{r},t)$ and ℓ_j^- at $(\mathbf{u},s)$ are* predictively equivalent*, $(\ell_i^-,(\mathbf{r},t)) \sim (\ell_j^-,(\mathbf{u},s))$, if they have the same conditional distributions for FLCs, i.e.,*

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

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

$$\epsilon(\ell^-,(\mathbf{r},t)) \equiv \left[\ell^-\right]. \quad (2)$$

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

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^-,*

$$(\ell^-,(\mathbf{r},t)) \sim (\ell^-,(\mathbf{u},s)) \quad (3)$$

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

Assumption 2.2 enables inference from a single realization of the process (as opposed to multiple independent realizations). For readability, we encode the space-time index $(\mathbf{r},t)$ by a single index $i = 1,\ldots,N$.[1]

The future is independent of the past given the predictive state (Shalizi, 2003),

$$\mathbb{P}\left(L_i^+ \mid \ell_i^-, \epsilon(\ell_i^-)\right) = \mathbb{P}\left(L_i^+ \mid \epsilon(\ell_i^-)\right) \quad (4)$$

and, by construction,

$$\mathbb{P}\left(L_i^+ \mid \ell_i^-\right) = \mathbb{P}\left(L_i^+ \mid \epsilon(\ell_i^-)\right). \quad (5)$$

Thus the prediction problem simplifies to estimating $\mathbb{P}\left(X_i \mid \epsilon(\ell_i^-)\right)$ for each i.

2.2 A Statistical Predictive-States Model

The focus on predictive distributions in constructing the predictive states does not prevent us from using them to model the joint distribution.

Proposition 2.3. *The joint pdf of $X_1,\ldots,X_{\tilde{N}}$ satisfies*

$$\mathbb{P}\left(X_1,\ldots,X_{\tilde{N}}\right) = \mathbb{P}\left(\mathbf{M}\right)\prod_{i=1}^{\tilde{N}}\mathbb{P}\left(X_i \mid \ell_i^-\right), \quad (6)$$

where **M** *is the* **margin** *of the spatio-temporal process, i.e., all points in the field that do not have a completely observed PLC.*

Proof. See Supplementary Material, §A. □

Proposition 2.3 shows that the focus on conditional predictive distributions does not restrict the applicability of the light cone setup to forecasts alone, but is in fact a generative representation for any spatio-temporal process.

[1] Appendix A in the SM gives an explicit rule (Eq. (31)) to map each $(\mathbf{r},t)$ to a unique i.

2.2.1 Minimal Sufficient Statistical Model

Given $\epsilon(\ell_i^-)$ Eq. (6) simplifies to (omitting $\mathbb{P}(\mathbf{M})$)

$$\mathbb{P}(X_1, \ldots, X_N \mid \mathbf{M}, \epsilon) = \prod_{i=1}^{N} \mathbb{P}\left(X_i \mid \ell_i^-; \epsilon(\ell_i^-)\right) = \prod_{i=1}^{N} \mathbb{P}\left(X_i \mid \epsilon(\ell_i^-)\right), \quad (7)$$

using (4). Any particular ϵ implicitly specifies the number of predictive states K, and all K predictive distributions $\mathbb{P}\left(X_i \mid \epsilon(\ell_i^-)\right)$. However, in practice only X_i and ℓ_i^- are observed; the mapping ϵ is exactly what we are trying to estimate.

2.3 Predictive States as Hidden Variables

Since there is a one-to-one correspondence between the mapping ϵ and the set of equivalence classes / predictive states $\epsilon(\ell^-)$ which are its range, Shalizi (2003) and Goerg and Shalizi (2012) do not formally distinguish between them. Here, however, it is important to keep distinction between the predictive state space $\mathcal{S}$ and the mapping $\epsilon : \ell_i^- \to \mathcal{S}$. Our EM algorithm is based on the idea that the predictive state is a hidden variable, S_i, taking values in the finite state space $\mathcal{S} = \{s_1, \ldots, s_K\}$, and the mixture weights of PLC ℓ_i^- are the soft-threshold version of the mapping $\epsilon(\ell_i^-)$. It is this hidden variable interpretation of predictive states that we use to estimate the minimal sufficient statistic ϵ, the hidden state space $\mathcal{S}$, and the predictive distributions $\mathbb{P}(X_i \mid S_i = s_j)$.

Introducing the abbreviation $\mathbb{P}_j(\cdot)$ for $\mathbb{P}(\cdot \mid S_i = s_j)$, the latent variable approach lets (7) be written as

$$\prod_{i=1}^{N} \mathbb{P}(X_i \mid S_i) = \prod_{i=1}^{N} \sum_{j=1}^{K} \mathbf{1}(S_i = s_j)\, \mathbb{P}_j(X_i). \quad (8)$$

Eq. (8) is the pdf of a K component mixture model with *complete data*, and $\mathbf{1}(S_i = s_j)$ is a randomized version of $\epsilon : \ell_i^- \mapsto S_i$.

