FAST ALGORITHM FOR ESTIMATING MUTUAL INFORMATION, ENTROPIES AND SCORE FUNCTIONS

Dinh Tuan Pham

Laboratoire de Modelling and Computation, CNRS, IMAG
BP. 53, 38041 Grenoble cedex, France

ABSTRACT

This papers proposes a fast algorithm for estimating the mutual information, difference score function, conditional score and conditional entropy, in possibly high dimensional space. The idea is to discretise the integral so that the density needs only be estimated over a regular grid, which can be done with little cost through the use of a cardinal spline kernel estimator. Score functions are then obtained as gradient of the entropy. An example of application to the blind separation of post-nonlinear mixture is given.

1. INTRODUCTION

Blind source separation consists in extracting independent sources from their mixtures, without relying on any specific knowledge of the sources and the mixing operation. Therefore the underlying principle is to construct the separating system such that the recovered sources are as independent as possible. An independence measure which is well known and widely used is the mutual information. It is directly related to the (Shannon) entropy: let $Y_1, \ldots, Y_K$, be random variables with densities $p_{Y_1}, \ldots, p_{Y_k}$, their mutual information can be computed as

$$I(Y_1, \ldots, Y_K) = \sum_{k=1}^{K} H(Y_k) - H(\mathbf{Y}) \qquad (1)$$

where $\mathbf{Y} = [Y_1 \ \cdots \ Y_K]^T$ and $H(Y_k) = -\mathrm{E} \log p_{Y_k}(Y_k)$ and $H(\mathbf{Y}) = -\mathrm{E} \log p_{\mathbf{Y}}(\mathbf{Y})$ are the entropies of Y_k and of $\mathbf{Y}$, E denoting the expectation operator and $p_{\mathbf{Y}}$ the density of $\mathbf{Y}$. (We also refer to $p_{\mathbf{Y}}$ and $H(\mathbf{Y})$ as the joint density and joint entropy of $Y_1, \ldots, Y_k$.) Thus a separation method based on minimizing the mutual information criterion would require estimating the entropies, which would involve the estimation of densities.

Density estimation in high dimensional space can be costly and subjected to large statistical fluctuation. Therefore it could be a good thing that the estimation of $H(\mathbf{Y})$ can be avoided. Indeed, since $\mathbf{Y}$ is the output of the separation system, which is linked to the observations $\mathbf{X}$ through the de-mixing operation $\mathbf{Y} = g(\mathbf{X})$, one has $H(\mathbf{Y}) = H(\mathbf{X}) + \mathrm{E} \log \det g'(\mathbf{X})$, provided that the transformation g is one-to-one and admits a continuous matrix of derivatives g'. Therefore, since $H(\mathbf{X})$ is fixed, minimizing $I(Y_1, \ldots, Y_K)$ is the same as minimizing $\sum_{k=1}^{K} H(Y_k) - \mathrm{E} \log \det g'(\mathbf{X})$ and estimating $\mathrm{E} \log \det g'(\mathbf{X})$ is easy because it is an expectation. However, there are cases when one would minimize the mutual information directly.

- The transformation g is not one-to-one. This happens for example in the convolutive mixture case. The demixing transformation is then a convolution of the observation sequence $\{\mathbf{X}(n)\}$ with some filter, and for obvious practical reasons, one would consider only finite segments of the reconstructed sources, hence g is not one-to-one as it maps a whole sequence to a vector of finite dimension. Even if one uses a FIR (finite impulse response) filter, the mapping is still not one-to-one as the inverse filter cannot be FIR.
- Density estimation is biased, which entails bias in the estimated entropies. By working with the mutual information, the bias effect of the marginal entropies $H(Y_k)$ can be canceled by that of the joint entropy $H(\mathbf{Y})$ if some care are taken (see below). By contrast, avoiding the estimation of $H(\mathbf{Y})$ as above can lead to large bias in the separating system, in the nonlinear mixture case.

