Chapter 2 : Music and Piecewise Gaussian Models

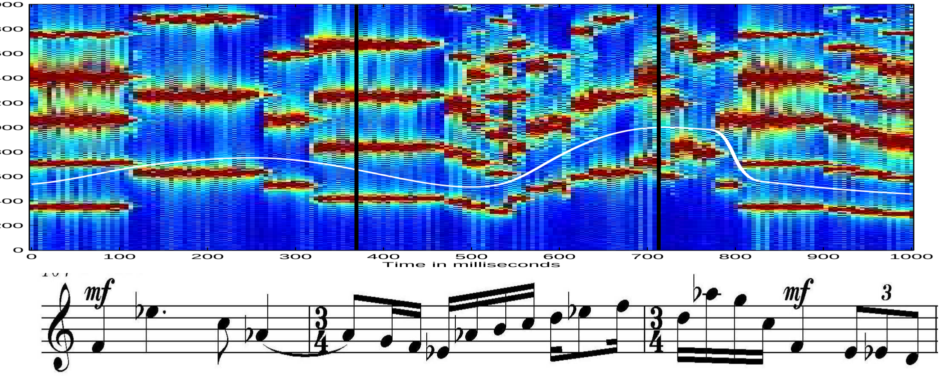


FIGURE: Three bars from an oboe playing ‘Winter711’ by Jan Beran : on the bottom, the score as humans write it (Figure courtesy of Chris Raphael), on the top the *spectrogram*. The spectrogram shows the distribution of power across frequencies as time progresses. The oboe has a rich set of harmonics : below the white line is the fundamental, in which the score is easily recognized, though stretched and squeezed a bit.

We are interested in constructing a stochastic model for the signal $s(t)$ which represents air pressure as a function of time while music is being played.

This relatively simple example will provide a good introduction into the basic ideas of Pattern Theory for variables with continuous values.

A typical piece of data might be given with a sampling interval $\Delta t = 1/8000$ seconds, so that if 5 seconds of data are considered, we have a sequence $s_k = s(k\Delta t), 1 \leq k \leq 40,000$ of real numbers and we want a stochastic model for this finite-dimensional piece of data.

We need to put in the model extra **hidden random variables** which represent the patterns. In this case, the main pattern consists in what is usually called the “musical score”. We need :

1. the number of notes m,
2. the times $t_i = k_i\Delta t$ where new notes begin, $1 < k_1 < k_2 < .... < k_m < N$,
3. the frequency ω_i of the i^{th} note in hertz (or its approximate integral period $p_i \approx 1/(\Delta t \cdot \omega_i) \in \mathbb{Z}$).

To construct this model, we will define a probability density $p(\vec{s}, m, \vec{t}, \vec{p})$ in **all** the variables. We can sample from this model to see whether it sounds like any music known to mankind (the simple model we will give will fail this test – but variants can improve it greatly).

But more significantly, we can use this probability distribution to reconstruct the score from an observed signal $\vec{s}_{\text{obs}}$. We recover the hidden variables m, $\vec{t}$ and $\vec{p}$ by maximizing the conditional probability

$$p(m, \vec{t}, \vec{p} \mid \vec{s}_{\text{obs}}) = \frac{p(\vec{s}_{\text{obs}}, m, \vec{t}, \vec{p})}{\sum_{m', \vec{t}', \vec{p}'} p(\vec{s}_{\text{obs}}, m', \vec{t}', \vec{p}')}.$$

When you use this general method, you have 3 problems :

- the first one is the construction of the model,
- the second one is finding an algorithm to maximize $p(m, \vec{t}, \vec{p} \mid \vec{s}_{\text{obs}})$ with respect to the variables m, $\vec{t}$ and $\vec{p}$,
- the third one is optimizing the parameters of the model to make it fit the data as well as possible.

Basics : Gaussian distributions

Definition

Let $\vec{x} = (x_1, ..., x_n)$ denote a vector in $\mathbb{R}^n$, we then define a **Gaussian distribution** *on $\mathbb{R}^n$ by its density*

$$p(\vec{x}) = \frac{1}{Z} e^{-(\vec{x}-\vec{m})^t Q(\vec{x}-\vec{m})/2},$$

where $\vec{m} \in \mathbb{R}^n$, Q is a $n \times n$ symmetric positive definite matrix, and Z is a constant such that $\int p(\vec{x}) d\vec{x} = 1$.

Gaussian distributions are very important in Probability Theory, particularly because of the following theorem.

Theorem

(Central Limit Theorem) *If $\vec{\mathcal{X}}$ in $\mathbb{R}^n$ is any random variable with mean 0 and finite second moments, and if $\vec{\mathcal{X}}^{(1)}$, ...,$\vec{\mathcal{X}}^{(N)}$ are independent samples of $\vec{\mathcal{X}}$ then the distribution of $\frac{1}{\sqrt{N}} \sum_{k=1}^{N} \vec{\mathcal{X}}^{(k)}$, tends, as $N \to +\infty$, to a Gaussian distribution with mean 0 and the same second moments as $\vec{\mathcal{X}}$.*

Proposition

The Gaussian distribution p defined above has the following properties :
a) $Z = (2\pi)^{\frac{n}{2}} (\det Q)^{-\frac{1}{2}}$.
b) $\int (\vec{x} - \vec{m}) p(\vec{x}) d\vec{x} = 0$, which means that $\vec{m}$ is the mean of p denoted by $\mathbb{E}_p(\vec{x})$.
c) If $C_{ij} = \int (x_i - m_i)(x_j - m_j) p(\vec{x}) d\vec{x}$ is the covariance matrix, then $C = Q^{-1}$.

Corollary

A Gaussian distribution is uniquely determined by its mean and covariance. And conversely, given any vector $\vec{m}$ and any positive definite symmetric matrix C, there exists a Gaussian distribution with $\vec{m}$ and C as its mean and covariance.

Let $p(\vec{x}) = \frac{1}{Z} e^{-(\vec{x}-\vec{m})^t Q(\vec{x}-\vec{m})/2}$ be a Gaussian distribution on $\mathbb{R}^n$. It is very helpful for our intuition to interpret and contrast in simple probability terms what $C_{ij} = 0$ and $Q_{ij} = 0$ mean :

• Fix $i < j$, then $C_{ij} = 0$ means that "the marginal distribution on (x_i, x_j) makes x_i and x_j independent".