2.4 Log-likelihood of ϵ

From (8) the complete data log-likelihood is, neglecting a $\log \mathbb{P}(\mathbf{M})$ term,

$$\ell(\epsilon; \mathcal{D}, S_1^N) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{K} \mathbf{1}(S_i = s_j)\, \mathbb{P}\left(X_i \mid \epsilon(\ell_i^-) = s_j\right)\right) = \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbf{1}(S_i = s_j) \log \mathbb{P}\left(X_i \mid \epsilon(\ell_i^-) = s_j\right), \quad (9)$$

where $S_1^N := \{S_1, \ldots, S_N\}$ and the second equality follows since $\mathbf{1}(S_i = s_j) = 1$ for one and only one j, and 0 otherwise.

The "parameters" in (9) are ϵ and K; X_i and ℓ_i^- are observed, and S_i is a hidden variable. The optimal mapping $\epsilon : L^- \to \mathcal{S}$ is the one that maximizes (9):

$$\epsilon^* = \operatorname*{argmax}_{\epsilon} \ell(\epsilon; \mathcal{D}, S_i). \quad (10)$$

Without any constraints on K or ϵ the maximum is obtained for $K = N$ and $\epsilon(\ell_i^-) = \ell_i^-$; "the most faithful description of the data is the data".[2] As this tells us nothing about the underlying dynamics, we must put some constraints on K and/or ϵ to get a useful solution. For now, assume that $K \ll N$ is fixed and we only have to estimate ϵ; in Section 3.3, we will give a data-driven procedure to choose K.

2.5 Nonparametric Likelihood Approximation

To solve (10) with $K \ll N$ we need to evaluate (9) for candidate solutions ϵ. We cannot do this directly, since (9) involves the unobserved S_i. Moreover, predictive distributions can have arbitrary shapes, so we want to use nonparametric methods, but this inhibits direct evaluation of the component distributions $\mathbb{P}\left(X_i \mid \epsilon(\ell_i^-) = s_j\right)$.

We solve both difficulties together by using a nonparametric variant of the expectation-maximization (EM) algorithm (Dempster et al., 1977). Following the recent nonparametric EM literature (Benaglia, Chauveau, and Hunter, 2011; Bordes, Chauveau, and Vandekerkhove, 2007; Hall, Neeman, Pakyari, and Elmore, 2005), we approximate the $\mathbb{P}\left(X_i \mid \epsilon(\ell_i^-) = s_j\right)$ in the log-likelihood with kernel density estimators (KDEs) using a previous estimate $\widehat{\epsilon}^{(n)}$. That is we approximate (9) with $\widehat{\ell}^{(n)}(\epsilon; \mathcal{D}, S_i)$:

$$\sum_{i=1}^{N} \sum_{j=1}^{K} \mathbf{1}(S_i = s_j) \log \widehat{f}\left(X_i \mid \widehat{\epsilon}^{(n)}(\ell_i^-) = s_j\right), \quad (11)$$

where an equivalent version of $\widehat{f}\left(X_i \mid \widehat{\epsilon}^{(n)}(\ell_i^-) = s_j\right)$ is given below in (20).

3 EM Algorithm for Predictive State Space Assignment

Since ϵ maps to $\mathcal{S}$, the hidden state variable S_i and the "parameter" ϵ play the same role. This in turn results in similar E and M steps. Figure 2 gives an overview of the proposed algorithm.

[2] On the other extreme is a field with only $K = 1$ predictive state, i.e. the iid case.

3.1 Expectation Step

The E-step requires the expected log-likelihood

$$Q(\epsilon \mid \epsilon^{(n)}) = \mathbb{E}_{\mathcal{S}|\mathcal{D};\epsilon^{(n)}} \ell(\epsilon; \mathcal{D}, S_i), \tag{12}$$

where expectation is taken with respect to $\mathbb{P}\left(S_i = s_j \mid \mathcal{D}; \epsilon^{(n)}\right)$, the conditional distribution of the hidden variable S_i given the data $\mathcal{D}$ and the current estimate $\epsilon^{(n)}$. Using (9) we obtain

$$Q(\epsilon \mid \epsilon^{(n)}) = \sum_{i=1}^{N}\sum_{j=1}^{K} \mathbb{P}\left(S_i = s_j \mid X_i, \ell_i^-; \epsilon^{(n)}(\ell_i^-)\right) \times \log \mathbb{P}\left(X_i \mid \epsilon^{(n)}(\ell_i^-) = s_j\right). \tag{13}$$

As for $\ell(\epsilon; \mathcal{D})$ we use KDEs and obtain an approximate expected log-likelihood

$$\widehat{Q}^{(n)}(\epsilon \mid \epsilon^{(n)}) = \sum_{i=1}^{N}\sum_{j=1}^{K} \mathbb{P}\left(S_i = s_j \mid X_i, \ell_i^-; \epsilon^{(n)}(\ell_i^-)\right) \times \log \widehat{f}\left(X_i \mid \widehat{\epsilon}^{(n)}(\ell_i^-) = s_j\right) \tag{14}$$

The conditional distribution of S_i given its FLC and PLC, $\{X_i, \ell_i^-\}$, comes from Bayes' rule,

$$\begin{aligned}\mathbb{P}\left(S_i = s_j \mid X_i, \ell_i^-\right) &\propto \mathbb{P}\left(X_i, \ell_i^- \mid S_i = s_j\right) \mathbb{P}\left(S_i = s_j\right) \\ &= \mathbb{P}_j\left(X_i\right) \mathbb{P}_j\left(\ell_i^-\right) \mathbb{P}\left(S_i = s_j\right),\end{aligned} \tag{15}$$

the second equality following from the conditional independence of X_i and ℓ_i^- given the state S_i.

