Estimating the Entropy of a Signal with Applications

Jean-François Bercher and Christophe Vignat

***Abstract*—In this paper, we present a new estimator of the entropy of continuous signals. We model the unknown probability density of data in the form of an AR spectrum density and use regularized long-AR models to identify the AR parameters. We then derive both an analytical expression and a practical procedure for estimating the entropy from sample data. We indicate how to incorporate recursive and adaptive features in the procedure. We evaluate and compare the new estimator with other estimators based on histograms, kernel density models, and order statistics. Finally, we give several examples of applications. An adaptive version of our entropy estimator is applied to detection of law changes, blind deconvolution, and source separation.**

***Index Terms*—AR processes, entropy estimation, parametric methods, regularization, spectrum analysis.**

I. Introduction

SINCE Shannon's work [1], entropy is used as a major tool in information theory. However, this tool is rarely used in signal processing, except in theoretical frameworks, because it appears difficult to compute or estimate the entropy from a set of real data. Interesting approaches involving direct use of entropy for signal processing applications can be found in [2]–[4]. In many applications, a measure of complexity of underlying probability density functions, or a measure of dependence between components or signals, allows the design of an optimal processing scheme, possibly in nonstationary contexts. Examples of such situations are plentiful:

- source separation;
- blind deconvolution;
- source coding;
- image alignment;
- detection of abrupt changes;

and so on. Thus, entropy-based approaches might be useful for such problems.

The entropy $H(X)$ of a random variable X with continuous probability density function (PDF) $p_X(x)$ is defined as

$$H(X) = -E_X[\log_2 p_X] = -\int_{-\infty}^{+\infty} p_X(x)\log_2 p_X(x)\,dx. \tag{1}$$

In the discrete case, where X takes values x_i with probabilities p_i, $H(X) = -\sum_i p_i \log_2 p_i$. Basic estimates can be built, using any raw estimates of the p_i in the preceding formula. More sophisticated entropy estimates, based on coding theorems that are specific to the discrete case, can be found in [5] and [6]. In signal processing, data usually have continuous PDF because of contamination by continuous noise. In this continuous case, two main approaches exist. First, the PDF can be approximated by an element of a parameterized set, whose entropy is known in term of the parameters [7]. Second, entropy estimators are based on a prior estimation of underlying PDF's (or cumulative distribution functions) using methods such as histograms [4], [8], order statistics [9], [10] (see [11] for a comparative study), or kernel methods [2], [3], [8].

In this paper, we derive and apply to signal processing problems a new estimator of entropy for all continuous PDF's with bounded support. This estimator can be implemented in recursive schemes and has tracking capabilities in nonstationary contexts. Furthermore, our approach provides a convenient estimation procedure of PDF's.

The main contributions of this work are

i) a new presentation and improvements of the approach of [12]–[14] for PDF estimation;
ii) derivation of an analytical close-form formula for the estimation of entropy;
iii) presentation of a practical procedure for entropy estimation, including recursive and adaptive features;
iv) evaluation and comparisons with other methods;
v) examples of application of the entropy estimator to signal processing problems.

This paper is organized as follows. In Section II, we discuss the relevance of AR modeling of PDF's and introduce regularized long-AR models. In Section III, we give the theoretical expression of the entropy associated with AR-PDF's and a practical procedure for estimating the entropy. In Section IV, we give examples of PDF's estimation. Then, we analyze and compare the behavior of the new estimator with other methods. Finally, in Section V, we give some applications of this estimate to signal processing problems, namely, detection of PDF changes, blind equalization, and source separation.

II. AR Modeling of PDF's

A. Introduction

Our approach consists of estimating the unknown PDF $p_X(x)$ as the power density spectrum $S_W(x)$ of some unit variance AR process $W(\omega, n)$. Applications of spectral estimation methods to PDF estimation were first introduced in the context of nonlinear signal processing in [12], [15]. AR-PDF estimation was also discussed in [13] and [14]. We discuss here the relevance of this model for PDF estimation, recall the link between spectral matching and linear prediction, and then propose to use a

Manuscript received May 17, 1999; revised December 10, 1999. The associate editor coordinating the review of this paper and approving it for publication was Prof. Jian Li.

J.-F. Bercher is with the Laboratoire Signaux et Télécoms, Groupe ESIEE, Noisy-le-Grand, France (e-mail: bercherj@esiee.fr).

C. Vignat is with the Laboratoire Systèmes de Communications, Université de Marne la Vallée, Noisy-le-Grand, France (e-mail: vignat@univ-mlv.fr).

Publisher Item Identifier S 1053-587X(00)04069-1.

long-AR regularized approach in order to obtain stable and accurate estimates. In the sequel, we will show that this model leads to an easy procedure for computing the associated entropy and that recursivity and adaptivity can be introduced in the procedure.

We suppose that the observation consists in samples of a process $X(\omega,n)$, identically distributed according to a continuous PDF $p_X(x)$ with bounded support, say, $[-(1/2),+(1/2)]$. This hypothesis, although restrictive, is usual in the context of PDF estimation.

B. PDF Estimation Using AR Modeling

We look for an estimate $\hat{p}_X(x)$ of the true (unknown) PDF $p_X(x)$ parameterized by a set of coefficients $\{a_k\}_{1\leq k\leq p}$ in the form of a power spectrum density $S_W(x)$

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

where σ_ϵ^2 is chosen such that $\int_{-(1/2)}^{+(1/2)} S_W(f)\,df=1$.

