Nonextensive maximum-entropy-based formalism for data subset selection

L. Rebollo-Neira
NCRG, Aston University, Birmingham B4 7ET, United Kingdom

A. Plastino
Instituto de Física La Plata (IFLP), Universidad Nacional de La Plata and CONICET CC 727, 1900 La Plata, Argentina

(Received 2 July 2001; revised manuscript received 2 October 2001; published 20 December 2001)

A method for data subset selection, which is based on the $q=\frac{1}{2}$ maximum information measure formalism, is proposed. The method evolves iteratively by selecting, at each iteration, the measure yielding a $q=\frac{1}{2}$ distribution capable of making predictions minimizing the Euclidean distance to the available data.

DOI: 10.1103/PhysRevE.65.011113 PACS number(s): 05.20.−y, 02.50.Tt, 02.30.Zz, 07.05.Kf

I. INTRODUCTION

We have shown in previous efforts on the maximum-entropy-based (MaxEnt) [1–3] that, in situations in which a density distribution is to be determined from measurements collected as a function of a variable parameter, *only a subset* of them is normally *relevant* to be employed for constraining the corresponding optimization process. This is true, of course, in the absence of noise (random errors). Otherwise, redundancy does yield the desired effect of reducing noise.

In [1,2] a MaxEnt formalism is advanced that selects relevant data from an available set. Such methodology, however, is marred by the limitation of assuming the selected data to be noiseless, in the sense that the resultant distribution is forced to *exactly* account for them. Other MaxEnt methods [4–10], which are not affected by this limitation, fail to provide information as to just *which* the relevant data are.

This paper aims at achieving the best of both worlds. On the one hand, we wish to use all the available data so as to determine the density distribution. On the other hand, we wish to be in a position to identify a subset of relevant data. The framework we are going to propose for achieving such a goal is based on the nonextensive information q measure advanced by Tsallis [11–18].

We consider the particular instance $q=\frac{1}{2}$. Such a case gives rise to a generalized $p^{1/2}$ distribution, which has been analyzed in [19]. The physical significance of this distribution is illustrated in Boghosian's work [20], while some mathematical applications are reported in [21,22]. Here we use the $p^{1/2}$ distribution in order to construct a sound mathematical scheme for data subset selection.

The paper is organized as follows. In Sec. II we introduce the notation, together with some considerations on the nonextensive maximum information measure distribution for the case of interest, i.e., $q=\frac{1}{2}$. In Sec. III we discuss the estimation of such a distribution from a given set of measurements and a (assumed to be known) relevant subset of them. In Sec. IV a selection criterion and an iterative algorithm for determining such a subset of relevant data is proposed. Some conclusions are drawn in Sec. V.

II. PRELIMINARY CONSIDERATIONS

Consider that we are given the M pieces of data $f_1^o, f_2^o, \ldots, f_i^o, \ldots f_M^o$, each of which is the expectation value (EV) of a random variable that, for a suitable set of N states labeled as $n=1,\ldots,N$, takes the values $f_{i,n}$; $n=1,\ldots,N$. The EVs are computed using a generalized distribution $p_n^{1/2}$; $n=1,\ldots,N$. Thus, the data model is expressed in terms of M equations of the form

$$f_i^o=\sum_{n=1}^{N} p_n^{1/2} f_{i,n}, \quad i=1,\ldots,M, \tag{1}$$

that, adopting a Dirac's vectorial notation, are recast as

$$|f^o\rangle=\hat{A}|p^{1/2}\rangle, \tag{2}$$

where $|p^{1/2}\rangle$ is represented in terms of the *standard basis* $|n\rangle$ $n=1,\ldots,N$ of $\mathcal{R}^N$

$$|p^{1/2}\rangle=\sum_{n=1}^{N}|n\rangle\langle n|p^{1/2}\rangle=\sum_{n=1}^{N} p_n^{1/2}|n\rangle, \tag{3}$$

while the data vector $|f^o\rangle$ is represented in terms of the standard basis $|i\rangle$; $i=1,\ldots,M$ of $\mathcal{R}^M$

$$|f^o\rangle=\sum_{i=1}^{M}|i\rangle\langle i|f^o\rangle=\sum_{i=1}^{M} f_i^o|i\rangle. \tag{4}$$

The operator $\hat{A}:\mathcal{R}^N\to\mathcal{R}^M$ is given by the matrix elements $\langle i|\hat{A}|n\rangle=f_{i,n}$, $i=1,\ldots,M$, $n=1,\ldots N$. Thus, by defining vectors $|f_n\rangle\in\mathcal{R}^M$ in such a way that $\langle i|f_n\rangle=f_{i,n}$ the operator $\hat{A}$ is expressed as

$$\hat{A}=\sum_{n=1}^{N}|f_n\rangle\langle n|. \tag{5}$$

It is shown in [21] that using this notation the nonextensive MaxEnt q distribution is of the form

$$p_j^q=z[1-(1-q)\langle j|\hat{A}^\dagger|\lambda\rangle]^{q/(1-q)}; \quad (j=1,\ldots,N),$$
$$q\in\mathcal{R}, \tag{6}$$

where $\hat{A}^\dagger$ stands for the adjoint of $\hat{A}$, z is a normalization constant, and $|\lambda\rangle$ is a vector in $\mathcal{R}^M$ whose entries are the Lagrange multipliers λ_i, $i=1,\ldots,M$ accounting for M

given constraints. For the particular value of q we are here considering, i.e., $q=\frac{1}{2}$, the corresponding distribution can be recast in the form [19,21]

