Constructive approximations to the $q=1/2$ maximum entropy distribution from redundant and noisy data

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, Casilla de Correo 727, 1900 La Plata, Argentina

The problem of constructing the $q=1/2$ nonextensive maximum entropy distributions from redundant and noisy data is considered. A strategy is proposed, which evolves through the following steps. (i) Independent constraints are first preselected by recourse to a data-independent technique to be discussed here. (ii) The data are *a posteriori* used to determine the parameters of the distribution by a previously introduced forward approach. (iii) A backward approach is proposed for reducing the parameters of such distribution. The previously introduced forward approach is generalized here in order to make it suitable for dealing with very noisy data.

I. INTRODUCTION

Among the generalized nonextensive maximum entropy distributions, which are defined in terms of a parameter q [1–3], the one corresponding to the value $q=1/2$ has played a particular role in diverse contexts [4–9].

In this paper we focus on developing strategies for constructing the $q=1/2$ distribution which is involved in a very special type of inverse problem: the problem of constructing such a distribution on the basis of redundant and noisy data (by noise we mean errors resulting from the random process associated with the experimental measurement procedure).

It is appropriate to start by discussing why we shall restrict consideration to the particular value $q=1/2$.

The problem of determining a p^q probability distribution maximizing the entropy

$$S_q = \frac{\sum_{n=1}^{N} p_n^q - \sum_{n=1}^{N} p_n}{1-q}$$

with constraints

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

$$1 = \sum_{n=1}^{N} p_n^q$$

has been shown in Ref. [6] to be numerically equivalent to determining the probability distribution $\tilde{p}$ minimizing

$$\|\tilde{p}\|_{1/q}^{1/q} = \sum_{n=1}^{N} \tilde{p}_n^{1/q}$$

with constraints

$$f_i^o = \sum_{n=1}^{N} \tilde{p}_n f_{i,n}, \quad i=1,\ldots,M,$$

$$1 = \sum_{n=1}^{N} \tilde{p}_n.$$

Since $\tilde{p}_n > 0$, it is true that $\|\tilde{p}\|_{1/q}$ is the $(1/q)$-norm of $\tilde{p}$. Thus, the problem of choosing the parameter q is equivalent to deciding which norm one wants to minimize as preserving the 1-norm of the distribution. In order to analyze the situation further let us join all constraints together by defining a $(M+1)\times N$ matrix $\tilde{A}$ of elements $\tilde{A}_{i,n}=f_{i,n}(i=1,\ldots,M;n=1,\ldots,N)$ and $\tilde{A}_{M+1,n}=1(n=1,\ldots,N)$. Hence, the constraints are expressed in the form

$$f^o = \tilde{A}p,$$

where f^o is a vector of $(M+1)$ components $f_1^o,\ldots,f_M^o,1$. It is well known from linear algebra that the general solution to this underdetermined linear system can be expressed as

$$\tilde{p} = \tilde{A}'^{-1}f^o + p',$$

where $\tilde{A}'^{-1}$ is the pseudoinverse of $\tilde{A}$, and p' a vector in the null space of matrix $\tilde{A}$. Consequently, the problem of deciding on the q parameter is tantamount to just choosing a vector p' in the null space of $\tilde{A}$. In particular, the choice $q=1/2$ (which as already discussed is equivalent to minimizing the 2-norm of the $\tilde{p}$ distribution) implies to set $p'=0$. This follows from the fact that, since vector $\tilde{A}^{-1}f^o$ and vector p' are orthogonal with each other, one has

$$\|\tilde{p}\|_2^2 = \|\tilde{A}^{-1}f^o\|_2^2 + \|p'\|_2^2.$$

Hence, by setting $p'=0$ the solution of minimum 2-norm is obtained. For a number of reasons, which are listed below,

we believe that this leads to the most suitable choice for the parameter q in relation to our problem.

(1) The underdetermined problem we have to solve is of the following special nature: We have less independent equations than unknowns, but there is a large number of redundant equations and a number of irrelevant ones [7]. If the data were noiseless, the role of such equations would be simply to verify the ability of the distribution to make correct predictions. Since the data are noisy we use all the equations with the purpose of reducing the effect of the noise, but not as independent constraints (in most cases the number of Lagrange multipliers is much less than the actual number of available constraints). Our task is to identify a subset of such independent constraints. The predictive power of our solution is assessed *a posteriori* by its capability of predicting the denoised data.

(2) The constraints typically represent measurements obtained as a function of some variable parameters: Intensity versus diffraction angle, magnetization versus magnetic field, etc. [12,13]. It is then natural to represent such measurements as linear functionals on the identical vector. Each linear functional provides a projection on the particular parameter value which is specified by the measurement instrument state [12]. It is clear then that in the space of the data it is appropriate to define a distance through the norm induced by the inner product. In our formalism both the space of the data and the space of the system are assumed to be Hilbert spaces. The only $1/q$ norm induced by a Hilbert space is the one corresponding to $q = 1/2$.

(3) As mentioned above, to choose a value of q other than $q = 1/2$ would imply to let the corresponding distribution have a component in the null space of the transformation generated by the constraints. In the type of problem described in (2) above such a null space is of a 'chaotic' nature (in the sense that arbitrarily small numerical perturbation on any of the elements of matrix $\widetilde{A}$ would produce an enormous distortion in the solution). We certainly wish to avoid this.

Unfortunately, in our context deciding on the appropriate q value of the distribution we wish to construct does not solve the problem of its optimal construction. While it is true that the problem of determining the $q = 1/2$ distribution from a fixed set of constraints is a simple linear problem [5], the problem becomes highly nonlinear when this distribution is to be determined optimally from a subset of constraints which are taken out of a much larger set of possible ones.

