Time-series information and learning

Daniil Ryabko
INRIA Lille, France
daniil@ryabko.net

***Abstract*—Given a time series $X_1, \dots, X_n, \dots$ taking values in a large (high-dimensional) space $\mathcal{X}$, we would like to find a function f from $\mathcal{X}$ to a small (low-dimensional or finite) space $\mathcal{Y}$ such that the time series $f(X_1), \dots, f(X_n), \dots$ retains all the information about the time-series dependence in the original sequence, or as much as possible thereof. This goal is formalized in this work, and it is shown that the target function f can be found as the one that maximizes a certain quantity that can be expressed in terms of entropies of the series $(f(X_i))_{i\in\mathbb{N}}$. This quantity can be estimated empirically, and does not involve estimating the distribution on the original time series $(X_i)_{i\in\mathbb{N}}$.**

I. Introduction

Given a stationary sequence $X_1, \dots, X_n, \dots$ where X_i belong to a large (continuous, high-dimensional) space $\mathcal{X}$, we are looking for its compact representation $f(X_1), \dots, f(X_n), \dots$ where $f(X_i)$ belong to a small (low-dimensional or finite) space $\mathcal{Y}$. Moreover, we require from our representation that it preserves all, or as much as possible of, the time-series dependence. Such problems arise in a variety of applications, starting from speech or hand-written text recognition, to the analysis of video or network data. Often in such applications there are no examples of correct or good representations; or such examples are available only for a small portion of the data. Thus, we consider the problem of finding representations in what is known as *unsupervised* manner, meaning that there is no complementary "training" sequence for which a good representation is provided.

To formalize the problem, let us first consider the "ideal" situation, which is as follows. There exists a function $f : \mathcal{X} \to \mathcal{Y}$ (a representation) such that for each $i \in \mathbb{N}$, X_i is independent of the rest of the sequence $X_1, X_2, \dots$ *given* $f(X_i)$. That is, all the time-series dependence is in the sequence $f(X_1), \dots, f(X_n), \dots$, and, given this sequence, the original sequence $X_1, \dots, X_n, \dots$ can be considered as noise, in the sense that X_i are conditionally independent. It is shown in this work that in this "ideal" situation the function f maximizes the following information criterion

$$I_\infty(f) := h(f(X_1)) - h_\infty(f(X)) \tag{1}$$

where $h(f(X_1))$ is the Shannon entropy of the first element and h_∞ is the entropy rate of the (stationary) time series $f(X_1), \dots, f(X_n), \dots$. This means that for any other function $g : \mathcal{X} \to \mathcal{Y}$ we have $I_\infty(f) \geq I_\infty(g)$, with equality if and only $(X_i)_{i\in\mathbb{N}}$ are also conditionally independent given $(g(X_i))_{i\in\mathbb{N}}$. It is worth noting that $I_\infty(f)$ can be also expressed as $I(f(X_0); f(X_{-1}), f(X_{-2}), \dots)$ (see Lemma 1 below); thus, what we are trying to maximize can be called *time-series information*. However, the form (1) is more suitable for empirical estimation.

This allows us to pass to the non-ideal situation, in which there is no function f that satisfies the conditional independence criterion. Given a set of functions mapping $\mathcal{X}$ to $\mathcal{Y}$ the function that preserves the most of the time-series dependence can be defined as the one that maximizes (1). Furthermore, it can be shown that the time-series information (1) can be estimated empirically. Estimating this quantity does not involve estimating the distribution of X_i (on a large, possibly continuous and high-dimensional, space $\mathcal{X}$): one can work only with representations $f(X_i)$, that is, with distributions on a small (finite) space $\mathcal{Y}$. Moreover, readily available methods for estimating the entropy and entropy rate (for example, those based on data compressors) can be used to estimate the time-series information (1).

Thus, we propose a method that, given a finite set $\mathcal{F}$ of representation functions, selects the one that preserves most of the time-series information. This method is distribution-free and is asymptotically consistent for stationary ergodic distributions.

Prior work. The considered problem is a variant of the *dimensionality reduction* problem, which in, its various forms, appears in almost all modern applications of statistical and machine-learning methods. What characterizes the specific problem considered is that we are concentrating on time-series dependence (which is often discarded or simplified in other approaches), as well as in that we are addressing an *unsupervised* problem, that is, no examples of "good" performance are provided.

