2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Estimating a generating partition from a time series: Symbolic shadowing

Yoshito Hirata and Kevin Judd

Centre for Applied Dynamics and Optimization, School of Mathematics and Statistics,
The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, AUSTRALIA
Email: yoshito@sat.t.u-tokyo.ac.jp, kevin@maths.uwa.edu.au

Abstract—We propose a deterministic algorithm for approximating a generating partition from a time series using tessellations. Using data generated by the Hénon and Ikeda maps, we demonstrated that the proposed method produces partitions that uniquely encodes all the periodic points up to some order.

1. Introduction

In nonlinear time series analysis, symbolic approaches have been sometimes used, for example, in evaluating surrogate data [1] and parameter fitting [2]. In these applications, a big problem is how to define a partition for generating a symbolic sequence. The most preferable partition is a generating partition, which assigns a point on the attractor a unique infinite symbol sequence. There are few methods known for finding a generating partition for a time series. Recently, Kennel and Buhl [3] proposed a method for estimating a generating partition, which assigns symbols so as to avoid topological degeneracies.

This paper intends to propose a method for estimating a generating partition from a time series [4]. It is significantly different from that of Kennel and Buhl [3] in that the proposed method will minimize the discrepancy between the original time series and a time series defined by its symbolic dynamics.

2. Algorithm

First we define some technical works to introduce the algorithm. A *partition* is a set of disjoint subsets of a state space M whose union covers the whole state space. Each point in the state space belongs to just an element of the partition because they are complete and disjoint. If we assign a symbol to each element of the partition, each point in the state space can be labeled with a symbol over alphabet $\mathcal{A}$. In the same way, we can generate, from a time series $\cdots x_{t-1}, x_t, x_{t+1}\cdots$ of states, a symbol sequence $\cdots X_{t-1}, X_t, X_{t+1}\cdots$.

If the time series is generated from a deterministic system of an invertible map, the information that X_t is a certain symbol locates the partition element that x_i belongs to. If we know that $X_{t-m}X_{t-m+1}\cdots X_{t+n}$ is a certain set of symbols, or *substring*, then x_t can be located more finely by considering the intersection of images and pre-images of the partition elements [5]. For convenience, we use "." to show the *center* of substrings, that is, $X_{t-m}X_{t-m+1}\cdots X_{t-1}.X_t\cdots X_{t+n}$ and call this the *substring* for x_t. We also write it as $X_{[t-m,t+n]}$ if it is obvious. Generally, a longer substring localizes a state better. Let $\Phi_{[-m,n]}(M)$ be the set of all the admissible substrings of length $(m+n+1)$.

We can a partition *generating* if any two states can be distinguished by some finite length substring, except possibly for a set of states of measure zero [6]. This means that with probability one, a state, or a trajectory, is uniquely identified using an infinite symbol sequence, which is a bi-directional extension of the substrings. If the partition is generating, the dynamical system is (almost everywhere) equivalent to the corresponding symbolic dynamics.

If a partition is close to being generating, one can expect that states with the same substring are close to each other, and they are well represented by a single point. Let us assign to each substring $S \in \Phi_{[-m,n]}(M)$ a point r_S, and call it *representative* for S. Typically a representative is a center of the sub-partition, which is the intersection of the corresponding images and pre-images.

Representatives can be used for two purposes. The first purpose is to roughly locate a point via a substring. The second purpose is to approximate the partition. This second purpose needs to be explained more.

If we tessellate the state space using the representatives, we can obtain a good approximation of the partition. For each substring S, we define its tile T_S to be the set of points in M whose nearest representative is r_S, that is,

$$T_S = \{x \in M : \|x-r_S\| \le \|x-r_{S'}\|, \forall S' \in \Phi_{[-m,n]}(M)\}. \quad (1)$$

Then for each $A \in \mathcal{A}$ we collect all tiles T_S for $S = S_{-m}\cdots S_n$, satisfying $S_0 = A$, to form a set,

$$B_{A,[-m,n]} = \cup\{T_S : S \in \Phi_{[-m,n]}(M), S_0 = A\}. \quad (2)$$

This $B_{A,[-m,n]}$ is a good approximation of A.

