Neuronal algorithms for full information spectral analysis

Vincent Vigneron, Alexandre Fayolle, Jean-Marc Martinez, Anne-Catherine Simon, Robert Junca, Jean-Louis Szabo, Christine Fuche

Neuronal algorithms for full information spectral analysis.

Vincent Vigneron[1,2], Alexandre Fayolle[1], Jean Marc Martinez[1], Anne Catherine Simon[1], Robert Junca[1], Jean-Louis Szàbo[1], Christine Fuch[2]

[1] CEA, Centre d'Etudes de Saclay, 91191 Gif-sur-Yvette, France
[2] CREL, 161, rue de versailles, 78180 Le Chesnay, France

Abstract. *In* Pattern Recognition *the input items have to be identified under various transformations of their representations. Contemporary neural-networks research concentrates mostly on decision making systems, whereas the fundamental functions associated with the preprocessing of observations have often been ignored. This paper is a step toward theories that are expected to help the emergence of invariant-features.*
key-words : data structure analysis, modelling, feature extraction, fractal dimension.

1 Introduction

Pattern recognition and *discrimination* is about guessing the unknown nature of an observation. This observation is just the collection of some numerical measurements such as an image, time series, or a digitized signature. More formally, an *observation* is a d-dimensional vector x upon which the statistician, the biologist, the physicist, the neuro-scientist design a decision-making process.

Most pattern recognition research has been concentrated on complicated mechanisms for decision-making processes (*Bayes classifier, nearest neighbor rule*, ...) [**?**], which are not the purpose of this article. We present here some advanced mathematical tools necessary for simplifying/solving them. We concentrate on realistic problems such as $X-$ray fluorescence analysis and uranium enrichment measurements. Although many classical approaches have been proposed to formulate the decision-making process, they are outside the scope of this publication.

2 The Learning Theory point of view on feature extraction

Up to recent days, feature extraction has been mostly considered a *supervised* process of (linear filters) mapping the original measurements into more effective features so as to minimize a criterion, assuming that the variables are already selected and given [**?**]. From an experimental device, the physicist gets some measurements, spoiled by noise and some determinist distortions. The "problem" is then to seek the "good" values of a "number" of "interesting" parameters. But, neither "good", nor the "number", nor "interesting" are clearly defined notions. Frequently, the physicist is unable to write the mathematical equations of the observed phenomenon. He hopes that usual recipes called *Fourier transform, deconvolution, least squares,...* will produce shining revelations[**?**]. Of course, these recipes are well-known and their honorability well established, sometimes with a name of a mathematician as a quality-label. Moreover, they save time, which can give a major advance on the other challengers. The author wants to underline thatin order to undertake experimental data, anyone has to consider

1. the equations first,
2. the exhaustive list of the hypotheses,
3. never letting the implicitely chosen method define the problem.

The reader will not see amount of mathematical equations, most of the mathematical background necessary to make use of the contents of this article is in [**?**]. Our basic knowledge and way of thinking is in terms of probabilities, which provide ways to (1) intuitively capture the most informative representation of the signal, (2) validate hypotheses, (3) contradict or not the experimental verdict.

In this context, the *Learning Theory approach* offers a great potential for archieving optimal solutions of complex real world problems, because it deals with *undefined* knowledge which is in the mind of the physicist before he carries out the experiment : non-linear correlations, hidden dependencies,... The latter suggestion is not a revelation, anyone can read the proof of it in the recent litterature concerning Learning Theory [**?**, **?**, **?**, **?**]. In particular, it was shown in Statistical Learning Theory that by taking into account the size of sample one can obtain better solutions to many problems than by using classical methods [**?**, **?**, **?**] These questions are complex and very problem-dependant, but we focuse on a specific one : *ill-conditioned problems* when the physicist has not a sufficient amount of experimental data.

