ENTROPY AND KULLBACK-LEIBLER DIVERGENCE ESTIMATION BASED ON SZEGÖ'S THEOREM

David Ramírez, Javier Vía and Ignacio Santamaría

University of Cantabria
Department of Communications Engineering
Santander, 39005, Spain
email:{ramirezgd,jvia,nacho}@gtas.dicom.unican.es

Pedro Crespo

CEIT and Tecnun (University of Navarra)
Manuel Lardizabal 15,
San Sebastián, 20018, Spain
email:pcrespo@ceit.es

ABSTRACT

In this work, a new technique for the estimation of the Shannon's entropy and the Kullback-Leibler (KL) divergence for one dimensional data is presented. The estimator is based on the Szegö's theorem for sequences of Toeplitz matrices, which deals with the asymptotic behavior of the eigenvalues of those matrices, and the analogy between a probability density function (PDF) and a power spectral density (PSD), which allows us to estimate a PDF of bounded support using the well-known spectral estimation techniques. Specifically, an AR model is used for the PDF/PSD estimation, and the entropy is easily estimated as a function of the eigenvalues of the autocorrelation Toeplitz matrix. The performance of the Szegö's estimators is illustrated by means of Monte Carlo simulations and compared with previously proposed alternatives, showing a good performance.

1. INTRODUCTION

Shannon's entropy and Kullback-Leibler divergence are two important measures in information theory [1], which have proven to be useful in many applications, from source coding to machine learning and signal processing. In this paper, we address the problem of estimating both quantities given a finite set of samples drawn from a continuous distribution of bounded support. This problem, which is much harder than that of estimating these quantities for discrete random variables, has been studied, for instance, in [2, 3, 4] for the entropy, and in [5, 6] for the Kullback-Leibler divergence.

In this paper, a new estimator for both information-theoretic measures is proposed based on the two following ideas. Firstly, we exploit the analogy between a probability density function (PDF) and a power spectral density (PSD). Both functions are non-negative and have finite area. Exploiting this analogy a PDF can be estimated using spectral estimation techniques [7]. For instance, in [8] the authors proposed to use non-parametric estimators of the PSD to estimate the PDF and in [4, 9] the PDF is modeled as an autoregressive (AR) process. Moreover, this idea has also been used in other problems like blind source separation [10] and blind channel equalization [11]. Secondly, we use the Szegö's theorem for sequences of Toeplitz matrices [12,13] which, in its simpler formulation, states that, as the size of the Toeplitz matrix tends to infinite, the arithmetic mean of the eigenvalues is equal to the integral of the Fourier transform of the sequence that generates the matrix. Moreover, the Szegö's theorem can be extended, for instance, to the product of Toeplitz matrices [13] and to block Toeplitz matrices [14]. Finally, it is important to point out that the analogy between the PDF and PSD was also used in [4] to estimate the entropy by direct application of the Plancherel-Parseval theorem. However, in this paper, we propose to use the Szegö's theorem.

This work was supported by the Spanish Government, Ministerio de Ciencia e Innovación (MICINN), under project MultiMIMO (TEC2007-68020-C04-02 and TEC2007-68020-C04-03), project COMONSENS (CSD2008-00010, CONSOLIDER-INGENIO 2010 Program) and FPU grant AP2006-2965.

2. PREVIOUS BACKGROUND

2.1 Analogy between PDF and PSD

It is well known that the power spectral density (PSD) of a discrete-time wide sense stationary random process has similar properties to a probability density function (PDF) of a continuous random variable (RV): both are non-negative and have finite area. This idea allows us to apply the well-known spectral estimation techniques to problems that involve the PDF. For instance, this analogy is exploited in [8,9] to estimate the PDF and in [4] to estimate the entropy of a RV based on parametric and nonparametric PSD estimators.