For brevity, let $w_{ij} := \mathbb{P}\left(S_i = s_j \mid X_i, \ell_i^-\right)$, an $N \times K$ weight matrix $\mathbf{W}$, whose rows are probability distributions over states. This $\mathbf{w}_i$ is the soft-thresholding version of $\epsilon(\ell_i^-)$, so we can write the expected log-likelihood in terms of $\mathbf{W}$,

$$\widehat{Q}^{(n)}(\mathbf{W} \mid \widehat{\mathbf{W}}^{(n)}) = \sum_{i=1}^{N}\sum_{j=1}^{K} w_{ij} \cdot \log \mathbb{P}\left(X_i \mid \widehat{\mathbf{W}}_j^{(n)}\right) \tag{16}$$

The current $\widehat{\mathbf{W}}^{(n)}$ can be used to update (conditional) probabilities in (15) by

$$\widehat{w}_{ij}^{(n+1)} \propto \widehat{f}(x_i \mid S_i = s_j; \widehat{\mathbf{W}}^{(n)}) \tag{17}$$

$$\times \mathcal{N}\left(\ell_i^- \mid \widehat{\boldsymbol{\mu}}_j^{(n)}, \widehat{\boldsymbol{\Sigma}}_j^{(n)}; \widehat{\mathbf{W}}^{(n)}\right) \tag{18}$$

$$\times \frac{\widehat{N}_j^{(n)}}{N}, \tag{19}$$

where i) $\widehat{N}_j^{(n)} = \sum_{i=1}^{N} \widehat{\mathbf{W}}_{ij}^{(n)}$ is the effective sample size of state s_j, ii) $\widehat{\boldsymbol{\mu}}_j^{(n)}$ and $\widehat{\boldsymbol{\Sigma}}_j^{(n)}$ are weighted mean and covariance matrix estimators of the PLCs using the jth column of $\widehat{\mathbf{W}}^{(n)}$, and iii) the FLC distribution is estimated with a weighted[3] KDE (wKDE)

$$\widehat{f}(x \mid S = s_j; \widehat{\mathbf{W}}^{(n)}) = \frac{1}{\widehat{N}_j^{(n)}} \sum_{r=1}^{N} \widehat{\mathbf{W}}_{rj}^{(n)} K_{h_j}(\|x_r - x\|), \tag{20}$$

Here the weights are again the jth column of $\widehat{\mathbf{W}}^{(n)}$, and K_{h_j} is a kernel function with a state-dependent bandwidth h_j. We used a Gaussian kernel, and to get a good, cluster-adaptive bandwidth h_j, we pick out on those x_i for which $\text{argmax}_k\, w_{ik} = j$ (hard-thresholding of weights; cf. Benaglia et al. 2011) and apply Silverman's rule-of-thumb-bandwidth (`bw.ndr0` in the R function `density`). After estimation, we normalize each $\widehat{\mathbf{w}}_i$ in (17), $\widehat{w}_{ij}^{(n+1)} \leftarrow \frac{\widehat{w}_{ij}^{(n+1)}}{\sum_{j=1}^{K} \widehat{w}_{ij}^{(n+1)}}$.

Ideally, we would use a non-parametric estimate for the PLC distribution, e.g., forest density estimators (Chow and Liu, 1968; Liu, Xu, Gu, Gupta, Lafferty, and Wasserman, 2011). Currently, however, such estimators are too slow to handle many iterations at large N, so we model state-conditional PLC distributions as multivariate Gaussians. Simulations suggest that this is often adequate in practice.

3.2 Approximate Maximization Step

In parametric problems, the M-step solves

$$\epsilon^{(n+1)} = \underset{\epsilon}{\text{argmax}}\, \widehat{Q}^{(n)}(\epsilon \mid \epsilon^{(n)}), \tag{21}$$

to improve the estimate. Starting from a guess $\epsilon^{(0)}$, the EM algorithm iterates (12) and (21) to convergence.

In nonparametric problems, finding an $\epsilon^{(n+1)}$ that increases $\widehat{Q}^{(n)}(\epsilon \mid \epsilon^{(n)})$ is difficult, since wKDEs with non-zero bandwidth are not maximizing the likelihood; they are not even guaranteed to increase it. Optimizing (14) by brute force isn't computationally feasible either, as it would mean searching K^N state assignments (cf. Bordes et al. 2007).

However, in our particular setting the parameter space and the expectation of the hidden variable coincide, since $\widehat{\mathbf{w}}_i$ is a soft-thresholding version of $\epsilon(\ell_i^-)$. Furthermore, none of the estimates above requires a deterministic ϵ mapping; they are all weighted MLEs or KDEs. Thus, like Benaglia, Chauveau, and Hunter (2009), we take the weights from the E-step, $\widehat{\mathbf{W}}^{(n+1)}$, to update each component distribution using (20).

