Probability density function estimation using Multi-layer perceptron

Touba Mostefa Mohamed[1], Abdenacer Titaouine[1], Touba Sonia[1], Ouafae Bennis[2]

[1]Biskra University, B.P 145 RP, 07000 Biskra, Algeria mm.touba@gmail.com

[2]PRISME Institute, University of Orléans, 21 rue Loigny La Bataille, 28000 Chartres, France

ouafae.bennis@univ-orleans.fr

Abstract : The problem of estimating a probability density function (pdf) can easily be encountered in many areas of experimental physics (high energy, spectroscopy, etc.) and other fields. The standard procedure is to bin the space and approximate the pdf by the ratio between the number of events falling inside each bin over the total and normalized to the bin volume. In this paper we estimate the univariate pdf using an MLP (Multi-Layer Perceptron) where the inputs are based on the exponential model. The proposed method is very effective and estimated densities are too close to some theoretical pdfs. Our method has been integrated in the famous steepest descent algorithm for marginal score functions estimation where two linearly mixed sources were successfully separated.

Key words: Probability density estimation, Neural networks, Multilayer perceptron, BSS, Score function.

Introduction

The probability density function (pdf) is a central concept in statistical data analysis, and the most popular instruments for pdf estimation are: histograms and kernel density estimation. More information about pdf estimation can be found in (Silverman, 1986). The reader can also be referred to (Vogt, 2007) for some basic analysis techniques.
In the following we give a short introduction to some density estimation methods.

Histograms:

It is the oldest and most widely used density estimator. The data range is divided into a set of successive and non-overlapping intervals (bins). The bins of the histogram are defined as the intervals [$x_0 + mh$, $x_0 + (m+1)h$] for m positive and negative integers, x_0 is the origin and h is the bin width. For a set of n observed data points supposed to be a sample of an unknown density function p_X . The histogram is defined by:

$$\hat{p}_X(x) = \frac{number\ of\ observations\ in\ the\ same\ bin\ as\ x}{nh} \qquad (1)$$

The histogram can be generalized by allowing the bin widths to vary. Then the estimate becomes:

$$\hat{p}_X(x) = \frac{number\ of\ observations\ \ in\ the\ same\ bin\ as\ x}{n(width\ of\ bin\ containing\ x)} \qquad (2)$$

However, there are some drawbacks in using histograms:

-The histogram is not continuous so trouble arises when derivatives are required (score functions in blind source separation)
-Choice of origin may have an effect in the interpretation
-Representing multivariate data by histogram is difficult

The naive estimator:

The pdf can be defined as a probability density as:

$$p_X(x) = \lim_{h \to 0} P(x - h \leq X \leq x + h) \qquad (3)$$

Thus,

$$\hat{p}_X(x) = \frac{number\ of\ observations\ falling\ into\ [x-h,x+h]}{2hn} \quad (4)$$

By this way, $\hat{p}_X(x)$ does no longer depend on the origin of the chosen data range discretisation.
The naïve estimator can be defined clearly by a weight function as follows:

$$\hat{p}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} \omega\left(\frac{x-x_i}{h}\right) \quad (5)$$

Where

$$\omega(x) = \begin{cases} \frac{1}{2} & if \quad -1 < x < 1 \\ 0 & if \quad otherwise \end{cases} \quad (6)$$

This means that rectangular boxes of width 2h and height $\frac{1}{2hn}$ are placed around each datum and then summed up to get the estimate $\hat{p}_X(x)$.
But this estimator also has got some drawbacks:

$\hat{p}_X$ is not continuous but has jumps at the points $x_i \pm h$ and has zero derivative everywhere else

The kernel estimator:

This estimator is obtained by replacing the weight function in the expression of the naïve estimator by a kernel function $K(x)$ which satisfies:

$$\int_{-\infty}^{+\infty} K(x)\,dx = 1$$

Then the kernel estimator is given by

$$\hat{p}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \quad (7)$$

Here h is the smoothing parameter. It controls the trade-off between the statistical significance of the pdf estimate and its effective resolution.