Let us start by assuming that the support of the PDF lies in $[-1/2, 1/2]$, or at least that it is bounded. Viewing the density function as a spectrum, we can compute its autocorrelation sequence as the inverse discrete Fourier transform of the PDF

$$\phi_x[k] = \mathcal{F}^{-1}\left(p(x)\right) = \int_{-1/2}^{1/2} e^{j2\pi xk} p(x)dx,$$

where $p(x)$ is the PDF of the RV[1] x and $\mathcal{F}^{-1}(\cdot)$ denotes the inverse discrete Fourier transform. Obviously, this autocorrelation sequence is nothing but samples of the characteristic function of the RV, as can be seen in the following equation

$$\phi_x[k] = E_p\left[e^{j2\pi xk}\right], \quad (1)$$

where $E_p[\cdot]$ denotes the expectation operator with respect to $p(x)$. Based on (1), Kay proposed to use the

[1] With some abuse of notation, in this paper we use x to denote both the RV and the argument of the PDF.

sample moment estimator of $\phi_x[k]$ from N available samples x_i $(i = 0, \ldots, N-1)$, which is given by

$$\hat{\phi}_x[k] = \frac{1}{N} \sum_{i=0}^{N-1} e^{j2\pi x_i k}. \quad (2)$$

Here, it is important to point out that this estimator also ensures that the associated estimate of the PDF integrates to one.

2.2 Szegö's theorem for sequences of Toeplitz matrices

In this subsection, we review the Szegö's theorem for sequences of Toeplitz matrices [12,13]. Consider a $K \times K$ Toeplitz matrix given by

$$\mathbf{R}_K = \begin{bmatrix} r[0] & r[-1] & \cdots & r[-K+1] \\ r[1] & r[0] & \cdots & r[-K+2] \\ \vdots & \vdots & \ddots & \vdots \\ r[K-1] & r[K-2] & \cdots & r[0] \end{bmatrix},$$

the Szegö's theorem deals with the behavior of the eigenvalues of $\mathbf{R}_K$ as K goes to infinite. Concretely, it states that under some mild assumptions on $r[k]$ (mainly the continuity of its Fourier spectrum $R(\nu)$ [14]), the eigenvalues of $\mathbf{R}_K$ are related to the Fourier transform of $r[k]$ as follows

$$\lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} \lambda_{K,i} = \int_{-1/2}^{1/2} R(\nu) d\nu,$$

where $\lambda_{K,i}, i = 1, \ldots, K$ are the eigenvalues of $\mathbf{R}_K$ and $R(\nu)$ is the Fourier transform of $r[k]$. A more general form of the theorem states that

$$\lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} F(\lambda_{K,i}) = \int_{-1/2}^{1/2} F(R(\nu))\, d\nu,$$

where $F(\cdot)$ is any continuous function on the range of $R(\nu)$. Additionally, it is also possible to extend the Szegö's theorem to matrix operations. For instance, an interesting result, which will be applied in Section 4, is related to the eigenvalues of the product of functions of Toeplitz matrices[2]. That is, given two $K \times K$ function matrices $g(\mathbf{R}_K)$ and $v(\mathbf{S}_K)$, then the eigenvalues $\{\beta_{K,i}\}_{i=1}^{K}$ of the matrix product $g(\mathbf{R}_K) \cdot v(\mathbf{S}_K)$ verify

$$\lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} \beta_{K,i} = \int_{-1/2}^{1/2} g(R(\nu)) v(S(\nu)) d\nu,$$

where $\mathbf{R}_K$ and $\mathbf{S}_K$ are Toeplitz matrices formed by the sequences $r[k] = \mathcal{F}^{-1}(R(\nu))$ and $s[k] = \mathcal{F}^{-1}(S(\nu))$, respectively [14].

[2]Let $\mathbf{A}$ be a $K \times K$ diagonalizable matrix $\mathbf{A} = \mathbf{U}\text{diag}(\lambda_1(\mathbf{A}), \ldots, \lambda_K(\mathbf{A}))\mathbf{U}^{-1}$. If g is a complex function on the set $\{\lambda_1(\mathbf{A}), \ldots, \lambda_K(\mathbf{A})\}$, the $K \times K$ function matrix $g(\mathbf{A})$ is defined as $g(\mathbf{A}) = \mathbf{U}\text{diag}(g(\lambda_1(\mathbf{A})), \ldots, g(\lambda_K(\mathbf{A})))\mathbf{U}^{-1}$.