The marginal distribution on x_i, x_j is defined by

$$p^{(i,j)}(x_i, x_j) = \int p(x_1, ..., x_n) dx_1 ... \widehat{dx_i} ... \widehat{dx_j} ... dx_n,$$

where the notation $dx_1 ... \widehat{dx_i} ... \widehat{dx_j} ... dx_n$ means that we integrate on all variables except x_i and x_j. When $C_{ij} = 0$, the covariance matrix for (x_i, x_j) by themselves is diagonal and there exist constants Z_{ij}, α_i and α_j such that the marginal distribution on x_i, x_j has the following expression

$$p^{(i,j)}(x_i, x_j) = \frac{1}{Z_{ij}} e^{-\alpha_i (x_i - m_i)^2 - \alpha_j (x_j - m_j)^2} = p^{(i)}(x_i) \cdot p^{(j)}(x_j).$$

- Fix $i < j$, then $Q_{ij} = 0$ means that “the conditional distribution on (x_i, x_j) fixing the other variables makes them independent”.

For $k \neq i, j$, fix $x_k = a_k$, then since $Q_{ij} = 0$ there exist constants b_0, b_i, b_j, c_i, c_j (depending on the a_k’s) such that the conditional distribution on x_i, x_j is

$$\begin{aligned} p(x_i, x_j | x_k = a_k \text{ for all } k \neq i, j) &= \text{cnst.} p(a_1, .., x_i, .., x_j, .., a_n) \\ &= \frac{1}{Z} e^{-(b_0 + b_i x_i + b_j x_j + c_i {x_i}^2 + c_j {x_j}^2)} \\ &= \frac{1}{Z} e^{-c_i (x_i - {m'}_i)^2 - c_j (x_j - {m'}_j)^2} \\ &= p(x_i | x_k = a_k \text{ for all } k \neq i, j) \cdot p(x_j | x_k = a_k \text{ for all } k \neq i, j). \end{aligned}$$

Differential entropy

The notion of entropy can be extended to continuous probability distributions like the Gaussian ones. In the first chapter, we defined the entropy of a probability distribution p on a finite (or countable) set $\Omega = \{\omega_i\}$ by $H(p) = \sum_i p_i \log_2(1/p_i)$, where $p_i = p(\omega_i)$. We now extend this definition to a probability density $p(x)$ in which case it is called the "differential entropy".

Definition

Let $p(x)$ be a probability density on $\mathbb{R}^n$, i.e. $dP = p(x_1, \cdots, x_n)dx_1 \cdots dx_n$ is a probability measure. The **differential entropy** *of P is defined by*

$$H_d(p) = \int_{\mathbb{R}^n} p(x) \log_2 \frac{1}{p(x)} dx = \mathbb{E}_P\left(\log_2(1/p)\right).$$

Rk : Unlike the case of entropy of a probability distribution on a finite space, the differential entropy can be negative. For example, when $p(x) = U_a(x) = \frac{1}{a}\mathbb{1}_{[0,a]}(x)$, then

$$H_d(p) = H_d(U_a) = \log_2 a,$$

which is < 0 when $a < 1$.

In the case of a finite probability space Ω with N elements, we already saw in the first chapter that, among all probability distributions on Ω, the uniform distribution (for all i, $p_i = 1/N$) is the one with maximal entropy. Now, we will see that in the continuous case, the Gaussian distribution has maximal differential entropy among all distributions with given mean and variance.

Proposition

Let $p(x)$ be a probability density on $\mathbb{R}$ with mean $\overline{x}$ and variance σ^2, and let $g(x)$ be the Gaussian distribution with same mean $\overline{x}$ and same variance σ^2 :

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\overline{x})^2/2\sigma^2},$$

then

$$H_d(g) = \log_2 \sqrt{2\pi e} + \log_2 \sigma \geq H_d(p).$$

In particular, g has maximal differential entropy among all distributions with given mean and variance.

This result can be easily extended to dimension n. Let g be the Gaussian distribution with mean $\overline{x} \in \mathbb{R}^n$ and covariance matrix C. Then the differential entropy of g is

$$H_d(g) = n \log_2 \sqrt{2\pi e} + \frac{1}{2} \log_2(\det C),$$

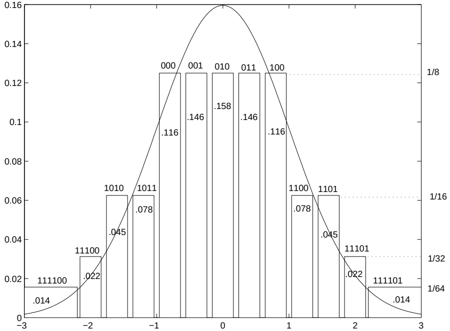
and it is maximal among all distributions with given mean and covariance matrix.

Link between entropy and differential entropy :

Let p be a continuous probability density on $\mathbb{R}$. Divide $\mathbb{R}$ into small bins B_i of size Δx, and denote $P_i = \int_{B_i} p(x)\,dx \simeq p(x_i)\Delta x$, where $x_i \in B_i$. Then

$$H_d(p) \simeq H(P) + \log_2 \Delta x.$$

Example : The discrete coding of a standard normal (which means Gaussian with mean 0 and variance 1) random variable with a prefix code when the range is divided into 11 bins, 9 with size $\Delta x = 0.4$, plus two tails. On the same figure, we plot both the density function $g(x).\Delta x$ and 2^{-l_i}, where l_i is the code length of the i^{th} message. Inside each bar i, we give the probability P_i of that bin, and above the bar, we give the corresponding codeword. In this case, the expected number of bits $\sum_i P_i l_i$ turns out to about 3.42. Compare this to the ideal coding length found with differential entropy, which is $H_d(g) + \log_2(1/\Delta x) \approx 3.37$.



Basics : Fourier Analysis

a) If f is a function in $L^2(\mathbb{R})$ then one goes back and forth between f and its Fourier transform $\widehat{f}$ via :

$$\boxed{\widehat{f}(\xi) = \int_{\mathbb{R}} e^{-2\pi i x\xi} f(x)dx; \qquad f(x) = \int_{\mathbb{R}} e^{2\pi i x\xi}\widehat{f}(\xi)d\xi.}$$

In this definition the variable x might represent a time (e.g. in seconds) and then the variable ξ represents a frequency (e.g. in hertz).

b) If $(f_n) \in l^2$ is a sequence, then the Fourier transform of (f_n) is the 1-periodic function $\widehat{f}$ related to f by :

$$\boxed{\widehat{f}(\xi) = \sum_{n=-\infty}^{+\infty} e^{-2\pi i n\xi} f_n; \qquad f_n = \int_0^1 e^{2\pi i n\xi}\widehat{f}(\xi)d\xi.}$$

c) If f is a periodic function of x with period 1, then the Fourier coefficients $\widehat{f_n}$ of f for $n \in \mathbb{Z}$ and the inversion formula are :

$$\boxed{\widehat{f_n} = \int_0^1 f(x)e^{-2\pi i n x}dx; \qquad f(x) = \sum_{n=-\infty}^{+\infty} \widehat{f_n}e^{2\pi i n x}.}$$

d) The finite Fourier Transform : if $(f_0, \ldots, f_N)$ is a finite sequence of length N then its discrete Fourier Transform is

$$\boxed{\widehat{f_m} = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} e^{-2\pi i \frac{nm}{N}} f_n; \qquad f_m = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} e^{+2\pi i \frac{nm}{N}}\widehat{f_n}.}$$

Some properties :

In the following, we denote $f^-(x) = f(-x)$.