[3] We also tried a hard-threshold estimator, but we found that the soft-threshold KDE performed better.

This in turn can then be plugged into (11) to update the likelihood function, and in (17) for the next E-step.

The wKDE update does not solve (21) nor does it provably increase the log-likelihood (although in simulations it often does so). We thus use cross-validation (CV) to select the best $\widehat{\mathbf{W}}^*$, and henceforth do not rely on an ever-increasing log-likelihood as the usual stopping rule in EM algorithms (see Section 5 for details).

3.3 Data-driven Choice of K: Merge States To Obtain Minimal Sufficiency

One advantage of the mixture model in (8) is that predictive states have, by definition, comparable conditional distributions. Since conditional densities can be tested for equality by a nonparametric two-sample test (or using a distribution metric), we can merge nearby classes. We thus propose a data-driven automatic selection of K, which solves this key challenge in fitting mixture models: 1) start with a sufficiently large number of clusters, $K_{\max} < N$; 2) test for equality of predictive distribution each time the EM reaches a (local) optimum; 3) merge until $K = 1$ (iid case) – step 5 in Fig. 2; 4) choose the best model $\widehat{\mathbf{W}}^*$ (and thus K^*) by CV.

0. **Initialization:** Set $n = 0$. Split data $\mathcal{D} = \{X_i, \ell_i^-\}_{i=1}^N$ in $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$. Initialize states randomly from $\{s_1, \dots, s_K\} \to$ Boolean $\widehat{\mathbf{W}}^{(0)}$.
1. **E-step:** Obtain updated $\widehat{\mathbf{W}}^{(n+1)}$ via (17).
2. **Approximate M-step:** Update mixture pdfs $\mathbb{P}(x_i \mid S_i = s_j)$ via (20) with $\widehat{\mathbf{W}}^{(n+1)}$.
3. **Out-of-sample Prediction:** Evaluate out-of-sample MSE for $\widehat{\mathbf{W}}^{(n+1)}$ by predicting FLCs from PLCs in $\mathcal{D}_{test}$. Set $n = n + 1$.
4. **Temporary convergence:** Iterate 1 - 3 until
$$\|\widehat{\mathbf{W}}^{(n)} - \widehat{\mathbf{W}}^{(n-1)}\| < \delta \tag{22}$$
5. **Merging:** Estimate pairwise distances (or test)
$$\widehat{d}_{jk} = \text{dist}\left(\widehat{f}_j^{(n)}, \widehat{f}_k^{(n)}\right) \forall j, k = 1, \dots, K. \tag{23}$$
   (a) If $K > 1$: determine $(j^{(\min)}, k^{(\min)}) = \arg\min_{j \neq k} \widehat{d}_{jk}$ and merge these columns
   $$\mathbf{W}_{j^{(\min)}}^{(n)} \leftarrow \mathbf{W}_{j^{(\min)}}^{(n)} + \mathbf{W}_{k^{(\min)}}^{(n)} \tag{24}$$
   Omit column $k^{(\min)}$ from $\mathbf{W}_{k^{(\min)}}^{(n)}$, set $K = K - 1$, and re-start iterations at 1.
   (b) If $K = 1$: return $\widehat{\mathbf{W}}^*$ with lowest out-of-sample MSE.

Figure 2: Mixed LICORS: nonparametric EM algorithm for predictive state recovery.

4 Forecasting Given New Data

The estimate $\widehat{\mathbf{W}}^*$ can be used to forecast $\tilde{X}$ given a new $\tilde{\ell}^-$. Integrating out S_i yields a mixture

$$\mathbb{P}\left(\tilde{X} \mid \tilde{\ell}^-\right) = \sum_{j=1}^K \mathbb{P}\left(\tilde{S} = s_j \mid \tilde{\ell}^-\right) \cdot \mathbb{P}_j\left(\tilde{X}\right). \tag{25}$$

As $\mathbb{P}_j\left(\tilde{X}\right) = \mathbb{P}\left(\tilde{X} = x \mid \tilde{S} = s_j\right)$ is independent of $\tilde{\ell}^-$ we do not have to re-estimate them for each $\tilde{\ell}^-$, but can use the wKDEs in (20) from the training data.

The mixture weights $\tilde{w}_j := \mathbb{P}\left(\tilde{S} = s_j \mid \tilde{\ell}^-\right)$ are in general different for each PLC and can again be estimated using Bayes's rule (with the important difference that now we only condition on $\tilde{\ell}^-$, not on $\tilde{X}$):

$$\begin{aligned}\widehat{\mathbf{w}}_j(\tilde{\ell}^-; \widehat{\mathbf{W}}^*) &\propto \widehat{\mathbb{P}}\left(\tilde{\ell}^- \mid \tilde{S} = s_j; \widehat{\mathbf{W}}^*\right) \times \widehat{\mathbb{P}}\left(\tilde{S} = s_j; \widehat{\mathbf{W}}^*\right) \\ &= \mathcal{N}\left(\tilde{\ell}^-; \widehat{\boldsymbol{\mu}}_{(j)}^*, \widehat{\boldsymbol{\Sigma}}_{(j)}^*; \widehat{\mathbf{W}}^*\right) \times \frac{\widehat{N}_j^*}{N}. \end{aligned} \tag{26}$$