3 Extremely ill-posed problems

Let us consider a physical problem with a set of p samples $\{x_i\}_{i=1}^{p}, x \in \Re^N$, N being the dimension of these samples. The *extremelly ill-posed problem* occurs when $p \ll N$. Suppose you want to undertake the global information contained in each sample. If the model is overly restrictive, it cannot "capture the rule", and, on the other hand, if it has a too high capacity, the model will likely be unable to generalize. Learning Theory tells us that the most economical model – in terms of free parameters – is often to be the best, from the performance point of view [**?**]. In neural networks applications, one often faces such singular learning problem where the data set consist of a relatively small number of high-dimensional input vectors. The basic idea is to transpose the problem from the high-dimensional input space to a low-dimensional "signal space". In this very case, the success of the transformation depends on a weak hypothese of strong dependencies between the components of the input vector. But we have shown for toy-problems [**?**] that this approach can be generalized to other cases. In particular, we propose 3 variants of feature extraction using Learning Theory concepts:

1. non-orthonormalized PCA (called *Sanger's rule*) [**?**],
2. *sharing weights* Hansen's neural network [**?**],
3. VQP (Vector Quantization Projection) [**?**].

These schemes are *unsupervised learning* methods ; for the pre-cited reasons, the results included may varied for problem to an other.

4 Mathematical tools for ill-posed problems

4.1 Non-orthonormalized PCA

A classical method for addressing the problem of *linear* feature extraction is the well-known statistical tool, *Principal Component Analysis* (PCA), which had been introduced first by Karhunen-Loeve. Two main aspects capture our attention :

- first, the decorrelation of the covariance matrix by PCA is equivalent to linear extraction of statistically independant features,
- second, PCA provides the orthonormal basis which plays a important role in the data *reconstruction/compression* problem.

The *Sanger's rule* (see Fig. 1) generalizes the standard PCA method based on the diagonalization of covariance matrix by formulating the problem in a more general way, i.e. by using a cost function. The minimization of the criterion leads to the desired PCA solution. It establishes the connection between statistical properties and Learning Theory ideas.

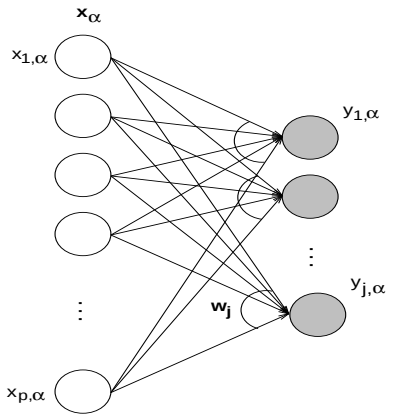

Fig. 1. A PCA network is a one-layer feed-forward neural network which is able to extract the eigencomponents of the stream of input vectors. It can induce the emergence of *invariant-features*. The non-linear caracteristic of the activation functions explain the (asymptotic) stability and strength of the solution.

The weights of the Sanger network are the eigenvectors (non-orthonormalized basis) and the output the projection of the learning example vector.

4.2 *Sharing weights* Hansen's neural network

Discussing PCA, therefore, it is wise (and in this case practical) to spend some time thinking a correct reformulation of the learning problem. Unreasonnable results may come from a violation of the natural physical *a priori* knowledge. This could be avoided by using *constraints* or by changing the parameters used to describe the phenomenon :

both strategies are combined with constrained weights networks by considering the linear subspace spanned by the actual inputs of the training set. The weights vector w are restricted to fall within the signal space $\mathcal{S}$ writing

$$w = \sum_{i=1}^{p} \gamma_i x_i.$$

with suitable parameters γ_i. Assuming the inputs are linearly independent, the coefficients are unique and computed iteratively using classical training scheme like *Backpropagation*[**?**], based on cost function[3], for

[3] depending the nature of the statistical noise, i.e. poissonian, gaussian, gamma, binomial [**?**].