Traditionally and statistically, the pdf is constructed by locating a Gaussian kernel at each observed datum, e.g., the fixed-width kernel density estimator (FKDE) and the adaptive kernel density estimator (AKDE). Although the FKDE, which constructs a density by placing fixed width kernels at all of the observed data, is widely used for nonparametric density estimation, this method normally suffers from several practical drawbacks (Silverman, 1986).

Neural networks for pdf estimation

To overcome the problem of high cost in computation and memory storage of the kernel estimator, a clustered radial basis function (RBF) based kernel density estimator, named RBF network, can be used (Hwang, Lay and Lippman,1993, Popat and Picard, 1993; Popat and Picard, 1994). The RBF network uses a reduced number of radial basis kernels, with each kernel being representative of a cluster of training data, to approximate the unknown density function. This method is often referred as mixture (Gaussian) modeling (Rabiner, 1989). These networks are also widely used in regression and classification applications (Moody & Darken, 1989).

The use of feedforward neural networks (Svozil, Kvasnicka & Pospichal, 1997) with sigmoid hidden units called multilayer percceptrons (MLPs) for pdf estimation was proposed in (Modha & Fainman, 1994), where the training approach based on the minimization of the negative log-likelihood is described. However, the pdf approximation capabilities of general multilayer feedforward neural networks have been established by White (1992).

It is well known that the gaussian mixture approach encounters difficulties in approximating the uniform distribution. This is not the case for the MLP model. Likas (2001) have presented an approach of pdf estimation based on the use of feedforward multilayer neural networks with sigmoid hidden units. The method is based on numerical integration technique.

In this paper we estimate the univariate pdf using an MLP (Multi-Layer Perceptron) where the inputs are based on the exponential model. The proposed method is very effective compared to some theoretical pdfs.

Problem formulation

For problems defined in $\mathcal{R}^p$, the network architecture (Fig.1) consisted of p input units, one hidden layer with H hidden units having the logistic activation function and of one output unit with exponential activation function (Modha & Fainman, 1994) :

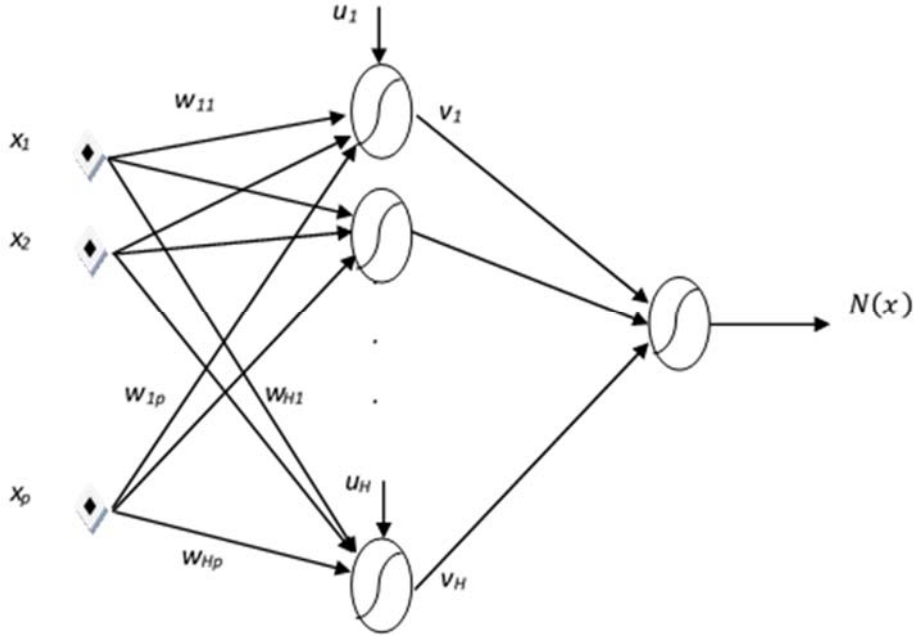


Figure 1: Basic MLP approach to pdf estimation

Let $x(k) \in \mathcal{R}^p$, $(k = 1, \dots, n)$ be a set of *n* data points drawn independently according to an unknown density $f(x)$ that we want to approximate, and let's define the model of pdf with parameter $\boldsymbol{\theta}$ by the function

$$p_N(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{N(\boldsymbol{x}, \boldsymbol{\theta})}{\int_{\mathcal{R}^p} N(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{y}} \tag{8}$$

