Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Fusion 5 (2004) 119–129
INFORMATION FUSION
www.elsevier.com/locate/inffus

Relative information of multi-rate sensors

Omid S. Jahromi *, Bruce A. Francis, Raymond H. Kwong

Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, 10 Kings College Road, Toronto, Ont., Canada M5S 3G4

Received 19 November 2003; received in revised form 26 January 2004; accepted 26 January 2004

Abstract

Fusion of the information provided by distributed sensors is important in the future *pervasive computing* systems. The objective of this paper is to quantify the relative amount of information delivered by individual sensors in a distributed sensor system.

To formalize our approach, we consider the problem of fusing the statistical information obtained using low-rate sensor recordings $v_i(n)$ in order to find the power spectrum of a high-rate random signal $x(n)$. It turns out that this problem is inherently ill-posed since, in general, low-rate observations are not sufficient for specifying a unique solution. We use the Maximum Entropy principle to resolve this issue. It is known that a complete statistical description of a Gaussian wide-sense stationary process $x(n)$ is provided by its power spectrum. This leads us to a measure of *informativity* for multirate sensors based on the qualities of the maximum entropy solution. We study the main properties of the proposed measure and provide computational methods for its calculation. Finally, we illustrate the concepts and methods discussed in the paper using a detailed simulation example.

Keywords: Multirate systems; Filter banks; Statistical information; Entropy; Multichannel measurement; Multiresolution measurement; Array processing

1. Introduction

In recent years, the concept of *pervasive computing*, where computing is embedded in the environment and comes about through a large number of small and unobtrusive sensors, has attracted considerable interest. Pervasive computing devices communicate with each other through wired or wireless networks. However, an integral part of the system is a multi-sensor information fusion system which allows the system to "talk" to the user while he or she is moving freely around the office space [1]. There are numerous technologies and products that need to be combined in order for pervasive-computing to become viable. Among these, multi-microphone information fusion technologies that allow robust speech localization, enhancement, and recognition are particularly important [2–4].

To assess the reliability and quality of any sensor fusion system, it is important to know the causal relationship between the sensors' information and the resulting fused output in a clear way. In other words, the relative influence of the different input information components (from different sensors) on the fused result must be known clearly [5]. This requires a quantitative measure of information contained in each sensor's data. We are going to introduce one such measure in this paper.

The information measure introduced here (called *statistical information*) is important in the sense that it provides an *objective measure* of the performance of individual sensors used in an array. It can be used for designing and optimizing a large variety of distributed sensor systems such as microphone arrays, geophone systems used for geophysical explorations and so on. Multi-resolution sensor fusion has a great appeal in optical engineering applications as well [6].

This paper deals with the fusion of multi-sensor information where the sensors are allowed to sample and transmit their data at different sampling rates. This is because small and unobtrusive sensing nodes are

* Corresponding author. Tel.: +416-946-7893/8813; fax: +416-978-4425.
E-mail addresses: omidj@control.toronto.edu (O.S. Jahromi), francis@control.toronto.edu (B.A. Francis), kwong@control.toronto.edu (R.H. Kwong).

1566-2535/$ - see front matter
doi:10.1016/j.inffus.2004.01.003

inherently constrained in computation and communication capabilities. Furthermore, price, power consumption, and network data rate limitations might prohibit individual sensor nodes from transmitting high-rate measurement data. Effective fusion of such sensors, thus, requires multi-rate signal processing techniques whereby a unified high-resolution measurement is produced from the low-resolution data communicated by individual sensor nodes. A basic account of such techniques can be found in the recent paper [7]. In this paper, we complement the works in [7,8] by introducing a quantitative measure of information gained by various sensors in a distributed sensor array.

In modern distributed arrays, the sensors are usually equipped with a sampling mechanism so that the output data provided by the array are discrete-time (Fig. 1). Under mild technical conditions, a discrete-time (sampled-data) sensor array can be considered as a *natural* analysis filter bank whose output data $v_i(n)$ are low-rate components of a non-observable high-rate signal $x(n)$ (Fig. 1(c)). The measurement signals $v_i(n)$ have a sampling rate which is lower than that of the desired (but not available) measurement signal $x(n)$ and possibly different from each other. This is modelled by the decimator blocks with down-sampling ratios N_i shown in the detailed diagram of Fig. 1(c). The linear filters $H_i(z)$ in Fig. 1(c) model the analog filters $H_i(s)$ which, in turn, model the inherent bandwidth limitations of each sensor.

In this paper, we assume that the aim of measurement is to specify the *statistics* of the time series $x(n)$ rather than its actual sample values. The signal $x(n)$ is the original object of measurement buy not available to the observer. Thus, statistical inference must be performed indirectly by examining available low-resolution measurement signals $v_i(n)$.

It is well-known that a complete statistical description of a Gaussian, zero-mean, wide-sense stationary (WSS) process is provided by its autocorrelation sequence (ACS) $R_x(k) \triangleq E\{x(n)x(n+k)\}$ or, equivalently, by its power spectral density (PSD)

$$P_x(e^{j\omega}) = \sum_{k=-\infty}^{\infty} R_x(k)e^{-j\omega k}.$$

This is because if we know $P_x(e^{j\omega})$, we can calculate the probability density functions that govern the statistical dependence among any number of samples of $x(n)$.

It is straightforward to calculate the observable signals $v_i(n)$ once the signal $x(n)$ and the sensors' model (i.e. $H_i(z)$ and N_i) are known. However, the inverse problem (that is, finding $x(n)$ given $v_i(n)$ and the observers' model) is not easily tractable. This is mainly due to the fact that the single-input multi-output linear system specified by the multi-rate model in Fig. 1(c) is

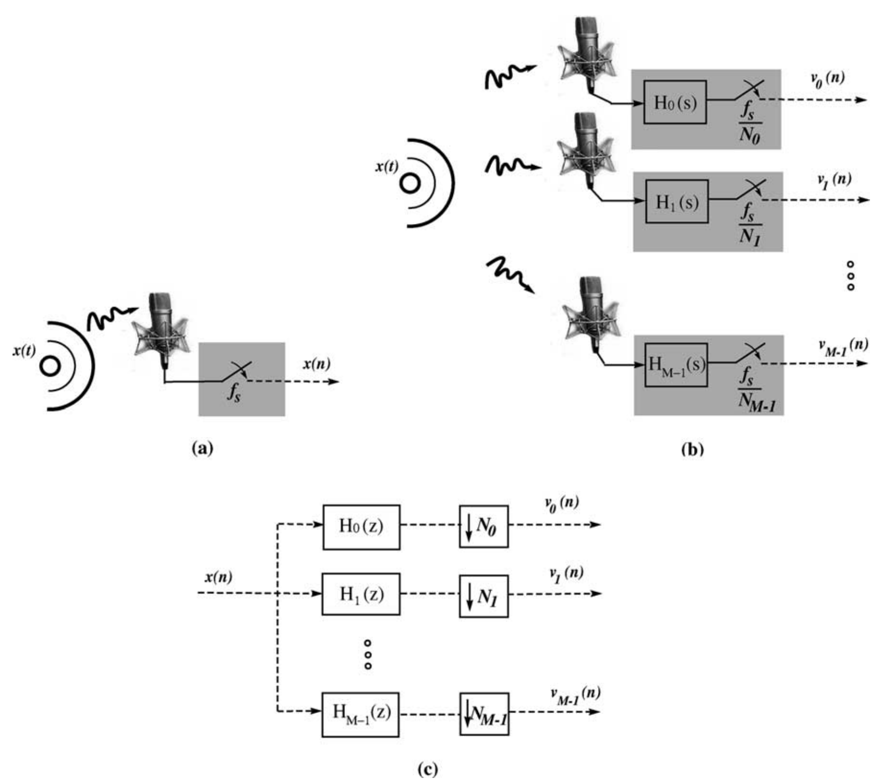

Fig. 1. A direct high-sampling-rate measurement (a) is emulated by a collection of low-sampling-rate measurements (b). The relation between the high-rate signal $x(n)$ (which is the object of measurement but not available) and the available low-rate measurements $v_i(n)$ can be modelled using a multi-rate filter bank as shown in (c).

not, in general, left-invertible. In other words, the measurements $v_i(n)$ may not be sufficient information for specifying $x(n)$ uniquely.