Consider that from a set of M constraints we want to select a subset of k ones and associate a parameter (Lagrange multiplier) to each equation. Let us indicate as $p^{(1/2)(k)}$ the distribution associated with the corresponding k equations. Hence the problems we have to face are the following (a) the selection of the optimal k constraints and (b) the estimation of the corresponding k parameters determining the distribution. In order to address these problems, let us specify the meaning of "optimal selection" in our context: we say that a selection is optimal if it yields a distribution capable of satisfactorily predicting all the available data involving the minimum number of parameters. Unfortunately the search for such an optimal selection is not in general possible, as it poses a NP-hard problem, i.e., unreachable in polynomial time with classical computers [10,11]. Hence we are forced to ascertain suitable suboptimal strategies, which also pose an open problem because there is not a unique way of constructing suboptimal solutions.

In some recent publications we have introduced a suboptimal iterative strategy, which is only optimal at each iteration step [7,8]. Such an approach is a forward data-dependent approach for subset selection. At each iteration, the indices obtained in the previous steps are fixed, and a new index is chosen in such a way that the distance between the observed data and the ones predicted by the physical model is minimized. Since the selection is only optimal at each step, the selected set of indices is, of course, not optimal in the above specified sense. Some indices that are relevant at a particular step may become much less relevant at the end of the process. It is then natural to try and eliminate the parameters corresponding to such indices. Again, the process of reducing parameters in an optimal way is, in general, a NP-problem and we need to address it by suboptimal strategies. Here we propose a strategy for reducing parameters that we call backward selection. This approach provides both the criterion for selecting the parameters to be deleted and the technique for properly modifying the ones to be retained. An approach for selecting independent constraints in the absence of data will also be advanced here, with the aim of designing a suboptimal strategy consisting of the following steps.

(i) Before the experiment is carried out we select a subset of indices corresponding to independent constraints.

(ii) The forward selection approach proposed in Ref. [8] is then applied for selecting indices, from the preselected set, in order to construct the distribution when the data are available.

(iii) Finally the backward selection approach is applied in order to reduce further the number of parameters of the distribution. Such backward selection is made possible in a fast and efficient way by means of a backward adaptive biorthogonalization technique.

Before advancing the above described strategy we would like to discuss how it is possible to adapt the strategy of Ref. [8] so as to make it suitable when dealing with very noisy data. This is achieved by introducing a vectorial space with inner product defined with respect to a measure depending on the experimental data, or their corresponding statistics.

The paper is organized as follows. The generalization of the previous approach, to turn it suitable when dealing with very noisy data, is introduced in Sec. II. Section III discusses the criteria for selecting relevant constraints. First, the selection criterion proposed in Ref. [7] is generalized and a numerical experiment is presented in order to illustrate the advantage of such a generalization. We then discuss a data-independent selection criterion. In Sec. III we introduce a backward procedure for eliminating constraints and, consequently, for properly adapting the concomitant parameters of the distribution. Section III C provides the foundations of the strategy that we illustrate by a numerical example in Sec. III D. The conclusions are drawn in Sec. IV.

II. GENERALIZING THE PREVIOUS APPROACH

Let us assume that we are given 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

of a random variable that takes values $f_{i,n}, n=1,\ldots,N$ according to the $q=1/2$ probability distribution $p_n^{1/2}, n=1,\ldots,N$ [7,8], i.e.,

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

The data $f_1^o, f_2^o, \ldots, f_i^o, \ldots, f_M^o$ will be represented as components of a vector $|f^o\rangle_\mu$ in a vector space, say $\mathcal{D}^M$. A central aim of this contribution is to allow for the possibility of assigning a different weight to each data. Accordingly, the inner product in $\mathcal{D}^M$, which we indicate as $_\mu\langle\cdot|\cdot\rangle_\mu$, is defined with respect to a measure $\mu(m)$ as follows.

For every f and g in $\mathcal{D}^M$,

$$_\mu\langle f|g\rangle_\mu = \sum_{i=1}^{M} \bar{f}_i g_i \mu_i, \tag{2}$$

where $\bar{f}_i$ indicates the complex conjugate of f. In the present situation we deal with real vectors, thereby, $\bar{f}_i \equiv f_i$. The data space, with the corresponding associated measure, will be denoted as $\mathcal{D}^M(\mu)$ and the standard orthogonal basis in $\mathcal{D}^M(\mu)$ will be represented by vectors $|i\rangle_\mu, i=1,\ldots,M$. The identity operator in $\mathcal{D}^M(\mu)$ is thus expressed as

$$\hat{I}_\mu = \sum_{i=1}^{M} |i\rangle_\mu \mu_{i\mu}\langle i| \tag{3}$$

with vectors $|i\rangle_\mu, i=1,\ldots,M$ satisfying the relations

$$\mu_{i\mu}\langle i|j\rangle_\mu = \delta_{i,j} \quad (\text{or } 0 \text{ if } \mu_i = 0). \tag{4}$$

Accordingly, vector $|f^o\rangle_\mu$ is expressed:

$$|f^o\rangle_\mu = \sum_{i=1}^{M} |i\rangle_\mu \mu_{i\mu}\langle i|f^o\rangle_\mu = \sum_{i=1}^{M} \mu_i f_i^o |i\rangle_\mu. \tag{5}$$

The measure μ, rendering a weighted distance between two vectors in $\mathcal{D}^M(\mu)$, will be chosen in relation to the observed data. For example, if the variances of the data are known and we denote by σ_i^2 the variance of data f_i^o, the choice $\mu_i = \sigma_i^{-2}$, gives rise to the square distance between $|f^o\rangle_\mu$ and $|g\rangle_\mu \in \mathcal{D}^M(\mu)$ as given by

$$\||f^o\rangle_\mu - |g\rangle_\mu\|^2 = _\mu\langle f^o - g|f^o - g\rangle_\mu = \sum_{i=1}^{M} (f_i^o - g_i)^2 \frac{1}{\sigma_i^2}. \tag{6}$$

The above distance is known to be optimal, in a maximum likelihood sense, if the data errors are Gaussian distributed [14].

