ESTIMATING THE ENTROPY OF A SIGNAL WITH APPLICATIONS

J.-F. Bercher

Équipe Communications Numériques, ESIEE
93 162 Noisy-le-Grand, FRANCE,
and Laboratoire Systèmes de Communications, UMLV
bercherj@esiee.fr

C. Vignat

Laboratoire Systèmes de Communications
Université de Marne la Vallée
93 166 Noisy-le-Grand, FRANCE
vignat@univ-mlv.fr

ABSTRACT

We present an estimator of the entropy of a signal. The basic idea is to adopt a model of the probability law, in the form of an AR spectrum. Then, the law parameters can be estimated from the data. We examine the statistical behavior of our estimates of laws and entropy. Finally, we give several examples of applications: an adaptive version of our entropy estimator is applied to detection of law changes, blind deconvolution and sources separation.

1. INTRODUCTION

Entropy is a major tool in information theory. However, this tool is not directly used in the context of signal processing, but in theoretical frameworks, because it is difficult to compute or even to estimate from a set of data. A few interesting attempts of direct use of entropy for signal processing applications can be found in [1, 2].

Let us recall that the entropy H_X of a random variable $X(\omega)$ with continuous probability law $f_X(x)$ is defined as:

$$H_X = -\int_{-\infty}^{+\infty} f_X(x) \log f_X(x)\, dx. \qquad (1)$$

This formula raises the two main difficulties that characterize the computation of the entropy:

- the first difficulty is due to the fact that the probability law $f_X(x)$ is generally unknown. However, there exist numerous methods for estimating probability laws, such as kernel methods. But these methods share the major drawback of slow convergence rate (typically $T^{-4/5}$).
- the second difficulty comes from the fact that, even if $f_X(x)$ is known, formula (1) requires numerical integration.

These two remarks illustrate the necessity for finding fast converging and accurate estimates of entropy. The underlying idea for the new estimate of entropy we propose here states as follows: the probability law is seeked in a set of functions that are characterized by two fundamental features:
(i) this set is characterized by a finite set of parameters (parametric model)
(ii) the functions are chosen in such a manner that computation of the entropy as defined by (1) reduces to a problem of AR process identification. Morever, this method inherits regularization techniques allowing integration of a priori informations about the law (such as its smoothness).

The first property induces obviously a small bias over the estimate entropy. The second property makes of this estimate a new and potentially powerful tool in signal processing.

This paper is organized as follows: in the first part, we present theoretical backgrounds of the estimation procedure. Then we present the statistical behaviour of the estimates. Finally, we exhibit sample applications of this method, such as detection of law changes, blind deconvolution and sources separation.

2. ESTIMATING ENTROPY

2.1. Construction of the estimate

The main features we wish for our estimate are:
- a fixed number of parameters, as opposed to nonparametric approaches,
- an easy method to select the best fitting parameters from the sole observation of samples of the random process $X(n,\omega)$,
- the capability of iteratively updating these parameters.

2.2. The AR identification problem

Given observations $w(n)$ of a process $W(n,\omega)$ and a fixed integer p, a set of parameters $\{a_i\}_{1\leq i\leq p}$ is seeked that best fits the following model :

$$w(n) = \sum_{i=1}^{p} a_i w(n-i) + \epsilon(n)$$

where $\epsilon(n)$ is a white noise with power σ_ϵ^2.

The exact solution of this problem is well known and requires the knowledge of the correlation function $R_w(k)$ of the process $W(n,\omega)$. This solution is given by

$$a = R_w^{-1} r_w,$$

with matrix $(R_w)_{i,j} = R_w(i-j)$ and vector $(r_w)_i = R_w(i)$. The spectrum of this AR process writes

$$S_w(f) = \frac{\sigma_\epsilon^2}{\left|1 - \sum_{k=1}^{p} a_k e^{-j2\pi kf}\right|^2}. \qquad (2)$$

2.3. Application to the estimation of entropy

2.3.1. The approach

Application of the AR identification tool to the entropy estimation problem is straightforward if we look for an estimate $\hat{f}_X(x)$ of the

probability law $f_X(x)$, for $x \in [-0.5, 0.5]$, under the form

$$\hat{f}_X(x) = \frac{\sigma_\epsilon^2}{\left|1 - \sum_{k=1}^{p} a_k e^{-j2\pi kx}\right|^2} \quad (3)$$

$$= S_w(x) \quad (4)$$

that is the restriction over interval $[-0.5, +0.5]$, of the power spectral density $S_w(x)$ of an AR process denoted as $W(n,\omega)$.

