Determinism and Sources of Predictability

Joan Serrà, Holger Kantz, Xavier Serra, and Ralph G. Andrzejak*

This is the Supplementary Material for the article entitled "Predictability of Music Descriptor Time Series and its Application to Cover Song Detection", submitted to IEEE Trans. on Audio, Speech and Language Processing.

In the article we find evidence that forecasts of music descriptor time series can still be performed at mid-term or relatively long horizons. In addition, we show that this ability becomes crucial for a real-world application, namely the automatic detection of cover songs. However, prediction errors are not as low as one would expect, even for short horizons. Thus, we here investigate in more detail which structures in our data are the sources of predictability.

Prediction relies on the assumption that similar initial conditions are mapped to similar states after a certain time step h. If h is so large that this property gets lost, then inaccuracy of initial conditions, inaccuracies of the model, and the lack of smoothness of the h-step map make predictions impossible in practice [1]. We assess this property for our data using a technique to quantify the expansion or contraction of sample neighborhoods as h increases, and by comparing the behavior of music descriptor time series to well-known processes.

Consider a time series $\mathcal{S}$ and its reconstructed state space $\mathcal{S}^*$. For each point $\mathbf{s}_n^*$ in $\mathcal{S}^*$, we select a neighborhood Ω_n of radius ϵ. As done with the locally constant predictor, we use a percentage of the average squared Euclidean norm between all points ($\epsilon_\kappa = 0.4$ for descriptors and noise, $\epsilon_\kappa = 0.01$ for the rest). In addition, samples with neighborhoods smaller than $\nu = 5$ elements are discarded from further processing. Once we have the neighborhood Ω_n, an average pairwise distance between all elements in Ω_n is computed. This computation is done by taking into account the evolution of these elements as a function of the prediction horizon h:

$$\vartheta_n(h) = \frac{1}{|\Omega_n|^2} \sum_{\mathbf{s}_{n'}^* \in \Omega_n} \sum_{\mathbf{s}_{n''}^* \in \Omega_n} \|\mathbf{s}_{n'+h} - \mathbf{s}_{n''+h}\| , \qquad (1)$$

where $\mathbf{s}_{n'}^*$ and $\mathbf{s}_{n''}^*$ are samples reconstructed by delay vectors which are found in the neighborhood Ω_n, and $\mathbf{s}_{n'+h}$ and $\mathbf{s}_{n''+h}$ are the future values at h steps ahead of the corresponding unreconstructed scalar samples of $\mathbf{s}_{n'}^*$ and $\mathbf{s}_{n''}^*$, respectively. The quantity given by Eq. (1) is normalized by an average pairwise distance between unreconstructed samples,

$$\theta = \frac{1}{(N-w-h)^2} \sum_{\mathbf{s}_{n'}^* \in \mathcal{S}^*} \sum_{\mathbf{s}_{n''}^* \in \mathcal{S}^*} \|\mathbf{s}_{n'} - \mathbf{s}_{n''}\| , \qquad (2)$$

such that $\zeta_n(h) = \vartheta_n(h)/\theta$. Finally, by averaging across all samples $n = w+1, \dots N-h$ of a time series, we obtain $\zeta(h) = \langle \zeta_n(h) \rangle$, an indicator of the expansion or contraction of the sample neighborhoods as the time evolves.

* See the main article for contact data and acknowledgments.

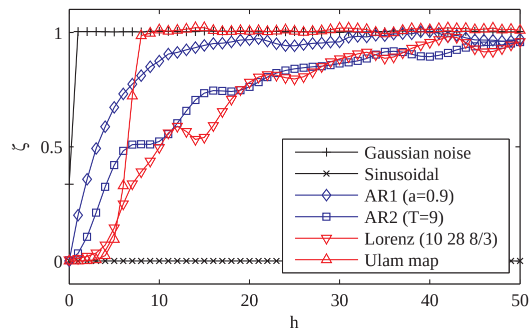

Fig. 1. Plots of ζ as a function of h for a number of time series obtained from known processes.