The relevance of this parameterization lies in the fact that any continuous spectrum density can be approximated, in the $\|\cdot\|_\infty$ sense, by an AR spectrum density. More precisely, if $S_Z(x)$ is a symmetric continuous spectral density on $[-(1/2),(1/2)]$, and $\delta>0$, then there exists an integer p and a real-valued causal AR(p) process $W(\omega,n)$ with innovation variance σ_ϵ^2 such that $\|S_W(x)-S_Z(x)\|_\infty<\delta$ (see [16, corol. 4.4.2. p. 132]). This result extends easily to the case of nonsymmetric, possibly one-sided, $S_Z(x)$; in this case, $W(\omega,n)$ is a complex-valued AR process.

Once the analogy between PDF's and power spectrum densities is stated, a natural question arises: Can we find a process $Z(\omega,n)$ whose spectrum is precisely the PDF of the random variable $X(\omega)$? It is easy to check that $Z(\omega,n)=e^{j(nX+\phi(\omega))}$ has this property if X is any sample of process $X(\omega,n)$, and $\phi(\omega)$ is uniformly distributed over $[0,2\pi]$ and independent of X. Indeed, its correlation function $R_Z(k)$ is nothing but the first characteristic function of X.

However, this "underlying process" $Z(\omega,n)$ of $X(\omega,n)$ is very likely not an AR process. Hence, to identify the parameters $\{a_k\}$ associated with PDF $\hat{p}_X$ in (2), we need to match a given spectrum $S_Z(x)=p_X(x)$ with an AR spectrum $S_W(x)=\hat{p}_X(x)$. A classical result about spectral matching [17] states that the best AR(p) model spectrum minimizing the integrated ratio of the two spectra $I(Z,W)=\int_{-(1/2)}^{+(1/2)} S_Z(x)/S_W(x)\,dx$ is nothing but the AR solution of the linear prediction problem whose parameters $\mathbf{a}=[a_1,\ldots,a_p]^T$ are such that $\mathbf{R}_Z\mathbf{a}=\mathbf{r}_Z$.[1] Matrix $(\mathbf{R}_Z)_{1\leq i,j\leq p}=R_Z(i-j)$ and correlation vector $(\mathbf{r}_Z)_{1\leq i\leq p}=R_Z(i)$ are built using correlation function $R_Z(k)$.

Thus, modelization of PDF p_X as an AR spectrum follows the two following steps: 1) estimation of the correlation sequence $R_Z(k)$, i.e., of the characteristic function of X, using the available data $\{x(n)\}_{1\leq n\leq N}$ from $X(\omega,n)$, as the statistical average correlation estimate

$$\hat{R}_Z(k)=\frac{1}{N}\sum_{n=1}^{N}e^{jkx(n)}$$

and 2) estimation of the coefficients $\{a_k\}$ of the AR process $W(\omega,n)$ by solving $\hat{\mathbf{R}}_Z\mathbf{a}=\hat{\mathbf{r}}_Z$.

C. Long AR Models and Regularization

As mentioned above, real PDF's are very likely not in the form of an AR spectrum. Hence, an accurate modelization of PDF's via AR techniques may require the use of long AR models. However, the counterpart of adopting a high number of coefficients is a loss in the stability of the estimate (e.g., spurious peaks). The exploitation of regularization techniques enables the use of long AR models, and thus, modeling of "non-AR" spectra, while preserving stability.

The idea is to use a long AR model with the addition of some prior knowledge about the "smoothness" of the spectrum. In [20], Kitagawa and Gersch defined the kth smoothness by

$$D_k=\int_0^1\left|\frac{\partial^k A(f)}{\partial f^k}\right|^2 df$$

with $A(f)=\sum_{k=1}^{p}a_k e^{j2\pi kf}$, and showed that $D_k\propto\mathbf{a}^t\boldsymbol{\Delta}_k\mathbf{a}$, where $\boldsymbol{\Delta}_k$ is the diagonal matrix with elements $[\boldsymbol{\Delta}_k]_{ii}=i^{2k}$.

The AR parameters are obtained as a regularized least-squares solution

$$\hat{\mathbf{a}}=(\hat{\mathbf{R}}_Z+\lambda\boldsymbol{\Delta}_k)^{-1}\hat{\mathbf{r}}_Z \tag{3}$$

where hyperparameter λ balances a fidelity to the data and a smoothness prior.

In [20] and [21], a Bayesian interpretation of this regularized least-squares is derived, which also leads to a selection rule for the hyperparameter λ, as the minimizer of the following marginal likelihood:

$$L(\lambda)=\log(\det(\hat{\mathbf{R}}_Z+\lambda\boldsymbol{\Delta}_k))-p\log(\lambda)-N\log(\sigma_Z^2) \tag{4}$$

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

Let us now turn to the problem of computing an estimate $\hat{H}(X)$ of the entropy $H(X)$ associated with $p_X(x)$. A natural approach at this step is to build the entropy estimate $\hat{H}(X)$ of the unknown PDF $p_X(x)$ as the entropy of the estimate PDF $\hat{p}_X(x)$.

III. Estimate of Entropy

In this section, we exhibit the analytical expression of the entropy associated with $\hat{p}_X(x)$. Then, we give an alternate and easier procedure for estimating the entropy. Finally, we show how to introduce recursivity and adaptivity in the procedure.

