Determinism and Stochasticity for Sound Modeling

Eric Métois
Information and Entertainment Group, Media Laboratory, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139

1.Introduction

"Let's flip a coin". As soon as these magic words reach the conditioned brain of a typical MIT student, the obscure notion of a discrete random variable with two equally likely outcomes pops out, ready to evaluate the next gambling situation. Is there an intrinsic truth behind the association of randomness to the outcome of this simple experiment or is it just the best we can do to model it? After all, some people may argue that quantum mechanics is the "living proof" of randomness' existence and that attaching probability distributions to our surrounding is not only convenient, but necessary. Still, another argument is that if I had access to a significant number of variables involved in the flip of this coin, the possible outcomes of the experiment would stop being equally likely. Realistically though, the performance (in terms of prediction) of a "more deterministic" model for that coin being flipped is probably not worth its complexity.

The purpose of this scenario is to illustrate the fact that stochasticity, when introduced in a model, can be often seen as a tradeoff between performance and complexity. It is easier to define performance rather than complexity but in the case of signal modeling, we will see how this notion can relate to the number of degrees of freedom in the model and to non-linearities in their inter-relationships.

2.The linear approach

2.1.The power spectrum/sonogram fascination

Early in the century the works of Helmholtz, inspired by Fourier's transform, gave birth to the **classical conception of timbre**. Helmholtz, and a large number of physicists after him (Bouasse, Olson,...), considered that the majority of the timbral information was contained within the quasi stationary part of a sound. Over this restricted temporal part, the signal can be written as a Fourier series and it was believed that the spectrum of this harmonic series would define timbre. The simplicity of this definition may seem attractive but when a more flexible recording technology became available, a few basic observations revealed its weaknesses. For instance, this spectrum can vary dramatically with the room's acoustical qualities or the listener's position and yet, the timbre of a given instrument is perceived unchanged. Another observation is the timbre change one can perceive when playing a sound backwards even though this process doesn't alter the harmonic spectrum given by this Fourier series. All these observations illustrated the fundamental role played by the temporal evolution of sound. The classic conception of timbre was not completely abandoned as it presented some interesting features in some cases and most of the works that dealt with timbre ever since still refer to it.

Among the most classic references in that domain are the works of Wessel, Grey, Slawson, Plomb and Risset. All of these make reference to some time/frequency representations (such as sonograms, pitch-synchronous analysis or more rarely wavelets) which add the notion of temporal evolution to the classic conception of timbre. Pitch synchronous analysis can be seen as a special case of a sonogram for which the signal is locally re-sampled at a multiple of its fundamental frequency and assumed to be perfectly periodic. Sonograms are nothing else but a short-time Fourier analysis applied to a sound. It consists in sliding a short temporal window on the signal and decompose this short chunk of observation with Fourier's tool as if it were a piece of an infinite support stationary time series. This representation offers a set of interesting features related to human perception of sound. The main reasons for referring to the Fourier transform of a sound are the following popular beliefs.

(i) The sine functions which are the basis on which Fourier decomposes a signal play a very important physical and perceptive role.

(ii) The spectrum of a sound is a set of n couples (amplitude A_n, frequency f_n) which leads to a multidimensional representation of timbre. Additionally, the spectral envelop (i.e. the series A_n) was shown to carry a lot of information in some cases (such as voice for instance).

(iii) The separation between amplitude and phase for each spectral component was confirmed by some studies on phase perception.

(iv) The perceptual notion of harmony within a complex sound can be interpreted through a model based on a set of distinct sine waves in a satisfying way.

Another argument for the use of a spectral representation is the fact that estimates of higher order statistics can seem to be computationally expensive and to require a prohibitive amount of data.

2.2.Assumptions behind spectral analysis

First of all let's recall that this part makes an assumption concerning the randomness of the signals. Indeed, we are viewing sampled sounds here as discrete time, wide-sense stationary stochastic processes x. The spectral distribution of a stochastic process x is traditionally defined as the Fourier transform of its autocorrelation function.

$$S_x(f) = \sum_{n\in Z} R_x(n).e^{-2i\pi nf}$$

In fact this definition is not completely correct as this series may not always converge. The correct notion of the spectral distribution relies on the concept of the spectral measure and the foundation of this notion is the Bochner theorem.

Theorem: Let $(r_n)_{n\in Z}$ be a sequence of elements of C. The sequence $(r_n)_{n\in Z}$ is positive semi definite (i.e the function $K(m,n) = r_{m-n}$ is positive semi definite) if and only if there is a positive measure μ on I such that

$$r_n = \int_I e^{2i\pi nf}\mu(df) \qquad \forall\, n \in Z$$

in this case the measure μ is uniquely defined.

If $x=(x_n)$ is a discrete-time stationary stochastic process and $(R_X(n))$ its autocorrelation, then this autocorrelation function is a positive semi definite sequence and one can define uniquely a positive measure $\mu_X(df)$ such that

$$R_x(n) = \int_I e^{2i\pi nf}\mu_x(df) \qquad \forall\, n \in Z$$

this measure is called the spectral measure of the process x. When this measure is continuous with respect to the Lebesgue measure, one can find a positive function $S_X(f)$ such that $\mu_X(df)=S_X(f)df$. The process is then said to be purely non-deterministic and this function defines the spectral density of the process x. Substituting this into the previous integral gives us a familiar relationship between the spectrum and the autocorrelation function:

$$R(n) = \int_I e^{2i\pi nf} S_x(f)df$$

but this approach allows us to realize that the notion of spectrum as a positive (and bounded) function can break down easily if the spectral measure is not continuous with respect to df.

In addition to the assumption that the analyzed sound is a wide sense stationary random process, the notion of spectrum relies also on the purely non-deterministic property of the process. We will see exactly what this means.

2.3.Deterministic / non-deterministic processes

Let $x=(x_n)$ be a discrete time wide sense stationary stochastic process and $(R_X(n))$ its autocorrelation, we recall that one can define uniquely its spectral measure $\mu_X(df)$ by

$$R_x(n) = \int_I e^{2i\pi nf}\mu_x(df) \qquad \forall\, n \in Z$$

Definition: Let's write $H_f(x)$ the vectorial subspace of finite linear combinations of the random variables x_n. $H(x)$ is the Hilbert subspace of $L^2(\mu_X)$ spread by the r.v $(x_n)_n$, i.e. the closure of $H_f(x)$. Also, $H_n(x)$ will designate the subspace spread by x_k for $k\le n$.