Let

$$R_{[-m,n]} = \{r_S : S \in \Phi_{[-m,n]}(M)\}. \quad (3)$$

Given a symbol sequence $X_1, X_2, \cdots, X_N = X_{[1,N]}$, we can generate a pseudo orbit $\{r_{X_{[t-m,t+n]}}\}_{t=m+1}^{N-n}$. We claim that the following minimization yields good estimates of the generating partition:

$$\min_{R_{[-m,n]},\mathcal{A}} \sum_{t=m+1}^{N-n} \|x_t - r_{X_{[t-m,t+n]}}\|^2. \quad (4)$$

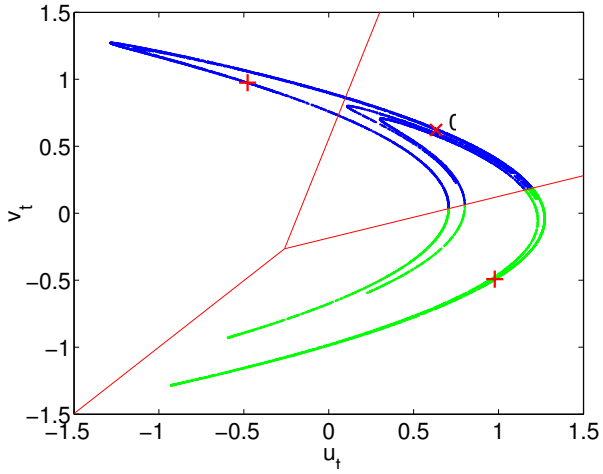

Figure 1: Initial partition for the Hénon map. The symbol × shows the fixed point, and + the periodic points of period 2. Blue and green points show the points initially labeled by symbols 0 and 1, respectively.

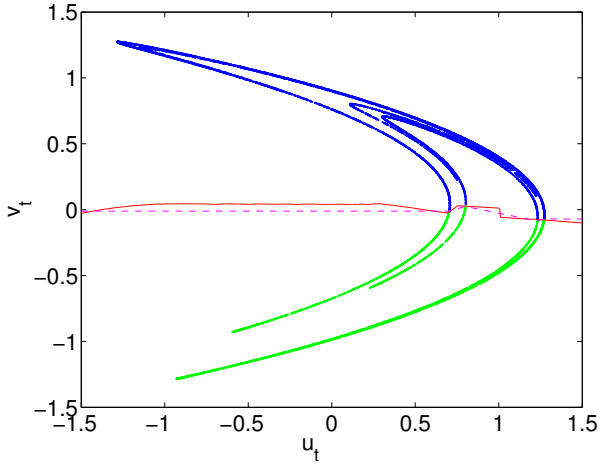

Figure 2: The obtained partition for the Hénon map. The solid line is the estimated partition and the dashed line is the generating partition conjectured by Grassberger and Kantz [12].

This is the minimization of the distance between the original time series and a pseudo orbit defined by the symbol sequence and representatives over the representatives and partition. However, this minimization is awkward as we need to specify the partition directly. The partition is necessary here for generating a symbol sequence. Therefore, we change the minimization in the following way:

$$\min_{R_{[-m,n]},X_{[1,N]}} \sum_{t=m+1}^{N-n} \|x_t - r_{X_{[t-m,t+n]}}\|^2. \tag{5}$$

We call $\sum_{t=m+1}^{N-n} \|x_t - r_{X_{[t-m,t+n]}}\|^2$ *discrepancy*.

To solve this minimization, we use an iterative algorithm:

1. Fixing $R_{[-m,n]}$, find a symbol sequence such that

$$\min_{X_{[1,N]}} \sum_{t=m+1}^{N-n} \|x_t - r_{X_{[t-m,t+n]}}\|^2. \tag{6}$$

2. Fixing $X_{[1,N]}$, find a set of representatives such that

$$\min_{R_{[-m,n]}} \sum_{t=m+1}^{N-n} \|x_t - r_{X_{[t-m,t+n]}}\|^2. \tag{7}$$

3. Go back to Step 1. until it converges.

Equation (6) is hard to minimize. Instead of solving it directly, we use the tessellation for approximating the symbol sequence.