One can now pose questions like these: How much statistical information about $x(n)$ can be gained by doing statistical experiments on $v_0(n)$? Which signal, $v_0(n)$ or $v_1(n)$, gives *more* statistical information about $x(n)$? If we know statistical properties of $v_0(n)$, how much *more* information about $x(n)$ will be gained by doing statistical experiments on $v_1(n)$?

To answer questions of the type mentioned above, we need to establish a *quantitative measure* of statistical information gained about $x(n)$ by statistical experiments performed on the low-rate measurements $v_i(n)$. The goal of this paper is to introduce such a measure and study its properties.

Notation. Vectors are denoted by capital or boldface lower-case letters. Boldface capital letters are reserved for matrices. The expected value of a random variable x is denoted by $E\{x\}$. The symbol $\triangleq$ is used to indicate that two quantities are equal by definition. In diagrams, solid lines are used to represent analog signals whereas dotted lines denote discrete-time signals. The symbol ☆ denotes the linear convolution operator. The binary relations $\leqslant$ and $\geqslant$ should be interpreted component-wise when applied to vector objects. The end of an example is marked using the symbol. $\diamond$

2. Background

A very important concept in statistical inference is the concept of information and its measures. In the following sections, we introduce the Kullback–Leibler divergence which is used as a measure of *information for discrimination* between two probability density functions. We then show the connection between Kullback–Leibler divergence and other important measures of information, namely *entropy* and *mutual information* introduced by Shannon. These measures of information are then generalized to WSS random processes.

2.1. The Kullback–Leibler divergence

Let $p_1(\cdot)$ and $p_2(\cdot)$ be two probability density functions (PDFs) defined on $\mathscr{C} \subset \mathbb{R}^N$. Also, let $\mathscr{H}_i$, $i = 1, 2$, be the hypothesis that a certain random variable is from the statistical population with PDF p_i, respectively. Kullback and Leibler [9] defined the logarithm of the likelihood ratio $\ln p_1(X)/p_2(X)$ as the information in X for discrimination in favor of $\mathscr{H}_1$ against $\mathscr{H}_2$. The mean value of this information, that is, *the mean information for discrimination in favor of* $\mathscr{H}_1$ *against* $\mathscr{H}_2$ *per observation from the PDF* p_1, is called the *Kullback–Leibler divergence* of p_1 from p_2:

$$D(p_1\|p_2) \triangleq \int_{\mathscr{C}} p_1(X) \ln \frac{p_1(X)}{p_2(X)}\, \mathrm{d}X. \tag{1}$$

Note that the base of the logarithm in the above definition is not important and only represents the choice of the unit of measurement.

Example 1. Let p_1 and p_2 be N-dimensional Gaussian PDFs with zero mean and covariance matrices $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively. The Kullback–Leibler divergence of p_1 from p_2 is given by [10, Chapter 9]

$$D(p_1\|p_2) = \frac{1}{2}\mathrm{Tr}(\mathbf{C}_1\mathbf{C}_2^{-1}) - \frac{1}{2}\ln\frac{|\mathbf{C}_1|}{|\mathbf{C}_2|} - \frac{N}{2}. \quad \diamond \tag{2}$$

One can verify that $D(p_1\|p_2)$ has the following properties:

(a) $D(p_1\|p_2) \geqslant 0$.
(b) $D(p_1\|p_2) = 0$ if and only if $p_1 = p_2$.
(c) In general, $D(p_1\|p_2) \neq D(p_2\|p_1)$.

The property (c) prevents $D(p_1\|p_2)$ from satisfying the conditions of a metric. However, it still has a strong appeal as a measure of distance between probability density functions [10,11].

2.2. Kullback–Leibler divergence for stochastic processes

The definition of Kullback–Leibler divergence given in (1) can be extended to zero-mean Gaussian WSS stochastic processes. As mentioned earlier, the statistics of such processes is completely determined by their PSD function or, equivalently, by their ACS. So, let $P_1(e^{j\omega})$ and $P_2(e^{j\omega})$ represent two hypotheses for the PSD of a zero-mean Gaussian stochastic process $x(n)$. Now, define X_N as the vector containing N samples of the process $x(n)$, that is

$$X_N \triangleq [x(0)\ x(1) \ldots x(N-1)]^{\mathrm{T}}. \tag{3}$$

Obviously, X_N is an N-dimensional Gaussian random vector with zero mean and an $N \times N$ covariance matrix which is uniquely determined by the PSD of $x(n)$. We denote by D_N the Kullback–Leibler divergence between the two possible PDFs of X_N. From, (2), we can then write

$$D_N = \frac{1}{2}\mathrm{Tr}(\mathbf{C}_1\mathbf{C}_2^{-1}) - \frac{1}{2}\ln\frac{|\mathbf{C}_1|}{|\mathbf{C}_2|} - \frac{N}{2}, \tag{4}$$

where $\mathbf{C}_1$ represents the covariance matrix of X_N if $P_1(e^{j\omega})$ is the true PSD and $\mathbf{C}_2$ represents the covariance matrix of X_N for the case that $P_2(e^{j\omega})$ is true. It turns out that under mild conditions $\lim_{N\to\infty} D_N$ exists. This limit, when it exists, is called the Kullback–Leibler divergence $D(P_1\|P_2)$ between the two power spectral densities $P_1(e^{j\omega})$ and $P_2(e^{j\omega})$.

Theorem 1 [12, Theorem 10.5.1, 11, Proposition 8.29]. *Let $P_1(e^{j\omega})$ and $P_2(e^{j\omega})$ be two candidate PSDs for the zero-mean Gaussian WSS random process $x(n)$. If $P_1(e^{j\omega})$ and $P_2(e^{j\omega})$ are essentially bounded from below,* [1] *then the Kullback–Leibler divergence of $P_1(e^{j\omega})$ from $P_2(e^{j\omega})$ exists and is given by*

$$D(P_1\|P_2) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{P_1(e^{j\omega})}{P_2(e^{j\omega})} - \ln\frac{P_1(e^{j\omega})}{P_2(e^{j\omega})} - 1\right)d\omega. \quad (5)$$

Note that if $P_1(e^{j\omega})$ and $P_2(e^{j\omega})$ are rational with no poles or zeroes on the unit circle, then the conditions of the above theorem are satisfied and $D(P_1\|P_2)$ exists. The Kullback–Leibler divergence between PSDs has important distance-like properties which can be used to introduce certain geometries on the space of PSD functions. The reader is referred to the excellent monograph by Amari and Nagaoka [13, Chapter 5] for an interesting discussion of this topic.