Isometry	$\| f \|_2 = \| \widehat{f} \|_2,$	$< f, g > = < \widehat{f}, \widehat{g} >$
Product/Convolution	$\widehat{fg} = \widehat{f} * \widehat{g},$	$\widehat{f * g} = \widehat{f}.\widehat{g}$
Symmetry	$\widehat{(f^-)} = \overline{\widehat{f}},$	$\widehat{\overline{f}} = \widehat{f}^-$
Translation	$f\widehat{(x-a)} = e^{-2\pi i a\xi}\widehat{f},$	$\widehat{e^{2\pi iax}}f = \widehat{f}(\xi - a)$
Scaling	$\widehat{f(ax)} = \frac{1}{a}\widehat{f}\left(\frac{\xi}{a}\right)$	
Derivatives	$\widehat{f'} = 2\pi i\xi\widehat{f}$	$\widehat{xf(x)} = \frac{i}{2\pi}\widehat{f}'$
Gaussian	$e^{\widehat{-x^2/2\sigma^2}} = \sqrt{2\pi}\sigma e^{-2\pi^2\sigma^2\xi^2}$	
Cauchy	$e^{\widehat{-2\pi\lvert x\rvert}} = \frac{1}{\pi(1+\xi^2)}$	$\widehat{\frac{1}{1+x^2}} = \pi e^{-2\pi\lvert\xi\rvert}$

One of the main signal processing properties of the Fourier Transform is the link between the **autocorrelation** and the **power spectrum**. Recall that the autocorrelation is given by

$$f * f^-(x) = \int f(y)f(y-x)dy$$

and so :

$$\boxed{(\widehat{f * f^-}) = |\widehat{f}|^2.}$$

where $|\widehat{f}|^2$ is the power spectrum.

Windowed Fourier Transform

Often a function has different oscillatory properties in different parts of its domain, and, to describe this, we need to define another variant of the Fourier transform, the **Windowed Fourier Transform**.
We consider a signal f defined for all $x \in \mathbb{R}$. We choose a window function w. Then the windowed Fourier Transform of f around point a and at frequency ξ is defined as

$$\widehat{f_a}(\xi) \quad \text{where} \quad f_a(x) = w(x-a)f(x).$$

Using the property of product/convolution conversion of the Fourier Transform, we get

$$\widehat{f_a} = w\widehat{(x-a)} * \widehat{f}.$$

To work out a simple case, assume that w is a Gaussian function,

$$w(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} = g_\sigma(x).$$

With such a choice for w, the size of the window is of the order of σ. Then, the Fourier Transform of w is

$$w\widehat{(x-a)}(\xi) = e^{-2\pi^2\sigma^2\xi^2}.e^{-2\pi i\xi a}.$$

And so the size of the “support” of $\widehat{w}$ is of the order of $1/\sigma$. So finally $\widehat{f_a} = g_\sigma\widehat{(x-a)} * \widehat{f}$ is a smoothing of $\widehat{f}$ with a kernel of width approximately $1/\sigma$.

It is important to understand the behavior of such a windowed Fourier Transform as σ goes to 0 and to infinity. As $\sigma \to 0$, you get a better resolution in time (the window is small), but you get less resolution in frequency (frequencies are spread out over other frequencies). Conversely, as $\sigma \to +\infty$, you get bad resolution in time, but very good resolution in frequency. Is it possible to have together good resolution in time and in frequency ? The answer to this question is no, and the reason for this is the following theorem (the Uncertainty Principle).

Theorem

Suppose that $f \in L^2$ is a real valued function such that $\int_{-\infty}^{+\infty} f^2(x)dx = 1$, which means therefore that f^2dx is a probability density in x. If $\overline{x} = \int xf^2dx$ is its mean, then the Standard Deviation of f^2dx is defined as usual by :

$$SD(f^2dx) = \sqrt{\int (x-\overline{x})^2 f^2 dx}.$$

Moreover, we have $\int |\widehat{f}(\xi)|^2 d\xi = 1$, which means that $|\widehat{f}(\xi)|^2 d\xi$ is also a probability density, but in the frequency variable ξ. Then

$$SD(f^2dx) \cdot SD(|\widehat{f}(\xi)|^2 d\xi) \geq \frac{1}{2\pi}.$$

The theorem says that $SD(f^2dx)$ and $SD(|\widehat{f}(\xi)|^2 d\xi)$ cannot be both be small which means that you cannot localize simultaneously in time and in frequency.

Aliasing phenomenon

To get a better feeling of how the different Fourier transforms interact, we look at the example where a simple periodic function of a real variable is sampled discretely and where the phenomenon of aliasing occurs.

For some $\omega > 0$, let $f(t) = e^{2i\pi\omega t}$ be a purely periodic signal with frequency ω and period $p = 1/\omega$. Notice that $\widehat{f}(\xi) = \delta_\omega(\xi)$ is the Dirac "function" at ω (there is only one frequency). Let Δt be a time interval and N a large integer so that $N\Delta t \gg p$. For $0 \leq k < N$, let $f_s(k) = f(k\Delta t)$ be discrete samples of f. Using the Discrete Fourier Transform, we get for l an integer :

$$|\widehat{f_s}(l)|^2 = \left| \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} e^{2i\pi\omega k\Delta t} e^{-2i\pi\frac{kl}{N}} \right|^2 .$$

Summing this geometric series, we get the following expression :

$$|\widehat{f_s}(l)|^2 = \frac{C}{\sin^2(\pi(\omega\Delta t - \frac{l}{N}))},$$

where C is a constant independent of l. We now distinguish 2 cases depending on whether the sampling is dense or sparse :

Dense sampling : $\Delta t \ll p$ or $\omega \Delta t \ll 1$. Then if l_0 is the nearest integer to $\omega N \Delta t$, we have $0 \leq l_0 < N$ and l_0 is the frequency of the signal on the discrete Fourier transform scale. The error between the true frequency ω and the frequency estimated on the discrete scale is small :

$$\left| \omega - \frac{l_0}{N\Delta t} \right| \leq \frac{1}{2N\Delta t}.$$

Moreover, the peak of $|\widehat{f_S}(l)|^2$ is found at $l = l_0$.