There could be two possible methods for preparing the initial condition. For the initial condition, we may use periodic points of low periods. When a partition is generating, periodic points should be uniquely encoded. Therefore, conversely we assign symbols for periodic points in a way that they are uniquely encoded. This technique is used in previous work [7, 8].

The second method for preparing the initial condition is using equiprobable bins along an axis. We pick up one of axes and find number A of equiprobable bins. Labeling each bin with a different symbol gives an initial symbol sequence, from which we can construct an initial set of representatives. Using the median for discretizing the time series is a common technique in surrogate data analysis [1].

Therefore, practically we use the following algorithm for the approximation.

1. We prepare an initial partition.

   - If we use unstable periodic points, first detect them from a time series. (We use the method of Ref. [9], but there are more sophisticated methods such as Ref. [10, 11].) Assign to each unstable periodic point a substring of length l of type $S_{-m}\cdots S_{-1}.S_0\cdots S_n (m = \lfloor l/2 \rfloor$ and $n = \lfloor (l-1)/2 \rfloor)$ over alphabet $\mathcal{A}$ so that the unstable periodic points are encoded uniquely. Let each unstable periodic point be the representative $r_{S_{[-m,n]}}$ of the substring $S_{[-m,n]} \in \Phi_{[-m,n]}(M)$.
   - If we use a set of equiprobable bins, pick up an axis and prepare the equiprobable bins. Let B_i be the i-th bin, and A_i, the corresponding symbol. If $x_t \in B_i$, then label X_t to be A_i and obtain the initial symbol sequence $X_{[1,N]}$. For $t = m+1, \cdots, N-n$, classify x_t depending on its substring $X_{[t-m,t+n]}$: Let $C_{S_{[-m,n]}} = \{x_t : m+1 \le t \le N-n, S_{[-m,n]} = X_{[t-m,t+n]}\}$. The set $C_{S_{[-m,n]}}$ is a set of points whose currently allocated substring is $S_{[-m,n]}$. Lastly find the representative r_S

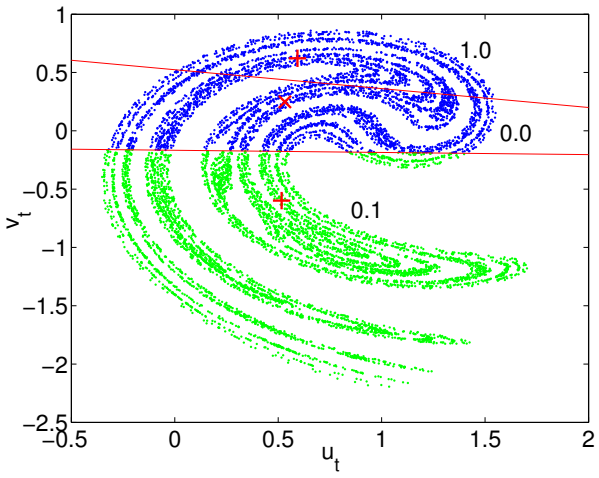

Figure 3: Initial partition for the Ikeda map. The symbol × shows the fixed point, and + the periodic points of period 2. Blue and green points show the points initially labeled by symbols 0 and 1, respectively.

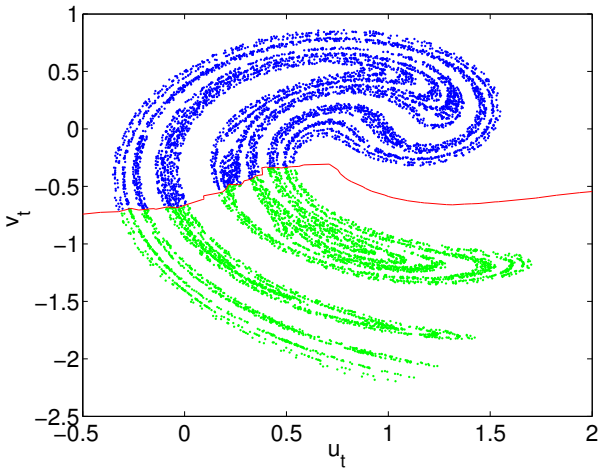

Figure 4: Partition estimated for the Ikeda map. The blue and green points show the points labeled with symbols 0 and 1, respectively.

for each substring S by

$$r_S = \sum_{y \in C_S} \frac{y}{|C_S|}. \tag{8}$$

2. For each observed point x_t, find its closest representative $r_{S_{-m} \cdots S_{-1} . S_0 \cdots S_n}$. Then make X_t to be S_0.