The space of the physical system is considered to be the Euclidean N-dimensional real space $\mathcal{R}^N$. The standard orthogonal basis in $\mathcal{R}^N$ will be indicated by vectors $|n\rangle, n=1,\ldots,N$, so that every vector $|r\rangle \in \mathcal{R}^N$ is represented as

$$|r\rangle = \sum_{n=1}^{N} \langle n|r\rangle|n\rangle = \sum_{n=1}^{N} r_n|n\rangle. \tag{7}$$

For any two vectors $|v\rangle$ and $|r\rangle$ in $\mathcal{R}^N$ the inner product is defined as

$$\langle v|r\rangle = \sum_{n=1}^{N} \langle v|n\rangle\langle n|r\rangle = \sum_{n=1}^{N} v_n r_n. \tag{8}$$

Using the adopted vector notation, Eqs. (1) are recast:

$$|f^o\rangle_\mu = \hat{A}_\mu|p^{1/2}\rangle \tag{9}$$

with

$$|p^{1/2}\rangle = \sum_{n=1}^{N} |n\rangle\langle n|p^{1/2}\rangle = \sum_{n=1}^{N} p_n^{1/2}|n\rangle, \tag{10}$$

and operator $\hat{A}_\mu: \mathcal{R}^N \to \mathcal{D}^M(\mu)$ given by

$$\hat{A}_\mu = \sum_{n=1}^{N} |f_n\rangle_\mu\langle n|. \tag{11}$$

Vectors $|f_n\rangle_\mu \in \mathcal{D}^M(\mu)$ are defined in such a way that $_\mu\langle i|f_n\rangle_\mu^\mu = f_{i,n}$, i.e.,

$$|f_n\rangle_\mu = \sum_{i=1}^{M} |i\rangle_\mu \mu_{i\mu}\langle i|f\rangle_\mu = \sum_{i=1}^{M} \mu_i f_{i,n}|i\rangle_\mu. \tag{12}$$

In the line of Ref. [7], in order to determine the maximum entropy $|p^{1/2}\rangle$ distribution we consider as constraint of the optimization process, a subset of k equations (1) labeled by indices $l_j, j=1,\ldots,k$. This leads to the following expression for the distribution:

$$|p^{(1/2)(k)}\rangle = \left(\frac{1}{N} - \frac{1}{N}\sum_{j=1}^{k} {}_\mu\langle g|l_j\rangle_{\mu\mu}\langle l_j|\lambda^k\rangle_\mu\right)\sum_{n=1}^{N} |n\rangle$$
$$+ \sum_{j=1}^{k} \hat{A}_\mu^\dagger|l_j\rangle_{\mu\mu}\langle l_j|\lambda^k\rangle_\mu \tag{13}$$

with

$$|g\rangle_\mu = \sum_{n=1}^{N} |f_n\rangle_\mu \equiv \sum_{n=1}^{N} \hat{A}_\mu|n\rangle. \tag{14}$$

The superscript k in $|p^{(1/2)(k)}\rangle$ given above indicates that the distribution is built out of k constraints. The Lagrange multiplier vector $|\lambda^{(k)}\rangle$ is determined by the requirement that $|p^{1/2(k)}\rangle$ predicts a complete data vector $|f^p\rangle_\mu = \hat{A}_\mu|p^{(1/2)(k)}\rangle \in \mathcal{D}^M(\mu)$ minimizing the distance to the observed vector $|f^o\rangle_\mu$. This is actually the prescription given in Ref. [7]. Nevertheless, the fact that here the distance is defined with respect to a measure, which we propose to be dependent on the experimental data, implies that the formalism of Ref. [7] needs to be adapted to this requirement. In subsequent sections, we discuss how this can be achieved in a straightforward manner by means of a recursive biorthogonalization technique for computing the Lagrange multipliers which determine $|p^{(1/2)(k)}\rangle$.

A. Determination of Lagrange multipliers

In order to estimate the Lagrange multipliers determining Eq. (13) we minimize the distance between the prediction

through the physical model and observed data. As discussed in Refs. [7,8] this entails to determine the Lagrange multipliers as

$$\sum_{j=1}^{k} |\alpha_{l_j}\rangle_{\mu\mu}\langle l_j|\lambda^{(k)}\rangle_\mu = \hat{F}_k|\lambda^{(k)}\rangle_\mu = \hat{P}_{V_k}|\tilde{f}^o\rangle_\mu, \qquad (15)$$

where we have denoted $\hat{F}_k = \sum_{j=1}^{k} |\alpha_{l_j}\rangle_\mu\langle l_j|$, with

$$|\alpha_{l_j}\rangle_\mu = \sum_{n=1}^{N} |f_n\rangle_{\mu\mu}\langle f_n|l_j\rangle - \frac{1}{N}|g\rangle_{\mu\mu}\langle g|l_j\rangle. \qquad (16)$$

Vector $|\tilde{f}^o\rangle_\mu$ is obtained from the data vector as $|\tilde{f}^o\rangle_\mu = |f^o\rangle_\mu - (|g\rangle_\mu/N)$ and $\hat{P}_{V_k}$ is the orthogonal projector onto the subspace spanned by $|\alpha_{l_j}\rangle_\mu, j = 1, \ldots, k$. Here we wish this projector to account for the different weights of the data. This will be achieved by recourse to a biorthogonalization technique [15] which, as applied in this context, produces biorthogonal vectors dependent on the weight assigned to each data.

Given a set of vectors $|\alpha_{l_n}\rangle_\mu, n = 1, \ldots, M$, we set $|\psi_1\rangle_\mu = |\alpha_{l_1}\rangle_\mu$ and inductively define vectors $|\tilde{\tilde{\psi}}_{k+1}\rangle_\mu$ as

$$|\tilde{\tilde{\psi}}_{k+1}\rangle_\mu = \frac{|\psi_{k+1}\rangle_\mu}{\||\psi_{k+1}\rangle_\mu\|^2} \qquad (17)$$

with

$$|\psi_{k+1}\rangle_\mu = |\alpha_{l_{k+1}}\rangle_\mu - \hat{P}_{V_k}|\alpha_{l_{k+1}}\rangle_\mu. \qquad (18)$$

The dual vectors $_\mu\langle\tilde{\alpha}_{l_n}^{k+1}|, n = 1, \ldots, k+1$, which are obtained from the recursive equations

$$_\mu\langle\tilde{\alpha}_{l_n}^{k+1}| = {}_\mu\langle\tilde{\alpha}_{l_n}^{k}| - {}_\mu\langle\tilde{\alpha}_{l_n}^{k}|\alpha_{l_{k+1}}\rangle_{\mu\mu}\langle\tilde{\tilde{\psi}}_{k+1}|, \quad n = 1, \ldots, k,$$

$$_\mu\langle\tilde{\alpha}_{l_{k+1}}^{k+1}| = \frac{{}_\mu\langle\psi_{k+1}|}{{}_\mu\langle\psi_{k+1}|\alpha_{l_{k+1}}\rangle_\mu} = \frac{{}_\mu\langle\psi_{k+1}|}{{}_\mu\langle\psi_{k+1}|\psi_{k+1}\rangle_\mu} = {}_\mu\langle\tilde{\tilde{\psi}}_{k+1}|,$$

$$(19)$$

satisfy the following properties.