In the paper of Modha & Fainman (1994), the parameter vector $\boldsymbol{\theta}$ is adjusted by minimizing the function:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{k=1}^{n} ln\{p_N(\boldsymbol{x}(k), \boldsymbol{\theta})\} \tag{9}$$

Replacing $p_N(\boldsymbol{x}, \boldsymbol{\theta})$ in (9) by its expression (8) we obtain:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= -\sum_{k=1}^{n} ln\{N(\boldsymbol{x}(k), \boldsymbol{\theta})\} + n\, ln\left\{\int_{\mathcal{R}^p} N(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x}\right\} \\ &= -\sum_{k=1}^{n} ln\{N(\boldsymbol{x}(k), \boldsymbol{\theta})\} + n\, ln(I_{\boldsymbol{\theta}}) \end{aligned} \tag{10}$$

With

$$I_{\boldsymbol{\theta}} = \int_{\mathcal{R}^p} N(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} \tag{11}$$

The key idea in the algorithm of . Likas (2001), is the numerical integration technique used to compute (11).

Our work consists of estimating the monovariate pdf modeled by an exponential law. Hence, the equation (8) becomes:

$$p_d(x,\delta) = \frac{N_e(x,\delta)}{\int_a^b N_e(y,\delta)dy} = \frac{exp(\delta_1 x + \cdots + \delta_d x^d)}{\int_a^b exp(\delta_1 y + \cdots + \delta_d y^d)dy} \quad (12)$$

Where

$[a,b] = [\min x, \max x]$, d is the model order ($d = 1, 2, \ldots$)

$(\delta_i, i = 1, \ldots, d)$ are the model parameters.

This model has the following advantage:

- Most common densities (normal, uniform, exponential …etc.) are well fitted by the exponential densities (Ould Mohamed, 2012, p.54)

In Fig. 2, inputs in Fig. 1 are changed according to the model order so that the MLP architecture becomes:

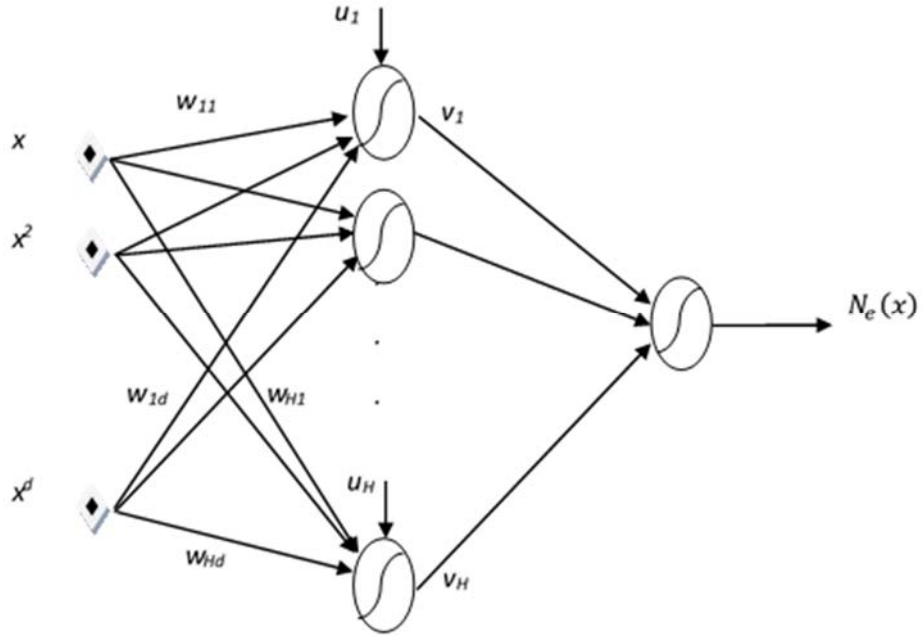


Figure 2: MLP architecture for exponential pdf modeling and estimation

A supervised training of the MLP is performed by constructing a training set using some non-parametric technique for pdf estimation. This means that, for the training set x_k, we have selected the histogram estimation method.

Application to blind source separation

Blind signal separation (BSS) and independent component analysis (ICA) are emerging techniques of array processing and data analysis that aim to recover unobserved signals or "sources" from observed mixtures (typically, the output of an array of sensors), exploiting only the assumption of mutual independence between the signals, more details can be found in (Hyvärinen, Karhunen and Oja, 2001; Jutten and Comon, 2007).

In instantaneous case, BSS becomes the problem of identifying the probability distribution of a vector $\boldsymbol{x} = \boldsymbol{As}$, given a sample distribution. In this perspective, the statistical model has two components: the mixing matrix $\boldsymbol{A}$ and the probability distribution of the source vector. the main idea is to find a matrix $\boldsymbol{B}$ (separating matrix) such that the components of the vector $\boldsymbol{y} = \boldsymbol{Bx}$ are mutually statistically independent.
Mutual information, $I(.)$, is a measuring criterion for designing a system which generates independent outputs.
If, $\boldsymbol{x} = \boldsymbol{As}$, where

$\boldsymbol{s} = (s_1, s_2, \ldots, s_p)^T$, source signals , $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$, observed signals

and

$\boldsymbol{A} = [a_{ij}]$, mixing matrix

Then

$\boldsymbol{B}$ is estimated by minimizing the mutual information $I(\mathbf{y})$ of $\boldsymbol{y} = \boldsymbol{Bx}$

$$I(\mathbf{y}) = \int_y p_{\mathbf{y}}(\mathbf{y}) ln\left\{\frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_{i=1}^{p} p_{y_i}(y_i)}\right\} d\mathbf{y} \quad (13)$$

Where,

$p_{\mathbf{y}}(\mathbf{y})$ is the joint pdf of vector $\mathbf{y}$, $p_{y_i}(y_i)$ is the marginal pdf of y_i

It is well-known that $I(y)$ is always non-negative and vanishes if and only if the y_i's are independent. Consequently, the parameters of the separating system can be calculated based on minimization of the mutual information of the outputs. It is very helpful to know an expression for the gradient of the mutual information. However, the gradient of the mutual information, $\frac{\partial I(\mathbf{y})}{\partial \boldsymbol{B}}$, can be expressed (Taleb and Jutten, 1999) by the following expression:

$$\frac{\partial I(\mathbf{y})}{\partial \boldsymbol{B}} = E\{\psi_{\mathbf{y}}(\mathbf{y})\boldsymbol{x}^T\} - \boldsymbol{B}^{-T} \quad (14)$$

where

$\psi_{\mathbf{y}}(\mathbf{y}) = \left(\psi_{y_1}(y_1), \ldots, \psi_{y_p}(y_p)\right)^T$ is the marginal score functions vector

and

$$\psi_{x_i}(x_i) \triangleq -\frac{d\, ln\left(p_{x_i}(x_i)\right)}{d\, x_i} = -\frac{\dot{p}_{x_i}(x_i)}{p_{x_i}(x_i)} \quad (15)$$

Then the steepest descent algorithm is applied on the parameter vector to search the minimum of $I(\mathbf{y})$:

$$\boldsymbol{B} \leftarrow \boldsymbol{B} - \mu \frac{\partial I(\mathbf{y})}{\partial \boldsymbol{B}} \quad (16)$$

μ is the step-size (positive constant)

We can see that in calculating $\frac{\partial I(\mathbf{y})}{\partial B}$, the pdf's of the components of $\mathbf{y}$ must be estimated, and the algorithm can be summarized in Fig. 3, where $\boldsymbol{I}_p$ denotes the identity matrix

Simulation results

In this step we have conducted experiments with data drawn independently from known distributions, which in turn we tried to approximate with the proposed approach. MLP training in the likelihood minimization was performed using gradient descent algorithm.