One observation is that $L^2(\mu_X)$ and $H(x)$ are isomorphic. In fact, one can find a unitary operator linking these two spaces. We'll skip the justification of this observation and introduce directly Kolmogorov's isomorphism.

Definition: The Kolmogorov isomorphism associated to the process x is the unitary operator V_X from $L^2(\mu_X)$ to $H(x)$ defined as follows:

$$\left(\sum_n a_n e^{2i\pi f}\right) \leftrightarrow \left(\sum_n a_n x_n\right) = V_x\left(\sum_n a_n e^{2i\pi f}\right)$$

(sums are intended to be finite)

The purpose of linear prediction theory is to evaluate the projection of the random variable x_n onto the spaces $H_{n-p}(x)$ spread by the random variables x_k for $k\le n-p$ (where $p\ge 1$). In general, it is fairly easy to evaluate the projection of a vector on a subspace for which we know an orthonormal basis. Therefore in our case, it would be ideal to extract a white random process ν of variance 1 which would be correlated with x and verify:

$$H_n(x) = H_n(\nu) \text{ for any } n$$

In this case, the process x could be written under the causal form:

$$x_n = \sum_{k=0}^{\infty} h_k \nu_{n-k} \quad \text{where} \quad \sum_{k=0}^{\infty} |h_k|^2 < \infty$$

because of the equality $H_n(x) = H_n(\nu)$, we'd get $x_n / H_{n-p}(x) = \sum_{k=p}^{\infty} h_k \nu_{n-k}$ and in particular for p=1, we would observe that $x_n - x_n / H_{n-1}(x) = h_0 \nu_n$. This brings us to the definition of the innovation process.

Definition: The innovation of a process x is the process I defined as $I_n = x_n - x_n / H_{n-1}(x)$.
x is said to be **deterministic** (resp. **non-deterministic**) when $I_n=0$ (resp. $I_n \neq 0$). When x is non-deterministic, we can define the normalized innovation process as $\nu_n = I_n \Big/ \left(E\left[I_n^2\right]\right)^{1/2}$.

Intuitively, the variance of the process I_n is a measure of the information brought by x_n after the random variables x_k had been observed for $k \leq n-1$. The bigger this variance is, the more x will behave as a white noise.

Remark: If x is deterministic, x_n will belong to $H_{n-1}(x)$; but x_{n-1} will belong to $H_{n-2}(x)$, so by substitution, x_n will belong to $H_{n-2}(x)$. Iterating this scheme will prove that x_n belongs to all the $H_k(x)$, which is to say that $H(x) = \cap_k H_k(x) = H_n(x)$.
If x is non-deterministic and ν is its normalized innovation, then by construction it is obvious that ν_n belongs to $H_n(x)$, which is to say that $H_n(\nu) \subset H_n(x)$.

The random process x will be said to be **purely non-deterministic** when $H_n(\nu) = H_n(x)$.

2.4.Wold's statement

Theorem: For any random process x, $H_n(x)$ has the following orthogonal decomposition:

$$H_n(x) = H_n(\nu) \oplus \cap_k H_k(x)$$

This decomposition is referred to as "Wold's decomposition". This result shows us that x is purely non-deterministic if and only if $\cap_k H_k(x) = \{0\}$.

Proof: To prove this statement, all we need to show is that a vector y of $H_n(x)$ is orthogonal to all the ν_k for $k \leq n$ if and only if it belongs to all the $H_k(x)$. By construction of the innovation, we clearly have $H_n(x) = H_{n-1}(x) \oplus \{\nu_n\}$. Therefore y is orthogonal to ν_n if and only if it belongs to $H_{n-1}(x)$. The same way, $H_{n-1}(x) = H_{n-2}(x) \oplus \{\nu_{n-1}\}$ and so forth. Iterating this process shows that y of $H_n(x)$ is orthogonal to all the ν_k for $k \leq n$ if and only if it belongs to all the $H_k(x)$ (i.e. $y \in \cap_k H_k(x)$).

Let's go back to the spectral measure of our process for a moment. As we saw, $\mu_x(df)$ doesn't have to be continuous with respect to Lebesgue's measure df but we can always decompose it uniquely as $\mu_x(df) = S_x(f)df + \mu_x^s(df)$, where $\mu_x^s(df)$ is a singular measure with respect to df (i.e. the support of $\mu_x^s(df)$ is of measure zero with respect to df). There is a strong relationship between Wold's decomposition, the decomposition of the spectral measure, and the notion of determinism.

Theorem: Let x be a non-deterministic process and ν its normalized innovation. Then the two random processes defined by

$$y_n = x_n / H_n(\nu) \quad \text{and} \quad z_n = x_n / \cap_k H_k(x)$$

are respectively purely non-deterministic and deterministic, moreover we have:

$$\mu_y(df) = S_x(f)df \quad \text{and} \quad \mu_z(df) = \mu_x^s(df)$$

A few words about the proof: We already know (from earlier) that any element of $H_n(\nu)$ will be purely non-deterministic and that any element of $\cap_k H_k(x)$ will be deterministic.
An element of $H_n(\nu)$ is a linear combination of a white noise, which can be seen as the output of a linear filter excited by a white noise. A recall to linear filtering will tell us that the power spectrum of such a process is the square of the absolute value of the transfer function of that filter and that therefore, it will be a nice continuous function.
Any element y of $\cap_k H_k(x)$, being linearly predictable will obey a relationship of the kind $y_n = \sum_k a_k y_{n-k}$ which can be interpreted as $G(z)Y(z)=0$ where $G(z) = 1 - \sum_k a_k z^{-k}$. This last form (seen as a linear filtering scheme) implies in the spectral domain that $\int_I \left|G(e^{2i\pi f})\right|^2 \mu_y(df) = 0$. The spectral measure of y being non-negative, it should equal zero everywhere except on the countable set of points on which $\left|G(e^{2i\pi f})\right|^2$ is zero. If we are not convinced that this set is indeed countable, we can define $S(z) = G(z)G^*(1/z)$ (which implies that $S(e^{2i\pi f}) = \left|G(e^{2i\pi f})\right|^2$) and then realize that, being factorable, it should verify the Paley-

Wiener condition which states that $\int_I \left|\ln\left(S(e^{2i\pi f})\right)\right| df < \infty$. This proves that the spectral density of Y will be singular with respect to df, being in the form: $\mu_y(df) = \sum_k c_k \delta(f - f_k)$.