The need to estimate the joint entropy also arises when one would like to exploit the temporal dependency of the source signal. In [1], blind separation of instantaneous mixtures of Markovian sources is considered and a maximum likelihood approach leads to a criterion based on the conditional entropy of $Y_k(n)$ given $Y_k(n-1), \ldots, Y_k(n-p)$, where $\{Y_k(n)\}$ denotes the sequence of reconstructed sources. For a stationary sequence of random variables $\{Y(n)\}$, the conditional entropy of $Y(n)$ given $Y(n-1)$, $\ldots$, $Y(n-p)$ is the same as that of $Y(p)$ given $Y(p-1)$, $\ldots$, $Y(1)$ and is defined as

$$H[Y(p)|Y(1:p-1)] = -\mathrm{E} \log p_{Y(p)|Y(1:p-1)}[Y(1:p)]$$

Thanks to the European Bliss Project for funding.

where the notation $Y(1:p)$ stands for the vector with components $Y(1),\ldots,Y(p)$ and

$$p_{Y_k(p)|Y_k(1:p-1)} = p_{Y_k(1:p)}/p_{Y_k(1:p-1)} \tag{2}$$

is the conditional density of $Y(p)$ given $Y(1:p-1)$, $p_{Y(1:p)}$ denoting the density of the vector $Y_k(1:p)$. Simple calculation shows that

$$H[Y(p)|Y(1:p-1)] = H[Y(1:p)] - H[Y(1:p-1)] \tag{3}$$

which again leads us to considering joint entropy.

This paper proposes a method for estimating the joint entropy of *moderately large number of variables with an affordable computational cost*. From the above discussion, we are in fact interested in the estimation of the mutual information and conditional entropy, that of the joint entropy is only an intermediate step. Further, we are more interested in the estimation of the (joint) score function, which arises when one computes the gradient of the (joint) entropy (see below). Again, when mutual information and conditional entropy are involved, one actually needs only the difference between certain score functions, the so called score difference function [2] and the conditional score function, defined in next section.

2. SCORE FUNCTIONS AS GRADIENT OF THE ENTROPY FUNCTIONAL

Let $\mathbf{Y}$ be a random vector with components $Y_1,\ldots,Y_K$ and density $p_{\mathbf{Y}}$. The score function $\psi_{\mathbf{Y}}$ of $\mathbf{Y}$ (also called the joint score function of $Y_1,\ldots,Y_K$ and denoted by $\psi_{Y_1,\ldots,Y_k}$) is defined as the gradient of $-\log p_{\mathbf{Y}}$. It can be viewed as the gradient of the entropy functional, in the sense that for a "small" random increment $\partial\mathbf{Y}$ of the vector $\mathbf{Y}$, one has

$$H(\mathbf{Y}+\partial\mathbf{Y}) - H(\mathbf{Y}) \approx \mathrm{E}[\psi_{\mathbf{Y}}^T(\mathbf{Y})\partial\mathbf{Y}] \tag{4}$$

up to the first order. A brief intuitive proof of this statement is given as follows (a more rigorous proof can be found in [3]). One writes $H(\mathbf{Y}+\partial\mathbf{Y}) - H(\mathbf{Y})$ as

$$\mathrm{E}\Big[\log\frac{p_{\mathbf{Y}}(\mathbf{Y})}{p_{\mathbf{Y}}(\mathbf{Y}+\partial\mathbf{Y})}\Big] + \mathrm{E}\Big[\log\frac{p_{\mathbf{Y}}(\mathbf{Y}+\partial\mathbf{Y})}{p_{\mathbf{Y}+\partial\mathbf{Y}}(\mathbf{Y}+\partial\mathbf{Y})}\Big].$$

But the last term can be written as

$$\mathrm{E}\Big[\log\frac{p_{\mathbf{Y}}(\mathbf{Y}+\partial\mathbf{Y})}{p_{\mathbf{Y}+\partial\mathbf{Y}}(\mathbf{Y}+\partial\mathbf{Y})} - \frac{p_{\mathbf{Y}}(\mathbf{Y}+\partial\mathbf{Y})}{p_{\mathbf{Y}+\partial\mathbf{Y}}(\mathbf{Y}+\partial\mathbf{Y})} + 1\Big]$$

and since $\log x = x-1-(x-1)^2/2+\ldots$, this expression can be expected to be of higher order than $\partial\mathbf{Y}$ and thus can be ignored. Then, by a Taylor expansion of $\log p_{\mathbf{Y}}(\mathbf{Y}+\partial\mathbf{Y})$, one gets the desired result.