After re-normalization of $\widehat{\tilde{\mathbf{w}}} = (\widehat{\tilde{w}}_1^*, \dots, \widehat{\tilde{w}}_K^*)$, the predictive distribution (25) can be estimated via

$$\widehat{\mathbb{P}}\left(\tilde{X} = x \mid \tilde{\ell}^-\right) = \sum_{j=1}^K \widehat{\tilde{w}}_j^* \cdot \widehat{f}\left(\tilde{X} = x \mid S_i = s_j; \widehat{\mathbf{W}}^{(*)}\right). \tag{27}$$

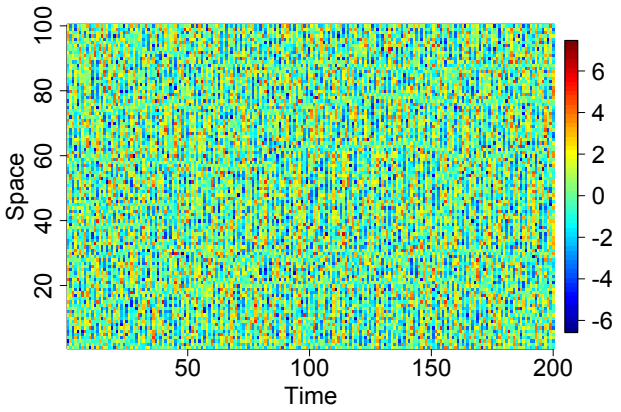
A point forecast can then be obtained by a weighted combination of point estimates in each component (e.g. weighted mean), or by the mode of the full distribution. In the simulations we use the weighted average from each component as the prediction of $\tilde{X}$.

5 Simulations

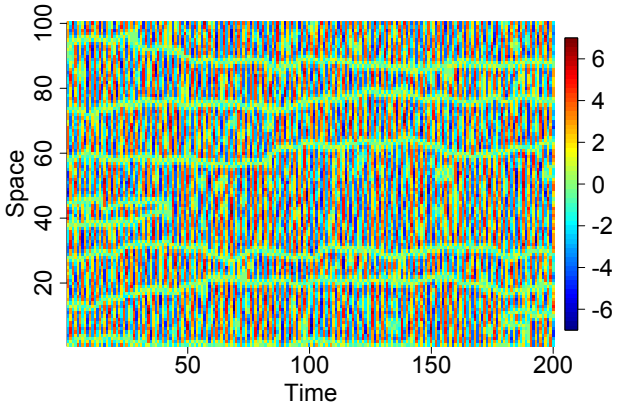
To evaluate the predictive ability of mixed LICORS in a practical, non-asymptotic context we use the following simulation. The continuous-valued $(1+1)D$ field $X(\mathbf{r}, t)$ has a discrete latent state space $d(\mathbf{r}, t)$. The observable field $X(\mathbf{r}, t)$ evolves according to a conditional Gaussian distribution,

$$\mathbb{P}\left(X(\mathbf{r}, t) \mid d(\mathbf{r}, t)\right) = \begin{cases} \mathcal{N}(d(\mathbf{r}, t), 1), & \text{if } |d(\mathbf{r}, t)| < 4, \\ \mathcal{N}(0, 1), & \text{otherwise,} \end{cases} \tag{28}$$

$$\text{and initial conditions: } X(\cdot, 1) = X(\cdot, 2) = \mathbf{0} \in \mathbb{R}^{|S|}. \tag{29}$$



(a) observed field $X(\mathbf{r}, t)$



(b) state-space $d(\mathbf{r}, t)$

Figure 3: Simulation of (28) – (30).

The state space $d(\mathbf{r}, t)$ evolves with the observable field,

$$d(\mathbf{r},t) = \left[\frac{\sum_{i=-2}^{2} X((\mathbf{r}+\mathbf{i}) \bmod |\mathbf{S}|, t-2)}{5} - \frac{\sum_{i=-1}^{1} X((\mathbf{r}+\mathbf{i}) \bmod |\mathbf{S}|, t-1)}{3} \right] \tag{30}$$

where $[x]$ is the closest integer to x. In words, Eq. (30) 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, (28) gives $h_f = 0$, making FLC distributions one-dimensional. As $d(\mathbf{r}, t)$ is integer-valued and the conditional distribution becomes $\mathcal{N}(0, 1)$ if $|d(\mathbf{r}, t)| > 4$, the system has 7 predictive states, $\{s_{-3}, s_{-2}, \ldots, s_2, s_3\}$, distinguished by the conditional mean $\mathbb{E}(X(\mathbf{r}, t) \mid s_k) = k$. Thus $X(\mathbf{r}, t) \mid s_k = \mathcal{N}(k, 1)$, $k = -3, -2, \ldots, 2, 3$.