The final step to the proof of this theorem is to use the uniqueness of the spectral measure's decomposition and the orthogonality of Wold's decomposition.

This is the essence of the relationship between the purely non-deterministic property of a process and the meaning of its power spectrum.

2.5.In the real world

When the stochastic process is an observed signal and when the analyzing tool is a computer, objects such as autocorrelation functions, measures and the Fourier transform loose most of their meaning. Any estimate of statistics relies on the ergodicity of the system, allowing averages over instances and averages over time on a single instance to be equivalent. Expectations become averages over a limited number of observations and the eventual singularities of the spectral measure become spikes in the periodogram (or other approximate measure of the spectrum). The boundary between determinism and non-determinism introduced earlier relies on a decision rule that detects spikes. But besides these limitations imposed by the nature of digital signal processing, the notions of deterministic and non-deterministic processes suffers from some more intrinsic limitations. They are biased by linear system theory.

Indeed, the definition of the innovation, as being a white noise uncorrelated with the past values of the process, only takes second order statistics in account. It is constructed on the orthogonality principle which finds its foundation in linear mean-square estimation and can be somehow related to the Wiener-Hopf equation (where the purpose is to match correlation functions). When innovation is related intuitively as a measure of the "additional" information brought by a new observation when the past is known, we ought to be skeptical. Being uncorrelated with the past doesn't imply being independent from it. In fact, a correct measurement of this "additional" information requires the ability to estimate the joint probability of an increasing number of successive $(x_n)_n$ and to compute the corresponding entropies. These entropies give birth to the notions of redundancy and information which are more likely to give a better answer to the "deterministic vs. stochastic" question. These objects will be presented in the following section.

The notion of a deterministic (or predictable) process that is introduced by Wold's decomposition characterizes only a subclass of deterministic systems: deterministic AND linear systems (the future is a linear combination of the past). A process which, through Wold's tool, may appear to be non-deterministic or even purely non-deterministic, is not guaranteed to be stochastic at all. It might be the chaotic output of a non-linear deterministic dynamical system. The estimate of the second order statistics of a deterministic but chaotic, system can be amazingly similar to the ones of a purely random white noise.

To come back to the notion of complexity we presented in the introduction, we now see that the linear approach to signal modeling will give up determinism as soon as the system presents some non-linearities. In that scheme, the number of degrees of freedom of the model is related to both the description of the deterministic part (i.e. the number of terms of a linear constant coefficient difference equation) and the resolution of the description of the purely non-deterministic part (such as the order of an AR, MA or ARMA model). Both can be numbers of lag values.

3.Entropy, Information, Redundancy

3.1.Entropy

Let's consider a particular random variable, S the set of its possible values, and p_i its probability distribution. If the random variable is of discrete-type, its distribution p_i implies a partition of S and the notion of entropy was introduced as a measure of the uncertainty of this partition. The precise definition of entropy of a partition was derived from a set of postulates imposed by our understanding of uncertainty and the uniqueness of a function verifying these postulates. The entropy of a discrete-type random variable is then defined as the entropy of this partition.

Definition: The entropy H(x) of a discrete-type random variable is defined as:

$$H(x) = -\sum_i p_i . \ln(p_i) = E\{-\ln(p(x))\}$$

In the case where x is a continuous-type random variable, we can't directly relate a partition to it any longer. We are reduced to forming a discrete-type random variable by rounding x as follows:

$$x_\delta = n\delta \quad \text{if} \quad (n-1)\delta < x \leq n\delta$$

which leads to

$$P\{x_\delta = n\delta\} = \int_{(n-1)\delta}^{n\delta} p_x(X).dX \approx \delta . p_x(n\delta)$$

but then

$$H(x_\delta) = -\sum \delta . p_x(n\delta).\ln\left(\delta . p_x(n\delta)\right)$$
$$= -\ln(\delta) - \sum \delta . p_x(n\delta).\ln\left(p_x(n\delta)\right) \xrightarrow[\delta \to 0]{} \infty$$

Therefore, H(x) can't be defined as the limit of $H(x_\delta)$ when δ get infinitely small, but as the limit of the sum :

$$H(x) = \lim_{\delta \to 0}\left[H(x_\delta) + \ln(\delta)\right] = -\int p_x(X).\ln\left(p_x(X)\right)dX$$

Definition: The entropy H(x) of a continuous-type random variable is defined as:

$$H(x) = -\int p_x(X).\ln\left(p_x(X)\right)dX = E\left\{-\ln\left(p_x(x)\right)\right\}$$

The generalization from the discrete to the continuous cases is not intuitive. This should be kept in mind when one wants to estimate the entropy of a continuous-type random variable from a sampled and quantized version of its observation. After all, a computer will only manipulate discrete and finite resolution data and a blind estimation of the entropy of this discrete data might be a heavily biased measure of the original random variable's entropy.
When x is a random vector, these definitions generalize in a straight forward way through the notion of **joint entropy** (or **block entropy**) as $H(x,y) = E\left\{-\ln\left(p_{x,y}(x,y)\right)\right\}$. We can also define conditional entropy using this expectation form.

3.2.Mutual information

Based on the notion of unions of partitions, the **mutual information** of two random variables x and y is defined from entropy as the function:

$$I(x,y) = H(x) + H(y) - H(x,y)$$

which yields $I(x,y) = E\left\{\ln\frac{p_{x,y}(x,y)}{p_x(x)p_y(y)}\right\}$

and also to $I(x,y) = H(x) - H(x|y) = H(y) - H(y|x)$
(using Bayes' rule)

When more than two random variables are involved, the mutual information can be generalized to the notion of **joint mutual information** as:

$$I(x_1,x_2,..,x_d) = \sum_{k=1}^{d} H(x_d) - H(x_1,x_2,..,x_d)$$

and another useful object is the **redundancy**, defined as the increment of the joint mutual information as the number of random variable grows:

$$R(x_1,x_2,..,x_d) = I(x_1,x_2,..,x_d) - I(x_1,x_2,..,x_{d-1})$$
$$= H(x_d) + H(x_1,x_2,..,x_{d-1}) - H(x_1,x_2,..,x_d)$$

3.3.Sampled strict sense stationary stochastic process

In order to become more familiar with the particular case that will interest us in what follows, let's look at these same objects when the random variables $(x_n)_n$ are samples of a strict sense stationary stochastic process x.
The strict sense stationarity of x states that all of its statistics (of any order) are invariant through time:

$$p\left(x(t_1),x(t_2),..,x(t_d)\right) = p\left(x(t_1-\tau),x(t_2-\tau),..,x(t_d-\tau)\right)$$
$$\forall (d,\tau) \in N \times \Re$$