$$|p^{1/2}\rangle=|z\rangle+\hat{A}^{\dagger}|\lambda\rangle, \tag{7}$$

where $|z\rangle=\Sigma_{n=1}^{N}\langle n|z\rangle|n\rangle=\Sigma_{n=1}^{N}z|n\rangle$ is the vectorial representation of the normalization constant z. As discussed in [19], if our information consists of independent pieces of data then rank$(\hat{A})=M$, the operator $\hat{A}\hat{A}^{\dagger}$ has an inverse, and the unique Lagrange multiplier vector (which determines $|p^{1/2}\rangle$) is obtained from Eq. (2) as

$$|\lambda\rangle=(\hat{A}\hat{A}^{\dagger})^{-1}|f^{o}\rangle-(\hat{A}\hat{A}^{\dagger})^{-1}\hat{A}|z\rangle. \tag{8}$$

On the other hand, if the pieces of data we have gathered are not independent, then rank$(\hat{A})<M$, the operator $\hat{A}\hat{A}^{\dagger}$ has no inverse, and the vector $|\lambda\rangle$ is not unique. However, as discussed in [21], one can still use the pseudoinverse of the operator $\hat{A}\hat{A}^{\dagger}$ in order to obtain an appropriate vector $|\lambda\rangle$, without affecting the uniqueness of the $|p^{1/2}\rangle$ distribution. Proceeding in such a way, however, we are unable to discern just which of the M data equations of our model contain relevant information. At this point it is necessary to specify the precise meaning that we would like the term "relevant" to be endowed with in the present context.

Definition. Given a set of M empirical expectation values [and the M associated equations of the form (1)], we refer to a subset of $K\leqslant M$ of these equations as being relevant, if the K equations provide us with independent constraints that give rise to a $|p^{1/2}\rangle$ distribution able to correctly predict the remaining (available) $(M-K)$ data.

The scheme outlined above associates to each equation of the system given in Eqs. (1) and (2): (i) a particular subindex $(1\leqslant i\leqslant M)$, (ii) a particular row belonging to the matrix representation of the operator $\hat{A}$ (i.e., a particular component of the vectors $|f_n\rangle$), (iii) the corresponding component of $|f^{o}\rangle$, and (iv) a Lagrange multiplier. Nevertheless, if $M-K$ of those equations are not really relevant as constraints, one can regard them as giving rise to components of the vector $|\lambda\rangle$ that have a null value. The central idea of the theoretical framework to be advanced here is that of appropriately using such a fact. We do so by recourse to the construction of an sparse Lagrange multiplier vector, whose nonzero entries identify a subset of relevant data.

III. ESTIMATING THE LAGRANGE MULTIPLIERS FROM NOISY DATA

Let us suppose that we are able to identify K relevant equations and let us relabel the corresponding subindexes as l_k; $k=1,\ldots,K$. Since, by hypothesis, $\langle i|\lambda\rangle=0$ for $i\neq l_k$, Eq. (7), that yields the $q=\frac{1}{2}$ distribution, becomes

$$|p^{1/2}\rangle=|z\rangle+\sum_{k=1}^{K}\hat{A}^{\dagger}|l_k\rangle\langle l_k|\lambda\rangle. \tag{9}$$

The constant z, which determines $|z\rangle$, is fixed by normalization. Here we will adopt the criterion advanced in Ref. [20], and require that $\Sigma_{n=1}p_n^{1/2}=1$. Consequently, one easily obtains

$$\begin{aligned} z&=\frac{1}{N}-\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle n|\hat{A}^{\dagger}|l_k\rangle\langle l_k|\lambda\rangle\\ &=\frac{1}{N}-\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle f_n|l_k\rangle\langle l_k|\lambda\rangle, \end{aligned} \tag{10}$$

so that, by introducing the vector

$$|g\rangle=\sum_{n=1}^{N}|f_n\rangle\equiv\sum_{n=1}^{N}\hat{A}|n\rangle \tag{11}$$

we can write

$$z=\frac{1}{N}-\frac{1}{N}\sum_{k=1}^{K}\langle g|l_k\rangle\langle l_k|\lambda\rangle. \tag{12}$$

Hence,

$$\begin{aligned} |p^{1/2}\rangle=&\left(\frac{1}{N}-\frac{1}{N}\sum_{k=1}^{K}\langle g|l_k\rangle\langle l_k|\lambda\rangle\right)\sum_{n=1}^{N}|n\rangle\\ &+\sum_{k=1}^{K}\hat{A}^{\dagger}|l_k\rangle\langle l_k|\lambda\rangle. \end{aligned} \tag{13}$$

If the K pieces of data $f_{l_k}^{o}$, $k=1,\ldots K$, that we are considering were to be known without uncertainty, we could use them to straightforwardly determine the ,K Lagrange multipliers $\langle l_k|\lambda\rangle$, $k=1,\ldots k$ from the corresponding ,K equations, as explained in Sec. II. Moreover, since we are working under the hypothesis that the remaining equations of the original system (1) are irrelevant, we would be in a position to accurately predict the complete data vector in the fashion

$$\begin{aligned} |f^{o}\rangle\equiv|f^{p}\rangle&=\hat{A}|p^{1/2}\rangle\\ &=\frac{|g\rangle}{N}-\frac{1}{N}\sum_{k=1}^{K}|g\rangle\langle g|l_k\rangle\langle l_k|\lambda\rangle\\ &\quad+\sum_{k=1}^{K}\hat{A}\hat{A}^{\dagger}|l_k\rangle\langle l_k|\lambda\rangle\\ &=\frac{|g\rangle}{N}-\frac{1}{N}\sum_{k=1}^{K}|g\rangle\langle g|l_k\rangle\langle l_k|\lambda\rangle\\ &\quad+\sum_{n=1}^{N}\sum_{k=1}^{K}|f_n\rangle\langle f_n|l_k\rangle\langle l_k|\lambda\rangle. \end{aligned} \tag{14}$$