In Fig. 1 we represent the $\zeta(h)$ curves obtained from time series of a number of known dynamical systems [1]–[3]: white Gaussian noise, a sinusoidal, an AR process of order 1 (AR1), an AR process of order 2 (AR2), a realization of the Lorenz system, and a realization of the Ulam map. The last two provide typical realizations of chaotic time series. For example time series and information about the processes that might generate them we refer to the citations above.

We see that all neighborhoods have a ζ value close to 0 for $h = 0$ (recall that they are selected from the reconstructed state space with a small radius ϵ). As soon as $h > 0$, ζ increases until it reaches 1, i.e. until ϑ reaches θ, the average pairwise distance between samples. The only exception is the sinusoidal signal, whose ζ remains close to 0. This is an indicator of the periodicity of this signal: points inside a neighborhood do not disperse with time. On the other extreme we have Gaussian uncorrelated noise, which already at $h = 1$ reaches $\zeta = 1$. This implies that the sample spread within one time step is already equal to the mean sample spread of the time series. Correlated systems lead to intermediate values and behaviors. Chaotic systems produce an exponential growth of ζ ($h \leq 5$, lines with triangles). We can also see the typical contractions of the Lorenz system at integer multiples of the half mean oscillation period. Finally, the shape of the ζ curves for the AR processes resemble more a square root or a logarithmic function. A further interesting fact is the successive plateaus of the AR2 curve, which are repeated at multiples of the main periodicity of the time series (blue squares)[1].

If now we perform the same analysis for the music descriptor time series and average across songs, we see three interesting facts (Fig 2). First, curves do not start at a ζ value close to 0 for $h = 0$. This characteristic is shared with

[1]It can be easily checked by iteration of the model coefficients that an AR process of order 2 reaches subsequent plateaus around multiples of its main period.

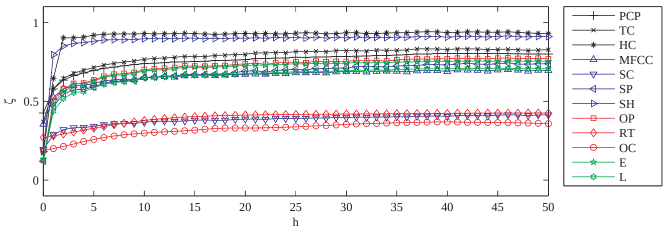

Fig. 2. Average ζ curves for all the considered descriptor time series.

multivariate Gaussian noise, and it is inherent to the difficulty of finding nearest neighbors for a noisy sample (thus having to set a larger ϵ).

Second, we see that curves do not reach $\zeta = 1$ for mid and relatively long time spans. This is related to the behavior we see in the self-prediction and cross-prediction figures of the article. We conjecture that this behavior is due to non-stationarity, resulting in an intermittent lack of recurrence.

Third, the shape of the ζ curves for all descriptors resemble those we find for the AR models (Fig. 1). Interestingly, the work by Meng et al. [4] cited in the article reports significative improvements in automatic music genre classification through the summarization (or compression) of music descriptor time series by means of AR coefficients. Furthermore, AR and TAR models yield here the highest performance, both in self-prediction and cross-prediction trials, being TAR models the best ones among all models tested. Notice that TAR models are adequate models for the abovementioned type of non-stationarity [5].

Based on all previous evidence, we can propose a hypothesis about the nature of the process governing the generation of music descriptor time series and, by extension, to general music dynamics. We could say that descriptor time series resemble to a certain extent a non-stationary piecewise AR processes with superimposed noise.

To provide further evidence, we implemented a simple algorithm for generating artificial descriptor time series following these guidelines. More concretely, we create artificial time series by concatenating, after an arbitrary number of iterations, the output of randomly chosen multivariate AR2 processes (with different means for each component) and subsequently superimpose white Gaussian noise with zero mean and unit variance (we then use a multiplicative factor to control the variance of the noise). The resultant time series generated by this algorithm closely resemble the real descriptor time series (Fig. 3). Moreover, ζ curves computed from these artificial time series are qualitatively similar to those generated for real descriptor time series from individual songs (Fig. 5). In particular, with an individual song analysis, we see some periodic plateaus or small contractions (Fig. 5a). This mimics the behavior of signals derived from AR2 processes, which have inherent periodicities (Fig. 1). Since we also use an AR2 process for the generation of the artificial time series, this behavior is also reflected in the artificial curves (Fig. 5b).