(a) They are biorthogonal with respect to vectors $|\alpha_{l_n}\rangle_\mu, n = 1, \ldots, k+1$, i.e.,

$$_\mu\langle\tilde{\alpha}_{l_n}^{k+1}|\alpha_{l_m}\rangle_\mu = \delta_{l_m, l_n}, \quad n = 1, \ldots, k+1, \quad m = 1, \ldots, k+1. \qquad (20)$$

(b) They provide a representation of the orthogonal projection operator onto V_{k+1} as given by

$$\hat{P}_{V_{k+1}} = \sum_{n=1}^{k+1} |\alpha_{l_n}\rangle_{\mu\mu}\langle\tilde{\alpha}_{l_n}^{k+1}| = \hat{P}_{V_{k+1}}^\dagger = \sum_{n=1}^{k+1} |\tilde{\alpha}_{l_n}^{k+1}\rangle_{\mu\mu}\langle\alpha_{l_n}|. \qquad (21)$$

The proof of (a) and (b) parallels that of Refs. [15,16], for the case of the standard Euclidean measure.

It follows from Eqs. (21) and (15) that the Lagrange multipliers yielding $|p^{(1/2)(k+1)}\rangle$ are obtained according to the recursive relation

$$_\mu\langle l_n|\lambda^{(k+1)}\rangle_\mu = {}_\mu\langle l_n|\lambda^{(k)}\rangle_\mu - {}_\mu\langle\tilde{\alpha}_{l_n}^k|\alpha_{l_{k+1}}\rangle_{\mu\mu}\langle l_{k+1}|\lambda^{(k+1)}\rangle_\mu,$$

$$n = 1, \ldots, k$$

$$_\mu\langle l_{k+1}|\lambda^{(k+1)}\rangle_\mu = {}_\mu\langle\tilde{\tilde{\psi}}_{k+1}|\tilde{f}^o\rangle_\mu \qquad (22)$$

with $_\mu\langle l_1|\lambda^{(1)}\rangle_\mu = {}_\mu\langle\alpha_{l_1}|\tilde{f}^o\rangle_\mu/\||\alpha_{l_1}\rangle_\mu\|^2$.

In writing down the above equations we confidently assume that the indices $l_n, n = 1, \ldots, k+1$ are given to us. Of course, we must choose them somehow. How? The question does not possess a unique suitable answer, though. We tackle this problem below.

III. SELECTION OF INDICES

The problem of deciding on the indices $l_n, n = 1, \ldots, k$ to be considered in the construction of the $|p^{1/2(k)}\rangle$ distribution is far from being a simple one. One would like, of course, to choose the smallest set of indices allowing to minimize the distance between the observed vector and the physical model. Unfortunately, as already mentioned the search for a global minimum is an NP-hard problem in most cases. A sensible simplification is obtained by resigning the goal of global minimization and accepting a less ambitious suboptimal solution which arises from the following iterative procedure: At each iteration the indices obtained in the previous steps are fixed, and a new index is chosen so as to minimize the distance between the data vector and the vector predicted by the physical model. This is basically the strategy of the forward selection approach proposed in Refs. [7,8]. Such strategy, useful indeed in many situations, is just one among the many possible suboptimal strategies that one can envisage. Here we advance an approach which is built out of two main ingredients: (i) a data-independent technique for selecting constraints to be discussed in Sec. III B; and (ii) a backward selection approach for reducing the number of parameters of a given distribution. To address the latter we need a technique evolving in the reverse direction with respect the forward technique of Refs. [7,8]. In this case the two challenges we have to face are the following.

(a) The one of deciding on the parameters to be eliminated.

(b) The one of appropriately modifying the parameters one wishes to retain.

These two points are addressed in Sec. III C by recourse to a backward birthogonalization approach. Before advancing our strategy we would like to illustrate how the forward selection approach of Refs. [7,8], can be adapted in a straightforward manner in order to make it suitable when dealing with very noisy data. This is the subject of Sec. III A.

A. Data-dependent selection criterion

As proposed in Refs. [7,8], a set of subindices $l_n, n = 1, \ldots, k+1$ can be iteratively determined by selecting, at iteration $k+1$, the index l_{k+1} corresponding to a vector $|\alpha_{l_{k+1}}\rangle_\mu$ [cf. Eq. (16)] that minimizes the norm of the residual resulting when approximating the observed data by the
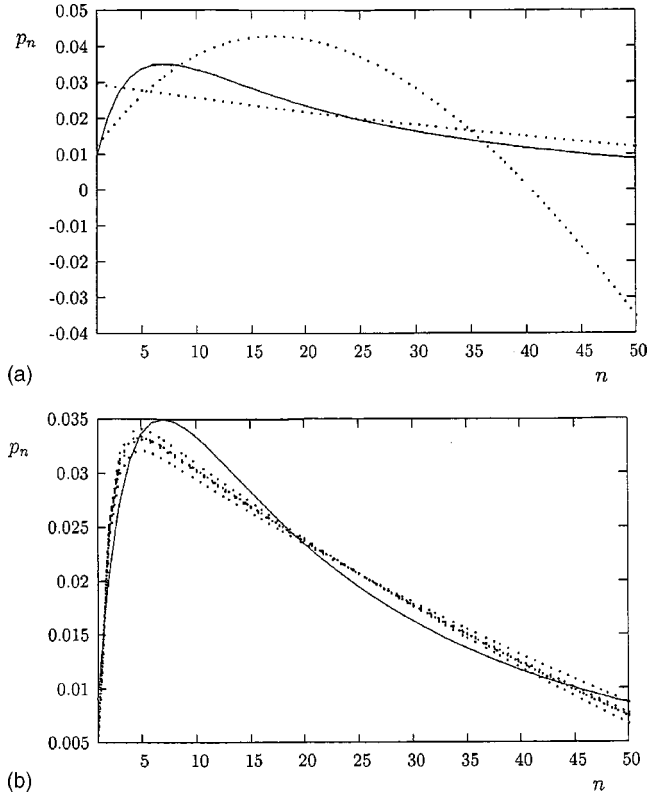
FIG. 1. (a) The theoretical distribution is represented by the solid line. Each dotted line corresponds to the approximation we obtain by using a uniform measure ($\mu=1$) for two different realization of the data. (b) The theoretical distribution is represented by the solid line. Each dotted line corresponds to the approximation we obtain (for five different realizations of the experiment) by weighting each data with a measure $\mu_i=\sigma_i^{-2}$.