3. Classify x_t depending on its substring $X_{[t-m,t+n]}$. Then we obtain a set C_S of points with each substring S.

4. For each substring $S \in \Phi_{[-m,n]}(M)$, update its representative:
$$r_S = \sum_{y \in C_S} \frac{y}{|C_s|}. \tag{9}$$

5. Return to Step 2. until the set of representatives and the symbol sequence no longer change, or they cycle.

6. Increase the length of substrings by $l \leftarrow l + 1$, $m \leftarrow \lfloor l/2 \rfloor$, and $n \leftarrow \lfloor (l-1)/2 \rfloor$. Return to Step (3) until a stopping criterion is achieved.

There could be several possible stopping criteria. In this paper, we use two stopping criteria. One is we stop the algorithm when $\sum_{t=m+1}^{N-n} \|x_t - r_{X_{[t-m,t+n]}}\|/(N-m-n)$ becomes sufficiently small. The other is we stop the algorithm when the substrings reach a certain length and the algorithm converges.

3. Examples

Now we apply the proposed algorithm to examples. The first example is the Hénon map: $(u_{t+1}, v_{t+1}) = (1 - au_t^2 + bv_t, u_t)$, where $(a, b) = (1.4, 0.3)$. The following calculations take $x_t = (u_t, v_t)$ and use a time series of 10 000 data points.

In this example, we decided to stop the algorithm when $\sum_{t=m+1}^{N-n} \|x_t - r_{X_{[t-m,t+n]}}\|/(N-m-n) < (0.05)^2$. The algorithm was initialized using the unstable periodic points of order up to 2. We assigned the initial substrings as shown in Fig. 1.

The algorithm stopped when the length of the substrings was 13 and 883 representatives were used. The obtained partition was shown in Fig. 2. It is very close to the generating partition conjectured by Grassberger and Kantz [12]. The difference between the two partitions is just 22 symbols out of 10 000 symbols. We confirmed that this partition can encode periodic points uniquely up to period 17.

The second example is the Ikeda map: $(u_{t+1}, v_{t+1}) = (1 + a(u_t \cos\theta - v_t \sin\theta), a(u_t \sin\theta + v_t \cos\theta))$, where $\theta = 0.4 - b/(1 + u_t^2 + v_t^2)$, $a = 6$, and $b = 0.9$. We generated a time series of length 10 000.

We initialized the algorithm using the unstable periodic points up to period 2. We assigned the initial substrings to the periodic points as shown in Fig. 3.

The algorithm was stopped when $\sum_{t=m+1}^{N-n} \|x_t - r_{X_{[t-m,t+n]}}\|/(N-m-n) < (0.05)^2$. Then the length of substrings was 11 and the partition contained 1386 representatives. The obtained partition is shown in Fig. 4. It was confirmed that the partition encodes the unstable periodic points uniquely up to period 8.

4. Comparison with Kennel and Buhl [3]

Recently Kennel and Buhl [3] also published a method for estimating a generating partition from a time series. They estimate a generating partition by reducing the topological degeneracies, the points which are close in the symbolic space but far apart in the state space.

The proposed method enjoys some advantages over that of Kennel and Buhl [3]. The first advantage is that our

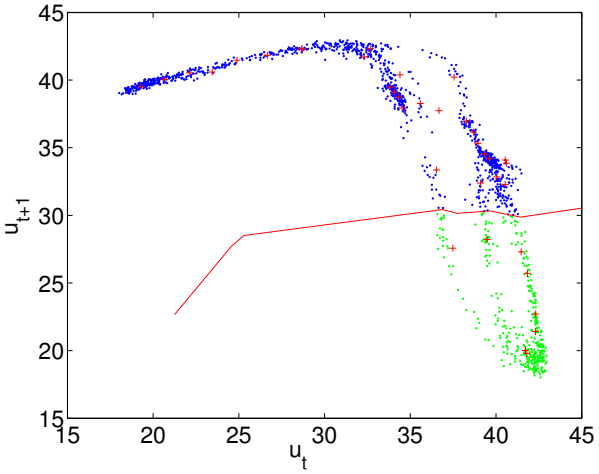

Figure 5: Partition obtained for the bubble data. Symbols + show the positions of the representatives.