2.3. Entropy rate, mutual information rate, and joint information rate of WSS random processes

Introductory expositions of the basic concepts of *entropy*, *mutual information* and *joint information* can be found in any standard text on information theory, e.g. [14–16]. For a more detailed exposition, including discussions on existence and interpretations, the reader is referred to [17]. The concept of entropy for random variables generalizes to the concept of *entropy rate* for WSS stochastic processes. Information theory of WSS stochastic processes has been studied in detail in Pinsker's classic work [12]. More recent books containing readable accounts on information rate include [16–18].

Let $x(n)$ be a Gaussian WSS random process with PSD $P_x(e^{j\omega})$. The entropy rate of this process is defined by

$$H(x) \triangleq \lim_{N\to\infty}\frac{1}{N}H(x(n), x(n+1), \ldots, x(n+N-1)).$$

It can be shown that the above limit exists and is given by [18, p. 568]

$$H(x) = \frac{1}{2}\ln 2\pi + \frac{1}{2} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\ln P_x(e^{j\omega})\,d\omega. \quad (6)$$

Since $H(\cdot)$ is a functional of P_x, we will use the notations $H(x)$ or $H(P_x)$ interchangeably for the entropy rate of a Gaussian WSS random process x. Now, let

$$S(n) = [x_0(n)\ x_1(n)\ldots x_{M-1}(n)]^{\mathrm{T}}$$

be an M-dimensional Gaussian jointly wide sense stationary (JWSS) random process. The entropy rate $H(S)$ of $S(n)$ is defined by a straightforward generalization of the definition used for the scalar case. One can show that [19, p. 45]

$$H(S) = \frac{M}{2}\ln 2\pi + \frac{M}{2} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\ln\det\mathbf{P}_S(e^{j\omega})\,d\omega, \quad (7)$$

where $\mathbf{P}_S(e^{j\omega})$ represents the *power spectral density matrix* corresponding to the vector process $S(n)$. [2] By definition, the quantity $H(S)$ is considered the *joint information rate* of the scalar processes $x_0(n), x_1(n), \ldots, x_{M-1}(n)$ that constitute $S(n)$. Finally, to define the *mutual information rate* $R(x_0; x_1; \ldots; x_{M-1})$ among scalar processes $x_0(n), \ldots, x_{M-1}(n)$, we consider them as constituting a vector process $S(n)$ and write

$$R(x_0; x_1; \ldots; x_{M-1}) \triangleq \sum_{i=0}^{M-1}H(x_i) - H(S). \quad (8)$$

It can be shown that mutual information rate is a non-negative quantity and is zero only when the processes $x_i(n)$ are mutually independent. For convenience, we may also use the short-hand notation $R(S)$ for $R(x_0; x_1; \ldots; x_{M-1})$ when it is clear that x_0 to x_{M-1} are components of the vector process $S(n)$.

3. Information gained by low-rate sensor observations

Assume that $x(n)$ (the desired but unavailable high-rate measurement signal in Fig. 1) is a zero-mean, Gaussian WSS process. We define "complete information" and "complete ignorance" as two extreme states of knowledge with regards to the statistics of $x(n)$. Naturally, we assign the information value zero to complete ignorance and reserve a positive value $I(x)$ for complete information.

Definition 1 (*Complete ignorance*). We say that the state of our knowledge regarding the statistics of $x(n)$ is "complete ignorance" if all we know about the statistics of $x(n)$ is its variance $\sigma_x^2 \triangleq E\{x(n)^2\}$. Furthermore, we assign the value zero to the quantity of information associated with this state of knowledge.

Without losing generality we assume that the (known) value for the high-rate signal's variance is one, i.e. $\sigma_x^2 = 1$.

Remark 1. The majority of measures of information (including entropy and the Kullback–Leibler divergence) are not invariant with respect to changes of scale. In other words, these measures depend on the unit of measurement used for the signal's amplitude. The

[1] This is to say, there is a positive constant ϵ such that $P_1(e^{j\omega})$ and $P_2(e^{j\omega})$ are both greater than ϵ for almost all ω.

[2] Power spectral density matrix is the generalization of the concept of power spectral density to multi-dimensional stochastic processes. See, for example, [19, Definition 3.5].

assumption that the input signal's variance is known is necessary to avoid the ambiguity caused by the scale-dependence of information measures.

Based on the principle of Maximum Entropy (which will be discussed later in Section 4.2), we assume, in the state of complete ignorance, that the power spectrum of $x(n)$ is a white (constant) spectrum with unit variance. In other words, in the state of complete ignorance we choose $P_x = \overline{P}_x$ where

$$\overline{P}_x(e^{j\omega}) \triangleq 1, \quad \omega \in [-\pi, \pi]. \tag{9}$$

Definition 2 (*Complete information*). We say that the state of our knowledge regarding the statistics of $x(n)$ is "complete information" if we know $P_x(e^{j\omega})$ exactly. Furthermore, the amount of information associated with this state of knowledge is given by

$$I(x) \triangleq D(P_x \| \overline{P}_x), \tag{10}$$

where $D(P_x \| \overline{P}_x)$ represents the Kullback–Leibler divergence of P_x from $\overline{P}_x$ and $\overline{P}_x$ is the white power spectrum defined in (9).

Recall from Section 2.3 that the entropy rate $H(x)$ of a Gaussian WSS random process $x(n)$ is given by

$$H(x) = \frac{1}{2}\ln 2\pi + \frac{1}{2} + \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln P_x(e^{j\omega})\,d\omega, \tag{11}$$

and that $D(P_2 \| P_1)$ is given by

$$D(P_2 \| P_1) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \left(\frac{P_2(e^{j\omega})}{P_1(e^{j\omega})} - \ln \frac{P_2(e^{j\omega})}{P_1(e^{j\omega})} - 1 \right) d\omega. \tag{12}$$

It follows from (9)–(12) that

$$H(x) + I(x) = c, \tag{13}$$

where $c \triangleq \frac{1}{2}\ln 2\pi + \frac{1}{2}$ is a constant. The statistical information $I(x)$ is always positive or zero while $H(x)$ can become negative or even, for a non-regular random process, $-\infty$. This is shown graphically in Fig. 2.

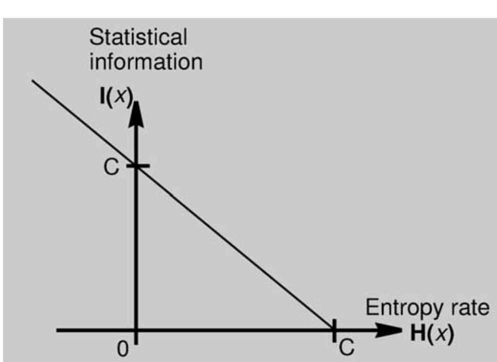

Fig. 2. Statistical information $I(x)$ and entropy rate $H(x)$ of a Gaussian WSS random process $x(n)$ add up to a constant if the variance of $x(n)$ is held fixed.

Now, let $\mathcal{Q}$ denote the set of PSDs with unit variance and let $\mathcal{S}^{(i;N)}$ denote the set of all PSDs which are consistent with N autocorrelation coefficients measured from the signals $v_0(n)$ to $v_i(n)$. In Section 4, we will develop a Maximum Entropy inference method to estimate $P_x(e^{j\omega})$ based on low-rate statistical experiments. By a statistical experiment on $v_i(n)$ we mean measuring its ACS $R_{v_i}(k)$ for $k \in \{0, 1, \ldots, N-1\}$. Let $P_x^{(i;N)}(e^{j\omega})$ denote the unique Maximum Entropy estimate of $P_x(e^{j\omega})$ which is in the constraint set $\mathcal{S}^{(i;N)} \cap \mathcal{Q}$.