Figure 3 shows one realization of (28) – (30) for $\mathbf{S} = \{1, \ldots, 100\}$ (vertical) and $t = 1, \ldots, T = 200$ (left to right), where we discarded the first 100 time steps as burn-in to avoid too much dependence on the initial conditions (29). While the (usually unobserved) state space has distinctive green temporal traces and also alternating red and blue patches, the observed field is too noisy to clearly see any of these patterns.[4]

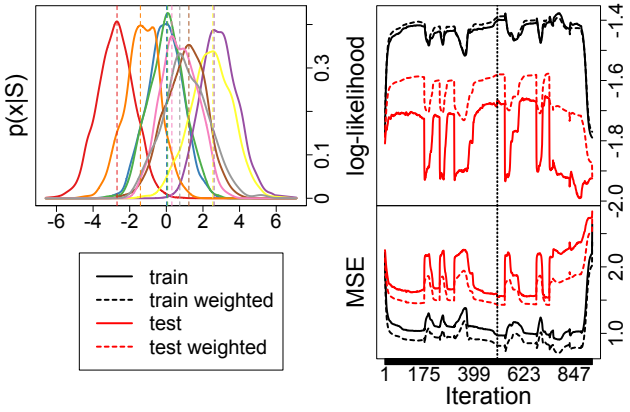


Figure 4: Mixed LICORS with $h_p = 2$ and $K_{\max} = 15$ starting states. Nonparametric estimates of conditional predictive distributions $\mathbb{P}(X = x \mid S = s_j)$ (top-left); trace plots for log-likelihood and MSE (right).

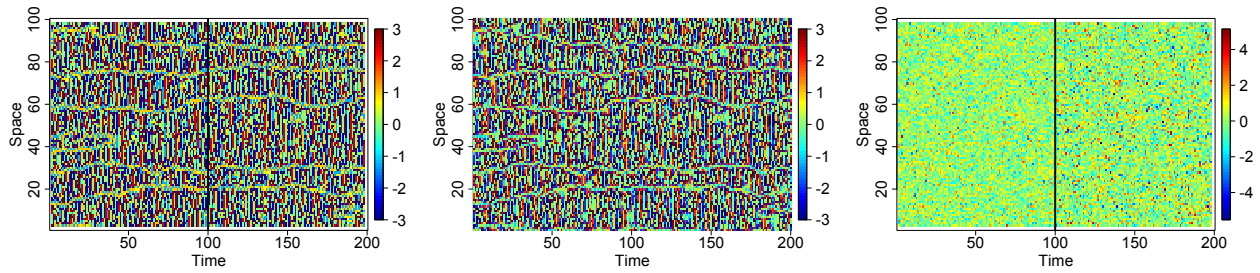
Figure 4 summarizes one run of mixed LICORS with $K = 15$ initial states, $h_p = 2$, and L_1 distance in (23). The first 100 time steps were used as training data, and the second half as test data. The optimal $\widehat{\mathbf{W}}^*$, which minimized the out-of-sample weighted MSE, occurred at iteration 502, with $\widehat{K} = 9$ estimated predictive states. The trace plots show large temporary drops (increases) in the log-likelihood (MSE) whenever the EM reaches a local optimum and merges two states. After merging, the forecasting performance and log-likelihood quickly return to — or even surpass — previous optima.

The predictions from $\widehat{\mathbf{W}}^*$ in Fig. 5a show that mixed LICORS is practically unbiased – compare to the visually indistinguishable true state space in Fig. 5b. The residuals in Fig. 5c show no obvious patterns except for a larger variance in the right half (training vs. test data).

5.1 Mixed versus Hard LICORS

Mixed LICORS does better than hard LICORS at forecasting. We use 100 independent realizations of (28) – (30) and for each one we train the model on the first half, and test it on the second (future) half. Lower

[4] All computation was done in R (R Development Core Team, 2010).



(a) estimated predictive states $\widehat{S}(\mathbf{r}, t)$ (b) true predictive states $S(\mathbf{r}, t)$ (c) residuals

Figure 5: Mixed LICORS model fit and residual check.

out-of-sample future MSEs are, of course, better.

We use EM as outlined in Fig. 2 with $K_{\max} = 15$ states, 10^3 maximum number of iterations. We keep the estimate $\widehat{\mathbf{W}}^*$ with the lowest out-of-sample MSE over 10 independent runs. The first run is initialized with a K-means++ clustering (Arthur and Vassilvitskii, 2007) on the PLC space; state initializations in the remaining nine runs were uniformly at random from $\mathcal{S} = \{s_1, \ldots, s_{15}\}$.

To test whether mixed LICORS accurately estimates the mapping $\epsilon : \ell^- \mapsto \mathcal{S}$, we also predict FLCs of an independently generated realization of the same process. If the out-of-sample MSE for the independent field is the same as the out-of-sample MSE for the future evolution of the training data, then mixed LICORS is not just memorizing fluctuations in any given realization, but estimates characteristics of the random field. We calculated both weighted-mixture forecasts, and the forecast of the state with the highest weight.