A. Theoretical Expression

The exact expression of entropy $\hat{H}(X)$, using $\hat{p}_X(x)$ defined as in (2) can be derived. Let us denote by $\{z_k\}$ the set of p supposedly simple poles of $\hat{p}_X(x)$ and by $\{\mu_k\}$ the set of associated

[1]Note that minimizing $\log I(Z,W)$ is equivalent to the minimization of $I(Z,W)$ and, in the case of a "good matching" [17], to the maximization of the Burg entropy. Hence, the general AR spectral matching method coincides with the "maximum entropy spectral estimation method," which was derived in the case of gaussian signals [18], [19].

residues. With $A(z) = \sum_{k=1}^{p} a_k z^{-k}$, a straightforward but tedious calculus (omitted here) yields (5), shown at the bottom of the page. Direct use of this analytical expression of entropy is obviously difficult since it requires computation of all poles of the AR model, together with their respective residues, which can be highly time consuming, particularly in the case of long AR models. Although we have here an explicit formula (5), it is desirable to find an equivalent formula for $\hat{H}(X)$ that provides an easier estimation procedure.

B. Easier Estimation Procedure

The entropy associated with $\hat{p}_X(x)$ as defined by (2) is $\hat{H}(X) = -\int_{-(1/2)}^{+(1/2)} S_W(x) \log_2 S_W(x)\, dx$. Hence, applying the Plancherel–Parseval formula to the right-hand side of the above relation yields

$$\begin{aligned}\hat{H} &= -\sum_{k=-\infty}^{+\infty} R_W(k) C_W^*(k) \\ &= -2\,\mathrm{Re}\left\{\sum_{k=0}^{+\infty} R_W(k) C_W^*(k)\right\} \end{aligned} \tag{6}$$

where $R_W(k)$ denotes the kth correlation coefficient of $W(\omega, n)$, and $C_W(k) = FT^{-1}[\log_2 S_W(x)]$ denotes the kth component of its cepstrum. Note that both $R_W(k)$ and $C_W(k)$ have Hermitian symmetry since $S_W(x)$ is real, which provides the right-hand side of (6).

At this step, we take advantage of the AR structure of process $W(\omega, n)$ since, for that particular type of process, both correlation and cepstrum functions obey recursive relations [22][2]:

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

$$C_W(k) = \begin{cases} \log \sigma_\epsilon^2, & \text{if } k = 0 \\ h(k) - \displaystyle\sum_{i=1}^{k-1} \left(\frac{i}{k}\right) C_W(i) h(k-i), & \text{if } k > 0 \end{cases} \tag{8}$$

with $h(k)$ the impulse response of the AR system, which is also computed recursively according to

$$h(k) = -\sum_{i=1}^{p} a_i h(k-i) + \delta(k). \tag{9}$$

The estimated entropy can thus be computed using (6)–(9), avoiding any numerical integration. Obviously, however, in practice, the infinite sum in (6) should be truncated, which leads to some truncation error. It is also important to note that (6) does not require the explicit estimation of the PDF but only of the first $(p+1)$ coefficients of the correlation sequence $R_W(k)$, which in turn enables the computation of the AR parameters and the cepstrum involved in (6).

As will be described below, as part of a practical method for implementing the method expressed by (6)–(9), it is possible to estimate recursively the correlation sequence.

C. Implementation in a Recursive Scheme

The first step consists of estimating the characteristic function R_W from the observation data $\{x_i\}_{1 \le i \le n+1}$. The statistical average correlation sequence can be estimated recursively using

$$\begin{aligned} R_W^{(n+1)}(k) &= \frac{1}{n+1} \sum_{i=1}^{n+1} e^{j2\pi k x_i} \\ &= \frac{n}{n+1} R_W^{(n)}(k) + \frac{1}{n+1} e^{j2\pi k x_{n+1}}. \end{aligned} \tag{10}$$

This empirical characteristic function is the inverse Fourier transform of the empirical distribution $\hat{p}_X^{(n+1)}(x) = 1/(n+1)\sum_{i=1}^{n+1} \delta(x - x_i)$. In kernel methods for density estimation, the empirical distribution is smoothed using a kernel $\phi(x)$. It is also possible to compute such an estimate recursively, as in (10); see [23] and references therein. For the characteristic function, with $\Phi(k) = \mathrm{FT}^{-1}(\phi(x))$, this leads to

$$R_W^{(n+1)}(k) = \frac{n}{n+1} R_W^{(n)}(k) + \frac{1}{n+1} \Phi(k) e^{j2\pi k x_{n+1}}. \tag{11}$$

The proposed method thus consists of the three following steps:

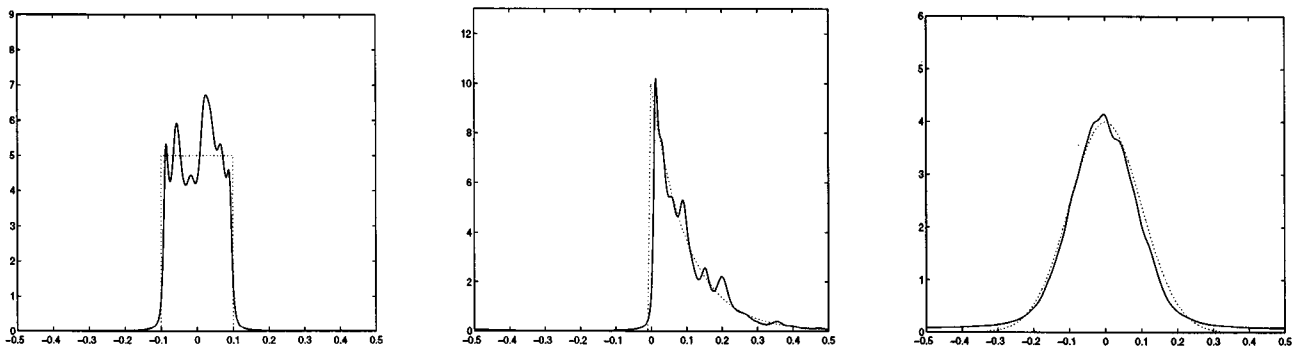
- *First step* — Estimate the $(p+1)$ correlation coefficients $R_W^{(n+1)}(k)_{0 \le k \le p}$ using the $(n+1)$ available samples using (10) or (11).
- *Second step* — The set of estimated correlations $R_W^{(n+1)}(k)_{0 \le k \le p}$ allows computation of parameters $a_{i\ 1 \le i \le p}^{(n+1)}$ and, thus, time series $R_W^{(n+1)}(k)$ and $C_W^{(n+1)}(k)$, using relations (7)–(9).
- *Third step* — Finally, application of (6) gives the estimated entropy of process $X(\omega, n)$ based on its $(n+1)$ first samples.