Unfortunately, data are never known without some uncertainty and, therefore, the prediction $|f^{p}\rangle$ will match the real data $|f^{o}\rangle$ only up to some error. Hence, by determining the Lagrange multipliers using only the K relevant equations we would introduce a bias, as a consequence of trying to reproduce a reduced set of data with precision higher than that of the remaining available data. We should adopt then an alternative criterion in order to determine $\langle l_k|\lambda\rangle$, $k=1,\ldots,K$. An equitable one would be to fix these figures as the values yielding a (predicted) data vector $|f^p\rangle$ such that one minimizes the distance to the observed vector $|f^o\rangle\in\mathcal{R}^M$. In order to discuss how this can be achieved, let us introduce the operator $\hat{F}:\mathcal{R}^K\to\mathcal{R}^M$, as given by

$$\hat{F}=\sum_{k=1}^{K}|\alpha_{l_k}\rangle\langle l_k|, \tag{15}$$

where

$$|\alpha_{l_k}\rangle=\sum_{n=1}^{N}|f_n\rangle\langle f_n|l_k\rangle-\frac{1}{N}|g\rangle\langle g|l_k\rangle. \tag{16}$$

Thus, we can express $|f^p\rangle$ in the convenient fashion

$$|f^p\rangle=\frac{|g\rangle}{N}+|v\rangle, \tag{17}$$

with

$$|v\rangle=\hat{F}|\lambda\rangle. \tag{18}$$

Notice that the component $|g\rangle/N$ of $|f^p\rangle$ given in Eq. (17) is the prediction that one would obtain on the basis of no data, i.e., from a constant, uniform distribution $p_n^{1/2}=1/N$; $n=1,\ldots,N$. The other component [the vector $|v\rangle$ given in Eq. (18)] belongs to a subspace $V_K=\text{range}(\hat{F})$, which is generated by vectors $|\alpha_{l_k}\rangle$, $k=1,\ldots,K$. The proposition below shows that the vector $|f^p\rangle$ of the form (17) that approaches $|f^o\rangle$ in the closest possible way is that obtained by letting the vector $|v\rangle$ to be the orthogonal projection of $[|f^o\rangle-(|g\rangle/N)]$ onto V_K.

Proposition 1. The unique vector $|f^p\rangle=(|g\rangle/N)+|v\rangle$, with $|g\rangle$ given in Eq. (11) and $|v\rangle\in V_K$, which minimizes the distance $\||f^o\rangle-|f^p\rangle\|$ is obtained as $|f^p\rangle=(|g\rangle/N)+\hat{P}_{V_k}[|f^o\rangle-(|g\rangle/N)]$.

Proof. Let $|f'\rangle$ be $(|g\rangle/N)+|v'\rangle$, where $|v'\rangle$ is an arbitrary vector in V_K, and let us write it as $|f'\rangle=(|g\rangle/N)+|v'\rangle-\hat{P}_{V_k}[|f^o\rangle-(|g\rangle/N)]+\hat{P}_{V_k}[|f^o\rangle-(|g\rangle/N)]$. If we calculate the squared distance $\||f^o\rangle-|f'\rangle\|^2$, since $|f^o\rangle-(|g\rangle/N)-\hat{P}_{V_k}[|f^o\rangle-(|g\rangle/N)]\in V_K^{\perp}$ (where $V_K^{\perp}$ denotes the orthogonal complement of V_K) we have

$$\begin{aligned}\||f^o\rangle-|f'\rangle\|^2&=\left\|\,|f^o\rangle-\frac{|g\rangle}{N}-|v'\rangle+\hat{P}_{V_k}\left(|f^o\rangle-\frac{|g\rangle}{N}\right)\right.\\&\qquad\left.-\hat{P}_{V_k}\left(|f^o\rangle-\frac{|g\rangle}{N}\right)\right\|^2\\&=\left\|\hat{P}_{V_k}\left(|f^o\rangle-\frac{|g\rangle}{N}\right)-|v'\rangle\right\|^2\\&\quad+\left\|\,|f^o-\frac{|g\rangle}{N}-\hat{P}_{V_k}\left(|f^o\rangle-\frac{|g\rangle}{N}\right)\right\|^2.\end{aligned}$$

Hence $\||f^o\rangle-|f'\rangle\|$ is minimized if $|v'\rangle\equiv\hat{P}_{V_k}[|f^o\rangle-(|g\rangle/N)]$, i.e., $|f'\rangle=(|g\rangle/N)+\hat{P}_{V_k}[|f^o\rangle-(|g\rangle/N)]$. □