2.3.2. An underlying process

Given law (3), it is possible to exhibit a random process $W(n,\omega)$ whose spectrum is precisely $S_w(f) = \hat{f}_X(f)$. If we define:

$$W(n,\omega) = e^{j(nX+\phi(\omega))},$$

where X is any sample of process $X(n,\omega)$ (for example $X = X(1,\omega)$) and $\phi(\omega)$ is a uniformly distributed phase, then $W(n,\omega)$ is a centered process whose correlation function $R_w(k)$ is:

$$R_w(k) = E[W^*(n)W(n+k)] = E\left[e^{jkX}\right] = FT^{-1}\left[\hat{f}_X(x)\right]$$

so that obviously $S_w(x) = \hat{f}_X(x)$. Note that this process is not ergodic.

2.3.3. Estimation of entropy

Although an analytical expression of entropy can be derived in terms of AR parameters [4], the particular choice expressed by (3) also leads to an easy and natural procedure for the estimation of entropy.

Denote by $\hat{\phi}_X(k)$ and $\hat{\psi}_X(k)$ the first and second characteristic functions of $X(\omega,n)$, defined respectively by $\hat{\phi}_X(k) = FT^{-1}\left\{\hat{f}_X(x)\right\}$ and $\hat{\psi}_X(k) = FT^{-1}\left\{\log \hat{f}_X(x)\right\}$, then

$$\begin{cases} \hat{\phi}_X(k) = R_w(k) & \forall k \\ \hat{\psi}_X(k) = C_w(k) & \forall k \end{cases} \quad (5)$$

where $C_w(k)$ denotes the cepstrum function of $W(n,\omega)$.

Application of the Parseval Plancherel identity to the definition (1) of the entropy writes, with the help of relations (5), as follows:

$$\hat{H}_X = -\sum_{k=-\infty}^{+\infty} \hat{\phi}_X(k)\hat{\psi}_X(k) = -\sum_{k=-\infty}^{+\infty} R_w(k)C_w(k). \quad (6)$$

The entropy H_X of $X(n,\omega)$ can thus be computed from both correlation and cepstrum functions of $W(n,\omega)$.

At this step, we can take advantage of the AR structure, exploiting the fact that correlation and cepstrum functions of an AR process verify

$$R_w(k) = \sum_{i=1}^{p} a_i R_w(k-i) + \sigma_\epsilon^2 \delta(k) \quad (7)$$

$$C_w(k) = \begin{cases} \rho_w(k) - \sum_{i=k+1}^{0} \left(\frac{i}{k}\right) C_w(i)\rho_w(k-i) & \text{if } k < 0 \\ 2\log R_w(0) & \text{if } k = 0 \\ \rho_w(k) - \sum_{i=1}^{k-1} \left(\frac{i}{k}\right) C_w(i)\rho_w(k-i) & \text{if } k > 0 \end{cases} \quad (8)$$

where function $\rho_w(k)$ is the normalized correlation (or correlation coefficient) of process $W(n,\omega)$, defined as $\rho_w(k) = \frac{R_w(k)}{R_w(0)}$.

The trick here lies in the fact that the first and second characteristic functions involved in formula (6), being identified with the correlation and cepstrum function of the AR process $W(n,\omega)$, can be computed using relations $(7,8)$. Thus the estimate $\hat{H}_X$ of the entropy as defined by (6) can be computed without any numerical integration, and from the sole observation of the data.

2.3.4. Implementation of the method

The proposed method thus consists in the three following steps:
first step: compute a raw estimate of the law of X, (equivalently of $S_w(x)$), involving samples $\{x_i\}_{1\le i\le n+1}$ as, for instance:

$$\hat{f}_X^{(n+1)}(x) = \frac{1}{n+1}\sum_{i=1}^{n+1} \delta(x - x_i) * h(x),$$

where function h is any kernel function. However, as we wish an iterative estimate, it is useful to notice that

$$\hat{f}_X^{(n+1)}(x) = \frac{n}{n+1}\hat{f}_X^{(n)}(x) + \frac{1}{n+1}h(x - x_{n+1})$$

so that by inverse Fourier transform:

$$R_w^{(n+1)}(k) = \frac{n}{n+1}R_w^{(n)}(k) + \frac{1}{n+1}H(k)e^{j2\pi k x_{n+1}} \quad (9)$$

where $H(k) = \mathrm{FT}^{-1}(h(x))$.
second step: the set of estimated correlations $R_w^{(n+1)}(k)_{0\le k\le p}$ allows to compute the set of parameters $a_i^{(n+1)}{}_{1\le i\le p}$ and thus both series $\hat{\psi}_X^{(n+1)}(k)$ and $\hat{\phi}_X^{(n+1)}(k)$, using (7,8).
third step: application of formula (6) gives the estimated entropy of process $X(n,\omega)$ computed with the $(n+1)$ first samples.