So when $(x_n)_n$ designate successive samples of x, this property implies:

$$H(x_n,x_{n-1},..,x_{n-d+1}) = H(x_k,x_{k-1},..,x_{k-d+1}) \quad \forall (d,n,k) \in N^3$$

which allows us to simplify our notations:

$$p(x_n,x_{n-1},..,x_{n-d+1}) \equiv p_d(\tau)$$
$$H(x_n,x_{n-1},..,x_{n-d+1}) \equiv H_d(\tau)$$
$$I(x_n,x_{n-1},..,x_{n-d+1}) \equiv I_d(\tau)$$
$$R(x_n,x_{n-1},..,x_{n-d+1}) \equiv R_d(\tau)$$

where τ is the corresponding sampling period.

Another aspect of a time-sampled observation is that it will always be quantized in amplitude. There is no such thing as an infinite resolution measurement and our observation will always look like the instance of a discrete-type random process. We know from before that there's no nice continuity between discrete-type and continuous-type processes when it comes to entropy measurement and so if N is the number of bins imposed by our finite resolution, it is only fair that N should be taken as an extra variable of our estimates:

$$H_d(\tau,N) \;;\; I_d(\tau,N) \;;\; R_d(\tau,N)$$

The last point concerns the estimation of the multi-dimensional statistics of the process. Indeed, one has very rarely access to several instances of the same process but rather a single observation in time. The best we can do is to estimate statistics by averaging measurements of this single observation in time and the validity of these estimated statistics requires the process to be **ergodic**.

4.The validity of lag spaces

4.1.State space and lag space

Let's consider a dynamical system described by its state variables $\underline{\mathbf{x}}$ related to each other in a general fashion:

$$\frac{d\underline{\mathbf{x}}}{dt} = f(\underline{\mathbf{x}}).$$

The evolution of the system from a given initial state can be monitored by the trajectory of the vector $\underline{\mathbf{x}}$ as time passes by. This vector $\underline{\mathbf{x}}$ lives in the **state space** and the observation of this trajectory can teach us a lot about the internal mechanism of this dynamic system (i.e. about the relationships f). However, the nature and even the number of these internal states (or degrees of freedom) are usually unknown and we only have access to a subset of them if not only one. Let's suppose the only observation we have is a single variable z. Even though the dimension of our observation is one, we can choose to build a vector of arbitrary dimension d by using lag values of z:

$$\underline{\mathbf{y}}(t) = (z(t), z(t-\tau), \ldots, z(t-(d-1)\tau))^T$$

This vector lives in the **lag space** in which it will draw another trajectory as time passes by.

4.2.The embedding theorem

Given two spaces X and Y, a **mapping** between them is a function f that associates every element x of X with the uniquely determined element y = f(x) of Y. The element y is then called the image of x and x the preimage of y. When the spaces X and Y are metric, the notions of continuity and smoothness can be introduced in that scheme. A C^k mapping that is bijective is called a **diffeomorphism**. A smooth mapping f that is injective is called an **immersion**. If we also want the mapping to preserve topological properties, it will have to be **proper**. A map is proper if the preimage of every compact set is a compact set. Finally, a proper immersion is called an **embedding**. If the details of the definition of an embedding seem a little tedious or obscure, one can think of an embedding as being a smooth local change of coordinates. It might disfigure a subset A of X but will keep its local properties and its fine structure intact.

The following result was stated and proved by Floris Takens in 1981 in his paper "Detecting strange attractors in turbulence" and we present it here in its original form.

Theorem: Let M be a compact manifold of dimension m. For pairs (φ, y), $\varphi: M \to M$ a smooth diffeomorphism and $y: M \to \Re$ a smooth function, it is a generic property that the map $\phi_{(\varphi,y)}: M \to \Re^{2m+1}$, defined by $\phi_{(\varphi,y)}(x) = \left(y(x), y(\varphi(x)), \ldots, y(\varphi^{2m}(x))\right)$ is an embedding; by "smooth" we mean at least C^2.

This result being a "generic" property means that it is not always true but that the set of cases for which it'll break is of probability 0. In other words, perturbing an unlucky case in an infinitely small way will make the result hold. In the same paper, Takens proves two other theorems of the same flavor that generalize this result to different types of map ϕ. The same year Mañé, in an independent study, came up with a similar result.

The final step is just to recognize that the relationship between the state space of a dynamical system and the lag space created from the observation of one of its state variable is a particular case of embedding theorem's hypotheses. In other words, the trajectory of a 1D observation in a d-dimensional lag space and the trajectory of the entire system in its state space differ only by a smooth local change of coordinates (given that d is big enough). This magical sounding result holds basically because we restrict our attention to the dynamics on a finite-dimensional attractor. The upper bound of this dimension is once again related to the complexity we're allowing our model to have.

In the context of modeling, classification or resynthesis, this result tells us that there is no need for "hidden variables" other than the lag values of the time series we are studying. Furthermore, the number of necessary lag values is directly related to the number of the system's degrees of freedom and this number can also be estimated by measuring statistics on the initial observation. We can also note that any invertible transformation of these lag vectors will also work. If this transformation were to be linear, this means that any linearly filtered version of the observation works just as well. This point could be an open door between linear systems theory and non-linear dynamical systems.

The state space of a system is an object we will never have access to whereas the lag space doesn't give us any such problem. This strong relationship between these two objects will allow us from now on to forget completely about the state space and to manipulate lag values of an observation as if they were directly state variables of our system.