Definition 3 (*Information content of low-rate measurements*). The statistical information contained in $v_0(n)$ to $v_i(n)$ about $x(n)$ is given by

$$I(v_0, \ldots, v_i) \triangleq \lim_{N\to\infty} D(P_x^{(i;N)} \| \overline{P}_x). \tag{14}$$

In words, $I(v_0, \ldots, v_i)$ measures the amount of statistical information about $x(n)$ that one can gain by performing statistical experiments on $v_0(n)$ to $v_i(n)$.

Lemma 1. *The limit used in Definition* 3 *exists.*

Proof. It follows from the definition of $\mathcal{S}^{(i;N)}$ that

$$(\mathcal{S}^{(i;N+1)} \cap \mathcal{Q}) \subset (\mathcal{S}^{(i;N)} \cap \mathcal{Q}). \tag{15}$$

Also, it is obvious that $P_x(e^{j\omega}) \in (\mathcal{S}^{(i;N)} \cap \mathcal{Q})$ for all values of i and N. Now, let $P_x^{(i;N)}(e^{j\omega})$ and $P_x^{(i;N+1)}(e^{j\omega})$ represent the Maximum Entropy estimates corresponding to the constraint sets $(\mathcal{S}^{(i;N)} \cap \mathcal{Q})$ and $(\mathcal{S}^{(i;N+1)} \cap \mathcal{Q})$, respectively. It follows from this fact and (15) that

$$H(P_x^{(i;N)}(e^{j\omega})) \geqslant H(P_x^{(i;N+1)}(e^{j\omega})) \geqslant H(P_x(e^{j\omega})). \tag{16}$$

Using (9), (11) and (12) we can then write

$$D(P_x^{(i;N)} \| \overline{P}_x) \leqslant D(P_x^{(i;N+1)} \| \overline{P}_x) \leqslant D(P_x \| \overline{P}_x). \tag{17}$$

The above relation shows that the sequence $D(P_x^{(i;N)} \| \overline{P}_x)$ is bounded from above and also non-decreasing in N. Thus, it has a limit. □

The above definition allows us to measure quantitatively the statistical information provided by one or a multitude of low-rate measurements. The following lemma shows that the information content of low-rate measurements increases (or in the worst case remains constant) as we include more channels. It also indicates that the information provided by multi-rate measurements cannot exceed that of the original signal.

Lemma 2. *If* $i \geqslant j$ *then*

$$I(v_0, \ldots, v_j) \leqslant I(v_0, \ldots, v_i) \leqslant I(x). \tag{18}$$

Proof. Using a procedure similar to the one used in the proof of the above lemma, one can show that

$$H(P_x^{(i;N)}(e^{j\omega})) \geqslant H(P_x^{(i+1;N)}(e^{j\omega})) \geqslant H(P_x(e^{j\omega})), \tag{19}$$

and then proceed to

$$D(P_x^{(i;N)}\|\overline{P}_x) \leqslant D(P_x^{(i+1;N)}\|\bar{P}_x) \leqslant D(P_x\|\overline{P}_x) \tag{20}$$

from which the asserted inequality follows in the limiting case when $N \to \infty$. □

It follows from the above lemma that, as we include more channels in our multi-rate measurement system, we can expect to gain more information. However, the total amount of information will eventually saturate to a limiting value below $I(x)$ which represents the "complete information" needed to specify statistics of $x(n)$. This is shown graphically in Fig. 3.

One should note, nonetheless, that the saturation of $I(v_0,\dots,v_i)$ as i increases does not mean that the new channels contribute a small amount of information. It means that the new channels contribute a small amount of *additional* information. In other words, the information provided by the new channels becomes more and more redundant as the number of channels increases. This observation motivates defining a quantitative measure for the *redundancy* of the statistical information provided by a multitude of low-rate signals.

To define redundancy, we first need to define the information content of a single low-rate signal $v_i(n)$. This can be done similar to what we did for multiple measurements before. Namely, we let $\mathscr{Q}$, as before, denote the set of PSDs with unit variance. We use the notation $\widetilde{\mathscr{S}}^{(i;N)}$ to denote the set of all PSDs which are consistent with N ACS values measured from the signal $v_i(n)$. We then define $\widetilde{P}_x^{(i;N)}(e^{j\omega})$ as the estimate of $P_x(e^{j\omega})$ provided by the MEIE from the constraint set $\widetilde{\mathscr{S}}^{(i;N)} \cap \mathscr{Q}$. The definition of the information content of the ith channel—independent of any other channels—is then made as follows.

Definition 4 (*Information content of a single measurement signal*). The statistical information contained in the signal $v_i(n)$ about $x(n)$ is given by

$$I(v_i) \triangleq \lim_{N\to\infty} D\left(\widetilde{P}_x^{(i;N)}\|\overline{P}_x\right). \tag{21}$$

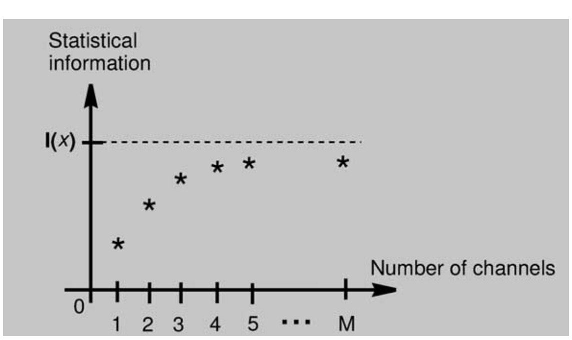

Fig. 3. A typical plot depicting the amount of statistical information $I(v_0,\dots,v_i)$ as the number of channels (i.e. $i+1$) is increased.

By using the above definition, the definition of redundancy follows naturally.

Definition 5 (*Redundancy*). The redundancy $R(v_0,\dots,v_i)$ of the statistical information contained in $v_0(n)$ to $v_i(n)$ about $x(n)$ is given by

$$R(v_0,\dots,v_i) \triangleq \sum_{k=0}^{i} I(v_k) - I(v_0,\dots,v_i). \tag{22}$$

The reader should easily notice the similarity between the definition of redundancy here and the classical definition of "mutual information rate" in information theory (Section 2.3).

4. Spectrum estimation using low-resolution sensor data

In this section we discuss in some detail the problem of estimating the power spectral density of $x(n)$ using statistics of the low-rate observable signals $v_i(n)$. Our analysis leads to an algorithm called the *Maximum Entropy Inference Engine*, or MEIE for short. MEIE accepts a finite number of autocorrelation coefficients (measured or somehow known) for $v_i(n)$ and produces a *unique* estimate $P_x^{(i;N)}(e^{j\omega})$ for the PSD of $x(n)$. The estimates provided by the MEIE are then used in the formulae developed in the previous section to measure the amount of information provided by the low-rate measurements $v_i(n)$.