2.4. Estimation of the AR parameters of the law

Accurate estimates of the law parameters $\{a_k\}$ have now to be derived from a finite number of samples $\{x_i\}_{1\le i\le N}$. We present three methods for solving this problem in our special context.

2.4.1. Normal equations

The AR parameters satisfy the well known normal, or Yule-Walker equations

$$R_w \begin{bmatrix} 1 \\ -a \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \end{bmatrix}$$

where R is the $(p+1)$ correlation matrix. The correlation coefficients are estimated using (9). However, some free parameters remain to be chosen : (i) the form and length of the kernel h, (ii) the order p of the AR model. The choice of these parameters should result, as usual in estimation problems, of a trade-off between bias and variance.

2.4.2. Long AR models and regularization

Accurate modelization of non-AR processes via AR techniques requires the use of long AR models. The counterpart of adopting a high number of coefficients is a loss in the stability of the estimate. The exploitation of regularization techniques enables to use long

AR models, and thus to model 'non-AR' spectra, without sacrifying stability.

The idea is to use a long AR model with the addition of some prior knowledge about the 'smoothness' of the spectrum. In [3], Kitagawa and Gersch defined the PSD k^{th} smoothness by

$$D_k = \int_0^1 \left| \frac{\partial^k A(f)}{\partial f^k} \right|^2 df \propto a^{\text{t}} \Delta_k a,$$

where Δ_k is the diagonal matrix with elements $[\Delta_k]_{ii} = i^{2k}$.

The corresponding regularized least squares estimate is

$$\hat{a} = (\hat{R}_w + \lambda \Delta_k)^{-1} \hat{r}_w. \tag{10}$$

The hyperparameter λ balances the fidelity to the data and the smoothness prior. In [3], a bayesian interpretation of this regularized least-squares is derived, that also leads to a selection rule for the hyperparameter λ, as the minimizer of the following marginal likelihood:

$$L(\lambda) = \log(\det(\hat{R}_w + \lambda \Delta_k)) - p \log(\lambda) - N \log(\sigma_\epsilon^2), \tag{11}$$

where σ_ϵ^2 is chosen such that the AR probability distribution is properly normalized.

2.4.3. Maximum a Posteriori estimation of the parameters

Another possible approach is to derive the parameters from the probabilistic model of the available data. Since the model of the probability density is given by (3), the probability law of a sample of data of length N is the product law (assuming, without any extra information, independence of the data). Prior knowledge regarding the smoothness of the law can be introduced in the form of the Kitagawa-Gersch gaussian prior for a, $f_A(a) \propto e^{-\lambda a^t \Delta_k a}$. This leads to the posterior distribution

$$f_{A|X}(a|x) \propto \prod_{i=1}^{N} \frac{\sigma_\epsilon^2}{|1 - \sum_{k=1}^{p} a_k e^{-j2\pi k x_i}|^2} e^{-\lambda a^t \Delta_k a}$$

The MAP estimate of the AR parameters can now be computed as the minimizer of

$$J(a, \lambda) = \sum_{i=1}^{N} \log \left| 1 - \sum_{k=1}^{p} a_k e^{-j2\pi k x_i} \right|^2 - N \log \sigma_\epsilon^2 + \lambda a^t \Delta_k a.$$

2.5. First simulation results

In this section, we give simulation results regarding the estimation of probability laws using our AR modelization approach. We also examine the accuracy of entropy estimates for several distributions. Experiments were performed on sequences of 500 samples[1]. The laws were modelized with an order $p = 50$. We tested the long AR method presented in § 2.4.2. We examined the cases of a uniform law $U_{\left[-\frac{1}{20}, \frac{1}{20}\right]}$ and a gaussian law $N(0, 0.01)$. The optimal regularization hyperparameter λ was selected using the minimization of a likelihood as in (11).

Typical results are given in Figs1 (a) and (b)

[1] Since the law is modelized as the restriction of the spectrum on interval $\left[-\frac{1}{2}, +\frac{1}{2}\right]$, the data had to be rescaled on this interval. This does not restrict our approach because the entropy of the rescaled variable differs from the original entropy only by a known additive term.