physical model. This process is tantamount to selecting the index l_{k+1} that maximizes the functionals [7]

$$e_n = \left| _\mu \langle \widetilde{\psi}_n | \widetilde{f}^o \rangle_\mu \right|^2, \quad n = 1, \dots, M \quad (23)$$

with $|\widetilde{\psi}_n\rangle_\mu = |\psi_n\rangle_\mu / \| |\psi_n\rangle_\mu \|$ and $|\psi_n\rangle_\mu = |\alpha_n\rangle_\mu - \hat{P}_{V_k}|\alpha_n\rangle_\mu$.

At this point, we would like to illustrate the advantage of allowing different weights for each data. We use the same example as in Ref. [7], i.e., the data are generated as

$$f_i^o = \sum_{n=1}^{50} p_n f_{i,n} + \epsilon_i, \quad i = 1, \dots, 100 \quad (24)$$

with p_n represented by the continuous line of Fig. 1 and $f_{i,n} = \exp(-n x_i), x_i = 0.01i, i = 1, \dots, 100, n = 1, \dots, 50$. This is an extremely bad conditioned problem. In order to have a good approximation of the distribution of Fig. 1, it was assumed in Ref. [7] that we know the data within an uncertainty of 0.1%. Here we consider the errors to be much larger. Each data is distorted by a zero mean Gaussian distributed random variable of variance σ_i^2 corresponding to 20% of the data value. If, as in Ref. [7], we consider a uniform measure ($\mu=1$), the approximation we obtain is represented by the dotted lines of Fig. 1(a) (for two different realizations of the data). As we clearly gather from Fig. 1(b),

by considering a nonuniform measure given as $\mu_i=\sigma_i^{-2}, i=1,\dots,100$ the approximation is enormously improved and becomes stable against different realization of the data.

B. Data-independent selection criterion

This alternative criterion for selecting indices is independent of the actual data. It is meant to speed up the posterior selection process and is grounded on the fact that redundant equations arise as a consequence of a physical model. Hence, redundancy can be detected without the actual realization of the experimental measurements. In our formalism, each constraint, the l_k one, say, is associated with a vector $|\alpha_{l_k}\rangle_\mu$. Hence the problem of discriminating linearly independent constraints is equivalent to the problem of discriminating linearly independent vectors. We address this problem by recourse to a recently introduced technique [17], which allows for a hierarchical selection giving rise to a stable inverse problem. The goal is achieved by selecting, at each step, the index l_k maximizing the ratios

$$r_n = \frac{\| |\psi_n\rangle_\mu \|^2}{\| |\alpha_n\rangle_\mu \|^2}, \quad n = 1, \dots, M. \quad (25)$$

This data-independent technique for eliminating redundancy makes the posterior data processing much faster, as the selection of indices for constructing the distribution can be carried out only on those indices rendering independent vectors. There is also room for different postprocessing strategies because, specially when the data are very noisy, the number of required Lagrange multipliers happens to be smaller than the number of indices rendering "numerical independence." One possibility is to apply the selection criterion discussed in the preceding section, but only on the preselected indices. Additional reduction of Lagrange multipliers is made possible by a backward strategy to be introduced in the following section.

C. Reducing Lagrange multipliers

As already discussed, the fact that Lagrange multipliers are associated with constraints that are selected on a step by step basis implies that at the end of the selection process, some Lagrange multipliers may have diminished relevance. To be in a position to eliminate Lagrange multipliers of little relevance, we need to develop an appropriate technique.

Consider that we wish to reduce the number k of Lagrange multipliers characterizing a $|p^{(1/2)(k)}\rangle$ distribution. Even if we know which particular parameter should be disregarded, the actual process of removing them yields a nonlinear problem. The nonlinearity follows from Eq. (15) where the Lagrange multipliers in the left-hand side of the equation are the coefficients of a linear superposition of nonorthogonal vectors. The right-hand side indicates that such a superposition is the orthogonal projection of the vector $|\widetilde{f}^o\rangle_\mu$ onto the subspace generated by vectors $|\alpha_{l_j}\rangle_\mu, j=1,\dots,k$. Thus, within the framework of this paper, the decision of eliminating some Lagrange multipliers comes along with the aim of leaving the vector orthogonal projection onto the reduced subspace. This entails that we must recalculate the

remaining Lagrange multipliers. The need for recalculating coefficients of a nonorthogonal linear expansion, when eliminating some others, is discussed in Ref. [18] where a backward biorthogonalization approach is advanced. Such a technique, which we describe next, has been devised in order to modify biorthogonal vectors so as to appropriately represent the orthogonal projector onto a reduced subspace.