4.1. Formulating the estimation problem

Consider the setup shown in Fig. 1(c) and recall that $x(n)$ is assumed to be a WSS random process. It is straightforward to show that the low-resolution observations $v_0(n)$ to $v_{M-1}(n)$ in Fig. 1(c) are also WSS processes [20]. The ACS values for these signals are given by

$$R_{v_i}(k) = R_{x_i}(N_i k), \tag{23}$$

where

$$R_{x_i}(k) \triangleq (h_i(k) \star h_i(-k)) \star R_x(k) \tag{24}$$

and $h_i(k)$ denotes the impulse response of $H_i(z)$. One can express $R_{v_i}(k)$ as a function of $P_x(e^{j\omega})$ as well. Let us define $G_i(z) \triangleq H_i(z)H_i(z^{-1})$. Now, (24) can be written as

$$R_{x_i}(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_x(e^{j\omega})G_i(e^{j\omega})e^{jk\omega}\,d\omega. \tag{25}$$

From (23) and (25), we can then write

$$R_{v_i}(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_x(e^{j\omega})G_i(e^{j\omega})e^{jN_i k\omega}\,d\omega. \tag{26}$$

Similar to Section 3, we use the notation $\widetilde{\mathscr{S}}^{(i,N)}$ to denote the set of all input PSDs which are consistent with the N autocorrelation values obtained from the signal $v_i(n)$. More precisely, we define

$$\widetilde{\mathscr{S}}^{(i,N)} \triangleq \left\{ P_x(e^{j\omega}) : \begin{array}{l} \frac{1}{2\pi}\int_{-\pi}^{\pi} P_x(e^{j\omega})G_i(e^{j\omega})e^{jN_ik\omega}d\omega \\ \quad = R_{v_i}(k),\ 0 \leqslant k \leqslant N-1; \\ P_x(e^{j\omega}) \in \mathbf{L}^1(-\pi,\pi); \\ P_x(e^{j\omega}) \geqslant 0 \end{array} \right\}. \tag{27}$$

We define $\mathscr{S}^{(i,N)}$ as in Section 3 as well. That is, $\mathscr{S}^{(i,N)}$ is the set of all input PSDs consistent with autocorrelation values known for *multiple* low-rate signals $v_0(n)$ to $v_i(n)$:

$$\mathscr{S}^{(i,N)} \triangleq \bigcap_{j=0}^{i} \widetilde{\mathscr{S}}^{(j,N)}. \tag{28}$$

Now consider the following problem.

Problem 1. Estimate $P_x(e^{j\omega})$ given that $P_x(e^{j\omega}) \in \mathscr{S}^{(i,N)}$. In other words, estimate $P_x(e^{j\omega})$ knowing a finite number of the autocorrelation coefficients $R_{v_i}(k)$ associated with the low-rate measurements $v_i(n)$.

In general, knowing $R_{v_i}(k)$ for some limited values of i and k is not sufficient for characterizing $P_x(e^{j\omega})$ uniquely. That is, given a finite set of ACS values $R_{v_i}(k)$ there usually exist infinitely many $P_x(e^{j\omega})$ which can generate those values. Thus, the spectrum estimation problem mentioned-above is ill-posed.

4.2. The maximum entropy principle

The Maximum Entropy (ME) principle, introduced in 1957 by Jaynes in the field of thermodynamics, provides an elegant way for dealing with a wide class of statistical ill-posed problems [21,22]. According to this principle, one should choose *the most random* solution that satisfies the known constraints. For the case of Problem 1, the ME principle asserts that $P_x(e^{j\omega})$ should be chosen such that (i) it is *consistent* with the known ACS values $R_{v_i}(k)$, and (ii) it associates with the *most random* signal $x(n)$.

Applying the ME principal to spectral estimation is credited to Burg [23] who used Jaynes' ME principle to estimate PSD of random signals given a limited number of their autocorrelation coefficients. Here, we will use the ME principle in a way similar to the way Burg did. However, Problem 1 is more challenging than the standard autocorrelation extrapolation problem considered by him.

Recall from Section 2.3 that the notion of entropy of a scalar random variable translates to the notion of entropy rate for WSS random signals. Recall, also, that $x(n)$ is assumed to be Gaussian. Thus, the entropy rate of $x(n)$ is given by

$$H(P_x) = \frac{1}{2}\ln 2\pi + \frac{1}{2} + \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln P_x(e^{j\omega})\,d\omega. \tag{29}$$

Now, according to the ME principle, we may seek a solution for Problem 1 by solving the following *constrained optimization* problem.

Problem 2 (*ME interpretation of Problem 1*). Find $P_x^{(i,N)}(e^{j\omega}) = \arg\max H(P_x)$ subject to $P_x \in \mathscr{S}^{(i,N)}$.

4.3. Solving the estimation problem

It has been shown in [8] that the solution to Problem 2, when it exists, is given by

$$P_x^{(i,N)}(e^{j\omega}) = \frac{1}{\sum_{j=0}^{i-1} G_j(e^{j\omega})F_j(e^{jN_j\omega})}, \tag{30}$$

where $F_j(z) \triangleq \sum_{k=-(N-1)}^{N-1} 2\lambda_{jk}z^{-k}$. The coefficients λ_{jk} of the transfer functions $F_j(z)$ in (30) are specified such that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{G_l(e^{j\omega})e^{jN_lk\omega}}{\sum_{j=0}^{i-1} G_j(e^{j\omega})F_j(e^{jN_j\omega})}\,d\omega = R_{v_l}(k) \tag{31}$$

holds for all values of $l = 0, 1, \ldots, i-1$ and $k = 0, 1, \ldots, N-1$.

Remark 2. The notation N denoting the number of ACS values measured per signal should not be confused with N_j which denotes the down-sampling ratio of each channel. Also, j as an index should not be confused with j in $e^{j\omega}$ which means $\sqrt{-1}$.

In general, maximizing the entropy functional (29) subject to equality constraints is a "partially-finite programing problem" in $\mathbf{L}^1$. The reader is referred to [24] for an analysis of the problem of existence of a ME solution in the general case. Here, we provide specific existence results that cover the particular case stated in Problem 2.

A set of given ACS values $\{R_{v_j}(k) \mid j = 0, 1, \ldots, i-1; k = 0, 1, \ldots, N-1\}$ associated with low-rate signals $v_0(n), v_1(n), \ldots, v_i(n)$ is called *admissible* if the set $\mathscr{S}^{(i,N)}$ (specified by these values through (27) and (28)) is non-empty. Using this definition, a sufficient condition for the existence of ME solution can be given as follows.

Lemma 3 [8]. *A solution of the form* (30) *exists for Problem* 2 *if*

(i) *The set of given ACS values is admissible*;
(ii) *the transfer functions* $G_i(e^{j\omega})$ *are bounded for* $\omega \in (-\pi, \pi)$.

Lemma 4 [8]. *When a solution of the form* (30) *exists*, *it is unique*.

4.4. Practical and computational issues

A practical computational algorithm which calculates the ME solution given by (30) and (31) is provided in [8].

This algorithm, called the Maximum Entropy Inference Engine (MEIE), accepts $H_j(z)$, N_j and $R_{v_j}(k)$ as the input, and produces an estimate $P_x^{(i;N)}(e^{j\omega})$ of the form (30) which satisfies (31) in the least-mean-square sense (Fig. 4).