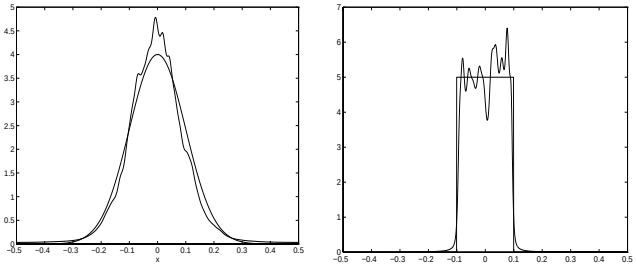

Figure 1: The AR estimates and theoretical gaussian (a) and uniform (b) laws.

We also tested the MAP approach as presented in § 2.4.3. From the practical point of view, the MAP estimation approach usually gives more accurate but slowly converging estimates, contrary to the long AR approach. As far as the accuracy of the estimate of the entropy is concerned, we evaluated the entropy of Gaussian and uniform data using a Monte Carlo type simulation over 100 trials. The results are given in the following table, where M_h is the mean of the estimates and σ_H their standard deviation.

	theoretical	M_h	σ_H
Gaussian $\sigma^2 = \frac{1}{8}$	1.09	0.9544	0.0485
Uniform $\left[-\frac{1}{6}, \frac{1}{6}\right]$	1.58	1.4786	0.0159

These results exhibit the good statistical behaviour of our estimate, that is a low bias and a small variance.

3. SAMPLE APPLICATIONS

3.1. Detecting law changes

3.1.1. An adaptive estimate

As our approach involves iterative evaluations of the empirical law and the corresponding correlation sequence, it is straightforward to derive an adaptive version of the entropy estimation. It suffices to introduce a forgetting factor μ in the updating formula (9) of the correlation sequence. Then, the AR parameters and entropy are evaluated for each new sample. The regularized least squares solution (10) can be computed recursively $a^{(n+1)} = a^{(n)} + \alpha \left(\left(\hat{R}^{(n)} + \lambda \Delta_k \right) a^{(n)} - \hat{r}^{(n)} \right)$.

An interesting application of this adaptive estimate consists in detecting law changes in signals. As an illustration, we consider a signal $x(n)$ that consists in 400 samples generated according to a mixture of two Gaussian distributions, followed by 400 uniformly distributed samples. First law has entropy $H_1 = -1.8$ whereas the second has entropy $H_2 = -1.18$.

Figure 2 below shows $x(n)$: it is difficult by a simple inspection to detect that there is a law change. Figure 3 shows the adaptive estimate of the negentropy on this test signal, using a forgetting factor $\mu = 0.98$.

The following points are of importance: (i) the law change appears clearly, (ii) the rupture time is properly revealed, (iii) the entropy is estimated with high accuracy and (iv) our adaptive estimate has a high tracking capability due to its fast convergence (typically in less than 100 samples).

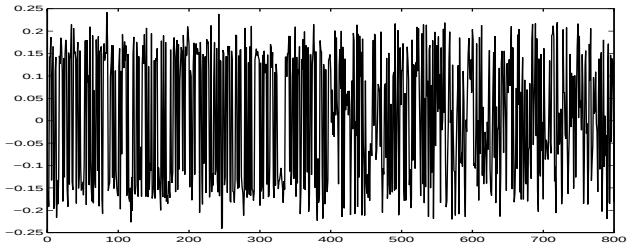

Figure 2: x(n) with abrupt law change at sample 400.

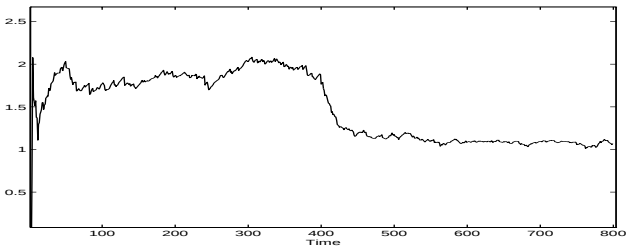

Figure 3: Adaptive estimate of the entropy.