4.3.Dimensions

The sufficient dimension d of the lag space and the number of degrees of freedom are now the same thing. Let's consider a deterministic system that produces the discrete time observation $x_n=x(n\tau)$. By "deterministic", we mean here that the past values of this observation allow the prediction of its future values with no error. This can be written as:

$$p(x_n | x_{n-1}, x_{n-2}, \ldots) = \delta(x_n - f(x_{n-1}, x_{n-2}, \ldots))$$

(Dirac distribution)

For this system to have a finite number of degrees of freedom, there has to be a dimension d such that:

$$p(x_n | x_{n-1}, x_{n-2}, \ldots) = p(x_n | x_{n-1}, x_{n-2}, .., x_{n-d}) = \delta(x_n - f_d(x_{n-1}, x_{n-2}, .., x_{n-d}))$$

If such a dimension exists, and given what we said before about sampled strict sense stationary stochastic processes, then we'd have:

$$\begin{aligned} p_{k+1}(\tau) &= p(x_n, x_{n-1}, x_{n-2}, .., x_{n-k}) \\ &= \delta(x_n - f_d(x_{n-1}, x_{n-2}, .., x_{n-d})).p(x_{n-1}, x_{n-2}, .., x_{n-k}) \\ &= \delta(x_n - f_d(x_{n-1}, x_{n-2}, .., x_{n-d})).p_k(\tau) \qquad \text{for any } k \geq d \end{aligned}$$

And therefore, for any k≥d, we'd have:

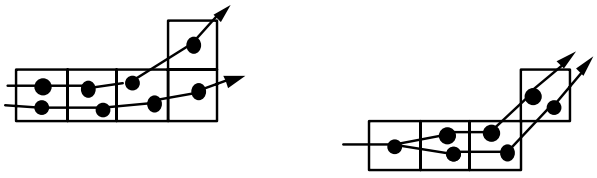
$$\begin{aligned} H_{k+1}(\tau) &= H_k(\tau) \\ I_{k+1}(\tau) &= (k+1)H_1(\tau) - H_{k+1}(\tau) = H_1(\tau) + I_k(\tau) \\ R_{k+1}(\tau) &= I_{k+1}(\tau) - I_k(\tau) = H_1(\tau) \end{aligned}$$

In that case, this dimension is referred to as the **embedding dimension**. A natural way to determine this embedding dimension is therefore by evaluating the observation's joint entropy for various successive dimensions and by watching its evolution as the dimension increases.

Having gone through the precise definition of entropy for continuous-type random variables, we are now aware that its estimation from a quantized (i.e. finite resolution) observation can be tricky. Considering our quantized observation as the instances of a discrete-type random variable can mislead us to a biased measure of the system's entropy (especially for chaotic systems for which the attractor has a fractal dimension). It seems necessary to estimate a parametric form of a continuous-type probability mass function (PMF) before computing the corresponding entropy. In a way, this step is related to the "histogram smoothing" action of Parzen's windows or to any kind of interpolating method.

4.4.Resolution

An evaluation of entropy based on the estimation of this parametric PMF promises to be more reliable but it is important to keep in mind the fact that the finite resolution of our observation can mislead us anyway. The following figure tries to illustrate two cases where such a thing happens. The points linked by a continuous line represents the "real" infinite resolution trajectory the observation in lag space whereas the boxes are the quantized version of the same trajectory.



(a) (b)

Case (a): We are locally observing two separate trajectories that don't cross. The system could very well be deterministic but if we only have access to the quantized observation, it will appear that the trajectory splits, which could mislead us to believe that the system has some randomness to it.

Case (b): The trajectory really splits but this split can not be detected through the low resolution of our observation. We might end up conclude that the system is deterministic whereas it's not.

The evolution (or should we say growth) of entropy with an increasing resolution is related to the **attractor's dimension**. Indeed, if d is big enough (i.e. at least the embedding dimension) then the object:

$$\lim_{N \to \infty} \frac{\sum_{x_i} p_d(x_i).\ln(p_d(x_i))}{-\ln(N)} = \lim_{N \to \infty} \frac{H_d(\tau, N)}{\ln(N)}$$

measures the dimension of the set of our system's solution (i.e. the attractor if the system were deterministic).

5.Modeling in the lag space

5.1.Probability Mass Function estimation

Given a sampled instance of a stationary ergodic stochastic process x and a dimension d, the problem here is to estimate a parametric form of the PMF $p_d(\tau)$. A natural object to think of is a histogram. Yet, as the dimension d increases, the internal representation of the histogram with a decent resolution requires an exploding amount of memory (N^d where N is the number of bins given by our resolution). Therefore a straight forward application of Parzen's idea becomes quickly unfeasible. Kris Popat and Rosalind Picard adapted Parzen's technique "by replacing the original data with a smaller set of representative points and by adapting the sizes and shapes of the kernel to match the statistics of the regions they represent". It's a very similar idea to radial basis functions (RBF) for data fitting. These most representative points are chosen by a standard clustering algorithm. In addition to this dramatic reduction of our problem's size, the kernels they propose are separable probability density function which will allow a recursive estimation of the conditional probabilities.

The proposed form for the estimate of the PMF is the following:

$$p(x_n, x_{n-1}, .., x_{n-d+1}) = \sum_{m=1}^{M} w_m . \prod_{k=0}^{d-1} K_{m,k} . e^{-(x_{n-k} - \mu_{m,k})^2 / (2\sigma_{m,k}^2)}$$

where M is the number of representative points, $w_m > 0$, and $\sum_{m=1}^{M} w_m = 1$.

Assuming that we now possess a fair parametric estimate of our observation's joint probability distribution for different dimensions, the question is to know what we'll use it for. We could use it to evaluate entropies for the successive dimensions in the hope of finding and embedding dimension. Once this embedding dimension is detected (or estimated) we could then consider that the system is deterministic and forget about these joint probabilities when we model it. This would be the **deterministic approach**. The other option is to use our estimates of these joint probabilities as a model of the system. Unless the parametric form of these probabilities have some Dirac distribution terms (which is very unlikely as the method is based on a "smoothing" idea) the model they will describe will carry some uncertainty. In other words, the system would be modeled as a random number generator which probability distribution is derived from the conditional probability distributions of the data. This would be the **stochastic approach**.

5.2.Deterministic approach

As we pointed out earlier, this approach assumes the existence of an embedding dimension. Having gone several times through the difficulties involved in entropy estimations, we know that the existence of such an object might not appear from our data as clearly as we'd like. This dimension d will be the result of a fairly arbitrary decision rather than an unquestionable observation. This assumption can be stated as:

$$p(x_n | x_{n-1}, x_{n-2}, .., x_{n-d}) = \delta(x_n - f(x_{n-1}, x_{n-2}, .., x_{n-d}))$$

$$\text{or simply } x_n = f(x_{n-1}, x_{n-2}, .., x_{n-d})$$

Our system is therefore entirely characterized by a set of d initial conditions and a representation of the function f(). In order for our model to generalize the behavior of our system with variations on the initial conditions for instance, the representation of f() should be defined on a wider set than the training data (i.e. our observation). Given also that we want to avoid a prohibitive size of this representation, the goal of this training should be to estimate a parametric form for f().