Furthermore, it is also straightforward to derive an adaptive version of this entropy estimation scheme. It suffices to introduce a forgetting factor μ in the updating formula (10) of the correlation sequence. For the correlation matrix, this gives

$$\hat{\mathbf{R}}_W^{(n)} = \frac{1}{n}\left[(n-1)\mu \hat{\mathbf{R}}_W^{(n-1)} + \mathbf{e}(n)\mathbf{e}(n)^+\right]$$

$$\hat{H}(X) = \frac{I_2}{I_1} + \log I_1, \qquad \text{with} \begin{cases} I_1 = \sigma_\epsilon^{-2} = \displaystyle\sum_{k=1}^{p} \mu_k^* \left(1 - A\left(\frac{1}{z_k^*}\right)\right)^{-1} \\ I_2 = 2\,\mathrm{Re}\left\{\displaystyle\sum_{k=1}^{p} \mu_k^* \left(1 - A\left(\frac{1}{z_k^*}\right)\right)^{-1} \log\left\{1 - A\left(\frac{1}{z_k^*}\right)\right\}\right\}. \end{cases} \tag{5}$$

[2]Relation (8) is derived using [22] and the fact that $R_W(k) = \sigma_\epsilon^2 h(k) * h(-k)^*$, where $h(k)$ is minimum phase with $h(0) = 1$.



(a) (b) (c)

Fig. 1. AR estimates and (a) theoretical uniform (b) exponential, and (c) Gaussian laws.

with $\mathbf{e}(n)^+ = [e^{-j(x_n+\phi)} \ \cdots \ e^{-j(px_n+\phi)}]$. Then, the AR parameters and entropy can be evaluated at each new sample using (3) and (6)–(9).

It is also possible to compute recursively the AR parameters, thus avoiding the matrix inversion required in (3). Indeed, the regularized least squares solution can be computed recursively, using a gradient approach [24]

$$\mathbf{a}^{(n+1)} = \mathbf{a}^{(n)} + \alpha \left[\left(\hat{\mathbf{R}}_W^{(n)} + \lambda \boldsymbol{\Delta}_k \right) \mathbf{a}^{(n)} - \hat{\mathbf{r}}^{(n)} \right]. \quad (12)$$

Finally, it is also possible to adopt a more simple "LMS-like" approach, such as

$$\mathbf{a}^{(n+1)} = \mathbf{a}^{(n)} + \alpha \left[(\mathbf{e}(n)\mathbf{e}(n)^+ + \lambda \boldsymbol{\Delta}_k)\mathbf{a}^{(n)} - e^{-j\phi}\mathbf{e}(n) \right]. \quad (13)$$

IV. Simulation Results and Comparisons

A. AR-PDF Estimation

In order to illustrate the versatility of the long AR approach for PDF estimation, experiments were performed on sequences of 250 samples distributed according to

a) a uniform PDF $U_{[-(1/10),(1/10)]}$;
b) an exponential PDF with parameter 0.1;
c) a Gaussian PDF $\mathcal{N}(0, 0.01)$.[3]

For these three PDF's, the parameters p and λ are, respectively,

a) $p = 32$, $\lambda = 5\ 10^{-5}$;
b) $p = 32$, $\lambda = 3\ 10^{-5}$;
c) $p = 20$, $\lambda = 8\ 10^{-4}$.

Results given in Fig. 1(a)–(c) show the relevance of this approach, which is able to approximate with accuracy different shapes of PDF's.

B. Entropy Estimation

In order to analyze the behavior of the AR entropy estimator, we performed a Monte Carlo study in the case of a uniform and a Gaussian PDF. We evaluated the mean and standard deviation σ_H of the AR entropy estimate over 50 realizations as a function of the length of available data. We compared these results with those obtained in the cases of histogram and kernel PDF approximation as well as the modified Vasicek's estimator recommended in [11].

- In the case of histograms, the entropy is estimated as $-\sum_{i=1}^{K} N_i/K \log_2(N_i/K)$, where N_i is the number of values in the ith bin, and K is the number of bins.
- In the kernel approach, the available samples $\{x_1, \ldots x_N\}$ are directly used for modeling the density as $\hat{p}_X(x) = (1/N)\sum_{i=1}^{N} \phi(x - x_i)$, where $\phi(x)$ is a smoothing kernel, which is usually chosen as a Gaussian kernel. In our experiments, we evaluated the density on a grid of $L = 1000$ points with a kernel width chosen to provide the best results. Finally, the entropy was evaluated as $-\sum_{i=1}^{L} \hat{p}_i \log_2 \hat{p}_i$.
- The Vasicek's estimator [9] relies on the remark that $H(X) = \int_0^1 \log_2(dF^{-1}(u)/du)\, du$. Then, the estimator is obtained by approaching the cumulative distribution function F with order statistics. The modified Vasicek's estimator [11] has the following form:

$$V_{m,n}(X) = \frac{1}{n}\sum_{i=1}^{n} \log_2 \left(\frac{n}{2m}\left(x_{(i+m)} - x_{(i-m)}\right) \right) + f(m,n)$$