In all problems we have considered a training set with $n = 2000$ data points drawn independently from the corresponding pdf to be approximated.

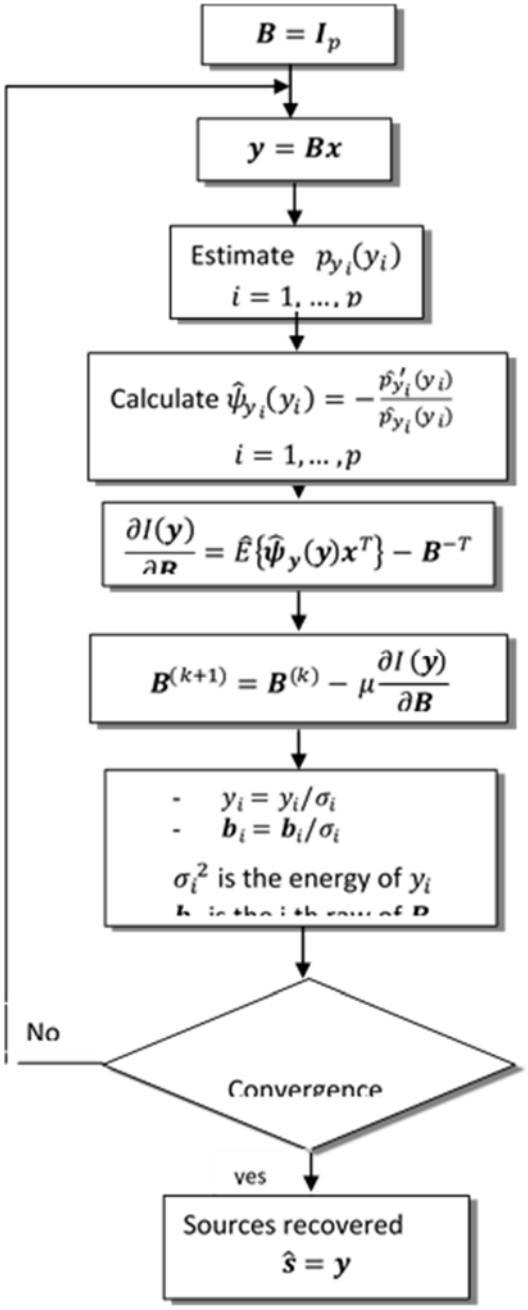
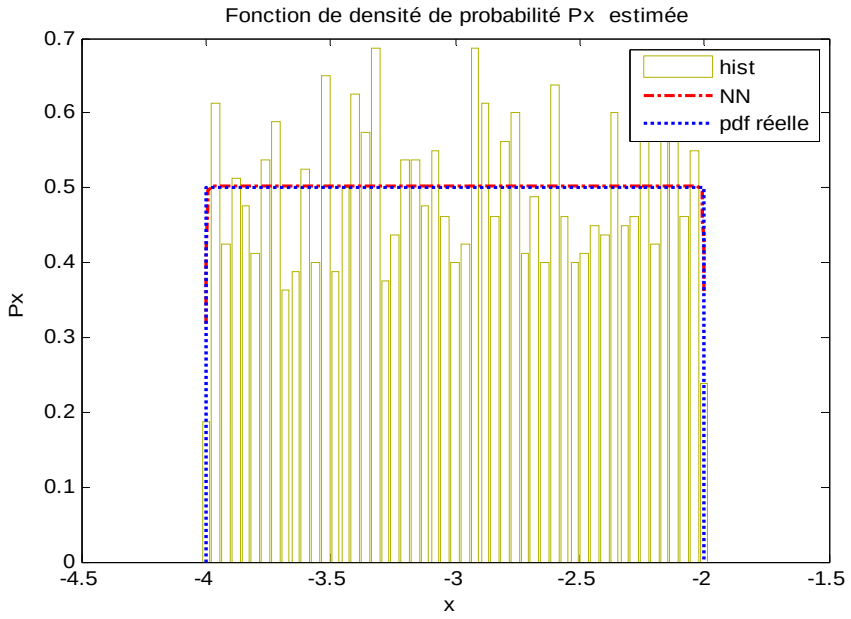