According to this proposition, the Lagrange multiplier vector $|\lambda\rangle$ must be determined through the requirement that $|f^p\rangle=(|g\rangle/N)+\hat{P}_{V_k}[|f^o\rangle-(|g\rangle/N)]$. Let $|\Delta f\rangle$ be the *residual vector* $|f^o\rangle-|f^p\rangle$. Thus,

$$|f^o\rangle=|f^p\rangle+|\Delta f\rangle=\frac{|g\rangle}{N}+\hat{F}|\lambda\rangle+|\Delta f\rangle, \tag{19}$$

so that, in order for $|f^p\rangle$ to be $(|g\rangle/N)+\hat{P}_{V_k}[|f^o\rangle-(|g\rangle/N)]$, we should require that $\hat{P}_{V_k}|\Delta f\rangle=0$ i.e., $|\Delta f\rangle$ should be orthogonal to every vector $|\alpha_{l_k}\rangle$, $k=1,\ldots,K$. We are thus led to the following equations:

$$\langle\alpha_{l_k}|\tilde{f}^o\rangle=\langle\alpha_{l_k}|\hat{F}|\lambda\rangle;\quad k=1,\ldots,K, \tag{20}$$

with

$$|\tilde{f}^o\rangle=|f^o\rangle-\frac{|g\rangle}{N}. \tag{21}$$

The left-hand side of Eq. (20) happens to give the components of vector $\hat{F}^{\dagger}|\tilde{f}^o\rangle\in\mathcal{R}^K$, whereas on the right-hand side we find the components of a vector $\hat{F}^{\dagger}\hat{F}|\lambda\rangle\in\mathcal{R}^K$. Thus, these equations can be recast in the form

$$\hat{F}^{\dagger}|\tilde{f}^o\rangle=\hat{F}^{\dagger}\hat{F}|\lambda\rangle. \tag{22}$$

Since the operator $\hat{F}^{\dagger}\hat{F}=\Sigma_{n=1}^{K}\Sigma_{k=1}^{K}|l_k\rangle\langle\alpha_{l_k}|\alpha_{l_n}\rangle\langle l_n|$ has an inverse, $|\lambda\rangle$ is readily obtained as

$$|\lambda\rangle=(\hat{F}^{\dagger}\hat{F})^{-1}\hat{F}^{\dagger}|\tilde{f}^o\rangle. \tag{23}$$

The predicted vector $|f^p\rangle$ minimizing the distance to the observed vector $|f^o\rangle$ is thereby given by

$$|f^p\rangle=\frac{|g\rangle}{N}+\hat{P}_{V_k}|\tilde{f}^o\rangle\equiv\frac{|g\rangle}{N}+\hat{F}(\hat{F}^{\dagger}\hat{F})^{-1}\hat{F}^{\dagger}|\tilde{f}^o\rangle. \tag{24}$$

So far we have assumed that the information indicating just which the relevant data are (i.e., the subindexes l_k; $k=1,\ldots,K$) is somehow accessible to us. This is far from being a realistic hypothesis, since, in practice, such an information is normally not available.

As a consequence, for the proposed method to be of practical interest, we need to tackle the problem of appropriately selecting the subindexes l_k; $k=1,\ldots,K$. We propose here that the selection be made by recourse to an iterative process. The corresponding procedure, as well as its pertinent foundations, constitutes the subject of the next section.

IV. SELECTING RELEVANT DATA

We propose here a "greedy" algorithm for selecting the above-mentioned subset of relevant equations. The concomitant selection is not static, but evolves iteratively. At each iteration, the kth one, say, an approximation to the predicted vector $|f_k^p\rangle$ is constructed that improves upon the previous one by choosing a new vector $|\alpha_{l_k}\rangle$, and, consequently, enlarging the dimension of the subspace $V_k = \text{range}(\hat{F}_k)$. Here the subscript k indicates that at iteration kth the operator $\hat{F}$ defined in Eq. (15) is constructed out of k vectors $|\alpha_{l_j}\rangle$, $j = 1,\ldots,k$. Starting with an initial subspace V_1 spanned by a single vector $|\alpha_{l_1}\rangle$ we build a sequence of subspaces V_k by considering $V_{k+1} = V_k \oplus |\alpha_{l_{k+1}}\rangle$. As already discussed, given V_{k+1}, we wish the component $|v\rangle$ of the predicted vector to be the orthogonal projection of $|\tilde{f}^o\rangle$ onto the subspace V_{k+1}, so as to minimize the distance between such vectors. Now, since $V_{k+1} = V_k \oplus |\alpha_{l_{k+1}}\rangle$, by fixing V_k in the previous iteration (the kth one) we aim at selecting the vector $|\alpha_{l_{k+1}}\rangle$ in such a way that the distance $|||\tilde{f}^o\rangle - |f^p\rangle||^2$ is minimized. According to the discussion of the preceding section this entails to look for the vector $|\alpha_{l_{k+1}}\rangle$ such that $|||\tilde{f}^o\rangle - \hat{F}_{k+1}(\hat{F}_{k+1}^\dagger \hat{F}_{k+1})^{-1}\hat{F}_{k+1}^\dagger|\tilde{f}^o\rangle||^2$ is minimal, which at first sight seems to demand a computationally expensive effort. However, the computational burden can be enormously reduced by making use of (i) the fact that $\hat{F}_{k+1}(\hat{F}_{k+1}^\dagger \hat{F}_{k+1})^{-1}\hat{F}_{k+1}^\dagger|\tilde{f}^o\rangle = \hat{P}_{V_{k+1}}|\tilde{f}^o\rangle$, and (ii) introducing an auxiliary representation for the operator $\hat{P}_{V_{k+1}}$. The following propositions are in order.

Proposition 2. The vectors $|\psi_k\rangle$, $k=1,\ldots,K$ defined as

$$|\psi_k\rangle = |\alpha_{l_k}\rangle - \hat{P}_{V_{k-1}}|\alpha_{l_k}\rangle, \tag{25}$$

are either zero or mutually orthogonal.