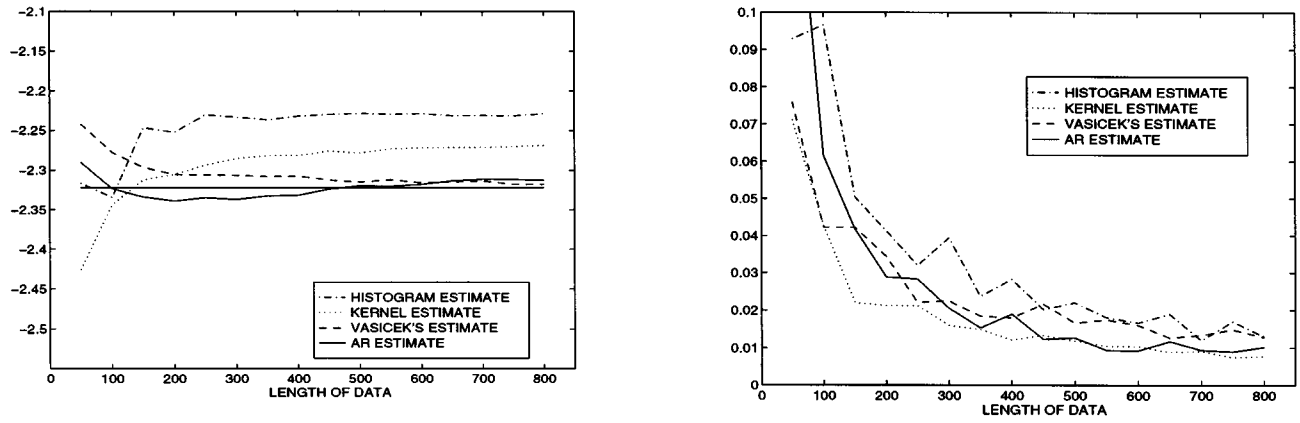
where

$\{x_{(i)}\}$ ordered set of samples $x(i)$ (with $x_{(i)} = x_{(1)}$ for $i < 1$ and $x_{(i)} = x_{(n)}$ for $i > n$);
m positive integer;
$f(m,n)$ function that accounts for a bias correction; see [11].

Figs. 2 and 3 give the results [mean (a) and standard deviation (b)] for a uniform density $U_{[-(1/10),(1/10)]}$ and for a normal density $\mathcal{N}(0, 0.01)$, with respectively theoretical entropy $H(X) = -2.3219$ bits and $H(X) = -1.2748$ bits. For the uniform density, we have chosen an AR order $p = 32$, 64 bins for histogram estimates, $\sigma^2 = 5$ (on the grid of 1000 points) for kernel estimates, and $m = 3$ for the modified Vasicek's estimator. Results for the Gaussian density were obtained with $p = 20$, 20 bins for histogram estimates, $\sigma^2 = 10$ for kernel estimates, and $m = 3$ for the modified Vasicek's estimator.

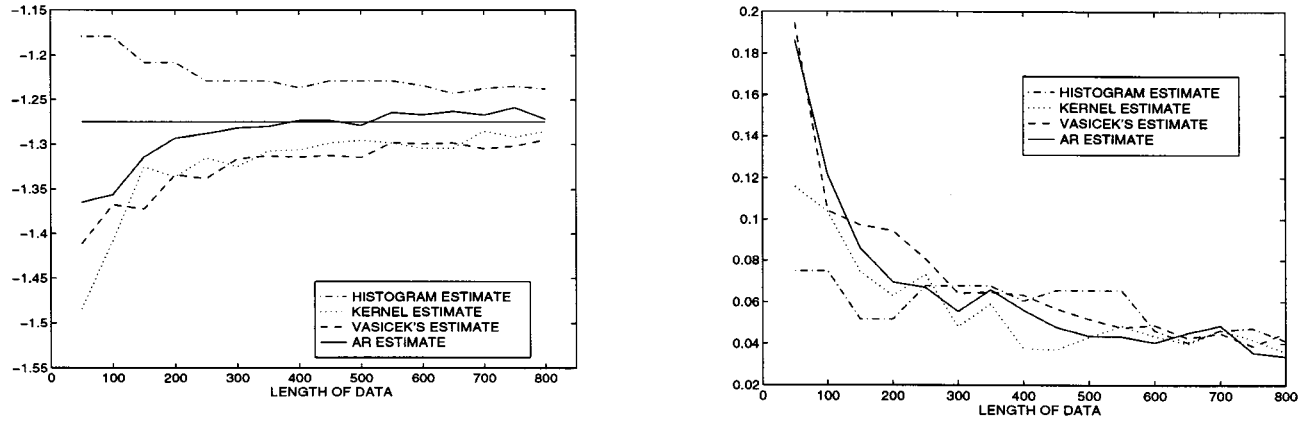
These results exhibit the good statistical behavior of our estimator, that is, a low bias and a small variance. For the uniform density, the AR-based estimator has about the same performance as the Vasicek's one concerning the bias but a lower standard deviation. The other estimators present a higher bias. In the case of the Gaussian density, the AR-based estimator clearly

[3]Since PDF's are modeled as power spectra on interval $[-(1/2), +(1/2)]$, the data had to be scaled on this interval. This does not restrict our approach because the entropy of the scaled variable differs from the original entropy only by a known additive term.