For this purpose, there are two basic sets of approaches, the global and the local approaches. A global approach assumes some fixed architecture for a closed form of the function f() on the entire set on which it's defined, and tunes the parameters of this architecture in order to fit the training data by minimizing some criteria. An example would be to fit a polynomial of dimension d and fixed order N to the observation with a least mean square criteria. A local approach will typically use the training data as the model and an interpolating method as the mean to generalize it. Local linear modeling is an example of a local approach. These approaches are not mutually exclusive as one can choose to take a local approach on a representative subset of the training data. Each one of these representative points carries some information about the system's behavior in the corresponding neighborhood to the rest of the model. These points are sometimes referred to as the anchor points of radial basis functions. If the number of anchor points is fixed, than there is an assumption concerning the closed form of the model even though it is based on interpolations between observed data. We could qualify this method as being "glocal" if we felt like inventing a word.

A local approach usually gives better performances as it doesn't constraint the model as much as a global method would. Yet, the representation it provides is just as big as the training data, which makes it heavy and rigid. By rigidity we mean here that the model often lacks a restricted number of knobs allowing mutations of the system. In that respect, global models are preferable because of their smaller size, their usually smaller computation requirements, and their limited fixed number of parameters (knobs).

5.3.Local linear modeling as an evaluation of the approach

Local linear modeling is a heavy approach to the estimation of the function f(). However, this method can be used to evaluate the validity of the deterministic approach for a given system. If k is the number of neighbors used to construct the local linear model around each possible lag vector of dimension d, the evolution of the performance (for prediction) of the local linear model with the number k can reveal important information about the system. When k is the smallest (i.e. k=1), the model is a lookup of the closest neighbor in lag space. When k is very large (i.e. k=number of observation), the model is a global linear model (AR). The plot of this evolution for different dimension was introduced by Casdagli in 1991 in terms of "deterministic vs. stochastic" plot (DVS).

If we restrict this study to a given dimension d for our lag space, the following figure illustrates what these plots can look like for three different systems. This figure is only an illustration, it is not the result of a particular analysis.

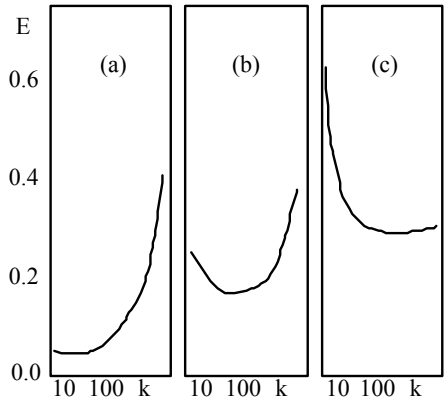


Fig. ? - Example of DVS plots for three systems.

Case (a): The prediction error reaches a low minimum for the dimension d with which he plot was made, which tells us that the system is fairly deterministic. The error increases with k and that means that the system in non-linear.

Case (c): The failure of a local linear model with a small k can be seen as overfitting. The flat part for large k tells us that the system is fairly linear. The globally poor performance tells us that the system is most probably stochastic.

Case (b): This case falls between the two previous. It could be a higher dimension non-linear chaotic system or a non-linear stochastic system.

This example illustrates the ambiguity between stochasticity and high dimension non-linearity of these plots. As we can recall, linear system theory could not differentiate between stochasticity and non-linearities (even for low dimension system) so this new ambiguity is still an improvement.

5.4.Relation between our PMF and a RBF approach

To recall the relationship between the method we've used to estimate the (d+1) dimension joint probability distribution of our observation and radial basis functions, let's have a closer look at this "glocal" approach. In fact, let's assume that we chose the anchor points of our RBFs through the same clustering method we've used for the PMF. Let's also imagine that d-dimensional RBFs we chose have the same shape than the (d+1) dimensional Parzen windows we used.

This means that our model for f() is:

$$x_n = f(x_{n-1}, x_{n-2}, .., x_{n-d}) = \sum_{m=1}^{M} \alpha_m . \mu_{m,0} . \prod_{k=1}^{d} b_{m,k}\left((x_{n-k} - \mu_{m,k})^2\right)$$

where the α_m are such

$$\text{that: } \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{d} b_{m,k}\left((x_{n-k} - \mu_{m,k})^2\right) = 1$$

but we also have:

$$p(x_n, x_{n-1}, x_{n-2}, .., x_{n-d}) = \sum_{m=1}^{M} w_m \prod_{k=0}^{d} b_{m,k}\left((x_{n-k} - \mu_{m,k})^2\right)$$

So by using Bayes rule, we can get an expression of the conditional PDF of x_n given the past d lag values.

$$p(x_n | X_{n-1}, .., X_{n-d}) = \frac{\sum_{m=1}^{M} w_m . b_{m,0}\left((x_n - \mu_{m,0})^2\right) \prod_{k=1}^{d} b_{m,k}\left((X_{n-k} - \mu_{m,k})^2\right)}{\sum_{m=1}^{M} w_m \prod_{k=1}^{d} b_{m,k}\left((X_{n-k} - \mu_{m,k})^2\right)}$$

and finally,

$$E\left[x_n | X_{n-1}, .., X_{n-d}\right] = \int_{X_n} X_n . p(x_n | X_{n-1}, .., X_{n-d}) . dX_n$$

$$= \frac{\sum_{m=1}^{M} w_m . \mu_{m,0} \prod_{k=1}^{d} b_{m,k}\left((X_{n-k} - \mu_{m,k})^2\right)}{\sum_{m=1}^{M} w_m \prod_{k=1}^{d} b_{m,k}\left((X_{n-k} - \mu_{m,k})^2\right)}$$

So if we pose $\alpha_m = \frac{w_m}{\sum_{m=1}^{M} w_m \prod_{k=1}^{d} b_{m,k}\left((X_{n-k} - \mu_{m,k})^2\right)}$

(which satisfies the relationship specified earlier) we will observe that

$$E\left[x_n | X_{n-1}, .., X_{n-d}\right] = \sum_{m=1}^{M} \alpha_m . \mu_{m,0} . \prod_{k=1}^{d} b_{m,k}\left((X_{n-k} - \mu_{m,k})^2\right)$$

$$= f(X_{n-1}, .., X_{n-d})$$

So this proves that our estimate of the d+1 dimensional joint probability distribution can be used to model the system in a deterministic fashion through the computation of these conditional expectations. In that case, the method is rigorously equivalent to a RBF approach for which the basis functions match the kernels we've used for the PMF's estimation.