From the above result and (1), the mutual information $I(Y_1,\ldots,Y_k)$ admits the following first order expansion

$$I(Y_1+\partial Y_1,\ldots,Y_k+\partial Y_k) - I(Y_1,\ldots,Y_k) \approx \sum_{k=1}^{K} \mathrm{E}\{[\psi_{Y_k}(Y_k) - \psi_{k,Y_1,\ldots,Y_K}(Y_1,\ldots,Y_K)]\partial Y_k\}$$

where $\psi_{k,Y_1,\ldots,Y_K}$ is the k-th component of the joint score function of $Y_1,\ldots,Y_K$. The functions $\psi_{Y_k} - \psi_{k,Y_1,\ldots,Y_K}$ have been introduced in [2] under the name of score difference functions (SDF). They can be viewed as the components of the gradient vector of the mutual information functional.

Similarly, by the above result and (3), for a sequence of random variables $\{Y(n)\}$, the conditional entropy of $Y(p)$ given $Y(1),\ldots,Y(p-1)$ admits the first order expansion

$$H[Y(p)+\partial Y(p)|Y(1:p-1)+\partial Y(1:p-1)] - H[Y(p)|Y(1:p-1)] \approx \mathrm{E}\{\psi^T_{Y(p)|Y(1:p-1)}[Y(1:p)]\partial Y(1:p)\}$$

where

$$\psi_{Y(p)|Y(1:p-1)} = \psi_{Y(1:p)} - \begin{bmatrix}\psi_{Y(1:p-1)}\\ 0\end{bmatrix} \tag{5}$$

The above function is no other than the gradient vector of $-\log p_{Y(p)|Y(1:p-1)}$, where $p_{Y(p)|Y(1:p-1)}$ is the conditional density of $Y(p)$ given $Y(1),\ldots,Y(p-1)$, defined in (2). This function will be referred to as the conditional score function of $Y(p)$ given $Y(1),\ldots,Y(p-1)$.

3. ESTIMATION METHOD

The main idea is to estimate first the entropy (joint, marginal and conditional) then taking their gradient as estimation of the score difference and conditional score functions, according to the relations described in previous section. This way, one gets both an estimated criterion for blind source separation and its gradient. An independent estimate of the score function would provided only an estimate of the gradient of the theoretical criterion, which often differs from the gradient of estimated criterion. One may take, for example, the negative of the logarithmic gradient of the estimated density as a score estimator, but this *is not the same* as our score estimator.

3.1. Estimation of entropy

To estimate the entropy, we need an estimator of the density. we shall adopt the kernel method, which estimates the density $p_{\mathbf{Y}}$ of a random vector $\mathbf{Y}$ from a sample $\mathbf{Y}(1),\ldots,\mathbf{Y}(N)$ as

$$\hat{p}_Y(\mathbf{y}) = \frac{1}{N}\sum_{n=1}^{N}\frac{\kappa[\mathbf{h}^{-1}(\mathbf{y}-\mathbf{Y}(n)]}{\det\mathbf{h}} = \hat{\mathrm{E}}\frac{\kappa[\mathbf{h}^{-1}(\mathbf{y}-\mathbf{Y}]}{\det\mathbf{h}},$$

where κ is a multivariate density and $\mathbf{h}$ is smoothing parameter matrix [4]. Here and in the sequel, the notation $\hat{\mathrm{E}}$ denotes the sampling mean operator. A natural estimator for $H(\mathbf{Y})$ is then $-\int \hat{p}_Y(\mathbf{y}) \log \hat{p}_Y(\mathbf{y}) d\mathbf{y}$ but it requires *multiple integration in a possibly high dimensional space*. Therefore we shall discretize the above integral to reduce it to a sum over some regular grid, yielding $-\sum_{\mathbf{i}} \hat{p}_Y(\mathbf{gi}) \log \hat{p}_Y(\mathbf{gi}) \det \mathbf{g}$, where the summation is done over all vectors $\mathbf{i}$ with signed integer components and $\mathbf{g}$ is a matrix which defines the size and orientation of the grid. Note that one can also avoid integration by estimating the entropy of $\mathbf{Y}$ as $(1/N) \sum_{n=1}^{N} \hat{p}_{\mathbf{Y}}[\mathbf{Y}(n)]$. But this method entails a computational cost of the order N^2, as each $\hat{p}_{\mathbf{Y}}[\mathbf{Y}(n)]$ itself requires a summation of N terms[1]. Our method, through a clever choice of grid and of kernel (with compact support), has a computational cost increasing only linearly with N, as it will be shown below. More importantly, it allows the canceling of bias mentioned in the introduction.