Figure 3: Steepest descent and neural pdf estimation for blind source estimation - linear instantaneous mixtures

Example 1

In this example we have generated samples using two simple distributions: Gaussian ($N(5,1)$) and uniform in the interval [-4,-2] ($U[-4,-2]$).
Fig. 4 illustrates the process of fitting the histograms of the two pdfs, and we can easily observe that the estimated pdfs coincide with the true ones.



(a)

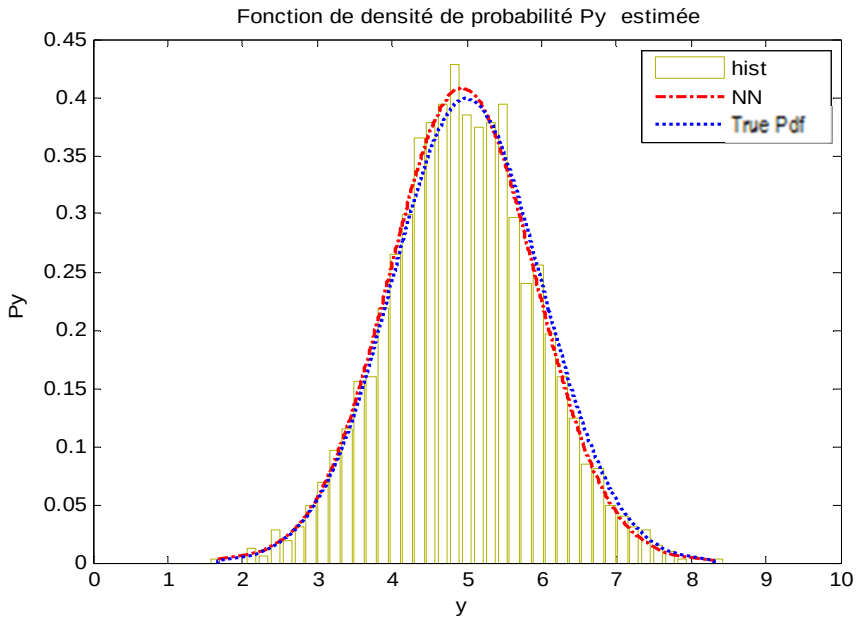


Figure 4: Estimated Pdfs, (a) uniform pdf , (b) Gaussian pdf – (*d*=2 and *H*=2)

Example 2

In this example, the unknown pdf, $g(x)$, was a mixture of the two pdfs used in the last example:

$$g(x) = 0.25\,U[-2,-1] + 0.25\,N(-7,0.25) + 0.25\,U[1,2] + 0.25\,N(7,0.25) \quad (17)$$

Fig. 5 is another illustration of the effectiveness of our method. It is also clear that the estimated pdf is very close to the true pdf and is a smooth function unlike the histogram estimator.