There is a vast body of literature on modelling and compressing time series, with different scopes and results. First, note that if in our "ideal" case we put an additional requirement that $f(X_i)$ form a Markov chain, then we get a so-called Hidden Markov model [1]. Hidden Markov models are used in many different applications, and estimating the hidden states is considered an important and difficult problem. From a different perspective, if X_i are independent and identically distributed and, instead of the time-series dependence (which is absent in this case), we want to preserve as much as possible of the information about another sequence of variables (labels) $Y_1, \dots, Y_n$, then one can arrive at the information bottleneck method [2]. The information bottleneck method can, in turn, be seen as a generalization of the rate-distortion theory of Shannon [3]. Applied to dynamical systems, the information bottleneck method can be formulated [4] as follows: minimize $I(\text{past; representation}) - \beta I(\text{representation; future})$, where β

is a parameter. A related idea is that of causal states [5]: two histories belong to the same causal state iff they give the same conditional distribution over futures.

What distinguishes the approach of this work from those described, is that we never have to consider the probability distribution of the input time series X_i directly — only through the distribution of the representations $f(X_i)$. Thus, modelling or estimating X_i is not required; this is particularly important for empirical estimates.

It should also be noted that the quantity (1) has been studied in a different context: [6] uses it to construct statistical test for the hypothesis that a time series consists of independent and identically distributed variables. In particular, [6] notes that this quantity can be estimated using universal codes (or data compressors). The conditional independence property also has been previously studied in a different context (classification) in [7]. Specifically, [7] shows that binary classification methods developed to work under the i.i.d. assumption actually only need a weaker assumption of conditional independence.

II. Preliminaries

Let $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ be measurable spaces. $\mathcal{X}$ can be thought of as a large (e.g., a high-dimensional Euclidean) space and $\mathcal{Y}$ as a small (low-dimensional or even finite) space. Time-series (or process) distributions are probability measures on the space $(X^{\mathbb{N}}, \mathcal{F}_{\mathbb{N}})$ of one-way infinite sequences (where $\mathcal{F}_{\mathbb{N}}$ is the Borel sigma-algebra of $X^{\mathbb{N}}$). We use the abbreviation $X_{l..k}$ for $X_l, \dots, X_k$. A distribution ρ is stationary if $\rho(X_{0..k} \in A) = \rho(X_{n..n+k} \in A)$ for all $A \in \mathcal{F}_k$, $k, n \in \mathbb{N}$ (with $\mathcal{F}_k$ being the sigma-algebra of $\mathcal{X}^k$).

A stationary distribution on $\mathcal{X}^{\mathbb{N}}$ can be uniquely extended to a distribution on $\mathcal{X}^{\mathbb{Z}}$ (that is, to a time series $\dots, X_{-1}, X_0, X_1, \dots$); we will assume such an extension whenever necessary. We assume that $f(X_1), \dots, f(X_n)$ have a density with respect to a fixed reference measure M_n on $\mathcal{X}^n$ for all functions $f : \mathcal{X} \to \mathcal{Y}$ considered.

If $\mathcal{Y}$ is finite, for a $\mathcal{Y}$-valued random variable Z denote $h(Z)$ its Shannon entropy $-\sum_{y \in \mathcal{Y}} P(Z = y) \log P(Z = y)$. If $\mathcal{Y}$ is continuous $h(Z)$ denotes the relative entropy $\int_{\mathcal{Y}} p \log p \, dM$ where p is the density of Z with respect to a fixed reference measure M; such densities are assumed to exist whenever we speak about entropies. For precise definitions and conditions for the existence of (conditional) relative entropies see [8]. Introduce the notation

$$h_0(f) := h(f(X_0)), \tag{2}$$

and $h_k(f)$ for the k-order entropy of the time series $(f(X_i))_{i \in \mathbb{N}}$:

$$h_k(f) := -\mathbb{E}_{X_0, \dots, X_{k-1}} h(f(X_k) | f(X_0), \dots, f(X_{k-1})) \tag{3}$$