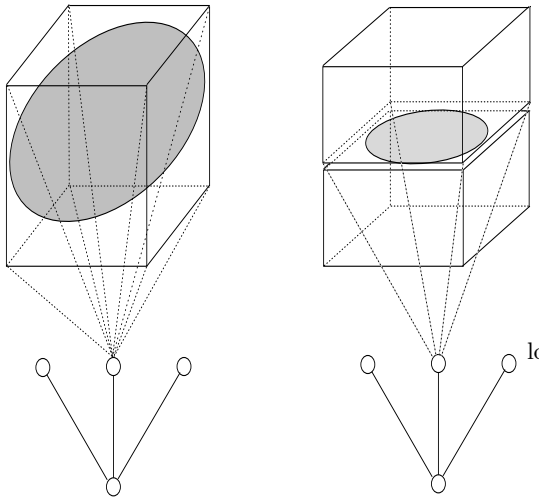

Fig. 2. *sharing weights* Hansen's network architecture (a 3-layered feed-forward perceptron with N inputs, and 3 neurons on the hidden layer).

example

$$\boldsymbol{E_\ell = \frac{1}{p}\sum_{i=1}^{p}(y_i - f(x_i, \gamma_i))^2}, \qquad (1)$$

where y_i is the desired output/prediction. This approach is *hybrid* in the sense the extraction of features is *supervised* by a teacher (i.e. $\{x_i, y_i\}_{i=1}^{p}$), but the weights structure is decided in an *unsupervised* way by the (linear combination of the) data. The results in Table 1, in the next section, illustrate the role of the signal space projection to facilitate prediction with neural networks. The network architecture is visualized in Fig. 2. The aim of current research concern the interpretation of such weights.

4.3 VQP

By using VQP, we sacrifice deliberately *linearity* and preserve the regularity (not obviously linear) of the patterns, disregarding the noise, oscillations,... Keep in mind that PCA is only linear[4], and fails to reduce data in a non-linear way. VQP is a sort of self-organizing mapper, as the famous *Kohonen's Self Organizing Maps*, the output taking automatically the relevant shape of the data. From a intuitive point of view, VQP *unfolds* the data structure towards a low dimensional space, which dimension is the (exact and unfortunately hidden) number of degrees of freedom of the observed phenomenon (see the application on the "toy-problem" in Fig. 4). This dimension can be determined through fractal analysis of the data set (Fig. 3). In order to illustrate the performance of VQP, we give the following application in fluorescence$-X$, for the concentration prediction of a mixture of components, i.e uranium and thorium.

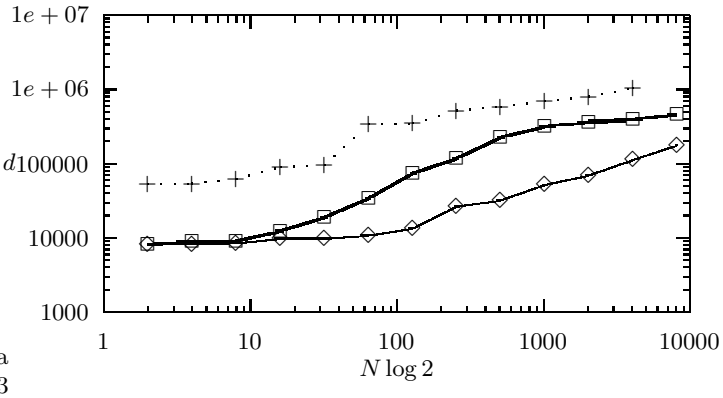

Fig. 3. Fractal dimension representation of sampling from the γ-spectra familly (where($\diamond$) is put for Nitrogene-Chlorine $\gamma-$spectra, ($\square$) is put for ^{235}U $\gamma-$spectra and ($+$) is put for $KX-$ray fluorescence spectra), computed by using $2^n (1 \leq n \leq 12)$ segments. The $\log - \log$ representation explain the real intrinsec dimension of the spectral distribution.

5 Benchmark with real-word problems

We shall give now two further examples of emergence of specific filters.

5.1 $KX-$ray fluorescence analysis problem

They are applied here to the automatic analysis of $X-$ray fluorescence spectra obtained with uranyl nitrate solutions irradiated by a collimated beam of photons emitted by a sealed source of iridium ^{192}Ir. The method could allow the surveillance of uranium quantities in nuclear fuel cycle in reprocessing plants [**?**]. The equipment now in service in the plant measures reprocessing solutions with concentration ranges of 1 to 400 $g.l^{-1}$ for uranium and 1 to 40 $g.l^{-1}$ for thorium/plutonium. These discretized spectra were used as training data. Fig 5 illustrates the dataset projection produced using VQP and PCA. The intrinsic dimensionality of the problem seems to be 2, which is precisely the number of unknowns.