3. SHANNON ENTROPY ESTIMATION

3.1 Development of the main idea

One of the most important measures in information theory is the Shannon's entropy [1]. In this section, we present a new technique to estimate the entropy of a continuous random variable of bounded support, i.e., the differential entropy, by means of the Szegö's theorem.

Given a random variable x with a probability density function $p(x)$, the Shannon's differential entropy (measured in bits) is given by [1]

$$H_p(x) = -E_p[\log_2 p(x)] = -\int p(x) \log_2 p(x) dx. \quad (3)$$

As can be seen from (3), the entropy is obtained as the integral of a function of the PDF or, equivalently, as the integral of a function of a PSD. Assuming that the support of x lies in the interval $[-1/2, 1/2]$, the entropy can be rewritten as

$$\begin{aligned} H_p(x) &= -\int_{-1/2}^{1/2} p(x) \log_2 p(x) dx \\ &= \lim_{K\to\infty} -\frac{1}{K} \sum_{i=1}^{K} \lambda_{K,i} \log_2 \lambda_{K,i}, \end{aligned}$$

where $\lambda_{K,i}$ are the eigenvalues of a Toeplitz matrix with its entries given by

$$[\mathbf{\Phi}_K]_{i,j} = \phi_x[i-j] = E_p\left[e^{j2\pi x(i-j)}\right], \quad i, j = 1, \ldots, K.$$

3.2 Practical implementation

From a practical standpoint, it is not possible to take the limit as K approaches infinite, and therefore, a finite version of the Szegö's theorem should be used

$$H_p(x) \approx -\frac{1}{K} \sum_{i=1}^{K} \lambda_{K,i} \log_2 \lambda_{K,i},$$

where K has to be large enough in order to obtain an accurate approximation of the Szegö's theorem.

Additionally, the autocorrelation sequence $\phi_x[k]$ must be estimated from the N available samples. A direct application of (2) to estimate $\hat{\phi}_x[k]$ for large k, would require a large number of samples. To overcome this limitation, we propose to use a parametric model for the PSD. Concretely, we use a regularized AR model as proposed in [4], and the model's order is chosen by means of the minimum description length (MDL) criterion [7]. To summarize, the following algorithm is proposed:

- Normalize the samples to ensure that they belong to the interval $[-1/2, 1/2]$ as follows

$$y_i = \alpha x_i.$$

- Estimate $p_{max} + 1$ lags of the autocorrelation using (2): $\hat{\phi}_y[0], \ldots, \hat{\phi}_y[p_{max}]$.
- Obtain the AR models of orders from 1 to p_{max}.

- Select the AR model which minimizes the MDL criterion (order p) and use it to extrapolate K ($K \gg p+1$) lags of $\hat{\phi}_y[k]$.
- Build the Toeplitz matrix $\hat{\mathbf{\Phi}}_K$ from the extrapolated autocorrelation sequence.
- Estimate the entropy of y as follows

$$\hat{H}_q(y) = -\frac{1}{K}\sum_{i=1}^{K} \hat{\lambda}_{K,i} \log_2 \hat{\lambda}_{K,i},$$

where $\hat{\lambda}_{K,i}$ are the eigenvalues of $\hat{\mathbf{\Phi}}_K$ and $q(x)$ is the PDF of the RV y.
- The entropy of x is given by

$$\hat{H}_p(x) = \hat{H}_q(y) - \log_2 \alpha.$$

4. KULLBACK-LEIBLER DIVERGENCE ESTIMATION

In this section, an estimator of the Kullback-Leibler (KL) divergence [1] is presented following a similar approach. Given two random variables x and y with probability density functions $p(x)$ and $q(x)$, the Kullback-Leibler divergence is defined (in bits) as

$$D(p||q) = E_p\left[\log_2 \frac{p(x)}{q(x)}\right] = \int p(x) \log_2 \frac{p(x)}{q(x)} dx.$$