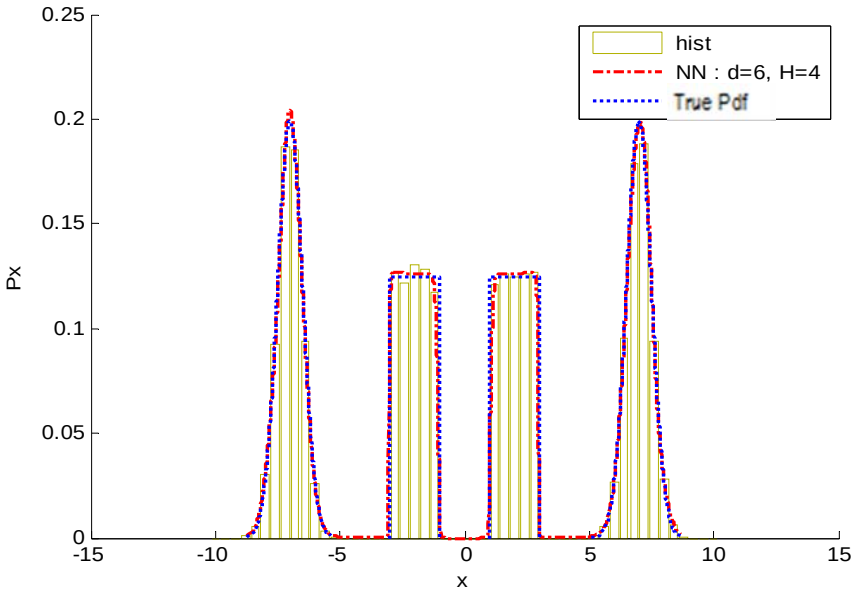


Figure 5: Estimated pdf (mixture of two pdfs) for two architectures of the MLP: (*d*=2,*H*=2) and (*d*=1,*H*=4)

Example 3

As sited in section (2.2), estimating pdf's in blind source separation is an essential step, and in some cases without this step, separation of the sources is impossible.

In this example, we apply our neural pdf estimation method to separate two linearly mixed independent sources.

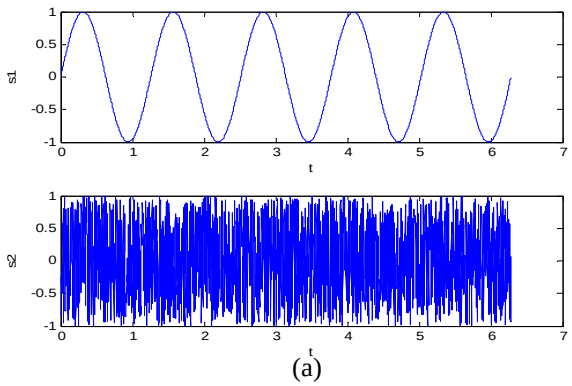
The independent sources are sine wave and uniformly distributed white noise in the interval [-1, 1]. These signals are linearly mixed with the mixture matrix

$$\boldsymbol{A} = \begin{bmatrix} -2.29 & 0.49 \\ 1.84 & 0.41 \end{bmatrix}$$

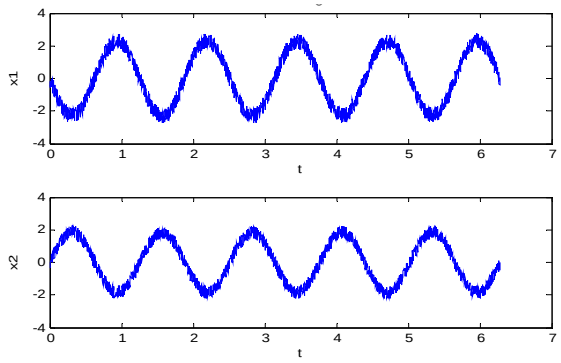
Fig. 6 shows the two sources and their mixtures.

Algorithm of Fig. 3 is used to separate the sources where marginal score functions are calculated from the estimated marginal pdf's and an MLP of two elements in the hidden layer (*H=2*, *d=2*) was used.

Outputs of the algorithm are shown in Fig. 7 and are good estimations of the source signals.