We shall need to take $\mathbf{g}$ proportional to $\mathbf{h}$. This also makes sense since $\mathbf{h}$ controls the amount of smoothing and the smoother $\hat{p}_{\mathbf{Y}}$ is the larger the grid size can be. The choice of the proportionality coefficient should take into account the calculation cost and the loss of accuracy due to the discretization. For the kernel considered below, we found that taking $\mathbf{g} = \mathbf{h}$ is most convenient and does not result in a too coarse grid. It is possible to take $\mathbf{g} = \mathbf{h}/m$ for some integer m, which reduces the grid size but increases the computation cost by a factor m^K. We will not consider this choice for simplicity and besides for K moderately large, m^K can be too large a factor even for $m = 2$.

Thus we are led to the estimator

$$\hat{H}(\mathbf{Y}) = -\sum_{\mathbf{i}} \hat{\pi}_{\mathbf{Y}}(\mathbf{i}) [\log \hat{\pi}_{\mathbf{Y}}(\mathbf{i}) - \log \det \mathbf{h}], \qquad (6)$$

where

$$\hat{\pi}_{\mathbf{Y}}(\mathbf{i}) = \frac{1}{N} \sum_{n=1}^{N} \kappa[\mathbf{i} - \mathbf{h}^{-1}\mathbf{Y}(n)] = \hat{\mathrm{E}}\kappa(\mathbf{i} - \mathbf{h}^{-1}\mathbf{Y}). \quad (7)$$

The $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$ may be viewed as the estimated probability that the random vector $\mathbf{h}^{-1}\mathbf{Y}$ belongs to a cell centered at $\mathbf{i}$ of unit volume.

In practice, the multivariate kernel κ is generated from a univariate kernel $\mathcal{K}$ by one of the two main methods [4]:

(i) tensor product: $\kappa = \mathcal{K}^{\times K}$ the K times tensor product of $\mathcal{K}$ with itself, defined as $\mathcal{K}^{\times K}(\mathbf{y}) = \sum_{k=1}^{K} \mathcal{K}(y_k)$, y_k denoting the components of $\mathbf{y}$.

(ii) spherical symmetry: $\kappa(\mathbf{y}) = C\mathcal{K}(||\mathbf{y}||)$ where C is a normalizing constant so that κ integrates to 1.

[1] even if one uses a kernel of compact support, so that only a limited number of terms in this sum is non zero, one still need to compute the pairwise differences $\mathbf{Y}(n) - \mathbf{Y}(m)$

Note that the Gaussian kernel is both a tensor product and spherically symmetric. But it does not have compact support. We shall used instead the tensor product of cardinal spline or third order. Recall that the cardinal spline of order r is the density of the sum of r independent uniform random variables in $[-1/2, 1/2]$. It tends to the Gaussian density (up to a scaling) as r increases, by the central limit Theorem. We choose the third cardinal spline because it is the simplest one which has continuous derivative (we need this condition for gradient calculation). Besides, it is already quite close to the Gaussian density. Explicitly, it is given by

$$\mathcal{K}(u) = \begin{cases} 3/4 - u^2, & |u| \le 1/2 \\ (3/2 - |u|)^2/2, & 1/2 \le |u| \le 3/2 \\ 0, & \text{otherwise} \end{cases} .$$