This divergence is finite if $p(x)$ is absolutely continuous with respect to $q(x)$, and zero if and only if $p(x) = q(x)$. A useful property, that will be used later, is that the Kullback-Leibler divergence is scale invariant. Specifically, consider the following RVs $v = \alpha x$ and $w = \alpha y$ ($\alpha > 0$) with PDFs $t(x)$ and $u(x)$, then the KL divergence is given by

$$D(t||u) = D(p||q).$$

For our purposes, it is useful to rewrite the KL divergence as follows

$$D(p||q) = -H_p(x) - \int p(x) \log_2 q(x) dx.$$

Finally, assuming that the support of both PDFs is constrained to the interval $[-1/2, 1/2]$ and taking into account the results from Section 2.2, it is possible to rewrite the KL divergence as

$$D(p||q) = \lim_{K\to\infty} \frac{1}{K}\sum_{i=1}^{K} \lambda_{K,i} \log_2 \lambda_{K,i} - \lim_{K\to\infty} \frac{1}{K}\sum_{i=1}^{K} \beta_{K,i}, \quad (4)$$

where $\lambda_{K,i}$ are the eigenvalues of $\mathbf{\Phi}_K^{(x)}$, and $\beta_{K,i}$ are the eigenvalues of the product of $\mathbf{\Phi}_K^{(x)}$ by $\log_2 \mathbf{\Phi}_K^{(y)}$, where the Toeplitz matrices $\mathbf{\Phi}_K^{(x)}$ and $\mathbf{\Phi}_K^{(y)}$ are given by

$$\left[\mathbf{\Phi}_K^{(x)}\right]_{i,j} = \int p(x) e^{j2\pi x(i-j)} dx, \quad i,j = 1,\ldots,K,$$

$$\left[\mathbf{\Phi}_K^{(y)}\right]_{i,j} = \int q(x) e^{j2\pi x(i-j)} dx, \quad i,j = 1,\ldots,K.$$

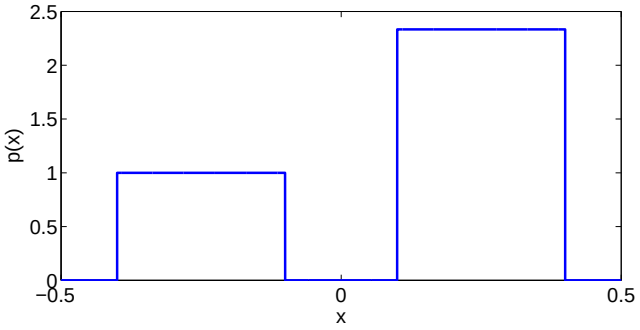

Figure 1: PDF of the uniform mixture

Finally, taking (4) into account, it can be seen that the KL divergence can also be expressed in terms of the eigenvalues of Toeplitz matrices. Therefore, we propose the following algorithm to estimate the KL divergence from N available samples:

1. Normalize the samples to ensure that the RVs are constrained to the interval $[-1/2, 1/2]$ as follows:

$$v_i = \alpha x_i, \quad w_i = \alpha y_i.$$

2. Obtain the best AR models for v_i and w_i as in the previous section and use these models to extrapolate K lags of both autocorrelation sequences.
3. Build the two $K \times K$ Toeplitz matrices: $\hat{\mathbf{\Phi}}_K^{(v)}$ and $\hat{\mathbf{\Phi}}_K^{(w)}$.
4. Estimate the Kullback-Leibler divergence as follows

$$\hat{D}(p||q) = \frac{1}{K}\sum_{i=1}^{K} \hat{\lambda}_{K,i}^{(v)} \log_2 \hat{\lambda}_{K,i}^{(v)} - \frac{1}{K}\text{trace}\left(\hat{\mathbf{\Phi}}_K^{(v)} \log_2 \hat{\mathbf{\Phi}}_K^{(w)}\right),$$

where $\hat{\lambda}_{K,i}^{(v)}$ are the eigenvalues of $\hat{\mathbf{\Phi}}_K^{(v)}$.