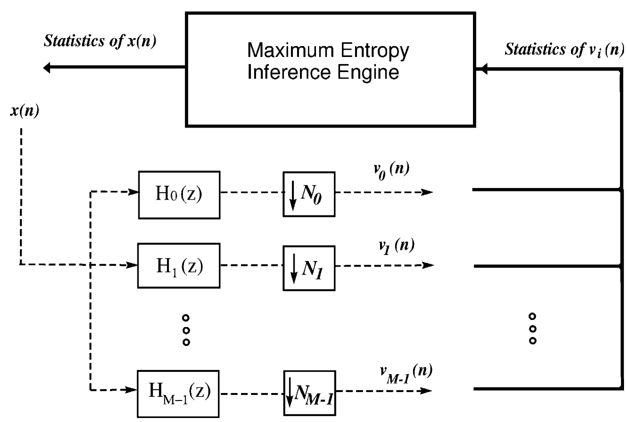

Fig. 4. MEIE is an algorithm that calculates an approximation to the ME estimate of $P_x(e^{j\omega})$ given autocorrelation coefficients of the low-rate observations.

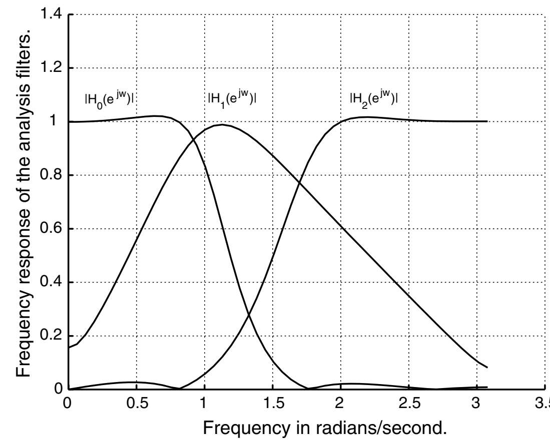

Fig. 5. Frequency response of the equivalent digital filters used in the simulation example.

It is not possible to feed an infinite number of ACS values to the MEIE. Thus, it is not possible, in practice, to calculate $P_x^{(i;N)}(e^{j\omega})$ or $\widetilde{P}_x^{(i;N)}(e^{j\omega})$ when $N \to \infty$. For this reason, we have to limit ourselves to a finite number of ACS values per channel and use, say, $P_x^{(i;4)}(e^{j\omega})$ as an

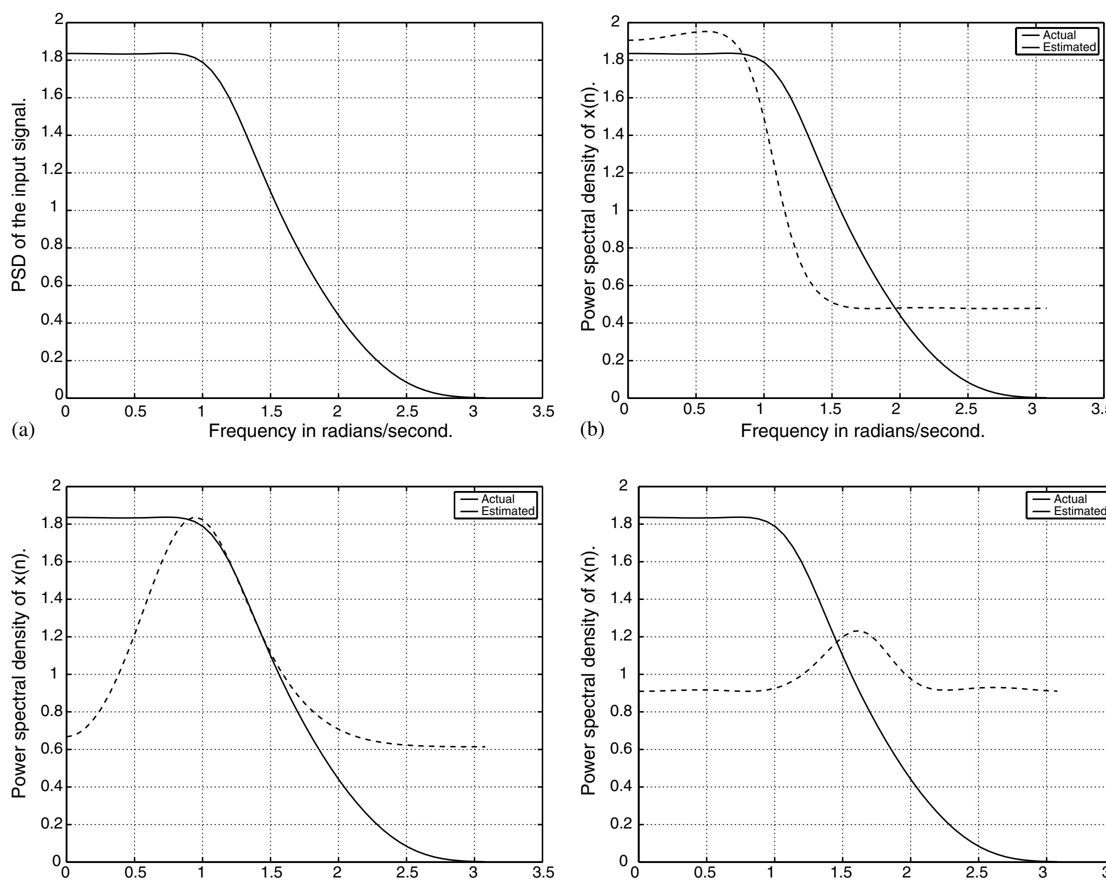

Fig. 6. (a) Actual power spectral density of the non-observable reference signal $x(n)$. The plots (b)–(d) show the PSD estimates $\widetilde{P}_x^{(i;4)}(e^{j\omega})$ obtained using N-MEIE for individual channels of the 3-channel low-resolution measurement system. The plot in (b) shows the estimate obtained using Channel 1; (c) shows the result obtained using Channel 2 and (d) shows the result obtained using Channel 3.

approximation to $P_x^{(i;\infty)}(e^{j\omega})$ in our information formulae.

Another practical difficulty in calculating the information formulae of Section 3 is that we have to force the MEIE to output a PSD with unit variance. That is, we have to make sure the solutions are in $\mathscr{S}^{(i,N)} \cap \mathscr{Q}$. This can be done by modifying the basic MEIE algorithm such that it considers the input variance as an additional constraint. The MEIE, however, cannot satisfy the constraints exactly which means that the solution provided by it is *approximately* a unit variance solution. To make the variance of the solutions as close as possible to one, we programmed a modified MEIE which emphasizes the variance constraint 10 times more than the constraints imposed by the measured ACS values. We call this modified algorithm the Normalized Maximum Entropy Inference Engine, or N-MEIE for short.