(a) (b)

Fig. 2. Comparison of histogram, kernel, Vasicek's, and AR-based estimates of entropy for a uniform density. (a) Mean values. (b) Standard deviations.



(a) (b)

Fig. 3. Comparison of histogram, kernel, Vasicek's, and AR-based estimates of entropy for a Gaussian density. (a) Mean values. (b) Standard deviations.

presents the lowest bias, whereas all have about the same standard deviation. Considering these two test cases, the AR-based estimator of entropy shows either comparable or lower bias and variance. Hence, it proves accurate and compares favorably with all other estimators considered.

As far as the complexity is concerned, the AR-based estimator also compares favorably with others since it suffices to estimate some correlation coefficients, find parameters using a gradient recursion, construct two times series, and compute their scalar product. The kernel approach requires a large amount of memory in order to evaluate and store the density estimate, even in the case of recursive kernel estimators (storage requirements can be reduced by using a coarser grid adjusted to the data range, but the bias increases with the length of intervals on the grid). Entropy evaluation requires L multiplications and L evaluations of $\log_2$. Vasicek's estimator requires storing and sorting the data and n evaluations of $\log_2$.

V. Sample Applications

A. Detecting PDF Changes

An interesting application of the adaptive estimates of Section III-C consists of detecting PDF changes in signals. As an illustration, we consider a signal $x(n)$ that is composed of 200 samples generated according to a mixture of two Gaussian distributions, with means ± 0.3 and standard deviation $\sigma = 0.06$, followed by 200 samples distributed uniformly on the interval $[-0.44, 0.44]$ and by 200 samples of the same gaussian mixture. First PDF has entropy $H_1 = -1.011$ bits, whereas the second has entropy $H_2 = -0.1844$ bits.

Fig. 4 shows signal $x(n)$; it is difficult, by a simple inspection, to diagnose that there are PDF changes. Fig. 5 shows the adaptive estimates [computed using (12) and (13)] of the negentropy of this test signal, using a forgetting factor $\mu = 0.98$.

The following points are of importance.

i) PDF changes appear clearly.
ii) Rupture points are properly revealed.
iii) The entropy is estimated with accuracy.
iv) The adaptive estimate has a good tracking capability.

B. Blind Deconvolution of AR Systems

The problem of blind deconvolution consists of recovering the input X and possibly the parameters of a filter from the sole observation of its output Y. The concept of entropy brings an interesting answer to this problem [25], relying on the following proposition.

Proposition 1: Let $Y(\omega, n)$ be the output of a unit norm filter whose input is a non-Gaussian i.i.d. sequence $X(\omega, n)$. Then, $H(Y) > H(X)$.

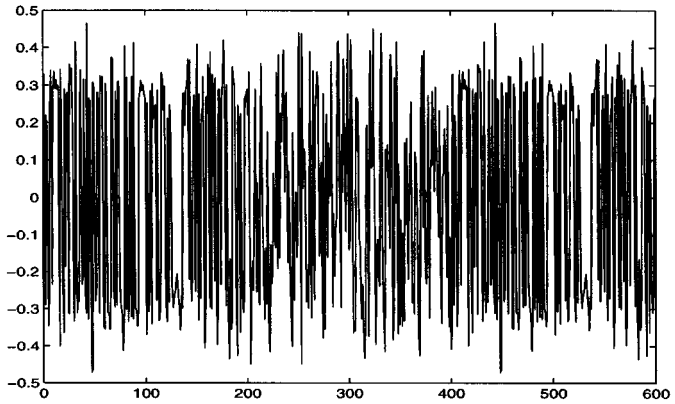


Fig. 4. Test signal for adaptive estimates: 200 samples from a gaussian mixture, 200 samples from a uniform distribution and 200 samples from a gaussian mixture.

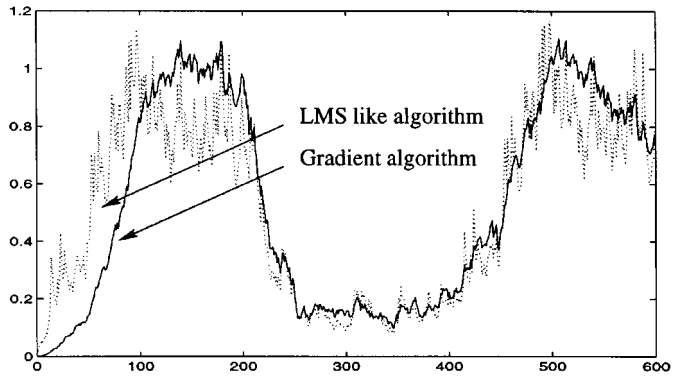


Fig. 5. Adaptive estimates of negentropy for signal in Fig. 4.

The intuitive reason behind this result is that $p_Y(y)$ is closer to a Gaussian distribution than $p_X(x)$, where the Gaussian distribution has the maximum entropy in the set of distributions of given variance; see [25].

Proof: The key of the proof is the entropy power inequality stated by Shannon [1, Th. 15 and App. 6]; see also [26].

Entropy power inequality: *If X and Y are two independent random variables with entropies $H(X)$ and $H(Y)$, then*

$$e^{2H(X+Y)} \geq e^{2H(X)} + e^{2H(Y)}$$

with equality if and only if X and Y are Gaussian variables.

The term "entropy power" comes from the fact that the power of a Gaussian variable X is proportional to $e^{2H(X)}$.