Proof. The proof stems from the fact that, for $k \leqslant n$, $\hat{P}_{V_{k-1}}\hat{P}_{V_{n-1}} = \hat{P}_{V_{k-1}}$. Thus, for $k \leqslant n$, one has

$$\begin{aligned}\langle\psi_k|\psi_n\rangle &= \langle\alpha_{l_k}|\alpha_{l_n}\rangle - \langle\alpha_{l_k}|\hat{P}_{V_{n-1}}|\alpha_{l_n}\rangle - \langle\alpha_{l_k}|\hat{P}_{V_{k-1}}|\alpha_{l_n}\rangle \\ &\quad + \langle\alpha_{l_k}|\hat{P}_{V_{k-1}}\hat{P}_{V_{n-1}}|\alpha_{l_n}\rangle \\ &= \langle\alpha_{l_k}|\alpha_{l_n}\rangle - \langle\alpha_{l_k}|\hat{P}_{V_{n-1}}|\alpha_{l_n}\rangle \end{aligned} \tag{26}$$

and, since for $k \leqslant n-1$, $\hat{P}_{V_{n-1}}|\alpha_{l_k}\rangle = |\alpha_{l_k}\rangle$, it follows that, for $k<n$, $\langle\psi_k|\psi_n\rangle = 0$. Moreover, the property $\langle\psi_k|\psi_n\rangle = \overline{\langle\psi_n|\psi_k\rangle}$ (where $\overline{\langle\psi_n|\psi_k\rangle}$ indicates the complex conjugate of $\langle\psi_n|\psi_k\rangle$) allows one to extend the relation $\langle\psi_k|\psi_n\rangle = 0$ to all $k \neq n$. For $n=k$ we have $\langle\psi_k|\psi_k\rangle = \langle\alpha_k|\alpha_k\rangle - \langle\alpha_k|\hat{P}_{V_{k-1}}|\alpha_k\rangle$ so that, since for $|\alpha_{l_k}\rangle \in V_{k-1}$ it holds that $\hat{P}_{V_{k-1}}|\alpha_k\rangle = |\alpha_k\rangle$, every $|\alpha_k\rangle \in V_{k-1}$ gives rise to a vector $|\psi_k\rangle$ of zero norm. Otherwise, $\langle\psi_k|\psi_n\rangle = \delta_{k,n}|||\psi_k\rangle||^2$, which expresses the orthogonality condition. □

Corollary 1. The dimension of the subspace S spanned by the vectors $|\alpha_i\rangle$, $i=1,\ldots,M$, is equal to the number of vectors given in Eq. (25) such that $|||\psi_k\rangle||^2 \neq 0$.

Proof. The proof is of an obvious character, since $|||\psi_k\rangle||^2 = 0$ implies $\hat{P}_{V_k}|\alpha_{l_{k+1}}\rangle = |\alpha_{l_{k+1}}\rangle$, which implies $V_{k+1} = V_k \oplus |\alpha_{l_{k+1}}\rangle \equiv V_k$. Thus, $S \equiv V_K$, where K is the number of nonzero vectors $|\psi_k\rangle$. □

The above corollary suggests the convenience of reordering the vectors $|\psi_k\rangle$ by setting $k+1=k$ if $|||\psi_k\rangle||^2 = 0$. The next proposition emphasizes the fact that the reordered family $|\psi_k\rangle$, $k=1,\ldots,K$, provides a representation for the orthogonal projector operator onto S.

Proposition 3. Let $S \equiv V_K$ be spanned by K linearly independent vectors $|\alpha_{l_k}\rangle$, $k=1,\ldots,K$. The orthogonal projection operator onto V_K can be expressed as

$$\hat{P}_{V_K} = \sum_{k=1}^{K} |\tilde{\psi}_k\rangle\langle\tilde{\psi}_k|, \tag{27}$$

where $|\tilde{\psi}_k\rangle = |\psi_k\rangle / |||\psi_k\rangle||$, $k=1,\ldots,K$.

Proof. The proof is achieved by showing the following:

$$\text{(a)}\quad \sum_{k=1}^{K} |\tilde{\psi}_k\rangle\langle\tilde{\psi}_k|g\rangle = |g\rangle, \quad \forall |g\rangle \in V_K.$$

$$\text{(b)}\quad \sum_{k=1}^{K} |\tilde{\psi}_k\rangle\langle\tilde{\psi}_k|g^\perp\rangle = 0, \quad \forall |g^\perp\rangle \in V_K^\perp.$$

(a) follows from the fact that every $|g\rangle \in V_K$ can be expressed as a linear combination of the K linearly independent vectors $|\alpha_{l_k}\rangle$, $k=1,\ldots,K$, i.e., $|g\rangle = \sum_{n=1}^{K} c_{l_n}|\alpha_{l_n}\rangle$. Hence,

$$\begin{aligned}\sum_{k=1}^{K} |\tilde{\psi}_k\rangle\langle\tilde{\psi}_k|g\rangle &= \sum_{k=1}^{K} \frac{|\psi_k\rangle}{|||\psi_k\rangle||^2} \left\langle \psi_k \middle| \sum_{n=1}^{K} c_{l_n}\alpha_{l_n} \right\rangle \\ &= \sum_{k=1}^{K} \frac{|\psi_k\rangle}{|||\psi_k\rangle||^2} \sum_{n=1}^{K} c_{l_n}\langle\psi_k|\psi_n + \hat{P}_{V_{n-1}}\alpha_{l_n}\rangle \\ &= \sum_{k=1}^{K} \frac{|\psi_k\rangle}{|||\psi_k\rangle||^2} \sum_{n=1}^{K} c_{l_n}\delta_{n,k}|||\psi_k\rangle||^2 + \sum_{n=1}^{K} c_{l_n}\hat{P}_{V_K}\hat{P}_{V_{n-1}}|\alpha_{l_n}\rangle \\ &= \sum_{n=1}^{K} c_{l_n}(|\psi_{l_n}\rangle + \hat{P}_{V_{n-1}}|\alpha_{l_n}\rangle) = \sum_{n=1}^{K} c_{l_n}|\alpha_{l_n}\rangle = |g\rangle. \end{aligned} \tag{28}$$