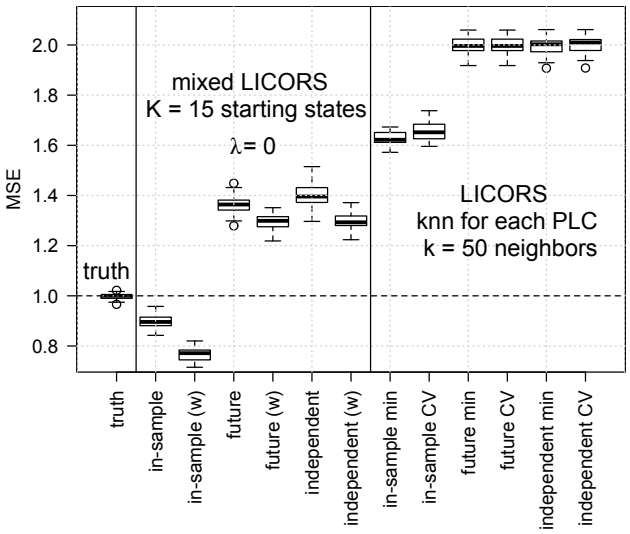


Figure 6: Comparing mixed and hard LICORS by forecast MSEs.

Figure 6 shows the results. Mixed LICORS greatly improves upon the hard LICORS estimates, with up to 34% reduction in out-of-sample MSE. As with hard LICORS, the MSE for future and independent realizations are essentially the same, showing that mixed LICORS generalizes from one realization to the whole process. As the weights often already are 0/1 assignments, weighted-mixture prediction is only slightly better than using the highest-weight state.

5.2 Simulating New Realizations of The Underlying System

Recall that all simulations use (29) as initial conditions. If we want to know the effect of different starting conditions, then we can simply use Eqs. (28) & (30) to simulate that process, since they fully specify the evolution of the stochastic process. In experimental studies, however, researchers usually lack such generative models; learning one is often the point of the experiment in the first place.

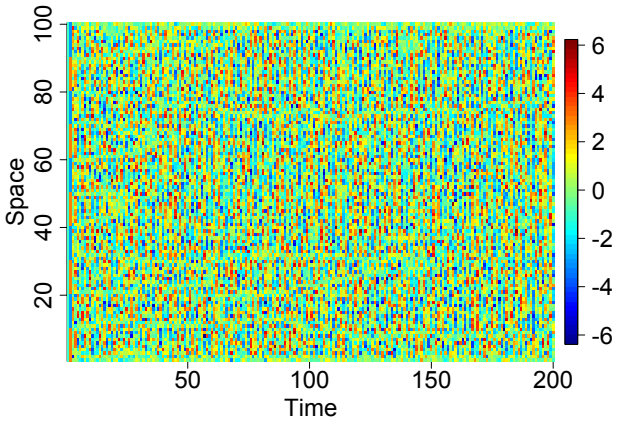
Since mixed LICORS estimates joint and conditional predictive distributions, and not only the conditional mean, it is possible to simulate a new realization from an estimated model. Figure 7 outlines this simulation procedure. We will now demonstrate that mixed LICORS can be used instead to simulate from different initial conditions *without* knowing Eqs. (28) & (30).

For example, Fig. 8a shows a simulation using the true mechanisms in (28) & (30) with starting conditions $X(\cdot, 1) = -\mathbf{1}$ and $X(\cdot, 2) = \pm\mathbf{3} \in \mathbb{R}^{|\mathbf{S}|}$ in alternating patches of ten times 3, ten times -3, ten times 3, etc. (total of 10 patches since $|\mathbf{S}| = 100$). The first couple of columns (on the left) are influenced by different starting conditions, but the initial effect dies out soon (since $h_p = 2$) and similar structures (left to right traces) as in simulations with (29) emerge (Fig. 3).

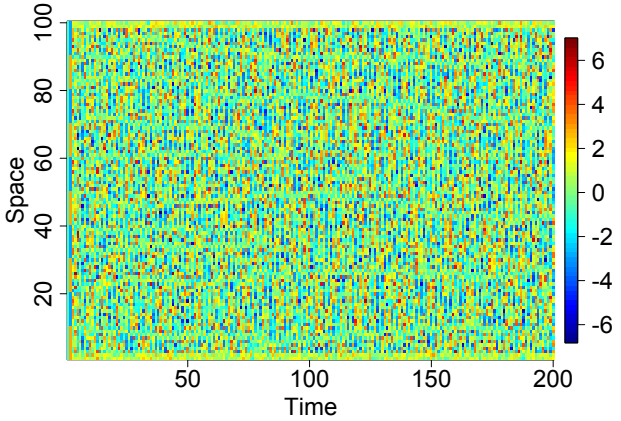
Figure 8b shows simulations solely using the mixed LICORS estimates in Fig. 4. While the patterns are quantitatively different (due to random sampling),

0. Initialize field $\{X(\cdot, 1-\tau)\}_{\tau=1}^{h_p}$. Set $t = 1$.
1. Fetch all PLCs at time t: $P_t = \{\ell^-(\mathbf{r}, t)\}_{\mathbf{r}\in\mathbf{S}^{\text{new}}}$
2. For each $\ell^-(\mathbf{r}, t) \in P_t$:
   (a) draw state $s_j \sim \mathbb{P}(S = s_j \mid \ell^-(\mathbf{r}, t))$
   (b) draw $X(\mathbf{r}, t) \sim \mathbb{P}(x \mid S = s_j)$
3. If $t < t_{\max}$, set $t = t + 1$ and go to step 1. Otherwise return simulation $\{X(\cdot, t)\}_{t=1}^{t_{\max}}$.