5. A simulation example

Consider a 3-channel multi-rate measurement system in a structure similar to the one shown in Fig. 1(c). Assume that the down-sampling rate is equal to four for all channels ($N_0 = N_1 = N_2 = 4$). Thus, each sensor works at only 25% of the sampling rate required to measure the signal $x(n)$ perfectly. Assume, also, that the equivalent digital filters $H_0(z)$, $H_1(z)$ and $H_2(z)$ are given as follows:

$$H_0(z) = \frac{0.0753 + 0.1656z^{-1} + 0.2053z^{-2} + 0.1659z^{-3} + 0.0751z^{-4}}{1.0000 - 0.8877z^{-1} + 0.6738z^{-2} - 0.1206z^{-3} + 0.0225z^{-4}},$$

$$H_1(z) = \frac{0.4652 - 0.1254z^{-1} - 0.3151z^{-2} + 0.0975z^{-3} - 0.0259z^{-4}}{1.0000 - 0.6855z^{-1} + 0.3297z^{-2} - 0.0309z^{-3} + 0.0032z^{-4}},$$

$$H_2(z) = \frac{0.1931 - 0.4226z^{-1} + 0.3668z^{-2} - 0.0974z^{-3} - 0.0405z^{-4}}{1.0000 + 0.2814z^{-1} + 0.3739z^{-2} + 0.0345z - 3 - 0.0196z^{-4}}.$$

The frequency response $|H_i(e^{j\omega})|$ for these filters is shown in Fig. 5. The reference signal $x(n)$ is chosen to be a Gaussian WSS process whose PSD $P_x(e^{j\omega})$ is shown in Fig. 6(a). In this example, we use $\widetilde{P}_x^{(i;4)}(e^{j\omega})$ and $P_x^{(i;4)}(e^{j\omega})$ as an approximation to $\widetilde{P}_x^{(i;\infty)}(e^{j\omega})$ and

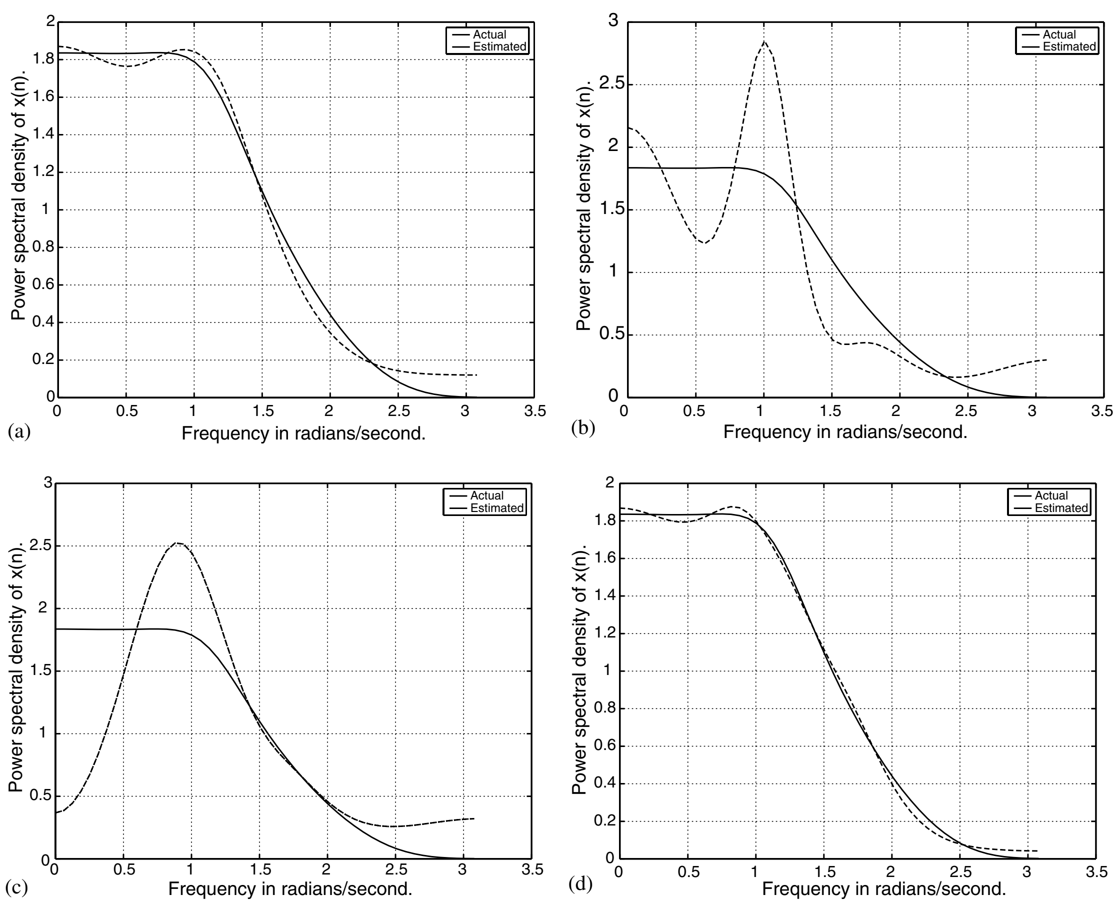

Fig. 7. Power spectral densities $P_x^{(i;4)}(e^{j\omega})$ estimated by N-MEIE by combining the low-resolution data provided by individual channels of the measurement system: (a) Channels 1 and 2, (b) Channels 1 and 3, (c) Channels 2 and 3 and (d) Channels 1, 2 and 3. Note that the scale in plots (b) and (c) has changed to accommodate the overshoots that exist in the estimated PSDs.

$P_x^{(i;\infty)}(e^{j\omega})$ in the information formulae (14), (21) and (22). The PSDs estimated by the N-MEIE using the data provided by each channel of the analysis system described above are shown in Fig. 6. Fig. 7 shows the PSDs estimated by the N-MEIE using the channels of this analysis system combined together. Given these estimates, one can easily calculate the quantity of information provided by each channel and the quantity of information provided by various combinations of the channels. These values, along with the value $I(x)$ corresponding to complete statistical information about $x(n)$ are plotted in Fig. 8. As can be seen in these plots, the collective quantity of information provided the measurement system increases as we include new channels but there exists a considerable amount of redundancy as well (Fig. 8(d)).

6. Concluding remarks

In this paper, we introduced a quantitative measure for the *statistical information* provided by multi-rate sensor arrays. Basically, we identified statistical information with the reduction in entropy of the unique PSD estimates obtained for $x(n)$ using the ME principle. It is illuminating, at this point, to recall the classical notion of "mutual information" and see how it compares to our definition of statistical information.

Let $X \in \mathbb{R}^N$ be a continuous random variable with PDF $p_X(X)$ and let $Y \in \mathbb{R}^M$, $M < N$, be a non-invertible function of X so that the conditional density function $p_{X\,|\,Y}(X\,|\,Y)$ exists. Then, the (average) information contained in Y about X is defined as the mutual information between X and Y, which, in turn, can be written as the difference between the entropy of X and the conditional entropy of X given Y [25, Eq. 2.2.17]. In other words, in classical information theory the "information contained in Y about X" is quantified by $R(X;Y) = H(X) - H(X\,|\,Y)$.

It is easy, using Eqs. (9)–(13), to show that $\lim_{N\to\infty} D(P_x^{(i;N)} \| \overline{P}_x)$ used in Definition 3 can be written as $H(\overline{P}_x) - \lim_{N\to\infty} H(P_x^{(i;N)})$. Thus, $I(v_0,\ldots,v_i)$ is similar to mutual information in the sense that it too measures the reduction in entropy of a random variable after a related variable is observed.