On the other hand, for all $|g^{\perp}\rangle \in V_K^{\perp}$ it is true that $\langle \alpha_{l_k}|g^{\perp}\rangle = 0$, $k=1,\ldots,K$ and hence $\sum_{k=1}^{K}|\tilde{\psi}_k\rangle\langle\tilde{\psi}_k|g^{\perp}\rangle = \sum_{k=1}^{K}|\psi_k\rangle\langle\alpha_k - \hat{P}_{V_{k-1}}\alpha_{l_k}|g^{\perp}\rangle/\||\psi_k\rangle\|^2 = 0$, which proves (b). □

We are ready now to establish a theorem that allows for the fast implementation of the proposed selection criterion.

Theorem 1. The vector $|\alpha_{l_k}\rangle$, that at iteration k minimizes the norm of the residual vector $|\Delta f\rangle$, is the one yielding the largest value of the functionals e_i, $i=1,\ldots,M$ given by

$$e_i = |\langle\tilde{\psi}_k|\tilde{f}^o\rangle|^2 = \frac{b_i}{d_i} = \frac{|\langle\alpha_i|\Delta f\rangle|^2}{\langle\alpha_i|\alpha_i\rangle - \sum_{l=1}^{k-1}|\langle\tilde{\psi}_l|\alpha_i\rangle|^2}; \quad b_i > 0. \tag{29}$$

Proof. According to Proposition 1, at iteration k the residue $|\Delta f\rangle$ of minimum norm should verify $|\Delta f\rangle = |\tilde{f}^o\rangle - \hat{P}_{V_k}|\tilde{f}^o\rangle$, so that $\||\Delta f\rangle\|^2 = \||\tilde{f}^o\rangle\|^2 - \langle\tilde{f}^o|\hat{P}_{V_k}|\tilde{f}^o\rangle$, and, since $\hat{P}_{V_k} = \hat{P}_{V_{k-1}} + |\tilde{\psi}_k\rangle\langle\tilde{\psi}_k|$,

$$\||\Delta f\rangle\|^2 = \||\tilde{f}^o\rangle\|^2 - \langle\tilde{f}^o|\hat{P}_{V_{k-1}}|\tilde{f}^o\rangle - |\langle\tilde{\psi}_k|\tilde{f}^o\rangle|^2. \tag{30}$$

The term $\langle\tilde{f}^o|\hat{P}_{V_{k-1}}|\tilde{f}^o\rangle$ is fixed in the preceding iteration. Therefore, it follows from Eq. (30) that, at iteration k, the norm of the residue $|\Delta f\rangle$ is minimized by the function $|\tilde{\psi}_k\rangle$ for which $|\langle\tilde{\psi}_k|\tilde{f}^o\rangle|^2$ takes its largest value. Now, by using Eq. (25),

$$\begin{aligned}|\langle\tilde{\psi}_k|\tilde{f}^o\rangle|^2 &= \frac{|\langle\alpha_{l_k}|\tilde{f}^o\rangle - \langle\alpha_{l_k}|\hat{P}_{V_k}|\tilde{f}^o\rangle|^2}{\||\psi_k\rangle\|^2}\\ &= \frac{|\langle\alpha_{l_k}|\tilde{f}^o - \hat{P}_{V_k}\tilde{f}^o\rangle|^2}{\||\psi_k\rangle\|^2},\end{aligned} \tag{31}$$

so that we can further write

$$|\langle\tilde{\psi}_k|\tilde{f}^o\rangle|^2 = \frac{|\langle\alpha_{l_k}|\Delta f\rangle|^2}{\||\psi_k\|^2} = \frac{|\langle\alpha_{l_k}|\Delta f\rangle|^2}{\langle\alpha_{l_k}|\alpha_{l_k}\rangle - \sum_{l=1}^{k-1}|\langle\tilde{\psi}_l|\alpha_{l_k}\rangle|^2}, \tag{32}$$

and the proof is completed. □

Theorem 1 guarantees that the recursive selection of vectors $|\alpha_{l_k}\rangle$ using the criterion (29) provides us, at the kth iteration, with (i) the vector $|\alpha_{l_k}\rangle$ that minimizes the norm of the residual error, and (ii) the Lagrange multiplier vector that approximates the available data $|f^o\rangle \in \mathcal{R}^M$ in the least square sense. Indeed, since every vector $|\alpha_i\rangle$ exhibiting a linear dependence on the the previously selected ones yields values of b_i and d_i equal to 0 [cf. Eq. (29)], according to the restriction $b_i > 0$ all the selected vectors are guaranteed to be linearly independent. Hence, the operator $\hat{F}^{\dagger}\hat{F}$ constructed out of, say K, selected vectors $|\alpha_{l_k}\rangle$, $k=1,\ldots,K$ does have an inverse, which allows for the computation of the Lagrange multiplier vectors $|\lambda\rangle \in \mathcal{R}^K$, as in Eq. (23). Moreover, since each index l_k, $k=1,\ldots,K$ represents an equation of the system (1) corresponding to a relevant piece of data, by selecting a subset of vectors $|\alpha_{l_k}\rangle$, $k=1,\ldots,K$ we are able to identify the subset of relevant data $f_{l_k}^o$, $k=1,\ldots,K$ that it was our goal to detect.