The fast computation of the $\hat{\pi}_{\mathbf{Y}}$ comes from the fact that it is evaluated on a regular grid and the kernel is a product kernel with support of length a multiple of the grid size. Indeed, let Y'_k be the components of $\mathbf{h}^{-1}\mathbf{Y}$, the term $\mathcal{K}^{\times K}[\mathbf{i} - \mathbf{h}^{-1}\mathbf{Y}(n)]$ in (7) can be non zero only if $i_k = \langle Y'_k(n) \rangle$ or $i_k = \langle Y'_k(n) \rangle \pm 1$, $k = 1, \ldots, K$, where i_k is the k-th component of $\mathbf{i}$ and $\langle y \rangle$ denotes the signed integer closest to y. Thus, one can compute the $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$ quickly by the following algorithm: One first initializes the $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$ to 0, then for $n = 1, \ldots, N$, updates them as

$$\begin{aligned} \hat{\pi}_{\mathbf{Y}}[\langle Y'_1(n) \rangle + i_1, \ldots, \langle Y'_K(n) \rangle + i_p] = \\ \hat{\pi}_{\mathbf{Y}}[\langle Y'_1(n) \rangle + i_1, \ldots, \langle Y'_K(n) \rangle + i_p] + \\ \frac{1}{N} \prod_{k=1}^{K} \mathcal{K}[i_k + \langle Y'_k(n) \rangle - Y'_k(n)], \quad i_k = -1, 0, 1. \end{aligned}$$

Note that for $u \in [-1/2, 1/2]$

$$\mathcal{K}(u) = 3/4 - u^2, \qquad \mathcal{K}(\pm 1 + u) = (1/2 \mp u)^2/2$$

and thus are very simple to compute.

The above algorithm requires a loop through all the data, updating 3^K probabilities at each step. Consequently the number of indices $\mathbf{i}$ for which $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$ is not zero cannot exceed $3^K N$ and in general is much less than that. Thus the cost for computing the $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$, as well as the entropy estimator, is $O(3^K N)$ which *grows linearly* with N.

The cardinal spline functions possess a nice properties called the partition of unity: $\sum_{i=-\infty}^{\infty} \mathcal{K}(u + i) \equiv u$ regardless of u. Therefore $\sum_{\mathbf{i}} \hat{\pi}_{\mathbf{Y}}(\mathbf{i}) = 1$. Thus the $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$ constitute a (discrete) probability distribution and the entropy estimator $\hat{H}(\mathbf{Y})$ is simply the entropy of this distribution plus the term $\log \det \mathbf{h}$. This estimator however has a small defect in that it is not translation invariant. Adding a constant to the random vector $\mathbf{Y}$ does not change its entropy, so one would like that the entropy estimator is unchanged too. Therefore we shall modify this estimator slightly by first centering the data, that is in computing the $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$, we take as $Y'_k(n)$ not the k-th component of $\mathbf{h}^{-1}\mathbf{Y}(n)$ but of $\mathbf{h}^{-1}[\mathbf{Y}(n) - \bar{\mathbf{Y}}]$, where $\bar{\mathbf{Y}} = \hat{\mathrm{E}}\mathbf{Y}$ is the sample mean of $\mathbf{Y}$.

3.2. Estimation of mutual information

An obvious way to estimate the mutual information is to subtract the estimated joint entropy to the sum of the estimated marginal entropies. However, to favor the canceling of bias, we need to chose $\mathbf{h}$ diagonal, with diagonal elements $h_1, \ldots, h_K$ where h_k is the smoothing parameter for estimation the marginal density of Y_k. Then, the probabilities $\hat{\pi}_{Y_k}(j)$ needed for estimating the entropy estimator $\hat{H}(Y_k)$, given by

$$\hat{\pi}_{Y_k}(j) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}[j - Y'_k(n)] = \hat{\mathrm{E}}\mathcal{K}(j - Y'_k),$$

would involve the same variables $Y'_k = [Y_k - \bar{Y}_k]/h_k$ as the one encountered in the computation of $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$ in subsection 3.1. Thus, by the partition of unity property of $\mathcal{K}$,

$$\hat{\pi}_{Y_k}(j) = \sum_{\mathbf{i}: i_k = j} \hat{\pi}_{\mathbf{Y}}(\mathbf{i}),$$

that is the $\hat{\pi}_{Y_k}(j)$ are simply the marginal probabilities of the $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$. The mutual information estimator is then

$$\hat{I}(Y_1, \ldots, Y_K) = \sum_{\mathbf{i}} \hat{\pi}_{\mathbf{Y}}(\mathbf{i}) \log \frac{\hat{\pi}_{\mathbf{Y}}(\mathbf{i})}{\prod_{k=1}^{K} \hat{\pi}_{Y_k}(i_k)}.$$