5. SIMULATION RESULTS

5.1 Entropy estimation

In this subsection, the performance of the proposed estimator is evaluated by means of Monte Carlo simulations. Concretely, we measure the mean value and the mean square error (MSE) of the estimator for a uniform density $x \sim U[-0.1, 0.1]$ and for a mixture of uniform densities $x \sim 0.3U[-0.4, -0.1] + 0.7U[0.1, 0.4]$, which is depicted in Fig. 1. The following parameters are selected for the proposed technique in all examples: the regularization parameter for the AR model is $\lambda = 10^{-5}$ [4], the size of the correlation matrix K is proportional to the model's order p, concretely, we used $K = 10p$ and $K = 40p$. We have compared the performance of the proposed estimator with those obtained by the following estimators:

- Kernel-based estimator: this estimate is based on the numerical integration (using a grid of $L = 1000$ points) of the kernel-based estimated PDF [15].

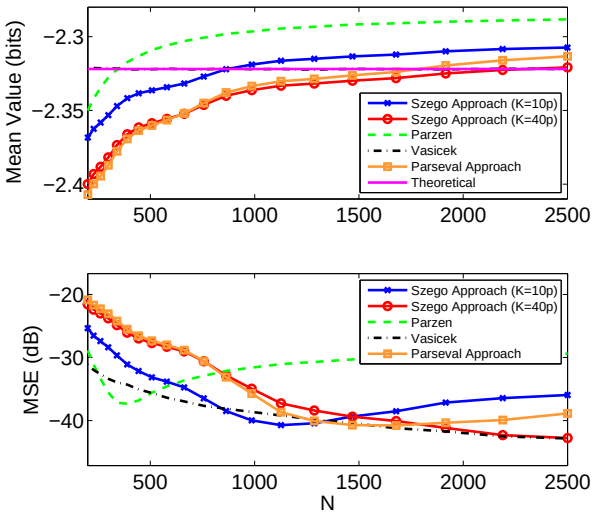

Figure 2: Mean value and MSE of the different estimators of the entropy for $x \sim U[-0.1, 0.1]$

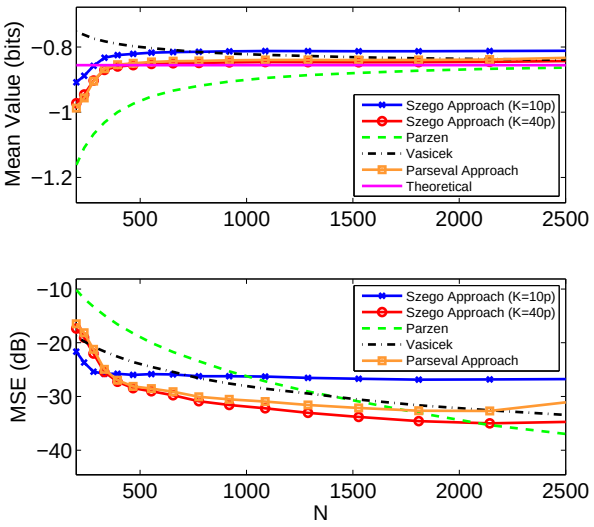

Figure 3: Mean value and MSE of the different estimators of the entropy for $x \sim 0.3U[-0.4, -0.1] + 0.7U[0.1, 0.4]$

Specifically, we have chosen a Gaussian kernel with kernel size $\sigma_k = 0.003$ and $\sigma_k = 0.002$ for the uniform and mixture of uniform densities, respectively.

- Vasicek's estimator [2], which relies on the fact that the entropy can be rewritten as

$$H_p(x) = \int_0^1 \log_2 \left(\frac{dF^{-1}(u)}{du} \right) du,$$

where $F(u)$ is the cumulative distribution function (CDF). The estimator is obtained by approaching the CDF with order statistics as follows

$$\hat{H}_p(x) = \frac{1}{N} \sum_{i=1}^{N} \log_2 \left\{ \frac{N}{2m} \left(x_{(i+m)} - x_{(i-m)} \right) \right\} + f(m, N),$$

where $\{x_{(i)}\}$ is the set of ordered samples x_i, m is the order spacing and $f(m, N)$ is a bias correction term [2]. For all examples, an order spacing of $m = 3$ is selected.