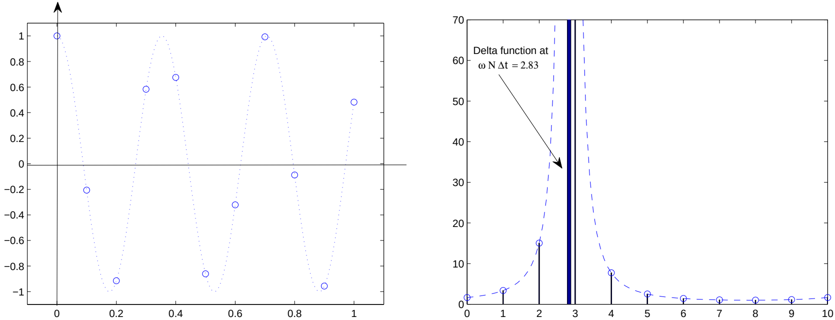


FIGURE: Example of dense sampling. On the left, we plot the real part of a purely periodic signal (dotted line) $t \to \exp(2i\pi\omega t)$ with $\omega = 2.83$, $p \approx .35$. It is sampled at points $k\Delta t$ with $\Delta t = 0.1$, and $0 \leq k < N$ with $N = 10$ in this toy example. On the right, we plot the sampled version of the power spectrum : since the sampling is dense ($\omega\Delta t < 1$), there is no aliasing. The peak is found at $l_0 = 3$ which is the closest integer to $\omega N \Delta t = 2.83$.

Sparse sampling : $\Delta t > p$ or $\omega \Delta t > 1$. Then instead of finding a peak of the discrete power near ω, you get the existence of 2 integers l_0 and n_0 such that $0 \le l_0 < N$ and $\omega \Delta t \approx n_0 + \frac{l_0}{N}$. And so the peak of $\widehat{f_s}$ is at $l = l_0$, which means that the peak of $\widehat{f}$ has been shifted far away from the true frequency ω. Notice that in this case the signal of frequency ω has the same samples as the signal of frequency $\omega - \frac{n_0}{\Delta t}$. That is what is called aliasing.

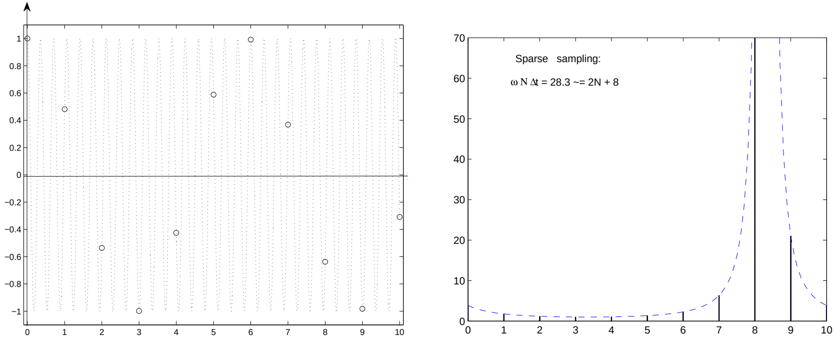


FIGURE: Example of sparse sampling. On the left, everything is the same as in the previous figure except that $\Delta t = 1$. Since the sampling is now sparse, we see aliasing. The peak of the power is found at $l_0 = 8$ which is the closest integer to $\omega N \Delta t = 28.3$ modulo N.

Gaussian models for stationary finite cyclic signals

In this part, we will combine the ideas of n-dimensional Gaussian distributions and of Discrete Fourier Transforms.

Let $\vec{s} = (s_1, \ldots, s_N)$ be a periodic signal ($s_{N+1} = s_1$). We will first take the Fourier Transform $\widehat{s}$ of $\vec{s}$. Usually we think of $\widehat{s}$ as a new vector in $\mathbb{C}^N$. But, instead we can regard $\widehat{s}$ simply as the coefficients of $\vec{s}$ when it is expanded in a new orthonormal basis, a rotated version of the standard unit vectors.

The usual canonical basis of $\mathbb{C}^N$ is $\vec{e}^{(1)}$, ..., $\vec{e}^{(N)}$ where $\vec{e}^{(k)} = (0, \ldots, 1, .., 0)$ (the 1 is at the k^{th} place). This basis is orthonormal. But instead we can choose another orthonormal basis : $\vec{f}^{(0)}$, ..., $\vec{f}^{(N-1)}$ where $\vec{f}^{(k)}$ is defined for $0 \leq k \leq N-1$ as

$$\vec{f}^{(k)} = \frac{1}{\sqrt{N}}(1, e^{2i\pi\frac{k}{N}}, \ldots, e^{2i\pi\frac{k(N-1)}{N}})$$

This basis is the Fourier basis. If $\vec{s}$ is the signal, in the canonical basis we have $\vec{s} = \sum_{k=1}^{N} s_k \vec{e}^{(k)}$ and in the Fourier basis (using the inverse Fourier Transform) we get

$$\vec{s} = \sum_{l=0}^{N-1} \widehat{s}_l \vec{f}^{(l)}.$$

Notice that if the signal $\vec{s}$ is real, then it has the property that $\widehat{s_{N-l}} = \overline{\widehat{s}_l}$. (This is analogous to the usual equivalence for the real Fourier transform : f is real iff $\widehat{f}$ satisfies $\widehat{f}(-\xi) = \overline{\widehat{f}(\xi)}$.)

Now, let us assume that $\vec{s}$ follows a Gaussian distribution with density :

$$p_Q(\vec{s}) = \frac{1}{Z} e^{-(\vec{s}-\vec{m})^t Q \overline{(\vec{s}-\vec{m})}/2}.$$

Here $\vec{s}$ may be a vector of $\mathbb{C}^n$, and in this case, we have to assume that the matrix Q (which may have complex entries) is a Hermitian positive definite matrix.

Definition

We say that the Gaussian distribution p_Q is **stationary** *if it satisfies for all integer l :*

$$p_Q(T_l \vec{s}) = p_Q(\vec{s}),$$

where $(T_l \vec{s})_k = s_{k-l}$ and $k-l$ means $(k-l) \mod N$.

Using the change to the Fourier basis, we get $\widehat{Q}$ and $\widehat{m}$ such that

$$p_Q(\vec{s}) = \frac{1}{Z} e^{-(\widehat{s}-\widehat{m})^t \widehat{Q} \overline{(\widehat{s}-\widehat{m})}/2}.$$

Theorem

p_Q is stationary iff $\vec{m}$ is a constant signal and Q is a Hermitian banded matrix (meaning $Q_{i,i+j}$ depends only on $j \mod N$, and if it is denoted by a_j then $a_j = \overline{a_{N-j}}$). Equivalently, in the Fourier basis, $\widehat{m} = (\widehat{m_0}, 0, \dots, 0)$ and $\widehat{Q}$ is a real positive diagonal matrix.