(a)



(b)

Figure 6: (a) Sources , (b) Mixture Signals

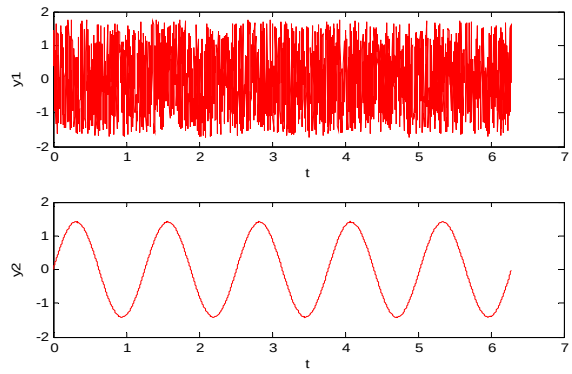


Figure 7: Estimated Sources (*d*=2, *H*=2, for pdf estimation)

Conclusions

As we mentioned before, this is a very important result about evaluation of the effectiveness of MLP in estimating probability density functions. We have modeled the data by exponential density law because most common densities (normal, uniform, exponential …etc.) are well fitted by the exponential densities (Ould Mohamed, 2012, p.54). We found that this method for one-dimensional problems has superior estimation capability compared to the widely used histogram approach. Our method has been integrated in the famous steepest descent algorithm for marginal score functions estimation where two linearly mixed sources were successfully separated.

Future research may focus on using our method to estimate the pdf for higher dimensions, and its application in separating nonlinear mixtures.

References

Silverman, B.W. (1986), " Density Estimation for Statistics and Data Analysis ", Chapman & Hall.

Vogt, J. (2007), " Basic Analysis Techniques & Multi-Spacecraft Data ", *6th COSPAR Capacity Building Workshop* (pp.4-16)., Sinaia.

Hwang, J. N., Lay, S. R., & Lippman A. , (1993), "Unsupervised learning for multivariate probability density estimation: Radial basis and projection pursuit," *IEEE Int. Conf Neural Networks* (pp. 1486-1491), San-Francisco, CA.

Popat, K. & Picard, R. W. (1993), "Novel cluster-based probability model for texture synthesis, classification,

and compression," *in Proc. SPIE Visual Commun. Image Processing'93*, Boston, MA.

Popat, K. & Picard, R. W. (1994), "Cluster-based probability model applied to image restoration and compression," *in Proc. ICASSP*, Adelaide, Australia.

Rabiner, L. R. (1989), "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2 (pp. 257-286).

Moody, J. & Darken, C. J. (1989), "Fast learning in networks of locally tuned processing units," *Neural Computation*, vol. 1, no. 3 (pp. 281-294).

Svozil, D., Kvasnicka, V. & Pospichal, J. (1997), " Introduction to Multi-Layer Feed-Forward Neural Networks ", *Chemometrics and Intelligent Laboratory Systems* Vol.39 (pp.43-62).

Modha, D.S. & Fainman, Y. (1994), "A learning law for density estimation," *IEEE Trans. On neural networks*, Vol.5, no.3 (pp.519-523).

White, H. (1992), "Mathematical perspectives on Neural Networks", M. Moser, D. Rumelhart (Eds).

Likas, A. (2001), "Probability density estimation using neural networks," *Computer Physics Communications*, Vol. 135 (pp. 167-175).

Ould Mohamed, M.S. (2012), "Contribution à la separation aveugle de sources par utilisation des divergences entre densités de probabilité : application à l'analyse vibratoire,'' thèse de doctorat de l'université de Reims Champagne – Ardenne.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001), "Independent Component Analysis." John Wiely & Sons.

Jutten, C. & Comon, P. (2007), "Séparation de sources – Tome2 : au-delà de l'aveugle et applications", chapitre 13 par Y. Deville. *Collection "Traité IC2, Information - Commande -Communication"*, Hermès - Lavoisier, Paris.

Taleb, A. & Jutten, C. (1999), "Source separation in post nonlinear mixtures." *IEEE Transactions on Signal Processing*, vol. 47, no. 10 (pp. 2807–2820).