- The entropy estimator of [4] which is also based on the analogy between a PDF and a PSD and the Plancherel-Parseval theorem.

Figures 2 and 3 show the results of the different estimators. As can be seen, Vasicek's estimator has a very low bias (for the first example, its mean value is indistinguishable from the theoretical value), although in terms of MSE its performance is degraded, especially for the mixture of uniform densities shown in Fig. 3. Regarding the Szegö's approach, we can see that, although it is biased, its overall MSE is rather small, especially for large values of K and N. On the other hand, the conducted simulations have shown that the bias depends on the size of the Toeplitz matrix and the number of available samples. Not surprisingly, when N is small, smaller matrices should be used to obtain a low bias; on the other hand, when N is large, larger Toeplitz matrices provide better results. Finally, the kernel-based approach does not provide in general accurate estimates and the estimator of [4] provides similar results. However, the proposed approach is more general in the sense that it can be applied to estimate other information-theoretic measures.

5.2 Kullback-Leibler divergence estimation

In this subsection, the performance of the proposed estimator for the KL divergence is presented. We have chosen the parameters of the previous subsection. The proposed estimator is compared to the estimator presented in [5], which is based on the M-th nearest neighbor (for the simulations $M = 5$ and $M = 10$ are selected) and is given by

$$\hat{D}(p||q) = \frac{1}{N} \sum_{i=1}^{N} \log_2 \frac{d_M(x_i)}{d'_M(x_i)} + \log_2 \frac{N}{N-1},$$

where $d_M(x_i)$ is the Euclidean distance from x_i to its M-th nearest neighbor in the set $\{x_l\}, l = 1, \ldots, i-1, i+i, \ldots, N$ and $d'_M(x_i)$ is the Euclidean distance from x_i to its M-th nearest neighbor in the set $\{y_l\}, l = 1, \ldots, N$.

Specifically, the mean and MSE values of the KL estimates $\hat{D}(p||q)$ and $\hat{D}(q||p)$, when $p(x) = \mathcal{N}(0, 1)$ and $q(x) = \mathcal{N}(0, 2)$, are shown in Figures 4 and 5, respectively. Here, we must point out that using the proposed technique with PDF of unbounded support can suffer from aliasing [8] in the implicit PSD estimation. However, as can be seen in the figures, the results are accurate. In particular, it can be seen that the proposed estimator provides very good MSE results, although its bias is larger than the KNN approach (obviously its variance is much lower), mainly when the number of available samples is small.

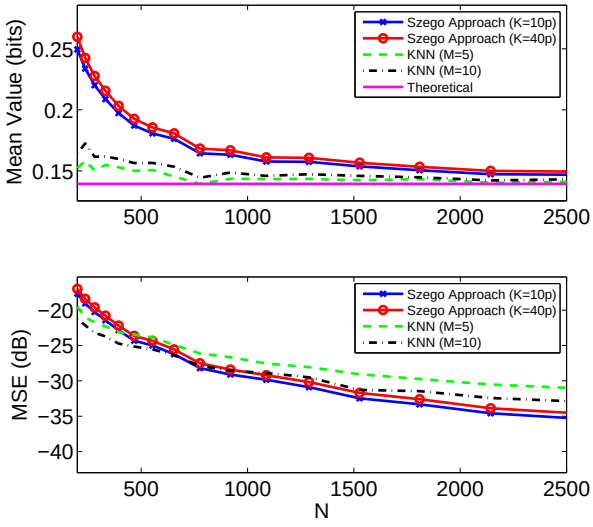

Figure 4: Mean value and MSE of the different estimators of the KL divergence $\hat{D}(\mathcal{N}(0,1),\mathcal{N}(0,2))$