White noise and colored noise

Such a stationary distribution can be written :

$$p(\vec{s}) = \frac{1}{Z} e^{-\sum_l |\widehat{s}_l|^2 / 2\sigma_l{}^2}.$$

Then the real and imaginary parts $\Re \widehat{s}_l$ and $\Im \widehat{s}_l$ are independent Gaussian random variables with mean 0 and standard deviation σ_l. Such a distribution is called **colored noise**.

The particular case $Q = I_N$, $\widehat{Q} = I_N$, and

$$p(\vec{s}) = \frac{1}{Z} e^{-\sum_l |\widehat{s}_l|^2 / 2},$$

is called **white noise**.

Notice that the differential entropy of colored noise is

$$\frac{N}{2} \log_2(2\pi e) + \sum_{l=0}^{l=N-1} \log_2(\sigma_l).$$

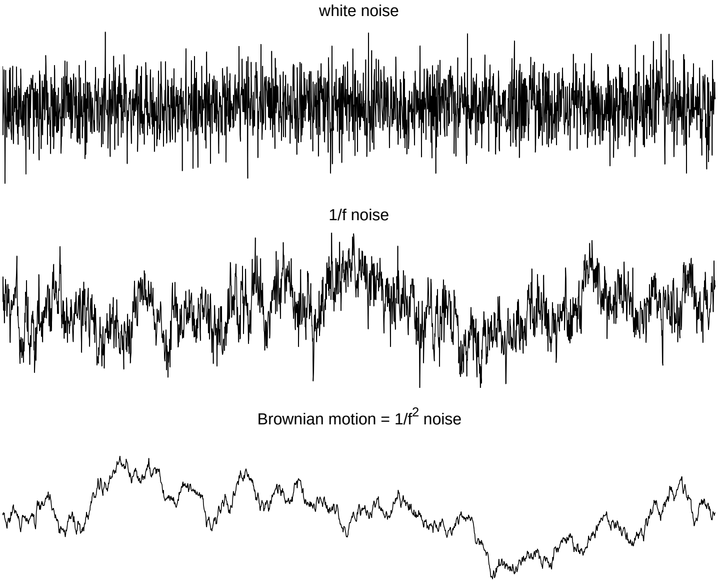


FIGURE: Three simulations of colored noise with the variance of the power at frequency f falling off as like $1/f^\alpha$. Such noises are usually called $1/f^\alpha$ noises.

The case of a musical note

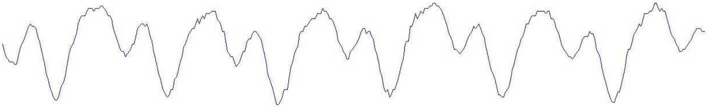
We return to the problem of finding a stochastic model for music. We first construct a Gaussian model of a single note. Let ω be the fundamental frequency of the note being played and $p = 1/\omega$ be its period. If the signal is $s(t)$ then

$$s(t+p) \simeq s(t),$$

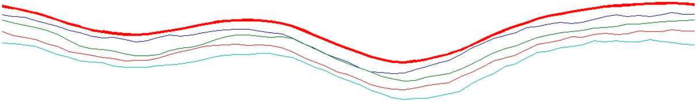
which means that the signal is close to be periodic, although in real music, there are always small residual variations.

See next figure for an example of such a signal taken from live music. Some deviations from perfect periodicity are shown in the second graph. With some averaging, we can make the signal periodic and then expand it in a Fourier series with frequencies n/p. Its n^{th} component is known as the n^{th} harmonic. In the Figure, all but three terms in the Fourier series are quite small.

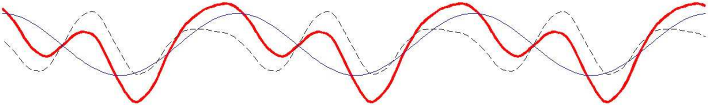
Six periods of a female voice singing the note sol



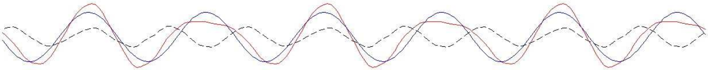
One period of the averaged detrended signal, compared to 4 samples



Three periods of a) the average signal (in red), b) its first harmonic (in blue) and c) the residual (dashed in black)



Three periods of a) the signal minus first harmonic (in red), b) its second harmonic (in blue) and c) the remaining residual (dashed in black)



How do we make a Gaussian model for this signal ?

We formalize the property $s(t+p) \simeq s(t)$ by assuming that the expected value of $\int (s(t+p)-s(t))^2 dt$ is quite small. We then constrain the expected total power of the signal by bounding $\int s(t)^2 dt$.

Take a discrete sample of the signal s and, for simplicity, we assume that s ‘wraps around’ at some large integer N, i.e. $s_{N+k} = s_k$, and that p is an integer dividing N. Let $q = N/p$, the number of cycles present in the whole sample. We’ll analyze the simplest possible Gaussian model for s which gives samples which are periodic plus some small residual noise. Its density is :

$$p_{a,b}(s) = \frac{1}{Z} e^{-a\sum_{k=0}^{N-1}(s(k)-s(k+p))^2/2 - b\sum_{k=0}^{N-1} s(k)^2/2} = \frac{1}{Z} e^{-\vec{s}^t Q \vec{s}/2}$$

where $a \gg b > 0$, $Q_{i,i} = b + 2a, Q_{i,i+p} = -a$, for $0 \leq i \leq N-1$ and otherwise 0.

Notice that Q is a positive definite quadratic form (if there is no term $b\sum_{k=0}^{N-1} s(k)^2/2$ then the quadratic form is only semi-definite). Then $p_{a,b}(s)$ is a stationary probability distribution, and so we can diagonalize the quadratic form in the Fourier basis.