5.2 Uranium-enrichment measurements

For further illustration of the ability of the VQP algorithm we show in Fig. 6 the training set and generalization set predictions. Traditional non-destructive

[4] *linearity* is rarely adapted to the shape of the parametric space to represent.

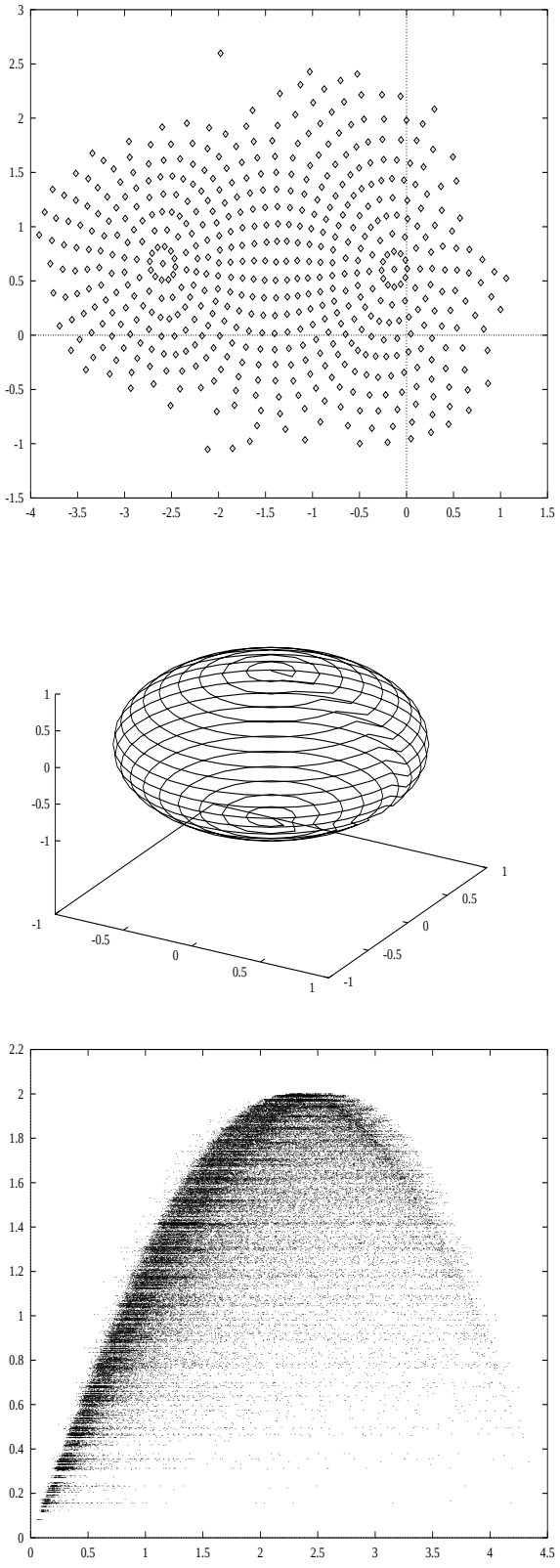

Fig. 4. "Toy-problem" of (b) *unfolding* of the spherical distribution (a). Notice that it is necessary to "cut" the surface to allow the unfolding. The "$dy-dx$" diagram (c) show that, locally, the topology is respected on contrast with medium distances, where the dispersion is important [**?**].