Sketch of the algorithm

Let us start by recalling that the vector $|\tilde{f}^o\rangle$ is obtained from the data vector $|f^o\rangle$ through Eq. (21) and the vectors $|\alpha_i\rangle$, $i=1,\ldots,M$, as given in Eq. (16). Beginning with $|\Delta f\rangle = |\tilde{f}^o\rangle$ and the inner products $\langle\alpha_i|\tilde{f}^o\rangle$, $i=1,\ldots,M$, the procedure evolves as follows.

(i) Initially set $k=1$, $|a_i\rangle = |\alpha_i\rangle$, $d_i = \langle\alpha_i|\alpha_i\rangle$, $i=1,\ldots,M$, and l_1 equal to the index i for which $e_i = |\langle\alpha_i|\tilde{f}^o\rangle|^2/d_i$ adopts the largest value as i ranges from 1 to M. Assign $|\psi\rangle = |\alpha_{l_1}\rangle$, $q = d_{l_1}$, and $\||\Delta f\rangle\|^2 = \||\Delta f\rangle\|^2 - e_{l_1}$.

(ii) For $i=1,\ldots,M$ compute the following:

$$|a_i\rangle = |a_i\rangle - \frac{|\psi\rangle\langle\psi|\alpha_i\rangle}{q},$$

$$b_i = \langle a_i|\tilde{f}^o\rangle,$$

$$d_i = d_i - \frac{|\langle\psi|\alpha_i\rangle|^2}{q},$$

$$\text{if } |b_i| = 0, \quad e_i = 0 \text{ otherwise } e_i = |b_i|^2/d_i.$$

(iii) Increase k to $k+1$ and set l_k equal to the index i for which e_i takes the largest value as i ranges from 1 to M. Assign $|\psi\rangle = |a_{l_k}\rangle$, $q = d_{l_k}$, and $\||\Delta f\rangle\|^2 = \||\Delta f\rangle\|^2 - e_{l_k}$.

(iv) Repeat steps (ii), and (iii). The algorithm is to be stopped when some convergence criterion is reached, e.g., when

$$\|\Delta f\|^2 \leq \delta^2, \tag{33}$$

where δ^2 is a square norm of the data error.

Let us assume that the given convergence criterion is reached at iteration K. At such stage the above algorithm has selected K indexes l_k, $k=1,\ldots,K$, and we are in a position to compute the inverse of operator $\hat{F}^{\dagger}\hat{F}$ (by simply evaluating the inverse of its matrix representation $\langle l_n|\hat{F}^{\dagger}\hat{F}|l_k\rangle = \langle\alpha_{l_n}|\alpha_{l_k}\rangle$, $n=1,\ldots,K$, $k=1,\ldots,K$). Hence, the Lagrange multiplier vector minimizing the distance to the available data is given by

$$|\lambda\rangle = (\hat{F}^{\dagger}\hat{F})^{-1}\hat{F}^{\dagger}|\tilde{f}^o\rangle, \tag{34}$$

and the corresponding $|p^{1/2}\rangle$ distribution by Eq. (13).

Finally we would like to stress that, according to the proposed scheme, the entire set of pieces of data $|f^o\rangle \in \mathcal{R}^M$ can be "encoded" into a vector of smaller dimension, namely, the Lagrange multiplier vector $|\lambda\rangle \in \mathcal{R}^K$. The reconstruction, interpolation, and extrapolation of the data are achieved via "predictions" of the $|p^{1/2}\rangle$ distribution. Indeed, the predicted values for the observed data are given by

$$|f^p\rangle=\hat{A}|p^{1/2}\rangle. \tag{35}$$

On the other hand, if x_n is a random variable representing a physical quantity that is not contained in our original data space, then the prediction of such a variable is to be computed as

$$\overline{x}=\sum_{n=1}^{N} p_n^{1/2}x_n. \tag{36}$$

V. NUMERICAL SIMULATION

We illustrate in this section the approach advanced in the present communication with a well-known example that deals with a highly unstable inverse problem, even for very small perturbations of the data.

The spaces $\mathcal{R}^N$ and $\mathcal{R}^M$ are chosen to be of dimension 50 and 100, respectively. The matrix elements of the operator $\hat{A}$ are given by the exponential decays

$$\langle i|\hat{A}|n\rangle=f_{i,n}=\exp(-nx_i), \quad x_i=0.01i,$$
$$i=1,\ldots,100, \quad n=1,\ldots,50. \tag{37}$$

The "true" data are generated as follows:

$$f_i=\sum_{n=1}^{50} p_n f_{i,n}, \quad i=1,\ldots,100 \tag{38}$$

with $f_{i,n}$ as in Eq. (37) and p_n given by

$$p_n=\frac{\exp\left(-\frac{[\ln(n)-\ln(7)]^2}{4\ln(2)}\right)}{\sum_{n=1}^{50}\exp\left(-\frac{[\ln(n)-\ln(7)]^2}{4\ln(2)}\right)}, \quad n=1,\ldots,50. \tag{39}$$

The "observed" data are simulated by distorting each data f_i with a 0.1% gaussian error.