If $(f(X_i))_{i \in \mathbb{N}}$ is stationary then we can define the entropy rate

$$h_\infty(f) := \lim_{k \to \infty} h_k(f).$$

III. Definitions and main results

Definition 1 (conditional independence given labels)**.** *Say that $(X_i)_{i \in \mathbb{N}}$ are conditionally independent given $(f(X_i))_{i \in \mathbb{N}}$, if for all n, k, and all $i_1, \dots, i_k \neq n$ X_n is independent of $X_{i_1}, \dots, X_{i_k}$ given $f(X_n)$:*

$$P(X_n | f(X_n), X_{i_1}, \dots, X_{i_k}) = P(X_n | f(X_n)) \text{ a.s.} \tag{4}$$

The main object of interest in this work is the *time-series information*:

Definition 2 (time-series information)**.** *The time-series information of a series $f(X_i)_{(i \in \mathbb{N})}$ with finite $h(f)$ is defined as*

$$I_\infty(f) := h_0(f) - h_\infty(f). \tag{5}$$

Equivalently, time-series information can be defined as the mutual information between $f(X_0)$ and $(f(X_{-1}), f(X_{-2}), \dots)$:

Lemma 1. *If the time series $(X_i)_{i \in \mathbb{Z}}$ is stationary and $h_0(f)$ is finite then $I_\infty(f) = I(f(X_0); f(X_{-1}), f(X_{-2}), \dots)$.*

Proof: Using the stationarity of $(X_i)_{i \in \mathbb{Z}}$ and (for the last equality) [8, Lemma 5.6.1] we derive

$$\begin{aligned} I_\infty(f) &= \\ &\lim_{k \to \infty} h(f(X_0)) - h(f(X_0) | h(f(X_{-1}), \dots, f(X_{-k})) \\ &= \lim_{k \to \infty} I(f(X_0); f(X_{-1}), \dots, f(X_{-k})) \\ &= I(f(X_0); f(X_{-1}), f(X_{-2}), \dots). \end{aligned}$$

■

We can also define k-order time-series information as follows

$$\begin{aligned} I_k(f) &:= h_0(f) - h_k(f) \\ &= I(f(X_k); f(X_0), \dots, f(X_{k-1})). \end{aligned} \tag{6}$$

Theorem 1. *Let $(X_i)_{i \in \mathbb{N}}$ be a stationary time series, and let $f, g : \mathcal{X} \to \mathcal{Y}$ be two functions such that the entropies $h_0(f)$ and $h_0(g)$ are finite. If $(X_i)_{i \in \mathbb{N}}$ are conditionally independent given $(f(X_i))_{i \in \mathbb{N}}$ then*

$$I_\infty(f) \geq I_\infty(g),$$

with equality if and only if $(X_i)_{i \in \mathbb{N}}$ are conditionally independent given $(g(X_i))_{i \in \mathbb{N}}$.

The proof of the theorem is given in section IV.

Thus, given a set $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$ with finite entropies $h(f(X_0))$ the function that is "closest" to satisfying the conditional independence property (4) can be defined as the one that maximizes the time-series information (5).

If the set $\mathcal{F}$ is finite, then it is possible to find the function that maximizes (5) given a large enough sample of the time series $(X_i)_{i \in \mathbb{N}}$, without knowing anything about its distribution (besides its stationarity), and without modelling or even estimating this distribution.

To see this, let us first consider the case of a finite set $\mathcal{Y}$. To have a consistent estimate of $I_\infty(f)$ it is enough to have a

consistent estimator for $h_0(f)$ and a consistent estimator for $h_\infty(f)$; for both tasks there are several solutions available. For example, one can simply use the empirical estimator for h_0, and $h_\infty(f)$ can be estimated as $\frac{1}{n}|\varphi(f(X_1),\dots,f(X_n))|$, where φ is any universal code and $|\cdot|$ denotes length. The latter approach has been used in [6], [9], [10] to solve various statistical problems concerning stationary time series.

Thus, as a corollary of Theorem 1 we obtain the following statement.