Note: If we recall Bayes' results in standard decision and estimation theory, we already knew that this conditional expectation was a very useful object. It is rigorously equal to the Bayes least square estimate of x_n based on the observation of $(X_{n-1}, .., X_{n-d})$.

$$\hat{x}_{BLS}(Y) = E[x|Y]$$

5.5.Stochastic approach

The last remark is an elegant introduction to the stochastic approach. As we saw that the conditional expectation of the (d+1) dimensional PMF is equivalent to a d-dimensional deterministic approach, it seems straight forward to introduce a model based on the same conditional PMF itself instead of its expectation.

Given the values $(X_{n-1}, .., X_{n-d})$ of the last d lag values, the prediction of the next lag value will be the

output of a random number generator whose PDF matches the conditional PDF:

$$p(x_n | x_{n-1},..,x_{n-d}) = \frac{p(x_n, x_{n-1},..,x_{n-d})}{\int_{X_n} p(X_n, x_{n-1},..,x_{n-d})dX_n}$$

Note: We recall here how easily one can create instances of an arbitrary PDF random variable from the instances of a uniformly distributed random variable. Let's consider two random variables x and y related to each other by x=g(y) were g() is a diffeomorphic function g:R->[0,1]. Let's suppose also that x is uniformly distributed on [0,1]. The relationship $p_X(X)dX=p_Y(Y)dY$ gives us:

$$\forall X \in [0,1],\ \frac{dy}{dx} = \frac{d}{dx}\left(g^{-1}(X)\right) = \frac{1}{p_y(Y)}$$

$$= \frac{1}{g'\left(g^{-1}(X)\right)} = \frac{1}{p_y(Y)}$$

$$= \frac{1}{g'(Y)} = \frac{1}{p_y(Y)}$$

and so $\forall Y \in R,\ p_y(Y) = g'(Y)$ i.e. $P_y(Y) = g(Y)$ (the cumulative function of y).

This tells us that if we posses a typical random number generator providing us with the instance X (in [0,1]) of x, we can create an instance Y of y (with arbitrary PDF $p_Y(Y)$) by applying the simple mapping $Y = P_y^{-1}(X)$.

It is important to note that such an approach to the system might not be an improvement over the deterministic approach. As an illustration, let's see what happens when the observation is the output of a simple 2D deterministic system. Let's even chose that system to be linear, namely:

$$x_n = (2\cos\theta)x_{n-1} - \lambda.x_{n-2} \quad (\text{where } \lambda \in]0,1[)$$

(x_n is a damping sine wave)

The next figure illustrates the estimate of the conditional probability distribution of x_n given x_{n-1}=y and x_{n-2}=x.

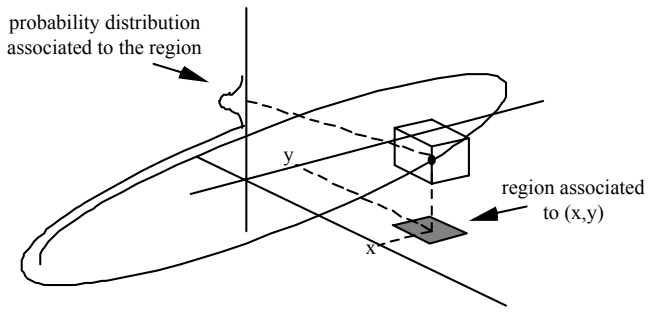


Fig.? - The "pessimism" of the PMF approach for modeling.

Chances are that given the finite number of "representative points" used for the estimation of the PMF, the variance of the estimate of this conditional probability distribution will not be zero (as it should be). This is why the figure is called "pessimism" of the PMF approach. The next figure illustrates the same point in terms of the system itself. (a) represents the true deterministic system, it's a simple plane, whereas (b) represents the model induced by the PMF approach (fuzzy plane).

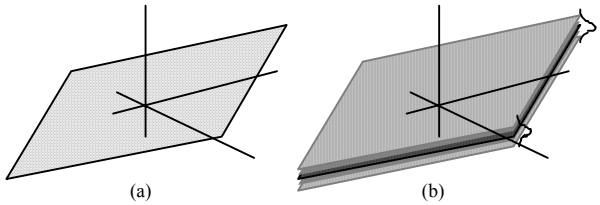


Fig.? - Stochastic model of a deterministic linear 2D system.

6.Implementation and results

6.1.The system

The data we chose to illustrate these approaches is a digital recording of the quasi-periodic part of a bowed violin string. The sampling frequency of the recording is 44.1 KHz and its resolution is 16 bits. A version of the PMF estimator presented earlier was implemented in C. We can explicitly tell the system the maximum number of clusters that it should use as well as the dimension of the lag space it should compute the PMF for. Once this analysis is completed, we can use this parametric PMF in order to resynthesize the original data. At this point, we can choose between a deterministic and a stochastic approach by using either conditional expectation or conditional probabilities.

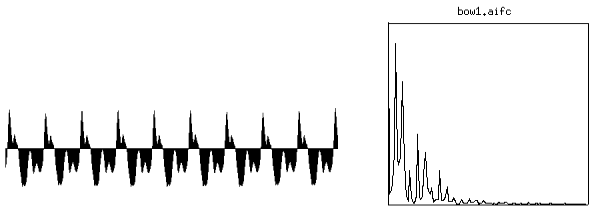


Fig.? - A chunk of the original data and its FFT-based spectrum estimation.

6.2.Conditional probabilities

In the context of our data and by experimenting with various numbers of clusters to be used for the PMF estimation, it appeared that 1000 clusters were doing a fairly good job for dimensions up to 5. The following figure presents some conditional probability distributions derived from the estimated PMF in dimension 5. As our

knowledge of the past increases (i.e. conditioning), we observe a dramatic reduction of the observation's randomness.

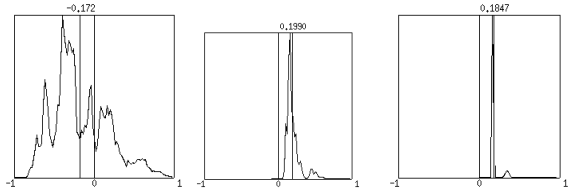


Fig.? - Estimated PMF with 1000 clusters. From left to right: p(x1), p(x2|X1), p(x3|X1,X2) where X1 and X2 where chosen arbitrarily.