Let us recall that $V_k = \text{span}\{|\alpha_{l_1}\rangle_\mu, \ldots, |\alpha_{l_k}\rangle_\mu\}$ and let $V_{k/\alpha_{l_j}}$ denote the subspace which is left by removing the vector $|\alpha_{l_j}\rangle_\mu$ from V_k, i.e,

$$V_{k/\alpha_{l_j}} = \text{span}\{|\alpha_{l_1}\rangle_\mu, \ldots, |\alpha_{l_{j-1}}\rangle_\mu, |\alpha_{l_{j+1}}\rangle_\mu, \ldots, |\alpha_{l_k}\rangle_\mu\}.$$
(26)

We have already discussed how to construct the orthogonal projector onto V_k [cf. Eq. (21)]. In order to represent the orthogonal projector onto the reduced subspace $V_{k/\alpha_{l_j}}$ the corresponding biorthogonal vectors $|\tilde{\alpha}_{l_n}^k\rangle_\mu$ need to be modified as established by the following theorem.

Theorem 1. Given a set of vectors $|\tilde{\alpha}_{l_n}^k\rangle_\mu, n=1, \ldots, k$ biorthogonal to vectors $|\alpha_{l_n}\rangle_\mu, n=1, \ldots, k$ and yielding a representation of $\hat{P}_{V_k}$ as given in Eq. (21), a new set of biorthogonal vectors $|\tilde{\alpha}_{l_n}^{k/j}\rangle_\mu, n=1, \ldots, j-1, j+1, \ldots, k$ yielding a representation of $\hat{P}_{V_{k/\alpha_{l_j}}}$ as given by

$$\hat{P}_{V_{k/\alpha_{l_j}}} = \sum_{\substack{n=1 \\ n\neq j}}^{k} |\alpha_{l_n}\rangle_{\mu\mu}\langle\tilde{\alpha}_{l_n}^{k/j}| = \sum_{\substack{n=1 \\ n\neq j}}^{k} |\tilde{\alpha}_{l_n}^{k/j}\rangle_{\mu\mu}\langle\alpha_{l_n}| \qquad (27)$$

can be obtained from vectors $|\tilde{\alpha}_{l_n}^k\rangle_\mu, n=1, \ldots, k$ through the following equations:

$$|\tilde{\alpha}_{l_n}^{k/j}\rangle_\mu = |\tilde{\alpha}_{l_n}^k\rangle_\mu - \frac{|\tilde{\alpha}_{l_j}^k\rangle_{\mu\mu}\langle\tilde{\alpha}_{l_j}^k|\tilde{\alpha}_{l_n}^k\rangle_\mu}{\||\tilde{\alpha}_{l_j}^k\rangle_\mu\|^2}, \quad n=1, \ldots, j-1,$$

$$j+1, \ldots, k. \qquad (28)$$

The proof of this Theorem, as well as the proof of the Corollary 2 below, are given in Refs. [16,18].

Corollary 1. Let the Lagrange multiplier vector $|\lambda^k\rangle_\mu$ satisfying Eq. (15) be given. Then, the Lagrange multiplier vector $|\lambda^{k/j}\rangle_\mu$ giving rise to the orthogonal projector onto the reduced subspace $V_{k/\alpha_{l_j}}$ is obtained from the previous $|\lambda^k\rangle_\mu$ as follows:

$$_\mu\langle l_n|\lambda^{k/j}\rangle_\mu = {}_\mu\langle l_n|\lambda^k\rangle_\mu - \frac{{}_\mu\langle\tilde{\alpha}_{l_n}^k|\tilde{\alpha}_{l_j}^k\rangle_{\mu\mu}\langle l_j|\lambda^k\rangle_\mu}{\||\tilde{\alpha}_{l_j}^k\rangle_\mu\|^2}. \qquad (29)$$

The proof trivially stems from Eq. (15) using Eq. (28) in Eq. (27), since $\hat{P}_{V_{k/\alpha_{l_j}}}|\tilde{f}^o\rangle_\mu = \sum_{\substack{n=1 \\ n\neq j}}^{k} |\alpha_{l_n}\rangle_{\mu\mu}\langle l_n|\lambda^{k/j}\rangle_\mu$ implies $_\mu\langle\tilde{\alpha}_{l_n}^{k/j}|\tilde{f}^o\rangle_\mu = {}_\mu\langle l_n|\lambda^{k/j}\rangle_\mu$.

Corollary 2. The following relation between $\|\langle\hat{P}_{V_k}|\tilde{f}^o\rangle_\mu\|$ and $\|\hat{P}_{V_{k/\alpha_{l_j}}}|\tilde{f}^o\rangle_\mu\|$ holds:

$$\|\hat{P}_{V_{k/\alpha_{l_j}}}|\tilde{f}^o\rangle_\mu\|^2 = \|\hat{P}_{V_k}|\tilde{f}^o\rangle_\mu\|^2 - \frac{|_\mu\langle l_j|\lambda^k\rangle_\mu|^2}{\||\tilde{\alpha}_k^l\rangle_\mu\|^2}. \qquad (30)$$

Corollary 1 gives us a prescription to modify the Lagrange multipliers characterizing a k-parameters distribution, if one of such multipliers is to be removed. Nevertheless, still the question has to be addressed as to how to choose the Lagrange multiplier to be disregarded. Corollary 2 suggests how the selection can be made optimal. The following proposition is in order.

Proposition 1. Let the Lagrange multipliers $_\mu\langle l_n|\lambda^k\rangle_\mu, n=1, \ldots, k$ and $_\mu\langle l_n|\lambda^{k/j}\rangle_\mu, n=1, \ldots, j-1, j+1, \ldots, k$ be obtained from Eqs. (15) and (29), respectively. The Lagrange multiplier $_\mu\langle l_j|\lambda^k\rangle_\mu$ to be removed for minimizing the norm of the residual error $|\Delta\rangle_\mu = \hat{P}_{V_k}|\tilde{f}^o\rangle_\mu - \hat{P}_{V_{k/\alpha_{l_j}}}|\tilde{f}^o\rangle_\mu$ is the one yielding a minimum value of the quantities

$$\frac{|_\mu\langle l_j|\lambda^k\rangle_\mu|^2}{\||\tilde{\alpha}_{l_j}^k\rangle_\mu\|^2}, \quad j=1, \ldots M. \qquad (31)$$