Overall, the outcome of these investigations, jointly with the results reported in the self-prediction and cross-prediction experiments of the article, suggest that the temporal evolution

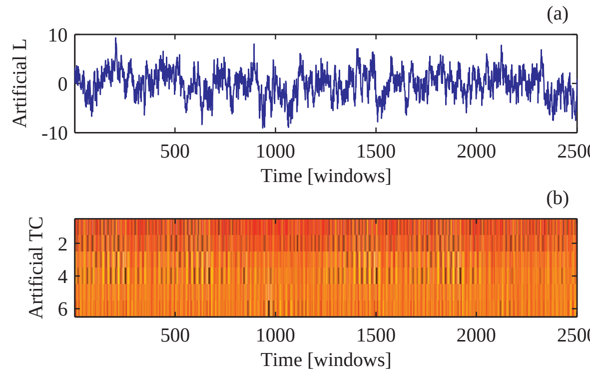

Fig. 3. Examples of artificially generated univariate (a) and a multivariate (b) time series. Compare these plots with the real descriptor time series of Fig. 4.

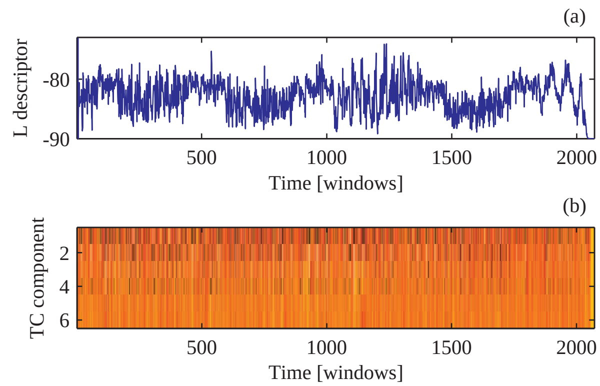

Fig. 4. Representative examples of a loudness (a) and a tonal centroid (b) descriptor time series. These correspond to a recording of the song “All along the watchtower” as performed by Jimi Hendrix.

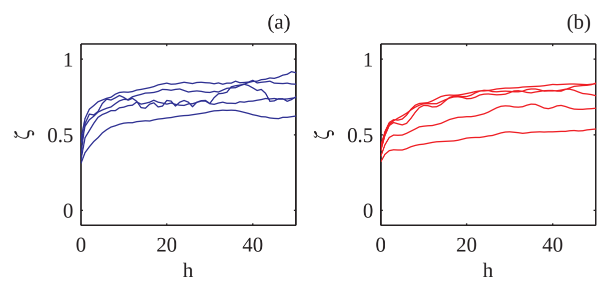

Fig. 5. Examples of individual ζ curves obtained with the PCP descriptor for 5 individual songs (a) and with 5 realizations of the algorithm for generating artificial descriptor time series (b).

of music descriptors might be explained by a concatenation of multiple autoregressive processes with superimposed noise. Since autoregressive models are linear stochastic models, this might imply that the irregularity of music is, to a large extent, due to random variations with more or less strong correlations. Evidently, non-stationarity is a very relevant element in music, and this is related to the succession of different sections in the course of a musical piece (introduction, verse, chorus, etc.).

REFERENCES

[1] H. Kantz and T. Schreiber, *Nonlinear time series analysis*, 2nd ed. Cambridge, UK: Cambridge University Press, 2004.

[2] A. S. Weigend and N. A. Gershenfeld, *Time series prediction: forecasting the future and understanding the past*. Boulder, USA: Westwiew Press, 1993.

[3] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, 3rd ed. Amsterdam, The Netherlands: Elsevier, 2007.

[4] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal feature integration for music genre classification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1654–1664, 2007.

[5] H. Tong and K. S. Lim, "Threshold autoregression, limit cycles and cyclical data," *Journal of the Royal Statistical Society*, vol. 42, no. 3, pp. 245–292, 1980.