Theorem 2. *Let $X_0,\dots,X_n$ be sampled from a stationary time series, let the set $\mathcal{Y}$ be finite and let a finite set $\mathcal{F}$ of functions $f:\mathcal{X}\to\mathcal{Y}$ be given. Set*

$$\hat{I}(f) := \hat{h}_0(f) - \frac{1}{n}|\varphi(f(X_1),\dots,f(X_n))|,$$

where $\hat{h}_0(f)$ is the empirical entropy of $f(X)$ and φ is any universal code. Then

$$\operatorname{argmax}_{f\in\mathcal{F}} \hat{I}(f) = \operatorname{argmax}_{f\in\mathcal{F}} I(f)$$

from some n on with probability 1.

Note that for the consistency statement in Theorem 2 it is not required that there exists a function $f\in\mathcal{F}$ that satisfies (4); if the latter property does not hold then the empirical estimator $\hat{I}(f)$ still finds the "best" solution, in the sense of maximizing $I(f)$.

A different approach to estimate the entropy rate $h_\infty(f)$ (and thus $I_\infty(f)$) for finite or countable spaces $\mathcal{Y}$ is to use so-called *match lengths*; this method requires the time series $(f(X_i))_{i\in\mathbb{N}}$ to satisfy a Doeblin condition [11], [12], [13].

The case of continuous spaces $\mathcal{Y}$ can be treated in a similar manner, using a universal codes on discretised versions of data, that are then combined as proposed in [10]: Assume that for each n the random variables $X_1,\dots,X_n$ have a density with respect to a measure M_n. Denote $Y_n := f(X_n)$ and let $[Y_i]^k$ denote Y_i truncated to k bits of precision (i.e., we are using a series of quantisations). Let μ_k be a universal measure associated with a universal code $\varphi(\cdot)$ applied on the alphabet $[\mathcal{Y}]^k := \{[y]^k : y\in\mathcal{Y}\}$. The combined measure is defined [10] as

$$R(Y_1,\dots,Y_n) := \sum_{k\in\mathbb{N}} w_k \mu_k([Y_1]^k,\dots,[Y_n]^k)/M_n([Y_1]^k,\dots,[Y_n]^k),$$

where $(w_k)_{k\in\mathbb{N}}$ are positive summable real weights.

Furthermore, define the entropy estimate $\hat{h}$ as

$$\hat{h}(Y_1,\dots,Y_n) := -\frac{1}{n}\log R(Y_1,\dots,Y_n).$$

Under some mild conditions (see [10]) this estimator is consistent; thus, using $\hat{h}$ together with any consistent estimator of $h(f(X_0))$ we get an analogue of Theorem 2 for the case of continuous spaces $\mathcal{Y}$.

Note that in this approach an estimator for the entropy rate is a byproduct of a method for time-series prediction (or estimating densities of time-series). Using the same reasoning, another methods for estimating the entropy rate for real-valued processes can be inferred from, for example, [14].

IV. Proof of Theorem 1

Proof: First note that from the definition (1) of conditional independence and using the chain rule for entropies, it is easy to show that for any $n,k,i_1,\dots,i_k\in\mathbb{N}$ we have

$$h(f(X_n)|f(X_{i_1}),g(X_{i_1}),\dots,f(X_{i_k}),g(X_{i_k})) = h(f(X_n)|f(X_{i_1}),\dots,f(X_{i_k})) \text{ a.s.,} \quad (7)$$

so that

$$h(f(X_n)|f(X_{i_1}),\dots,f(X_{i_k})) \le h(f(X_n)|g(X_{i_1}),\dots,g(X_{i_k})) \text{ a.s.} \quad (8)$$

Consider the following entropies and information (with straightforward definitions): $h_0(f,g)$, $h_k(f,g)$, $I_k(f,g)$ and $I_\infty(f,g)$. We will first show that

$$I_k(f,g) = I_k(f) \text{ and } I_\infty(f,g) = I_\infty(f). \quad (9)$$

The latter equality follows from the former and the definition of h_∞. To prove the former we will consider the case $k=1$; the general case is analogous. Introduce the short-hand notation $Y_i := f(X_i), Z_i := g(X_i)$, $i\in\mathbb{N}$. First note that

$$h_0(f,g) = h(Y_0) + h(Z_0|Y_0). \quad (10)$$

Moreover,

$$\begin{aligned} h_1(f,g) &= h(Y_0,Z_0|Y_{-1},Z_{-1}) \\ &= h(Y_0|Y_{-1},Z_{-1}) + h(Z_0|Y_0,Y_{-1},Z_{-1}) \\ &= h(Y_0|Y_{-1}) + h(Z_0|Y_0) \end{aligned} \quad (11)$$