Proof. Since on the one hand $\hat{P}_{V_k}\hat{P}_{V_{k/\alpha_{l_j}}} = \hat{P}_{V_{k/\alpha_{l_j}}}\hat{P}_{V_k} = \hat{P}_{V_{k/\alpha_{l_j}}}$ and on the other hand orthogonal projectors are idempotent we have

$$\|\hat{P}_{V_k}|\tilde{f}^o\rangle_\mu - \hat{P}_{V_{k/\alpha_{l_j}}}|\tilde{f}^o\rangle_\mu\|^2 = {}_\mu\langle\tilde{f}^o|\hat{P}_{V_k}|\tilde{f}^o\rangle_\mu - {}_\mu\langle\tilde{f}^o|\hat{P}_{V_{k/\alpha_{l_j}}}|\tilde{f}^o\rangle_\mu$$

$$= \|\hat{P}_{V_k}|\tilde{f}^o\rangle_\mu\|^2 - \|\hat{P}_{V_{k/\alpha_{l_j}}}|\tilde{f}^o\rangle_\mu\|^2. \qquad (32)$$

Making use of Eq. (30), we further have

$$\|\hat{P}_{V_k}|\tilde{f}^o\rangle_\mu - \hat{P}_{V_{k/\alpha_{l_j}}}|\tilde{f}^o\rangle_\mu\|^2 = \frac{|_\mu\langle l_j|\lambda^k\rangle_\mu|^2}{\||\tilde{\alpha}_{l_j}^k\rangle_\mu\|^2}. \qquad (33)$$

It follows then that $\|\hat{P}_{V_k}|\tilde{f}^o\rangle_\mu - \hat{P}_{V_{k/\alpha_{l_j}}}|\tilde{f}^o\rangle_\mu\|^2$ is minimum if $|_\mu\langle l_j|\lambda^k\rangle_\mu|^2/\||\tilde{\alpha}_{l_j}^k\rangle_\mu\|^2$ is minimum.

Successive applications of criterion (31) lead to an algorithm for recursive backward approximations of the distribution. Indeed, let us assume that at the first iteration we eliminate the jth constraint yielding a minimum of Eq. (31). We then construct the new reciprocal vectors (28) and the corresponding new Lagrange multipliers as prescribed in Eq. (29). The process is to be stopped if the approximated distribution fails to predict the observed data within the required margin.

D. Numerical example

We illustrate here a strategy consisting of the following steps
(i) We use the data-independent selection criterion for discriminating independent constraints.
(ii) We apply the data-dependent selection criterion on the previously selected indices.
(iii) The number of Lagrange multipliers obtained at step (ii) is reduced and the remaining multipliers recomputed.
We consider the example described below.
The physical model yielding the matrix elements $f_{i,n}$ is given by the Lorentzian decays:
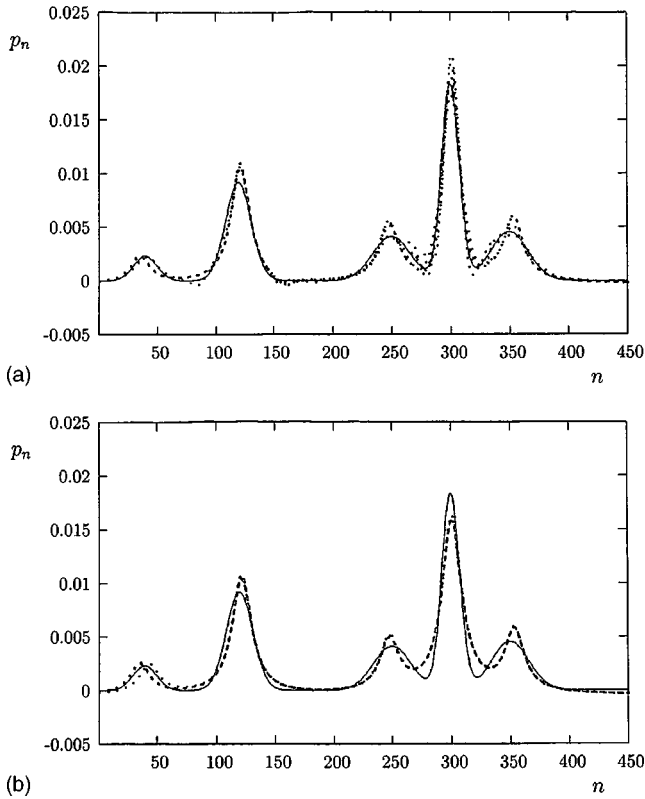
(a)

(b)

FIG. 2. (a) The theoretical distribution is represented by the solid line. The dotted lines correspond to the approximation we obtain for five different realizations of the data. Each line is constructed by iteratively selecting constraints out of the reduced set obtained by the data-independent technique. (b) The theoretical distribution is represented by the solid line. Each dotted line represents the approximation of the corresponding one in (a), after the elimination of some parameters.

$$f_{i,n} = \frac{1}{1 + 0.01(i - 100 - n)^2}, \quad i = 1, \ldots, 700,$$

$$n = 1, \ldots, 450. \tag{34}$$

We construct 700 vectors $|\alpha_n\rangle_\mu, n=1,\ldots,700$ as prescribed in Eq. (16) and select indices corresponding to the linearly independent vectors by the technique of Sec. III B for eliminating redundancy. Out of the redundant set of 700 vectors we found 100 linearly independent ones, up to a good precision, which is assessed by the biorthogonality quality of the corresponding basis and its reciprocal (dual).