On the one hand, we have

$$\sum_k (s(k) - s(k+p))^2 = \parallel s - T_{-p}(s) \parallel^2 = \parallel \widehat{s} - \widehat{T_{-p}(s)} \parallel^2 .$$

Using the fact that $\widehat{s}(l) - \widehat{T_{-p}(s)}(l) = \widehat{s}(l)(1 - e^{2i\pi \frac{pl}{N}})$, we get

$$\sum_k (s(k) - s(k+p))^2 = \sum_l |\widehat{s}(l)|^2 |1 - e^{2i\pi \frac{pl}{N}}|^2 = 4 \sum_l |\widehat{s}(l)|^2 \sin^2(\frac{\pi pl}{N}).$$

On the other hand, we have

$$\sum_k s(k)^2 = \sum_l |\widehat{s}(l)|^2.$$

So

$$p_{a,b}(s) = \frac{1}{Z} e^{-\sum_l (b + 4a\sin^2(\frac{\pi pl}{N}))|\widehat{s}(l)|^2/2} = \frac{1}{Z} \Pi_l e^{-(b + 4a\sin^2(\frac{\pi pl}{N}))|\widehat{s}(l)|^2/2}. \tag{1}$$

Then the expected power at frequency l is the mean of $|\widehat{s}(l)|^2$, which works out to be :

$$\mathbb{E}(|\widehat{s}(l)|^2) = \frac{1}{b + 4a\sin^2(\frac{\pi pl}{N})}.$$

Note that this has maxima $1/b$ if l is a multiple of N/p, that is all frequencies which repeat in each cycle ; and that all other powers are much smaller (because $a \gg b$).

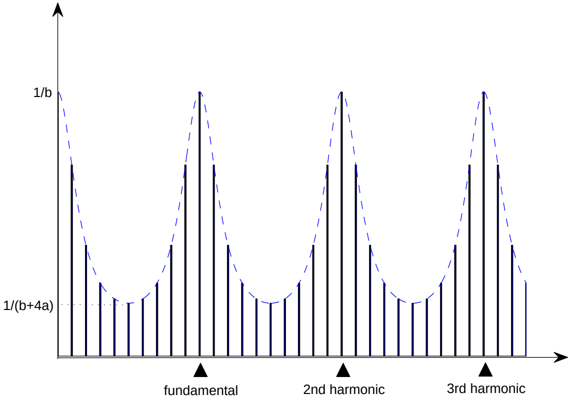


FIGURE: Expected power spectrum : $\mathbb{E}(|\widehat{s}(l)|^2) = 1/(b + 4a\sin^2(\frac{\pi pl}{N}))$.

Rk : This is not, however, an accurate model of real musical notes because the power in all harmonics (integer multiples of the fundamental frequency) is equally large. It is easy to change this and include extra parameters for the expected power of the various harmonics using the second expression in Equation (1).

Basics :Poisson processes

The simplest model we can choose for the set of discontinuities is a **Poisson process**. These processes are the precise mathematical description of what it means to throw down random points with a certain density.

We denote $\mathbb{D}$ the set of all countable discrete subsets of $\mathbb{R}$. We want to define a probability measure on $\mathbb{D}$ whose samples will be discrete sets of random points on $\mathbb{R}$ with two properties :
- First, they have a given density λ meaning that the expected number of points in every interval $[a, b]$ will be $\lambda(b - a)$.
- Second, for any two disjoint intervals I, J, the set of points in I and the set of points in J are to be independent of each other.

The key ingredient for the construction is knowing what to choose for the probability distribution on the random number $\mathcal{D}_{a,b} = |\mathcal{S} \cap [a, b]|$ of points in the interval $[a, b]$ from a random sample $\mathcal{S}$ from $\mathbb{D}$. To make things work, this must be chosen to be the Poisson distribution with mean $\lambda(b - a)$:

$$\mathbb{P}(\mathcal{D}_{a,b} = d) = e^{-\lambda(b-a)} \frac{(\lambda(b-a))^d}{d!}.$$

We then construct a random $\mathcal{S} \in \mathbb{D}$ in 3 steps :

1. First choose for every k in $\mathbb{Z}$, an integer d_k following the Poisson distribution with mean λ.
2. Then choose $x_1^{(k)}, ..., x_{d_k}^{(k)}$ independent random real numbers in $[k, k+1]$.
3. $\mathcal{S}$ is the union of all these sets : $\mathcal{S} = \{x_l^{(k)}\}$.

Rk1 : This construction uses a particular decomposition of the real line into intervals, namely $[k, k+1]$, but it does not affect the constructed random set $\mathcal{S}$.

Rk2 : Poisson processes can be generalized to other spaces, and other measures.

Can we be give a simple expression for the density of a Poisson process ?
This is easiest to do for a Poisson process on a bounded interval $[0, B]$ in $\mathbb{R}$ with density λ. Then a sample is determined by its cardinality m and the unordered set of points $\{x_1, \cdots, x_m\}$. With probability 1, the points are distinct and then there is a unique order for which $0 < x_1 < \cdots < x_m < B$. Thus the underlying space for the Poisson probability measure is the disjoint union over m, with $0 \leq m < \infty$, of the polyhedral subsets $\mathbb{D}_m$ of $\mathbb{R}^m$ defined by $0 < x_1 < \cdots < x_m < B$. Note that the cube in $\mathbb{R}^m$ given by $0 \leq x_i \leq B$ for all i is the disjoint union of the $m!$ copies of $\mathcal{S}_m$ gotten by permuting the coordinates. Thus the m-dimensional volume of $\mathbb{D}_m$ is $B^m/m!$. The probability density defining the Poisson measure is just

$$dP(m, x_1, ..., x_m) = e^{-\lambda B} \lambda^m dx_1 ... dx_m.$$

In fact, conditional on m, this is the uniform distribution on the x_i ; and, integrating out the x_i, we get $P(m) = (e^{-\lambda B} \lambda^m) \cdot (B^m/m!)$, the Poisson distribution on m with mean λB. In particular if we denote $S = \{x_1, ..., x_m\}$, still ordered, then the density of the probability has the simple form

$$dP(S) = Ae^{-\alpha |S|} \Pi dx_i,$$

where A and α are constants (depending on λ).

The model for music

We construct the model for music in two stages. We recall that for this model, we need : the sampled sound signal $\vec{s}$, and hidden random variables : the number of notes m, the times $\vec{t}$ where new notes begin, and the periods $\vec{p}$ of the notes.