From this form, one may expect that the bias in $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$ be more or less canceled by those in $\hat{\pi}_{Y_k}(i_k)$, as the latter themselves are computed from the $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$. More importantly, if the vector $\mathbf{Y}$ has independent components, $\hat{I}(Y_1, \ldots, Y_K)$ will converges to 0 as $n \to \infty$, *regardless the choice of* $\mathbf{h}$, since the limit of $\hat{\pi}_{\mathbf{Y}}(\mathbf{i})$ is the expectation of a product of independent random variables which equals the product of expectations. Thus one can chose $\mathbf{h}$ fairly large without worrying that this would invalidate $\hat{I}(Y_1, \ldots, Y_K)$ as an empirical independence criterion.

3.3. Estimation of score and score difference function

Let $\hat{H}(\mathbf{Y})$ be the entropy estimator defined in subsection 3.1. From the result of section 2, we define the score estimator $\hat{\psi}_{\mathbf{Y}}$ of $\mathbf{Y}$, at the data point $\mathbf{Y}(n)$, as the gradient of $N\hat{H}(\mathbf{Y})$ with respect to $\mathbf{Y}(n)$. This function is defined only on the data points, but it can be easily be extended to the whole space. Hence, by the definition of $\hat{\psi}_{\mathbf{Y}}$, an infinitesimal change $\partial\mathbf{Y}(n)$ in the $\mathbf{Y}(n)$ induces a corresponding change in $\hat{H}(\mathbf{Y})$:

$$\hat{H}(\mathbf{Y} + \partial\mathbf{Y}) - \hat{H}(\mathbf{Y}) = \frac{1}{N}\sum_{n=1}^{N} \hat{\psi}_Y^T[\mathbf{Y}(n)]\partial\mathbf{Y}(n)$$

which equals $\hat{\mathrm{E}}[\hat{\psi}_Y^T(\mathbf{Y})\partial\mathbf{Y})]$. This formula thus appears as the sample analogue of (4).

Before computing the gradient of $\hat{H}(\mathbf{Y})$, one should note that this estimator depends on the data only through $\mathbf{Y}'(n) = \mathbf{h}^{-1}[\mathbf{Y}(n) - \bar{\mathbf{Y}}]$, but the matrix $\mathbf{h}$ itself depends also on the data. Here we take $\mathbf{h}$ diagonal with k-th diagonal element being a constant multiple of the sample standard deviation $\hat{\sigma}_{Y_k}$ of Y_k. Let the vectors $\mathbf{Y}(n)$ be changed by an infinitesimal amount $\partial\mathbf{Y}(n)$, the corresponding change $\partial\mathbf{Y}'(n)$ in $\mathbf{Y}'(n)$ is

$$\mathbf{h}^{-1}\partial\mathbf{Y}(n) - \hat{\mathrm{E}}(\mathbf{h}^{-1}\partial\mathbf{Y}) - \mathbf{h}^{-1}\partial\mathbf{h}\mathbf{h}^{-1}[\mathbf{Y}(n) - \bar{\mathbf{Y}}]$$

where $\partial\mathbf{h}$ is the corresponding change in $\mathbf{h}$. Such change would induce a change $\mathrm{tr}(\mathbf{h}^{-1}\partial\mathbf{h})$ in $\log\det\mathbf{h}$, where tr denotes the trace. Therefore the corresponding change in $\hat{H}(\mathbf{Y})$, is, by (6) and (7),

$$\sum_{\mathbf{i}} \hat{\mathrm{E}}[\dot{\mathcal{K}}^{\times K}(\mathbf{i} - \mathbf{Y}')^T\partial\mathbf{Y}'] \log\hat{\pi}_{\mathbf{Y}}(\mathbf{i}) + \mathrm{tr}(\mathbf{h}^{-1}\partial\mathbf{h}),$$