Figure 7: Simulate new realization from spatio-temporal process on $\mathbf{S}^{\text{new}} \times \{1, \ldots, t_{\max}\}$ using true and estimated dynamics (use (26) in 2a; (20) in 2b).



(a) simulations from true dynamics



(b) simulations from estimated dynamics

Figure 8: Simulating another realization of (28) & (30) but with different starting conditions.

the qualitative structures are strikingly similar. Thus mixed LICORS can not only accurately estimate $S(\mathbf{r}, t)$, but also learn the optimal prediction rule (28) solely from the observed data $X(\mathbf{r}, t)$.

6 Discussion

Mixed LICORS attacks the problem of reconstructing the predictive state space of a spatio-temporal process through a nonparametric, EM-like algorithm, softly clustering observed histories by their predictive consequences. Mixed LICORS is a probabilistic generalization of hard LICORS and can thus be easily adapted to other statistical settings such as classification or regression. Simulations show that it greatly outperforms its hard-clustering predecessor.

However, like other state-of-the-art nonparametric EM-like algorithms (Bordes et al., 2007; Hall et al., 2005; Mallapragada, Jin, and Jain, 2010), theoretical properties of our procedure are not yet well understood. In particular, the nonparametric estimation of mixture models poses identifiability problems (Benaglia et al., 2009, §2 and references therin). Here, we demonstrated that in practice mixed LICORS does not suffer from identifiability problems, and outperforms (identifiable) hard-clustering methods.

We also demonstrate that mixed LICORS can learn spatio-temporal dynamics from data, which can then be used for simulating new experiments, whether from the observed initial conditions or from new ones. Simulating from observed conditions allows for model checking; simulating from new ones makes predictions about unobserved behavior. Thus mixed LICORS can in principle make a lot of expensive, time- and labor-intensive experimental studies much more manageable and easier to plan. In particular, mixed LICORS can be applied to e.g., functional magnetic resonance imaging (fMRI) data to analyze and forecast complex, spatio-temporal brain activity. Due to space limits we refer to future work.

Acknowledgments

We thank Stacey Ackerman-Alexeeff, Dave Albers, Chris Genovese, Rob Haslinger, Martin Nilsson Jacobi, Heike Jänicke, 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. GMG was supported by an NSF grant (# DMS 1207759). CRS was supported by grants from INET and from the NIH (# 2 R01 NS047493).

References

D. Arthur and S. Vassilvitskii. `k-means++`: The advantages of careful seeding. In H. 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.

T. Benaglia, D. Chauveau, and D. R. Hunter. An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18:505–526, 2009. URL http://sites.stat.psu.edu/~dhunter/papers/mvm.pdf.

T. Benaglia, D. Chauveau, and D. R. Hunter. Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures. In D. R. Hunter, D. S. P. Richards, and J. L. Rosenberger, editors, *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger*. World Scientific, Singapore, 2011. URL http://hal.archives-ouvertes.fr/docs/00/35/32/97/PDF/npEM_bandwidth.pdf.

L. Bordes, D. Chauveau, and P. Vandekerkhove. A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, 51:5429–5443, 2007. doi: 10.1016/j.csda.2006.08.015.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14: 462–467, 1968.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977. URL http://www.jstor.org/pss/2984875.

G. M. Goerg and C. R. Shalizi. LICORS: Light cone reconstruction of states for non-parametric forecasting of spatio-temporal systems. Technical report, Department of Statistics, Carnegie Mellon University, 2012. URL http://arxiv.org/abs/1206.2398.

P. Hall, A. Neeman, R. Pakyari, and R. Elmore. Nonparametric inference in multivariate mixtures. *Biometrika*, 92:667–678, 2005.

S. L. Lauritzen. Sufficiency, prediction and extreme models. *Scandinavian Journal of Statistics*, 1: 128–134, 1974. URL http://www.jstor.org/pss/4615564.

S. L. Lauritzen. Extreme point models in statistics. *Scandinavian Journal of Statistics*, 11:65–91, 1984. URL http://www.jstor.org/pss/4615945. With discussion and response.

H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman. Forest density estimation. *Journal of Machine Learning Research*, 12:907–951, 2011. URL http://jmlr.csail.mit.edu/papers/v12/liu11a.html.

S. Lloyd. Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129 – 137, 1982. ISSN 0018-9448. doi: 10.1109/TIT.1982.1056489.

P. K. Mallapragada, R. Jin, and A. K. Jain. Nonparametric mixture models for clustering. In E. R. Hancock, R. C. Wilson, T. Windeatt, I. Ulusoy, and F. Escolano, editors, *Structural, Syntactic, and Statistical Pattern Recognition [SSPR//SPR 2010]*, pages 334–343, New York, 2010. Springer-Verlag. doi: 10.1007/978-3-642-14980-1_32.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL http://www.R-project.org/. ISBN 3-900051-07-0.

C. R. 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.