where the first equality is by definition, the second is the chain rule for entropy and the third follows from (7) and conditional independence of X_i given $f(X_i)$. Thus, from (10), (11) and the definition of $I_1(f)$ we get

$$\begin{aligned} I_1(f,g) &= h_0(f,g) - h_1(f,g) \\ &= h(Y_0) + h(Z_0|Y_0) - h(Y_0|Y_{-1}) - h(Z_0|Y_0) \\ &= I_1(f) \end{aligned}$$

finishing the proof of (9).

To prove the theorem it remains to show that, if $(X_i)_{i\in\mathbb{N}}$ are not conditionally independent given $(g(X_i))_{i\in\mathbb{N}}$ then $I_\infty(f) > I_\infty(g)$. For that it is enough to show that

$$I_k(f) > I_k(g) \quad (12)$$

from some k on. Assume that $(X_i)_{i\in\mathbb{N}}$ are not conditionally independent given $(g(X_i))_{i\in\mathbb{N}}$, so that

$$P(X_n|g(X_n),X_{i_1},\dots,X_{i_k}) \neq P(X_n|g(X_n)) \quad (13)$$

for some k,n and $i_1,\dots,i_k\neq n$. By stationarity, we obtain from (13) that there exists $k\in\mathbb{N}$ such that

$$P(X_0|g(X_0),X_1,\dots,X_k,X_{-1},\dots,X_{-k}) \neq P(X_0|g(X_0)) \quad (14)$$

We will show that (12) holds for all k for which (14) holds. Clearly, if (14) holds for $k \in \mathbb{N}$ then it also holds for all $k' > k$. Thus, it is enough to consider the case $k = 1$; the general case is analogous. With this simplification in mind, and using our Y and Z notation, note that (14) implies that

$$P(Y_0|Z_0, X_1, X_{-1}) \neq P(Y_0|Z_0), \tag{15}$$

for otherwise we would get $P(X_0|Y_0, Z_0, X_1, X_{-1}) \neq P(X_0|Y_0, Z_0)$, contradicting conditional independence of X given $f(X)$. Finally, from (15) and (7) we get

$$P(Y_0|Z_0, Y_1, Y_{-1}) \neq P(Y_0|Z_0),$$

so that

$$h(Y_0|Z_0) - h(Y_0|Z_0, Y_1, Y_{-1}) > 0. \tag{16}$$

We will show that (16) implies that at least one of the following two inequalities holds

$$h(Y_1|Z_0, Y_{-1}) > h(Y_1|Y_0, Y_{-1}), \tag{17}$$

$$h(Y_1|Z_0) > h(Y_1|Y_0). \tag{18}$$

Indeed, if both (17) and (18) are false then from (7) and (8) we obtain

$$h(Y_1|Z_0, Y_{-1}) = h(Y_1|Y_0, Y_{-1}) = h(Y_1|Y_0, Z_0, Y_{-1}), \tag{19}$$

and

$$h(Y_1|Z_0) = h(Y_1|Y_0) = h(Y_1|Y_0, Z_0). \tag{20}$$

Thus, using the chain rule for entropy and (19) we derive

$$\begin{aligned} &h(Y_0|Z_0, Y_1, Y_{-1}) \\ &\quad = h(Y_0, Z_0, Y_1, Y_{-1}) - h(Z_0, Y_1, Y_{-1}) \\ &\quad = h(Y_1|Z_0, Y_0, Y_{-1}) + h(Z_0, Y_0, Y_{-1}) \\ &\qquad - h(Y_1|Z_0, Y_{-1}) - h(Z_0, Y_{-1}) \\ &\quad = h(Y_1|Z_0, Y_{-1}) + h(Y_0|Z_0, Y_{-1}) \\ &\qquad - h(Y_1|Z_0, Y_{-1}) = h(Y_0|Z_0, Y_{-1}). \end{aligned} \tag{21}$$