method works in a deterministic manner, it means, it is expected too run fast, while the method of Kennel and Buhl contains a stochastic optimization. The second advantage is that our method contains fewer parameters to specify in advance than that of Kennel and Buhl. The third advantage is that our method is justified with proofs. For the details of the proofs, see Refs. [4, 13].

In addition to these advantages, there is a difference in the performances. We applied the proposed algorithm to the bubble data, which is used in Kennel and Buhl [3]. After embedding the data in the two dimensional space (u_t, u_{t+1}), we initialized the algorithm using the median of the first coordinate. We stopped the algorithm when the length of substrings reached 8. Figure 5 shows the obtained partition. This partition is different from that obtained by the method of Kennel and Buhl as each element of the partition in Fig. 5 is contiguous.

There are some possibilities which caused this difference. One of the possibilities is that there exist some generating partitions for this system. Another possibility is that dynamical noise may make a difference between the two methods. The cause of this difference should be explained.

5. Conclusion

We proposed a method for estimating a generating partition from a time series. The partition is approximated by tessellating the state space using representatives, or points in the state space with a unique substring. Using this property, we state the problem of finding a generating partition as minimizing the discrepancy between the original time series and a time series specified with a symbol sequence and representatives. By solving this minimization problem approximately using an iterative algorithm, we estimated a generating partition.

We demonstrated that our method worked well with time series generated from the Hénon and Ikeda maps. We also discussed that the proposed method has advantages over that of Kennel and Buhl [3] as it runs potentially fast, it does not have many parameters to be specified in advance, and it has a mathematical proof.

References

[1] P. E. Rapp et al., "Phase-randomized surrogates can produce spurious identifications of random structure," *Phys. Lett A*, vol.192, pp.27–33, 1994.

[2] X. Z. Tang et al., "Symbol sequence statistics in noisy chaotic signal reconstruction," *Phys. Rev. E*, vol.51, pp.3871–3889, 1995.

[3] M. B. Kennel and M. Buhl, "Estimating good discrete partitions from observed data: symbolic false nearest neighbors," *Phys. Rev. Lett.*, vol.91, 084102, 2003.

[4] Y. Hirata, K. Judd, and D. Kilminster, "Estimating a generating partition form observed time series: Symbolic shadowing," *Phys. Rev. E*, vol.70, 016215, 2004.

[5] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding.* Cambridge University Press, Cambridge UK, 1995.

[6] D. J. Rudolph, *Fundamentals of measurable dynamics: ergodic theory on Lebesgue spaces.* Oxford University Press, Oxford UK, 1990.

[7] R. K. Davidchack et al., "Estimating generating partitions of chaotic systems by unstable periodic orbits," *Phys. Rev. E*, vol.61, pp.1353–1356, 2000.

[8] J. Plumecoq and M. Lefranc, "From template analysis to generating partitions I: Periodic orbits, knots and symbolic encodings," *Physica D*, vol.144, pp.231–258, 2000.

[9] D. Auerbach et al., "Exploring chaotic motion through periodic orbits," *Phys. Rev. Lett.* vol.58, pp.2387–2389, 1987.

[10] P. So et al., "Detecting unstable periodic orbits in chaotic experimental data," *Phys. Rev. Lett.*, vol.76, pp.4705–4708, 1996.

[11] P. So et al., "Extracting unstable periodic orbits from chaotic time series data," *Phys. Rev. E*, vol.55, pp.5398–5417, 1996.

[12] P. Grassberger and H. Kantz, "Generating partitions for the dissipative Hénon map," *Phys. Lett.*, vol.113A, pp.235–238, 1985.

[13] Y. Hirata, Ph.D. dissertation, School of Mathematics and Statistics, The University of Western Australia, Crawley, Australia, 2003.