Let g be the impulse response of the filter with input $X(\omega, n)$. Its output $Y(\omega, n)$ is $Y(\omega, n) = \sum_i g_i X(\omega, n-i)$. The classical result on the entropy of rescaled variables and the assumption of stationarity give

$$\begin{aligned} H(g_i X(\omega, n-i)) &= H(X(\omega, n-i)) + \log_2 |g_i| \\ &= H(X) + \log_2 |g_i|. \end{aligned}$$

Now, the entropy power inequality gives

$$\begin{aligned} e^{2H(Y)} &\geq \sum_i e^{2H(g_i X(\omega, n-i))} = \sum_i e^{2H(X) + 2\log_2 |g_i|} \\ &= e^{2H(X)} \frac{1}{\log_e 2} \sum_i |g_i|^2. \end{aligned}$$

Then, taking $\log_e$ of both sides, we have

$$H(Y) \geq H(X) + \frac{1}{2} \log_2 \sum_i |g_i|^2.$$

Finally, for a unit norm filter, that is, $\sum_i |g_i|^2 = 1$, the last relation reduces to $H(Y) \geq H(X)$. □

The deconvolution procedure then simply consists of adjusting the parameters θ of a filter F_θ, with input Y, such that its output $\hat{X}$ has minimum (estimated) entropy. If θ are the filter parameters, this becomes

$$\theta_{\text{opt}} = \arg\min_\theta \hat{H}(\hat{X})$$
$$\text{submitted to } \begin{cases} \hat{X}(f) = F_\theta(f) Y(f) \\ \|F_\theta\|^2 = 1. \end{cases}$$

Simulations were performed in the case of nonminimum–phase AR filters. They showed that the AR parameters can be identified very accurately and that 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, and the initial solution was chosen as a standard minimum-phase solution.

In the case of non-AR filters, experiments showed that the procedure suffers from local minima. However, the procedure may prove of value when used in a compound criterion that should be considered in the presence of observation noise, such as

$$[\theta_{\text{opt}}, \mathbf{X}_{\text{opt}}] = \arg\min_{\theta, \hat{\mathbf{X}}} \hat{H}(\hat{X}) + \alpha \|\mathbf{Y} - \mathbf{H}_\theta \hat{\mathbf{X}}\|^2$$

where $\mathbf{H}_\theta$ is a convolution matrix. Note also that the previous criterion can be used in a standard deconvolution context, where $\mathbf{H}_\theta$ is known. In this case, the entropy term will help to select the "right" solution among several equivalent solutions, as it occurs in ill-posed problems.

C. Source Separation

In the context of source separation, N signals $\mathbf{s}(n) = [s_1(n), \ldots, s_N(n)]$ are mixed by an unknown $N \times N$ matrix $\mathbf{A}$ to provide observed signals $\mathbf{x}(n) = [x_1(n), \ldots, x_N(n)]$. The problem consists of recovering the sources from the sole observation of signals $\mathbf{x}(n)$ using only an assumption on the mutual independence of sources. The objective is reached by designing a matrix $\mathbf{B}$ such that the reconstructed signal $\hat{\mathbf{s}}(n) = \mathbf{B}\mathbf{x}(n)$ has independent components. The information theoretic measure of independence is the mutual information I_M, that is, the Kullback–Leibler divergence between $p_{\hat{S}_1, \ldots, \hat{S}_N}(\hat{s}_1, \ldots, \hat{s}_N)$ and $\Pi_{1 \leq i \leq N}\ p_{\hat{S}_i}(\hat{s}_i)$. For the source separation problem, minimization of the mutual information reduces to the minimization of

$$C(\mathbf{B}) = -\log |\det \mathbf{B}| + \sum_{i=1}^{N} H(\hat{s}_i). \tag{14}$$

In classical approaches, as no estimate of entropy is available, $\mathbf{B}$ is chosen as the solution of the nonlinear decorrelation equations: $E[\hat{s}_i \psi_j(\hat{s}_j)] = 0$, for $i \neq j$, which express the stationarity condition of $C(\mathbf{B})$. Function $\psi_j(s_j)$, which is the so-called score function, is the log derivative of the density $p_{S_j}(s_j)$.

TABLE I
EXPERIMENTS ON SOURCE SEPARATION. RESULTING MATRIX $M = AB$ SHOULD BE THE IDENTITY.

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

Using our AR parameterization, we can i) either estimate the cost function $C(\mathbf{B})$ and minimize it using any standard optimization procedure or ii) estimate the solution of the decorrelation equations using the analytical expression (in terms of the AR parameters) of the score function $\psi(x) = \mathbf{a}^+\mathbf{T}\mathbf{a}/|\mathbf{a}^+\mathbf{e}|^2$, where $\mathbf{T}$ is the Toeplitz matrix with entries $\mathbf{T}_{kl} = -j2\pi(k-l)e^{-j2\pi(k-l)x}$, and $\mathbf{e}^+ = [1\, e^{j2\pi x} \ldots e^{j2\pi(p-1)x}]$.

We performed simulations using the first approach, using 500 samples of data in the case of the mixture of $N = 2$ sources. Table I presents, for several distributions of the sources, the resulting matrix $M = AB$, which should be the identity matrix, up to a scaling factor and a permutation. These results show that this approach enables proper separation of the input sources, with performances comparable with classical methods [3].

VI. CONCLUSION

The concept of entropy plays a central role in information theory. However, it is rarely used directly in signal processing applications. In this paper, we have presented an estimator of the entropy of a signal and illustrated its behavior through several motivating examples of signal processing applications.