The experimental measures were generated considering that the distribution characterizing the physical system is the sum of five Gaussian functions represented by the continuous line of Fig. 2. Each data was distorted by a random error of variance σ_i^2 corresponding to 10% of the data value. A realization of these data is shown in Fig. 3. The inversion problem in this example is much more stable than the one of the previous example so that the results do not vary much by weighting the data. Hence, in order to illustrate this strategy we use a uniform measure in all the involved procedures. Out of the preselected linearly independent vectors, by using the data-dependent strategy, we selected between 8 and 12
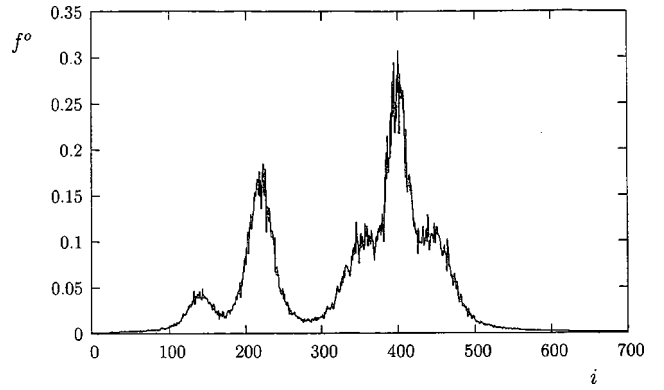

FIG. 3. The simulated data after distortion by random noise.

indices (depending on the particular realization of the data) to be able to predict the 700 pieces of data within the uncertainty up to which the data were generated, i.e., we require that $\||f^p\rangle_\mu - |f^o\rangle_\mu\|^2 < \||\epsilon\rangle_\mu\|^2$, where $|\epsilon\rangle_\mu$ is a vector of components $\epsilon_i = t\sigma_i$ where, in general, t is real number in the interval [1,3]. In this case we first set $t=1$. The approximation of the corresponding distribution is depicted by the dotted lines of Fig. 2(a) (for five different realizations of the data). We then increased the value of t up to $t=2$ and applied the proposed strategy for reducing Lagrange multipliers. In spite of the fact that the number of parameters was significantly reduced (only five were kept) as it can be seen in Fig. 2(b) the distribution is still a good approximation of the original one. The inference to the data by this distribution is also of great quality. As shown in Fig. 4, the predicted data are really close to the noiseless ones. Notice that, by recourse to our approach, we are able to denoise and compress 700 data by using only five Lagrange multipliers.

IV. CONCLUSIONS

In this paper, we have considered the problem of constructing the $q=1/2$ maximum entropy distribution from redundant and noisy data. A previously developed approach has been generalized here in order to be able to incorporate, in a straightforward manner, *information on the data errors*.
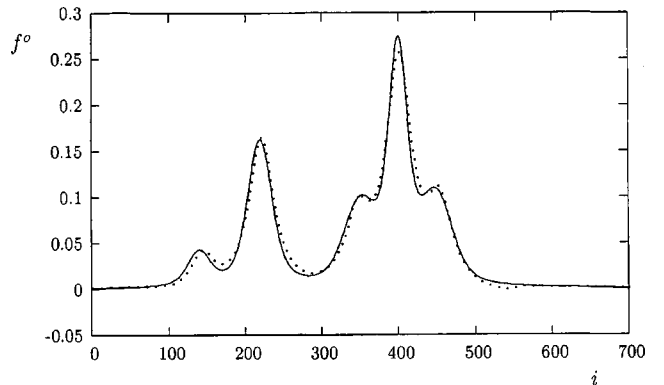

FIG. 4. The theoretical data are represented by the continuous line. The dotted line corresponds to the predictions obtained by means of the approximation of Fig. 2(b).

The advantage of this generalized approach, when dealing with very noisy data, has been illustrated by a numerical simulation.

Additionally, a strategy for selecting relevant constraints has been advanced. The corresponding implementation consists of two different steps. The first step is independent of the actual data, as it operates by discriminating independent equations *on the basis of the physical model*. The data are used, *a posteriori*, to reduce further the number of constraints. The latter process is carried out through a forward and backward procedure as follows: First the selection is made starting from an initial constraint and incorporating others, one by one, till the observed data are predicted within a predetermined precision. Afterwards, the number of parameters of the distribution is reduced further by applying a backward selection criterion for eliminating some of the Lagrange multipliers and recalculating the remaining ones. It should be stressed that the combination of the forward and backward procedures is not, in general, equivalent to stopping the forward approach at a corresponding earlier stage. The irreversibility of the process is a consequence of the fact that, due to the complexity of the problem, the implementation of a selection criterion aiming at global optimization is not possible. The strategies we have presented here are only optimal at each operational step. Hence, they do not generate reversible procedures.

Considering the complexity of the mathematical problem which is posed by the aim of constructing, in an optimal way, the $q=1/2$ maximum entropy distribution from redundant and noisy constraints, we believe that the well founded suboptimal strategies we have employed here should be of utility in a broad range of situations.

[1] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).

[2] C. Tsallis, Fractals **6**, 539 (1995), and references therein.

[3] A. R. Plastino and A. Plastino, Phys. Lett. A **177**, 177 (1993).

[4] B. M. R. Boghosian, Phys. Rev. E **53**, 4754 (1996).

[5] L. Rebollo-Neira, A. Plastino, and J. Fernandez-Rubio, Physica A **258**, 458 (1998).

[6] L. Rebollo-Neira, J. Fernandez-Rubio, and A. Plastino, Physica A **261**, 555 (1998).

[7] L. Rebollo-Neira and A. Plastino, Phys. Rev. E **65**, 011113 (2002).

[8] L. Rebollo-Neira and A. Plastino, Phys. Rev. E **66**, 032102 (2002).

[9] B. R. La Cour and W. C. Schieve, Phys. Rev. E **62**, 7494 (2000).

[10] S. Smale, Math. Intell. **20**, 7 (1998).

[11] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000).

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

[13] L. Rebollo-Neira, A. G. Constantinides, A. Plastino, A. Alvarez, R.Bonetto, and M. Iñiguez Rodriguez, J. Phys. D **30**, 2462 (1997).

[14] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory* (Wiley, Chichester, 1994).

[15] L. Rebollo-Neira, e-print mathph/0209026.

[16] L. Rebollo-Neira, *Progress in Mathematical Physics* (Nova Science, New York, to be published).

[17] L. Rebollo-Neira, IEE Proc. Vision Image Signal Process. **51**, 31 (2004).

[18] L. Rebollo-Neira, Int. J. Math. Math. Sci. **35**, 1843 (2004).