6.3.Deterministic approach

Having an estimate of our data's probability mass function in dimension 5, we can now build a deterministic model of dimension 4 as:

$$\hat{X}_n = f(X_{n-1}, X_{n-2}, X_{n-3}, X_{n-4}) = E[x|X_{n-1}, X_{n-2}, X_{n-3}, X_{n-4}]$$

After feeding the first four samples of the original data as initial conditions, we can iterate this deterministic function f() and try to reconstruct the bowed string sound of the violin. The following figure is the result of this reconstruction.

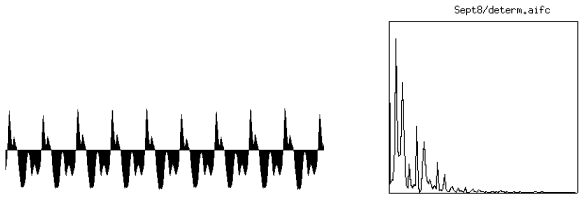


Fig.? - A chunk of the deterministic reconstruction and its FFT-based spectrum estimation.

The accuracy of this reconstruction is remarkable. Even after over 20000 iterations, the model is perfectly stable and accurate. The result of this reconstruction was saved in a sound file and it is very difficult to tell it apart from the original.

6.4.Stochastic approach

The same dimension 5 probability mass function can be seen as a stochastic model for our data. We can then synthesize an instance of this stochastic process by generating random numbers accordingly to the conditional probabilities:

$$p(x_n | X_{n-1}, X_{n-2}, X_{n-3}, X_{n-4})$$

Once again, the first four samples of the original data are fed as initial conditions and the rest of the reconstruction results from an iteration of the model. The following figure is the result of this process.

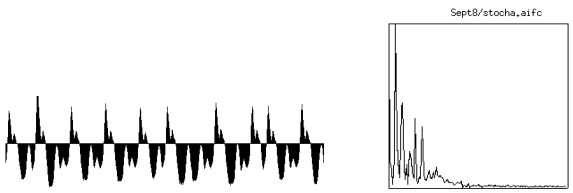


Fig.? - A chunk of the stochastic reconstruction and its FFT-based spectrum estimation.

In addition to being computationally very expensive, this approach doesn't lead to the accuracy we had with the deterministic approach. Ironically, this model could have been thought of as being more "complete" than the deterministic model which uses only the means of these conditionals.

7.Conclusion

7.1.What did we learn about the data?

The first information we get from the accuracy of the deterministic approach concerns the limited dimensionality of our data. It appears that this violin sound doesn't have more than 4 degrees of freedom.

The improvement to the deterministic model brought by an increase of the number of clusters is a first hint that the system possesses some non linearities. This observation becomes obvious when we look at the periodograms of the sound and of its reconstruction. Other than the fact that these two periodograms are very similar, they both show a fair number of peaks. One can count at least 12 or 13 harmonics in the periodogram of the sound produced by a system of dimension 4. A linear system of dimension 4 can not have more than 2 picks in its spectrum and we can therefore state with confidence that the estimated model is indeed non linear.

The failure of the stochastic approach has to do with the nature of our data. Being quasi-periodic, we could have anticipated the high predictability of our data. This example is a good illustration of what we introduced earlier as the "pessimism" of a stochastic approach.

Given that our data was quasi periodic to start with, there is no doubt we could have modeled it fairly accurately in a linear fashion. The aspect of its periodogram however suggests that such a linear model would have to be high dimensional (at least dimension 30 or 40). The reduction of the number of degrees of freedom to 4 was made possible by the introduction of non linearities in our model. In our case, it might seem like the price to pay was fairly high (1000 clusters).

7.2.Complexity v.s. Stochasticity

The point we've tried to illustrate through this paper and this example is the role of stochasticity in

modeling and its relationship with the complexity of the model. For both linear and non linear techniques, the assumption of randomness in our data can allow us to measure the information that it carries, providing us with some handles on our data before even making big assumptions concerning the architecture of a possible model. When it come to building the model itself, we then have the choice to consider this randomness as being part of the model or as being an artifact of our measurements. This choice is guided by the accuracy of the "deterministic" part of the estimated model and this accuracy is directly linked to the model's complexity.

As we saw, the complexity of a model can be though of in terms of dimension and non linearities of its deterministic part. The stochastic part of the model is often a desperate way to describe the deterministic part's error. When this error becomes very small, as for our example, the stochastic part of the model becomes unnecessary. In fact, it can even be a source of confusion because it is an artifact of our measurements. If no decent accuracy can be reached by the deterministic part of our model, either because we don't allow a high enough complexity or because the architecture we chose for the model is fundamentally wrong, then the introduction of a stochastic part in our model becomes a necessity. Whether this randomness is an estimate of the model's error or it translates true randomness in our data is a question that could theoretically be answered by entropy measurements. We still have to keep in mind the fact that these entropies are not deprived from measurement artifacts either so it seems that this fundamental question concerning the existence of randomness is still up in the air.

References

[1] W. M. Hartmann. *The frequency-domain grating*. Journal of the Acoustical Society of America 78(4), p.1421-1425, 1985.

[2] James A. Moorer. *Signal Processing Aspects of Computer Music - A Survey*. Computer Music Journal 1(1), p.4-37, 1977.

[3] Alan V. Oppenheim and Ronald W. Shafer. *Discrete-Time Signal Processing*. Prentice Hall, 1989. (Sampling, LCCDE, spectrum analysis)

[4] Kris Popat and Rosalind W. Picard. *Novel cluster-based probability model for texture synthesis, classification, and compression*. Proc. SPIE Visual Communications and Image Processing 1993, Boston.

[5] Rosalind W. Picard and Fang Liu. *A new Wold ordering for image similarity*. Proceedings of IEEE Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, April 1994.

[6] Xavier Serra and Julius O. Smith III. *Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition.* Computer Music Journal 14(4), p.12-24, 1990.

[7] Floris Takens. *Detecting strange attractors in turbulence*. Lecture Notes in Mathematics vol.898 (Springer, Berlin) p.366-381, 1981.

[8] Charles W. Therien. *Decision Estimation and Classification*. Wiley&Sons, 1989. (Nonparametric estimation and classification)

[9] Charles W. Therien. *Discrete Random Signals and Statistical Signal Processing*. Prentice Hall, 1992. (Stochastic modeling, Wold decomposition, spectrum analysis)