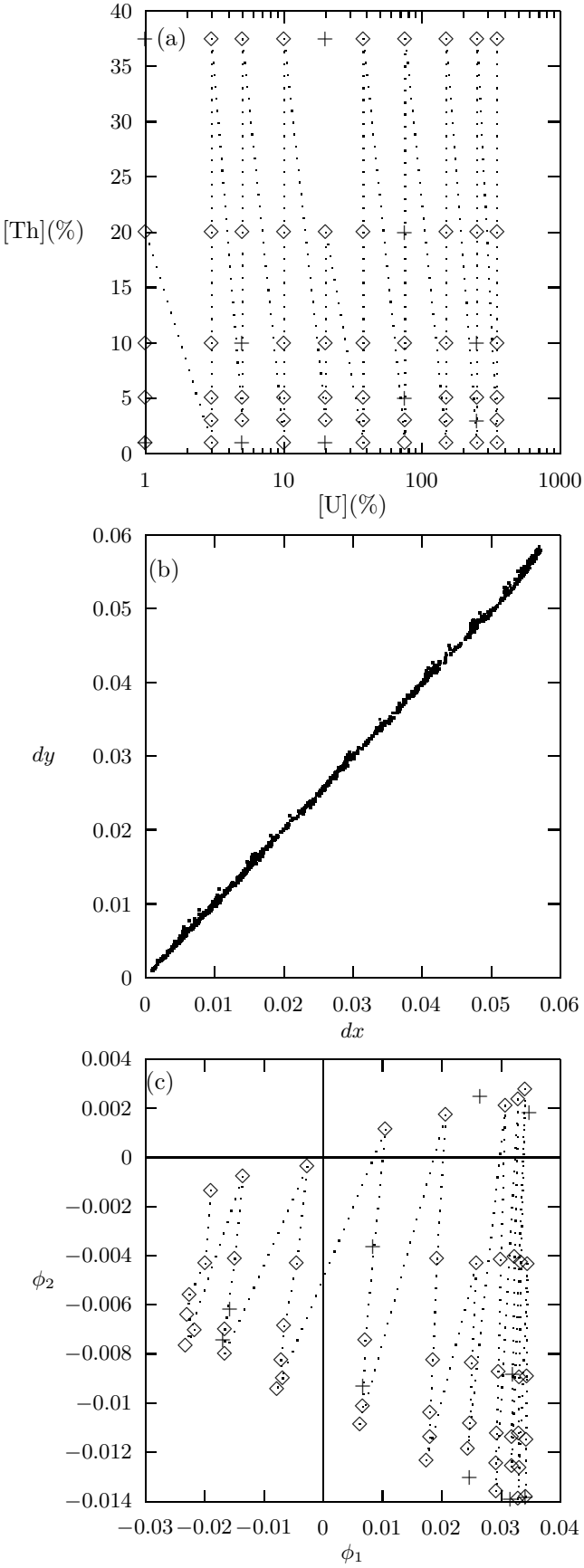

Fig. 5. Real-world problem : prediction of uranium-thorium concentration by $KX-$ray fluorescence analysis. Fig. (b) depicts the caracteristics of the 50 samples training-set, which is projected in a non-linear manner on a new mapping (a), repered by (ϕ_1, ϕ_2). (c) is the $dx - dy$ quality projection representation. The ($\diamond$) is put for a learning sample, the (+)for a test sample.

analysis during for uranium-enrichment measurement involves the use of several X- and γ-ray peaks, mainly in the 60 to 200 keV region. Most of these methods were developed more than 20 years ago, and are based on measurements of the full energy peak at 185.7 keV ([**?**, **?**, **?**]). This approach requires calibration of the system and the measurement conditions to remain constant. Other methods have been developed using several γ-ray peaks and calibration with a limited number of peaks [**?**, **?**].

Calibration procedures and matrix effects can be avoided by focusing the spectra analysis on the $K_\alpha X$ region (which contains the main uranium components) and by using infinitely thick samples. Such samples sufficiently thick that any further increase does not affect the $\gamma-$ray emissions.

The spectral processing of the $K_\alpha X$ region involves quantification of the photon emissions identified with ^{235}U, ^{238}U and X-ray fluorescence. This approach requires well-defined data for the emissions of photons, together with the characteristics and geometry of the detector.

Table 1. $X-$ray fluorescence analysis of Uranium-Thorium. Training and test results. Neuronal learning has been realised, in the very case of "VQPization" and "'PCA-ization' with a multilayer perceptron.