The probability distribution $p(\vec{s}, m, \vec{t}, \vec{p})$ can be decomposed in the following way :

$$p(\vec{s}, m, \vec{t}, \vec{p}) = \prod_{l=1}^{m} p\left((\vec{s}|_{I_l}) \,\middle|\, p_l, t_l, t_{l+1} \right) \cdot p(\vec{p}, \vec{t}, m), \quad I_l = \{t \mid t_l \leq t < t_{l+1}\}.$$

We have already constructed a Gaussian model for $p(s|_{[t_l,\ldots,t_{l+1}-1]} \mid p_l, t_l, t_{l+1})$:

$$p(s|_{[t_l,\ldots,t_{l+1}-1]} \mid p_l, t_l, t_{l+1}) = \frac{1}{Z} \exp\left(-a \sum_{n=t_l}^{t_{l+1}-p_l-1} (s_{n+p_l} - s_n)^2/2 - b \sum_{n=t_l}^{t_{l+1}-1} {s_t}^2/2 \right),$$

where $a \gg b$. Notice that this model is a value model because it constrains value patterns, and that $p(\vec{p}, \vec{t}, m)$ is a geometric model expressing how the domain $\{1, \cdots, B\}$ of the sampled sound is structured by the music. The simplest geometric model is gotten by taking the random variable $\vec{t}$ to be Poisson and each p_l to be independent of the other periods and uniformly sampled from the set of periods of all the notes the musical instrument is capable of producing (something like 'atonal' music). If *per* represents this set of periods, then this gives :

$$p(\vec{p}, \vec{t}, m) = \frac{1}{Z} e^{-Cm} \mathbb{1}_{\{\vec{p} \in per^m\}}, \text{ where } C = \log(|per|) + \log((1-\lambda)/\lambda),\ Z = (1-\lambda)^B.$$

The dogma of Pattern Theory is that we should sample from this model. It is not at all hard to create samples using Matlab. The results are not very convincing – they give a totally atonal, un-rhythmic 'music', but we can certainly construct various models for music of increasing sophistication which sound better. Possible elaborations and constraints include :

1. Tempered scale, which means that we impose $p_l \approx (\text{sampling rate})/447 \cdot 2^{-f_k/12}$ where $f_k \in \mathbb{Z}$,
2. Tempo : $t_{l+1} - t_l \approx a_l T_0$ where $a_l \in \mathbb{Z}$.
3. Get a better model for harmonics : e.g. an expected power spectrum like in Figure below. But for an instrument, we may also have a resonant frequency enhancing all nearby harmonics of the note, shown in the same figure.

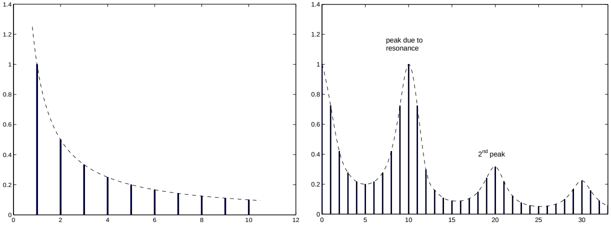


FIGURE: Better model for the power of harmonics : on the left, a simple decay of higher frequencies; on the right, resonances of the instrument enhance harmonics near specific frequencies.

Rk : Some really sophisticated models have been studied. See, for instance, the paper of Z. Ghahramani and M. Jordan *Factorial Hidden Markov Models* using Bach Chorales for some examples.

Finding the Best Possible Score via Dynamic Programming

Since music is a 1D signal, we can compute by **dynamic programming** the best possible score, i.e. the mode of the posterior probability distribution in the hidden variables $m, \vec{p}, \vec{t}$. One might make guesses about the note boundaries and periods based on local evidence, but this is often misleading and if you look at the past and the future you get less ambiguity.

The probability model for music is

$$p(\vec{s}, m, \vec{t}, \vec{p}) = Ae^{-Cm} \prod_{k=1}^{m} \frac{1}{Z_k} e^{-a(\sum_{t=t_k}^{t_{k+1}-p_k-1} (s(t+p_k)-s(t))^2/2) - b(\sum_{t=t_k}^{t_{k+1}-1} s(t)^2/2)}.$$

Then if we fix $\vec{s} = \vec{s}_o$ and define $E(m, \vec{t}, \vec{p}) = -\log p(\vec{s}_o, m, \vec{t}, \vec{p})$, we see that it is of the form $\sum_k f(t_k, t_{k+1}, p_k)$. We consider all possible scores on $[1, t]$ including a last note which ends at time t. The last note has a time of beginning $t' + 1 < t$ and a period p. Then for such scores we have :

$$E = E_1(\vec{s}_o|_{[0,t']}, \text{ notes up to } t') + E_2(\vec{s}_o|_{[t'+1,t]}, p) + E_3(\vec{s}_o \text{ from } t+1 \text{ on}).$$

Here E_1 assumes the last note ends at t', E_2 assumes there is one note extending from $t'+1$ to t (so it has no other Poisson variables in it) and E_3 assumes a note begins at $t+1$.

Using the algorithm of dynamic programming, we compute by induction on t the "best score" for the time interval $[0, t]$ *assuming* a note ends at t. Let $e(t') = \min E_1(\vec{s}_o|_{[0,t')}, \text{ notes up to } t')$ and assume by induction that we know $e(t')$ for all $t' < t$. We then find

$$e(t) = \min_{t'<t, p} [e(t') + E_2(\vec{s}_o|_{[t',t)}, p)],$$

and that continues the induction. Only at the end, however, do we go back and decide where the note boundaries are.

Other piecewise Gaussian Models

Many other types of 1D signals can be fit well by piecewise Gaussian signals. Another source of nice examples are 1D slices of 2D images of the world.

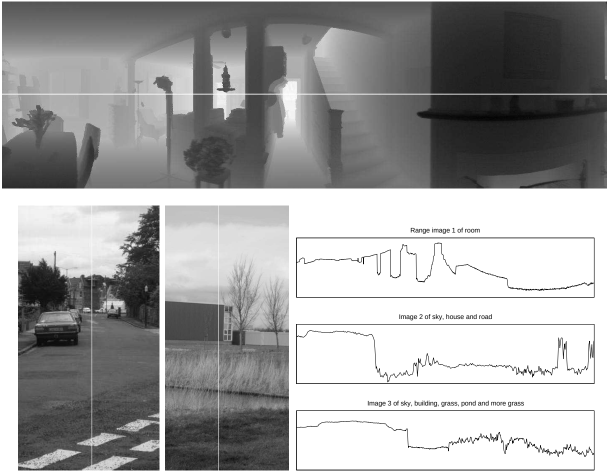


FIGURE: A range image of the interior of a house on top ; two usual images of the world on the bottom left ; the 1D slices shown in white in the images are plotted on the bottom right.

Example 1 : white noise with varying means, fixed variance. The random variables are $\vec{s}$, the signal ; $m, \vec{t}$, the number and values of the points of discontinuity ; and the means $\vec{\mu}$ in each interval. We need to put some distribution on the means : this is known as a *hyperprior*, a prior probability distribution on the parameters of the model. We simply assume the means are independent and uniformly distributed in some interval M. Then the full model is given by :

$$p(\vec{s}, m, \vec{t}, \vec{\mu}) = \frac{(1-\lambda)^{N-1}}{(\sqrt{2\pi}\sigma)^N} \cdot \prod_{l=0}^{l=m} e^{-\sum_{n=t_l}^{t_{l+1}-1}(s_n-\mu_l)^2/2\sigma^2)} \cdot \frac{1}{|M|^{m+1}} \left(\frac{\lambda}{1-\lambda}\right)^m ,$$

where $1 = t_0 < t_1 < \cdots < t_m < t_{m+1} = N + 1$.