Our estimator relies on a simple analogy between the problems of PDF estimation and power spectrum estimation. The problem of PDF estimation is tackled using an AR modelization, which is a well-known approach in signal processing. This parametric modelization enables the accurate description of a large class of PDF. Moreover, in order to obtain accurate and stable estimates, we have chosen to use the long AR approach of [20], where the problem of AR parameters estimation is regularized by a smoothness constraint. Relying on the AR modelization, we have presented an estimation procedure for the entropy in a recursive scheme. The corresponding estimator does not require the explicit estimation of the PDF but only of some samples of a correlation sequence. Thus, it is easy to derive adaptive versions of this estimator. As illustrated by a simulation study, the AR estimate of entropy proves accurate and compares favorably with other classical estimates.

Finally, we have given several examples of applications where the entropy-based approach provides valuable results. It is worth recalling that many other applications can be considered, as soon as they involve some measure of dependence between random variables or signals or a measure of complexity.

ACKNOWLEDGMENT

The authors wish to thank A. Pagès-Zamora for useful comments on an early version of this manuscript and sending us [12] and a preprint of [15]. Suggestions and comments of F. Barbaresco were helpful and are acknowledged. The authors also wish to thank the anonymous referees for their valuable comments and suggestions.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423; 623–656, July/Oct. 1948 [Online] Available http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.
[2] P. Viola, N. N. Schraudolph, and T. J. Sejnowski, "Empirical entropy manipulation for real-world problems," in *Advances in Neural Information Processing Systems 8*. Cambridge, MA: MIT Press, 1996.
[3] D. T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Trans. Signal Processing*, vol. 44, pp. 2768–2779, Nov. 1996.
[4] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Process.*, vol. 16, no. 3, pp. 233–246, 1989.
[5] I. Kontoyiannis *et al.*, "Nonparametric entropy estimation for stationary processes and random fields, with application to english text," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1319–1327, May 1998.
[6] D. S. Ornstein and B. Weiss, "Entropy and data compression schemes," *IEEE Trans. Inform. Theory*, vol. 39, pp. 78–83, Jan. 1993.
[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[8] P. Hall and S. Morton, "On the estimation of entropy," *Ann. Inst. Stat. Math.*, vol. 45, pp. 69–88, 1993.
[9] O. Vasicek, "A test of normality based on sample entropy," *J. R. Stat. Soc. Ser. B*, vol. 38, pp. 54–59, 1976.
[10] J. C. Correa, "A new estimator of entropy," *Commun. Stat.–Theory Methodol.*, vol. 24, pp. 2439–2449, 1995.
[11] R. Wieczorkowski and P. Grzegorzewski, "Entropy estimators—Improvements and comparisons," *Commun. Stat.–Simul. Comput.*, vol. 28, no. 2, pp. 541–567, 1999.
[12] A. Pagès-Zamora and M. A. Lagunas, "New approaches in nonlinear signal processing: Estimation of the PDF function by spectral estimation methods," in *Proc. IEEE-Athos Workshop Higher-Order Stat.*, June 1995, pp. 204–208.
[13] S. Kay, "Model-based probability density function estimation," *IEEE Signal Processing Lett.*, vol. 5, pp. 318–320, Dec. 1998.
[14] J.-F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," in *Proc. ICASSP*, Phoenix, AZ, Mar. 1999.
[15] A. Pagès-Zamora and M. A. Lagunas, "Fourier models for nonlinear signal processing," *Signal Process.*, vol. 76, no. 1, pp. 1–16, 1999.
[16] P. J. Brockwell and R. A. Davis, *Times Series: Theory and Methods*, 2nd ed. New York: Springer-Verlag, 1987.
[17] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
[18] J. P. Burg, "Maximum entropy spectral analysis," presented at the Proc. 37th Meet. Soc. Explor. Geophys., 1967.
[19] J. E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 230–237, Apr. 1981.
[20] G. Kitagawa and W. Gersh, "A smoothness priors long AR model method for spectral estimation," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 57–65, 1985.
[21] J.-F. Giovannelli, G. Demoment, and A. Herment, "A bayesian method for long AR spectral estimation: A comparative study," *IEEE Trans. Ultrason. Freq. Ferroelect.*, vol. 43, pp. 220–233, Mar. 1996.
[22] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.
[23] O. Hössjer and U. Host, "On-line density estimators with high efficiency," *IEEE Trans. Inform. Theory*, vol. 41, pp. 829–835, May 1995.

[24] S. Haykin, *Adptive Filter Theory*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[25] D. Donoho, "On minimum entropy deconvolution," in *Applied Time Series Analysis II*. New York: Academic, 1981, pp. 565–609.
[26] A. Dembo and T. M. Cover, "Information theoretic inequalities," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1501–1518, Nov. 1991.

Jean-François Bercher was born in Marseille, France, in 1965. He received the B.E.E. degree in 1990 from the Institut National Polytechnique de Grenoble, Grenoble, France. He then received the Ph.D. degree in physics from the Laboratoire des Signaux et Systèmes, Université de Paris-Sud, Orsay, France, in 1995.

He is now with the Laboratoire Signaux et Télécoms of the ESIEE, Noisy-le-Grand, France, where he is Associate Professor. His interests are in applications of information theory to inverse problems, signal processing, and telecommunications.

Dr. Bercher is affiliated with MGEN.

Christophe Vignat was born in France in 1965. He received the B.E.E. degree from the École Supérieure d'Électricité, Paris, France. He then received the Ph.D. degree in physics from the Laboratoire des Signaux et Systèmes, the Université de Paris-Sud, Orsay, France, in 1993.

Since 1995, he has been Associate Professor at the Université de Marne-la-Vallée, Noisy-le-Grand, France. After working on adaptive systems, his interests shifted toward applications of signal processing to communication systems.

Dr. Vignat is a member of CROUS.