3.2. Blind deconvolution of AR systems

The difficult problem of blind deconvolution lies in recovering the inputs and parameters of a filter from the sole observation of its output. The concept of entropy brings an interesting answer to this problem, relying on the following proposition:

Proposition 1 *Let $Y(n,\omega)$ be the output of a filter $G(f)$ normalized such that $\int_{-\infty}^{\infty} |G(f)|^2 df = 1$, whose input is a non-gaussian i.i.d. sequence $X(n,\omega)$. Then $H_Y > H_X$.*

The intuitive reason behind this result is that $f_Y(y)$ is closer to a gaussian distribution than $f_X(x)$, the gaussian distribution having the maximum entropy in the set of distributions of given variance, see [5]. The deconvolution procedure then simply consists in adjusting the filter such that the reconstructed input has minimum (estimated) entropy. If θ are the filter parameters, this writes

$$\theta_{opt} = \arg\min_{\theta} \hat{H}_X \quad \text{submitted to } X(f) = Y(f)/G_\theta(f)$$

Simulations were performed in the case of non-minimum phase AR systems. They show that the AR parameters can be identified with an outstanding accuracy, and the input can be perfectly reconstructed, even if the AR order is overestimated. These simulations were performed in the case of uniform and binary inputs, with 500 samples of data.

3.3. Source separation

In the context of source separation, N signals $s(n) = [s_1(n), ..., s_N(n)]$ are mixed by an unknown $N \times N$ matrix A to provide observed signals $x(n) = [x_1(n), ..., x_N(n)]$. The task is, from the sole observation of signals $x(n)$, to recover the sources assuming only their independence. This goal is reached by designing a matrix B such that the reconstructed signal $\hat{s}(n) = Bx(n)$ has independent components. The information theoretic measure of independence is the mutual information, that is the Kullback-Leibler divergence between $p_{S_1,...,S_N}(s_1, ..., s_N)$ and $\Pi_{1\le i\le N}\, p_{S_i}(s_i)$. In the source separation context, this reduces to minimizing the following cost function

$$C(B) = -\log|\det B| + \sum_{i=1}^{N} H(\hat{s}_i) \tag{12}$$

In classical methods, as no estimate of the entropy is available, B is chosen as the solution of :

$$E[\hat{s}_i \psi_j(\hat{s}_j)] = 0 \quad (i \neq j) \tag{13}$$

that expresses the stationarity condition of $C(B)$. Function $\psi_j(s_j)$ is the so-called score function, the log derivative of the density $p_{S_j}(s_j)$.

Using our AR parametrization, we can

(i) either estimate the cost function $C(B)$ and minimize it using any standard optimization procedure

(ii) or estimate the solution of (13) using an analytical expression (in terms of the AR parameters) of the score functions $\psi_i(s_i)$.

We performed simulations using the first approach, using 500 samples of data in the case of the mixture of $N = 2$ sources. The following table presents, for several distributions of the sources, the resulting matrix $M = AB$ that should be the identity matrix, up to a scaling factor and a permutation.

source 1	source 2	M
$U_{[-0.5,0.5]}$	$\frac{1}{2}\{N(0.3,0.2)$ $N(-0.3,0.2)\}$	$\begin{bmatrix} 1 & 0.0395 \\ 0.0011 & 1 \end{bmatrix}$
$U_{[-0.5,0.5]}$	$U_{[-0.5,0.5]}$	$\begin{bmatrix} 1 & 0.02 \\ 0.0163 & 1 \end{bmatrix}$
$binary$ $\left[-\frac{1}{2},\frac{1}{2}\right]$	$N(0,1)$	$\begin{bmatrix} 1 & 0.0354 \\ 0.02 & 1 \end{bmatrix}$
$U_{[-0.5,0.5]}$	$N(0,1)$	$\begin{bmatrix} 1 & 0.23 \\ -0.0491 & 1 \end{bmatrix}$

4. REFERENCES

[1] P. Viola, N. N. Schraudolph, T. J. Sejnowski, "Empirical Entropy Manipulation for Real-World Problems", *Advances in Neural Information Processing Systems 8*, MIT Press, 1996.

[2] D. T. Pham, "Blind Separation of Instantaneous Mixture of Sources via an Independent Component Analysis", *IEEE trans. on Signal Processing*, vol. 44, no 11, pp. 2768-2779, nov. 1996.

[3] G.Kitagawa and W.Gersh, "A smoothness priors long AR model method for spectral estimation", *IEEE trans. Automatic Control*, no. 1, pp. 57–65, 1985.

[4] J.-F. Bercher and C. Vignat, "Entropy Estimation with Applications to Signal Processing Problems", *LSC internal report no. 98-057*.

[5] D. Donoho, "On minimum entropy deconvolution", *Applied Time Series Analysis II*, pp. 565-609, Academic Press, 1981.