Continuing in the same way but using (20) instead of (19) we obtain

$$h(Y_0|Z_0, Y_1, Y_{-1}) = h(Y_0|Z_0)$$

contradicting (16). Thus, either (17) or (18) holds true; consider the former inequality — the latter one is analogous. We have

$$\begin{aligned} I_2(f) &= h(Y_1) - h(Y_1|Y_0, Y_{-1}) \\ &> h(Y_1) - h(Y_1|Z_0, Y_{-1}) \geq h(Y_1) - h(Y_1|Z_0, Z_{-1}) \\ &= I(Y_1; Z_0, Z_{-1}) = I(Z_0, Z_{-1}; Y_1) \\ &\geq I(Z_0, Z_{-1}; Z_1) = I(Z_1; Z_0, Z_{-1}) = I_2(g), \end{aligned}$$

where we have used the definition of I_k, (17), (8), the definition of mutual information and the symmetry thereof. This demonstrates (12) and concludes the proof. ∎

V. Conclusion

This paper takes some first step towards unsupervised representation learning for highly dependent time-series data. These steps are to define the objective and to show that it is attainable for finite sets of representations. The next steps are to see what results can be obtained for infinite sets of representations. As an extreme case, consider the set $\mathcal{F}$ of all possible representation functions $f : \mathcal{X} \to \mathcal{Y}$. Is it possible to find a function f that maximizes time-series information over $\mathcal{F}$, for arbitrary stationary (ergodic) time series? Clearly if this is possible, it is only in some appropriately weak (time-average) asymptotic sense. Perhaps stronger results can obtained smaller (yet infinite) sets $\mathcal{F}$. Finite-time analysis is possible if one makes some further assumptions on the time series X_i (beyond stationarity and ergodicity). Some steps in this direction are taken in [15], where the control problem is also considered. The next important step is to construct efficient algorithms for specific sets $\mathcal{F}$ of representations, and to test them on real applications.

Acknowledgments

This work was supported FP7/2007-2013 under grant agreements 270327 (CompLACS) and 216886 (PASCAL2), by the French National Research Agency (project Lampada ANR-09-EMER-007) and the Nord-Pas-de-Calais Regional Council and FEDER through CPER 2007-2013.

References

[1] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4 –16, jan 1986.

[2] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37th annual Allerton Conference on Communication, Control, and Computing*, 1999, pp. 368–377.

[3] C. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, 1959.

[4] F. Creutzig, A. Globerson, and N. Tishby, "Past-future information bottleneck in dynamical systems," *Phys. Rev. E*, vol. 79, p. 041925, 2009.

[5] C. R. Shalizi and J. P. Crutchfield, "Computational mechanics: Pattern and prediction, structure and simplicity," *Journal of Statistical Physics*, vol. 104, no. 3-4, pp. 817–879, 2001.

[6] B. Ryabko and J. Astola, "Universal codes as a basis for time series testing," *Statistical Methodology*, vol. 3, pp. 375–397, 2006.

[7] D. Ryabko, "Pattern recognition for conditionally independent data," *Journal of Machine Learning Research*, vol. 7, pp. 645–664, 2006.

[8] R. Gray, *Entropy and information theory*. Springer Verlag, 1990.

[9] B. Ryabko and V. Monarev, "Using information theory approach to randomness testing," *Journal of Statistical Planning and Inference*, vol. 133, no. 1, pp. 95–110, 2005.

[10] B. Ryabko, "Compression-based methods for nonparametric prediction and estimation of some characteristics of time series," *IEEE Transactions on Information Theory*, vol. 55, pp. 4309–4315, 2009.

[11] P. Shields, "Entropy and prefixes," *The Annals of Probability*, vol. 20, no. 1, pp. 403–409, 1992.

[12] I. Kontoyiannis and Y. Suhov, "Prefixes and the entropy rate for long-range sources," in *IEEE International Symposium On Information Theory*, 1994, pp. 194–194.

[13] A. Quas, "An entropy estimator for a class of infinite alphabet processes," *Theory of Probability & Its Applications*, vol. 43, no. 3, pp. 496–507, 1999.

[14] G. Morvai, S. Yakowitz, and L. Gyorfi, "Nonparametric inference for ergodic, stationary time series," *Ann. Statist*, vol. 24, no. 1, pp. 370–379, 1996.

[15] D. Ryabko, "Unsupervised model-free representation learning," arxiv, Tech. Rep. arXiv:1304.4806, 2013.