Feat. extract.	training error[a]	test error[b]	outputs
`Hansen`	9.3010^{-1}	4.6510^{-1}	4096
`VQP`	5.6910^{-2}	2.8410^{-2}	2
`ACP (Sanger)`	6.4810^{-2}	5.1510^{-2}	2

[a] Mean Square Error detailled in Eq. (1)
[b] idem.

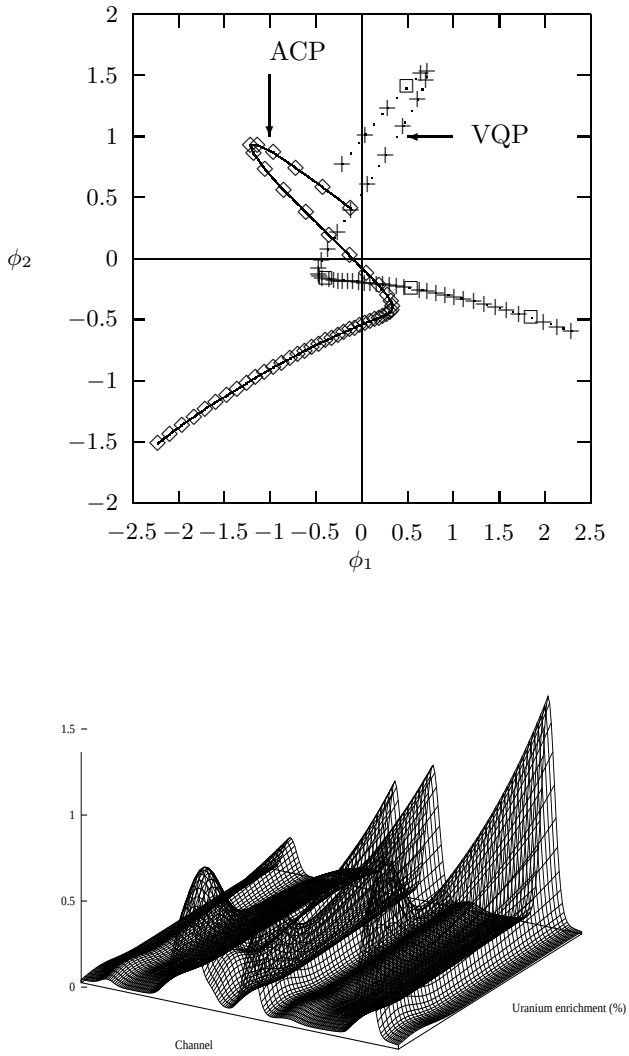

Fig. 6. (down) Learning dataset inside a more relevant mapping space (up) (ϕ_1, ϕ_2) obtained by VQP (+) or PCA ($\diamond$). The ($\square$) are obtained by *VQP-ization* of the test dataset. Notice that in the diagram the both projected dataset are *rotated* with respect to a fixed point, without translating the sampling lattice.

6 Conclusions

Learning is one solution to ill-posed problems. With such algorithms, no *a priori* law is introduced. The data are just normalized for numerical reasons. Physicist must just be aware of the asymptotic behaviour of the learning algorithm and to the exhibit features, even if it looks unreasonnable to the physicist. We have investigated the possibility to invoke neural networks for the analysis of non-linear ill-posed problems in *weights-sharing networks*, networks for so-called *PCA* and *VQP*. The proof is made it is advantageous to re-express the training set without loss of *information*. The present view is that recognition must be based on *features* that are more/less invariant with respect to different transformation.

Cognitive scientists can help the physicists to decide between both *supervised* or *unsupervised* learning, between *deterministic* or *probabilistic* networks, with a major motivation : *preserving the maximum amount of information* in *compression* or *prediction*[**?**].

As an example, we have illustrated trough this article a emerging idea[**?**]. The classical philosophy usually considers two types of inference :*deduction*, i.e. from general to particular, *induction*, i.e. from particular to general. We have described a new concept of inference : moving *from particular to partic-*

ular. Vapnik calls this type of inference *transductive inference*.

Acknowlegments
The authors wish to thank the staff of SAR and LPRI, and in particular J.L. Boutaine, J. Morel and M.C. Lepy for their support and useful discussions during these works.