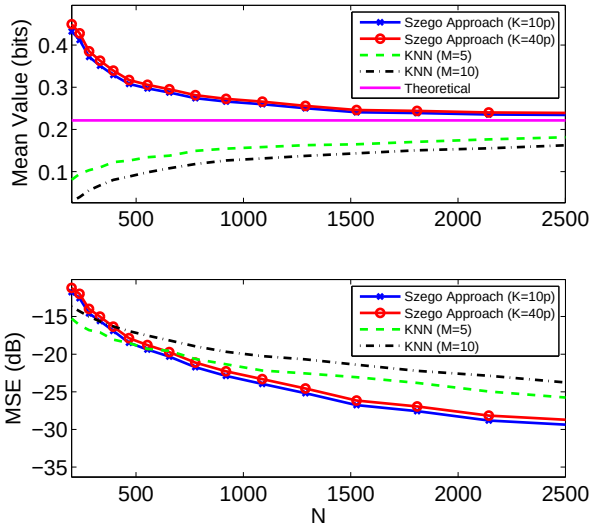

Figure 5: Mean value and MSE of the different estimators of the KL divergence $\hat{D}(\mathcal{N}(0,2),\mathcal{N}(0,1))$

6. CONCLUSIONS

In this work, a new estimator for the Shannon's entropy and the Kullback-Leibler divergence of one dimensional data is presented. The idea is based on the combination of the Szegö's theorem for sequences of Toeplitz matrices and the analogy between a probability density function (PDF) and a power spectral density (PSD). The performance of the proposed method has been compared to that of other techniques by means of numerical examples, which show that the proposed estimators provide very accurate results. Finally, we have to point out that the proposed technique could be extended to other information-theoretic measures (for instance, Renyi's entropy or Csiszar's divergence), as well as to multidimensional data. These extensions will be considered in future work.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[2] O. Vasicek, "A test of normality based on sample entropy," *J. R. Stat. Soc. Ser. B*, vol. 38, pp. 54–59, 1976.

[3] L. Kozachenko and N. Leonenko, "A statistical estimate for the entropy of a random vector," *Problems Infor. Transmiss.*, vol. 23, no. 2, pp. 9–16, 1987.

[4] J.-F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *IEEE Trans. on Signal Process.*, vol. 48, no. 6, pp. 1687–1694, Jun. 2000.

[5] F. Pérez-Cruz, "Kullback-Leibler divergence estimation of continuous distributions," in *IEEE Int. Symp. on Information Theory (ISIT)*, 2008.

[6] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. on Inf. Theory*, vol. 51, no. 9, pp. 3064–3074, Sept. 2005.

[7] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Prentice Hall, 2005.

[8] A. Pagés-Zamora and M. Lagunas, "New approaches in nonlinear signal processing: Estimation of the PDF function by spectral estimation methods," in *IEEE-Athos Workshop Higher-Order Stat.*, June 1995, pp. 204–208.

[9] S. Kay, "Model-based probability density function estimation," *IEEE Signal Process. Lett.*, vol. 5, pp. 318–320, Dec. 1998.

[10] L. Vielva, I. Santamaría, C. Pantaleón, J. Ibáñez, D. Erdogmus, and J. C. Principe, "Estimation of the mixing matrix for uderdetermined BSS using spectral estimation techniques," in *11th European Signal Process. Conf.*, Toulouse, France, September 2002.

[11] J. Vía, I. Santamaría, and M. Lázaro, "Blind restoration of binary signals using a line spectrum fitting approach," in *12th European Signal Process. Conf.*, Vienna, Austria, September 2004.

[12] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. Berkeley: Univ. Calif. Press, 1958.

[13] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*. Foundations and Trends in Communications and Information Theory, 2006, vol. 2, no. 3.

[14] J. Gutierrez-Gutierrez and P. M. Crespo, "Asymptotically equivalent sequences of matrices and hermitian block Toeplitz matrices with continuous symbols: Applications to MIMO systems," *IEEE Trans. on Inf. Theory*, vol. 54, no. 12, pp. 5671–5680, Dec. 2008.

[15] E. Parzen, "On estimation of a probability density function and mode," *Time Ser. Anal. Papers*, 1967.