In Fig. 1, the solid line represents the exact distribution p_n given in Eq. (39). The dotted curve corresponds to the solution $p_n^{1/2}$ that one obtains from five different realizations of the observed data. In each of them our algorithm selects four relevant data points x_{l_k}, $k=1,\ldots,4$ corresponding to the indexes l_k, $k=1,\ldots,4$, as listed below

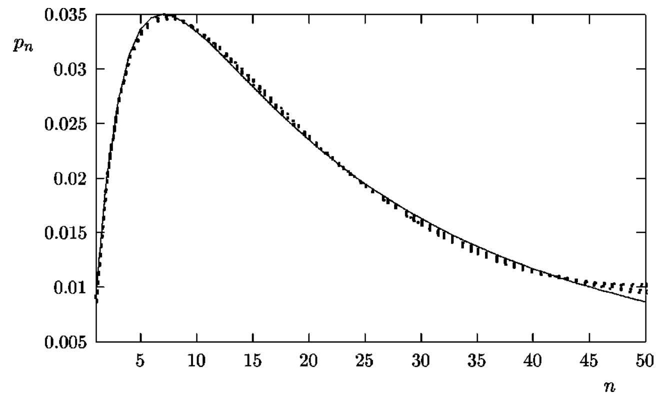

FIG. 1. Exact distribution (solid line), as given by Eq. (39), versus results that we obtain for five different realizations of the observed data (dotted curves).

1	69	2	35
1	68	2	19
1	70	2	27
1	69	2	24
1	69	2	30

Notice that the selection of two *consecutive* points (1 and 2) is effected in all cases.

The convergence criterion (33) is seen to yield stability of the approach against different realizations of data.

VI. CONCLUSIONS

A method for data subset selection, which is based on the $q=\frac{1}{2}$ nonextensive maximum information measure formalism, has been advanced.

The method proceeds iteratively by selecting, at each step, a measure endowed with information not contained in the previously selected measures. The selection is made optimal in the following sense: at each iteration the selected data gives rise to a $q=\frac{1}{2}$ distribution that effects predictions that minimize the Euclidean distance to all the available data.

Information relative to the question concerning just *which* the relevant data are, is to be stored as a set of indexes (integer numbers). Information on the data themselves is stored as parameters of the model (Lagrange multipliers). Out of this information one can reconstruct, interpolate, and extrapolate the original data via a $q=\frac{1}{2}$ nonextensive maximum information measure distribution.

[1] D. Guerín, A. Alvarez, L. Rebollo-Neira, A. Plastino, and R. Bonetto, Acta Crystallogr., Sect. A: Found. Crystallogr. **42**, 30 (1985).

[2] A. Plastino, L. Rebollo-Neira, and A. Alvarez, Phys. Rev. A **40**, 1644 (1989).

[3] A. Arvia, E. Custidiano, A. Plastino, and L. Rebollo-Neira, J. Electroanal. Chem. Interfacial Electrochem. **131**, 1 (1990).

[4] S. J. Gull and G. J. Daniel, Nature (London) **272**, 686 (1978).

[5] *Maximum Entropy and Bayesian Methods in Inverse Problems*, edited by C. Ray Smith (Kluwer Academic Publishers, Dordrecht, 1985).

[6] *Maximum Entropy and Bayesian Methods*, edited by J. Skilling (Kluwer Academic Publishers, Dordrecht, 1989).

[7] *Maximum Entropy and Bayesian Methods*, edited by A.

Mohammad-Djafari and G. Demoment (Kluwer Academic Publishers, Dordrecht, 1993).
[8] L. Rebollo-Neira, A. Constantinides, A. Plastino, F. Zyserman, A. Alvarez, R. Bonetto, and H. Viturro, Physica A **198**, 514 (1993).
[9] L. Rebollo-Neira and A. G. Constantinides, Signal Process. **56**, 135 (1997).
[10] L. Rebollo-Neira, A. G. Constantinides, A. Plastino, A. Alvarez, R. Bonetto, and M. Iñiguez Rodriguez, J. Phys. D **30**, 2462 (1997).
[11] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).
[12] C. Tsallis, Fractals **6**, 539 (1995), and references therein.
[13] E. M. F. Curado and C. Tsallis, J. Phys. A **24**, L69 (1991); **24**, 3187 (1991); **25**, 1019 (1992).
[14] A. R. Plastino and A. Plastino, Phys. Lett. A **177**, 177 (1993).
[15] A. R. Plastino and A. Plastino, Phys. Lett. A **174**, 384 (1993).
[16] A. R. Plastino and A. Plastino, Phys. Lett. A **193**, 140 (1994).
[17] A. Plastino and A. R. Plastino, Braz. J. Phys. **29**, 50 (1999).
[18] C. Tsallis, Braz. J. Phys. **29**, 1 (1999), and references therein. An updated bibliography can be found in http://tsallis.cat.cbpf.br/biblio.htm
[19] L. Rebollo-Neira, A. Plastino, and J. Fernandez-Rubio, Physica A **258**, 458 (1998).
[20] B. M. R. Boghosian, Phys. Rev. E **53**, 4754 (1996).
[21] L. Rebollo-Neira, J. Fernandez-Rubio, and A. Plastino, Physica A **261**, 555 (1998).
[22] B. R. La Cour and W. C. Schieve, Phys. Rev. E **62**, 7494 (2000).