Minimizing this, for fixed $\vec{s} = \vec{s}_o$, we find that μ_l is just the mean $\overline{s}_l$ of $\vec{s}_o \Big|_{[t_l, t_{l+1})}$ and this gives the problem of minimizing with respect to the $\vec{t}$'s :

$$E(\vec{t}) = \sum_{l=0}^{l=m} \sum_{n=t_l}^{n=t_{l+1}-1} (\vec{s}_o(n) - \overline{s}_l)^2 + am,$$

where $a = 2\sigma^2 \left(\log(|M|) + \log((1-\lambda)/\lambda)\right)$. Note that the inner sums are just the variances of the data on the intervals between the $\vec{t}$'s, i.e. if $\widehat{\sigma_l^2}$ is the variance of $\vec{s}_o \Big|_{[t_l, t_{l+1})}$, then :

$$E(\vec{t}) = \sum_{l=0}^{l=m} (t_{l+1} - t_l)\widehat{\sigma_l^2} + am.$$

Minimizing this for different a's gives the optimal ways of segmenting any signal into a piecewise constant function formed from the μ's plus white noise of various variances. As a varies, we get segmentations with larger and smaller numbers of segments.

Example 2 : white noise with varying means and variance. The random variables are $\vec{s}$, the signal ; $m, \vec{t}$, the number and values of the points of discontinuity ; and $\vec{\mu}, \vec{\sigma}$ the means and standard deviations in each interval. Now we need a hyperprior for the variance too : we will assume here that the variances are independent and uniformly distributed in the log domain, i.e. with respect to the measure $d\sigma/\sigma$. If S denotes the domain on which $\log \sigma$ is uniformly distributed, then we get the model :

$$p(\vec{s}, m, \vec{t}, \vec{\mu}, \vec{\sigma}) = \frac{(1-\lambda)^{N-1}}{(2\pi)^{N/2}} \cdot \prod_{l=0}^{l=m} \frac{1}{\sigma_l^{t_{l+1}-t_l+1}} e^{-\sum_{n=t_l}^{t_{l+1}-1} (s_n-\mu_l)^2/2\sigma_l^2} \cdot \frac{1}{(|S||M|)^{m+1}} \left(\frac{\lambda}{1-\lambda}\right)^m .$$

Note that we get one factor σ_l in the denominator for each s_n in the l^{th} interval *and* one more for the density $d\sigma/\sigma$ in the prior on σ_l itself.

Minimizing this for fixed $\vec{s} = \vec{s}_o$, a surprising thing occurs. The means and variances both become the maximum likelihood choices for the data on the l^{th} interval, namely the mean and variance[1] of the data itself there. So the sum in the exponent reduces to just $(t_{l+1} - t_l)/2$ and, taking out constants, the minimization reduces to that of :

$$E(\vec{t}) = \sum_{l=0}^{l=m} (t_{l+1} - t_l + 1) \log(\widehat{\sigma_l}) + am.$$

Here we must assume that $t_{l+1} - t_l \geq 2$ and that the data has some generic noise in it or we have trouble with 0 variances.

[1] Variance in the sense of $\sum_{i=1}^{i=N} (x_i - \bar{x})^2/N$, not in the sense of the unbiased estimator of the variance $\sum_{i=1}^{i=N} (x_i - \bar{x})^2/(N-1)$.

Example 3 : weak string model. Here we assume that the signal in each interval between discontinuities is a random walk, i.e. we assume $s_n - s_{n-1}$ are *iid* Gaussian when $n \neq t_l$ for any l. This wouldn't, however, make a proper probability distribution on the signal unless we also put some weak constraint on its actual values. As with music, we can do this by multiplying the probability density by $e^{-\epsilon\|\vec{s}\|^2}$ (giving us a modified random walk with restoring force). The model is :

$$p(\vec{s}, m, \vec{t}) = \frac{(1-\lambda)^{N-1}}{(2\pi)^{N/2}} \cdot \prod_{l=0}^{l=m} \sqrt{\det(Q_l)} e^{-a\sum_{n=t_l+1}^{t_{l+1}-1}(s_n - s_{n-1})^2/2 - \epsilon\sum_{n=t_l}^{t_{l+1}-1} s_n^2/2} \cdot \left(\frac{\lambda}{1-\lambda}\right)^m .$$

Here Q_l is the matrix of the quadratic form in the Gaussian, i.e.

$$Q_l = \begin{pmatrix} \epsilon + a & -a & 0 & \cdots & 0 & 0 \\ -a & \epsilon + 2a & -a & \cdots & 0 & 0 \\ 0 & -a & \epsilon + 2a & & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \epsilon + 2a & -a \\ 0 & 0 & 0 & \cdots & -a & \epsilon + a \end{pmatrix}$$

Minimizing the negative log of this probability gives us an optimal fit of the data to a set of Brownian walk segments separated by jumps.

Example 4 : weak string/noisy observation or $(u+v)$-model. A modification of the last model is the simplest example of a popular approach to signal enhancement via denoising. One assumes that the observations $\vec{s}$ are the sum of additive Gaussian white noise $\vec{u}$ and a true signal $\vec{v}$, which itself is piecewise continuous with jumps. In this approach, the hidden variables are *both* the jumps $\vec{t}$ and the smooth signal $\vec{v}$ between the jumps. One can model $\vec{v}$ using the previous example. If the model is used for deducing $\vec{t}$ and $\vec{v}$ with a fixed given $\vec{s} = \vec{s}_o$, then one can allow $\epsilon = 0$ and still get a proper posterior probability distribution on $\vec{v}$. The model is then :

$$p(\vec{s}_o, \vec{v}, m, \vec{t}) \propto e^{-b\sum_{n=1}^{n=N}(\vec{s}_o(n)-v_n)^2/2} \cdot \prod_{l=0}^{l=m} e^{-a\sum_{n=t_l+1}^{t_{l+1}-1}(v_n-v_{n-1})^2/2} \cdot \left(\frac{\lambda}{1-\lambda}\right)^m .$$

This is the Mumford-Shah model or the full weak string model of Blake and Zisserman in the 1D framework. If $S = \{t_1, \cdots, t_m\}$, then the energy $E = -2\log(p)$ of this model is just :

$$E(\vec{v}, S) = b\sum_{n=1}^{n=N}(\vec{s}_o(n) - v_n)^2 + a\sum_{n=2, n\notin S}^{n=N}(v_n - v_{n-1})^2 + c|S|.$$