There is, however, an important conceptual difference between mutual information and the quantity measured by $I(v_0,\ldots,v_i)$: The PDFs used in the classical formula

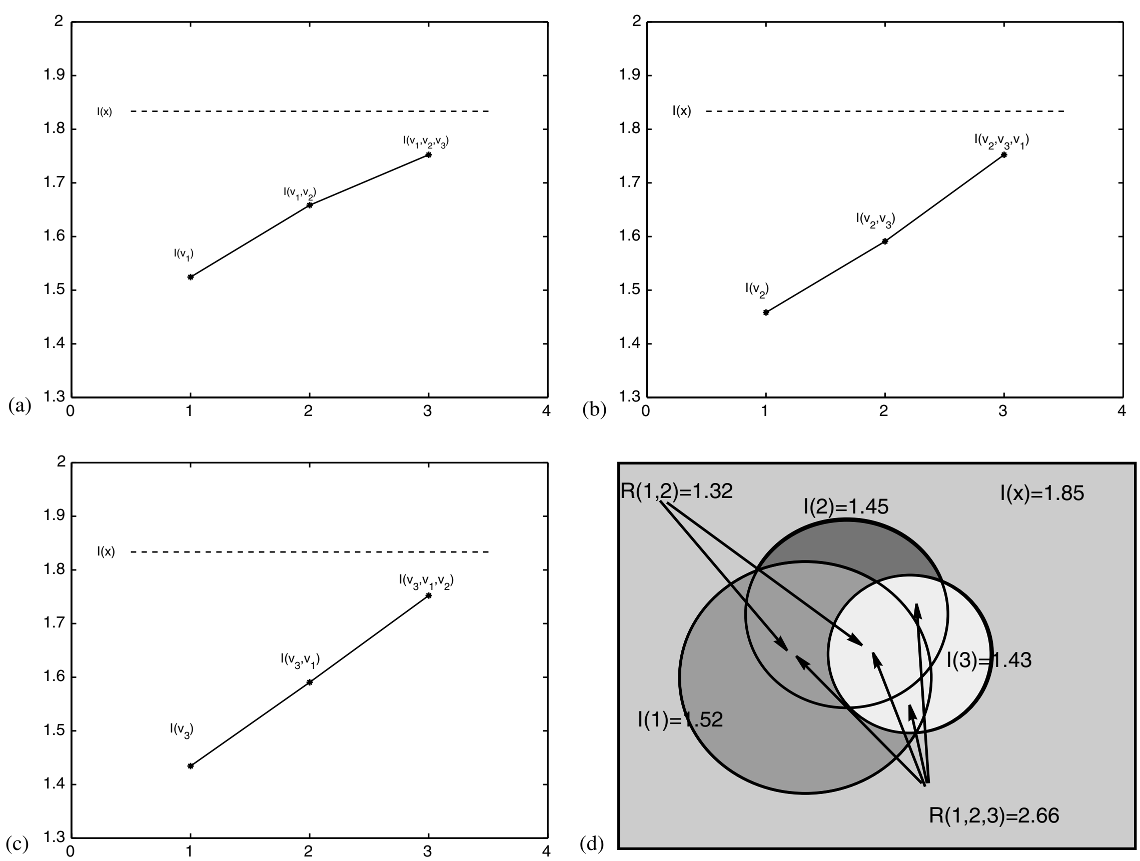

Fig. 8. Statistical information contained in certain combinations of the channels: (a) Channels 1, 1, 2 and 1, 2, 3; (b) Channels 2, 2, 3 and 2, 3, 1; (c) Channels 3, 3, 1 and 3, 1, 2 and (d) a symbolic (Venn) diagram showing the quantity of information provided by each channel and the redundancy among the channels. The figures in this diagram are approximate. Furthermore, it is deliberately drawn out of scale for clarity.

$H(X) - H(X \mid Y)$ are "actual" while the PSDs used in Definition 3 are "inferred". Recall that there is no deductive way to specify these PSDs even after doing statistical experiments on the low-rate signals.

Finally, we point out some aspects of our definition of statistical information that deem further attention and may be investigated in future research.

1. (*Cross-correlations are ignored.*) The PSD estimates $P_x^{(i;N)}(e^{j\omega})$ provided by the MEIE are based on the autocorrelation coefficients of the low-rate signals. Thus, any potential information contained in the cross-correlation among the low-rate measurements is ignored.
2. (*Entropy rate vs. statistical information.*) The notion of "statistical information" defined in this paper is not the same as "entropy rate" of $v_i(n)$. This is because $I(v_0, \ldots, v_i)$ depends both on the signals' statistics and on the characteristics of the measurement system, but this is not the case for $H(v_0, \ldots, v_i)$. Nonetheless, there are certainly connections between the mutual information rate of the low-rate signals $v_i(n)$ and their "redundancy" $R(v_0, \ldots, v_i)$ defined in Section 3. Investigating relations of this type is an open topic.

References

[1] P. Daukantas, Pervasive computing: the walls are listening, Washington Technology 16 (21) (2002).
[2] A. Pentland, Smart rooms, Scientific American 274 (4) (1996) 68–76.
[3] J. DiBiase, H. Silverman, M. Brandstein, Robust localization in reverberant rooms, in: M.S. Brandstein, D.B. Ward (Eds.), Microphone Arrays: Signal Processing Techniques and Applications, 2001.
[4] P. Aarabi, The fusion of distributed microphone arrays for sound localization, EURASIP Journal of Applied Signal Processing (Special Issue on Sensor Networks) 2003 (4) (2003) 338–347.
[5] B.V. Dasarathy, Elucidative fusion systems—an exposition, Information Fusion 1 (1) (2000) 5–15.
[6] B.V. Dasarathy, Adaptive fusion processor paradigms for fusion of information acquired at different levels of detail, Optical Engineering 35 (3) (1996) 634–639.
[7] O.S. Jahromi, P. Aarabi, Theory and design of multirate sensor arrays, IEEE Transactions on Signal Processing (submitted).
[8] O.S. Jahromi, B.A. Francis, R.H. Kwong, Spectral estimation using multirate measurements, IEEE Transactions on Signal Processing 52 (7) (July 2004).
[9] S. Kullback, R.A. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 79–86.
[10] S. Kullback, Information Theory and Statistics, John Wiley and Sons, New York, NY, 1954.
[11] I. Vajda, Theory of Statistical Inference and Information, Kluwer Academic Publishers, Boston, 1989.
[12] M.S. Pinsker, Information and Information Stability of Random Variables and Processes, Holden-Day, San Francisco, 1964.
[13] S.I. Amari, H. Nagaoka, Methods of Information Geometry, American Mathematical Society/Oxford University Press, 2000.
[14] F.M. Reza, An Introduction to Information Theory, McGraw-Hill, New York, 1961.
[15] I. Csiszár, J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, Akadémiai Kiadó, Budapest, 1981.
[16] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley Interscience, New York, 1991.
[17] Robert M. Gray, Entropy and Information Theory, Springer-Verlag, New York, 1990.
[18] A. Papoulis, Probability, Random Variables and Stochastic Processes, third ed., McGraw-Hill, New York, 1991.
[19] P.E. Caines, Linear Stochastic Systems, John Wiley, New York, 1988.
[20] V. Sathe, P.P. Vaidyanathan, Effect of multirate systems on the statistical properties of random signals, IEEE Transactions on Acoustics, Speech and Signal Processing 41 (1) (1993) 131–146.
[21] E.T. Jaynes, On the rationale of maximum entropy methods, Proceedings of the IEEE 70 (1982) 939–952.
[22] E.T. Jaynes, in: R.D. Rosenkrantz (Ed.), Papers on Probability, Statistics and Statistical Physics, Reidel, Dordrecht, 1983.
[23] J.P. Burg, Maximum entropy spectral analysis, in: 37th Annual International Meeting of Society of Exploration Geophysicists, Oklahoma City, Oklahoma, 1967.
[24] J.M. Browein, A.S. Lewis, Partially-finite programming in $\mathbf{L}_1$ and the existense of maximum entropy estimates, SIAM Journal of Optimization 3 (1993) 248–267.
[25] R.G. Gallager, Information Theory and Reliable Communication, John Wiley, New York, 1968.