where $\dot{\mathcal{K}}^{\times K}$ denotes the gradient of $\mathcal{K}^{\times K}$ (and thus $\sum_{\mathbf{i}} \dot{\mathcal{K}}^{\times K}(\mathbf{i}) = \mathbf{0}$, since $\sum_{\mathbf{i}} \dot{\mathcal{K}}^{\times K}(\mathbf{i} + \mathbf{y}) \equiv 1$). Plugging in the expression for $\partial\mathbf{Y}'(n)$, the last expression becomes

$$\begin{aligned}\hat{\mathrm{E}}\{[\tilde{\psi}_{\mathbf{Y}'}(\mathbf{Y}') - \hat{\mathrm{E}}\tilde{\psi}_{\mathbf{Y}'}(\mathbf{Y}')]^T\mathbf{h}^{-1}\partial\mathbf{Y}\} + \\ \mathrm{tr}\{\mathbf{h}^{-1}\partial\mathbf{h} - \hat{\mathrm{E}}[\mathbf{Y}'\tilde{\psi}_{\mathbf{Y}}^T(\mathbf{Y}')]\mathbf{h}^{-1}\partial\mathbf{h}\}\end{aligned} \quad (8)$$

where

$$\tilde{\psi}_{\mathbf{Y}'}[\mathbf{Y}'(n)] = \sum_{\mathbf{i}} \dot{\mathcal{K}}^{\times K}[\mathbf{i} - \mathbf{Y}'(n)] \log\hat{\pi}_{\mathbf{Y}}(\mathbf{i}). \quad (9)$$

Since $\mathbf{h}^{-1}\partial\mathbf{h}$ is a diagonal matrix, one may replace, in the above expression, the matrix $\hat{\mathrm{E}}[\mathbf{Y}'\tilde{\psi}_{\mathbf{Y}'}^T(\mathbf{Y})]$ by the diagonal matrix having the same diagonal, which we denote by $\mathbf{\Lambda}$ for simplicity. On the other hand, the k-th diagonal element of $\mathbf{h}^{-1}\partial\mathbf{h}$ is $N^{-1}\sum_{n=1}^{N}[Y_k(n) - \bar{Y}_k]\partial Y_k(n)/\hat{\sigma}_{Y_k}^2$. Therefore, we obtain from the above results

$$\begin{aligned}\hat{\psi}_{\mathbf{Y}}[\mathbf{Y}(n)] = \mathbf{h}^{-1}\{\tilde{\psi}_{\mathbf{Y}'}[\mathbf{Y}'(n)] - \hat{\mathrm{E}}\tilde{\psi}_{\mathbf{Y}'}(\mathbf{Y}')\} + \\ (\mathbf{I} - \mathbf{\Lambda})\sigma_{\mathbf{Y}}^{-2}[\mathbf{Y}(n) - \bar{\mathbf{Y}}],\end{aligned}$$

where $\hat{\sigma}_{\mathbf{Y}}$ denotes the diagonal matrix with diagonal elements $\hat{\sigma}_{Y_k}$.

The computation cost of $\hat{\psi}_{\mathbf{Y}}$ is still $O(3^K N)$ since the computation of $\tilde{\psi}_{\mathbf{Y}'}$, by (9) requires only a summation of 3^K terms for each data point

3.4. Estimation of the conditional entropy and conditional score

Since the conditional entropy $H[Y(p)|Y(1:p-1)]$ is the difference between two joint entropies, the above estimation method is still applicable. However, while in estimating the mutual information one is mostly interested in the case where the random variables are nearly independent, here

the sequence $\{Y(n)\}$ would be highly correlated. Hence it is desirable to perform a prewhitening first. Let $\mathbf{T}$ be the Cholesky factor of $\widehat{\mathrm{cov}}[Y(1:p)]$, the covariance matrix of $Y(1:p)$, that is $\mathbf{T}$ is the lower triangular matrix satisfying $\mathbf{T}\mathbf{T}^T = \widehat{\mathrm{cov}}[Y(1:p)]$. Prewhitening consists in replacing $Y(1:p)$ by the vector $\mathbf{T}^{-1}Y(1:p)$, which is then has uncorrelated components of unit variance. One would then using a smoothing parameter matrix $\mathbf{h}$ multiple of the identity, which amounts to working with the original data, but taking $\mathbf{h}$ a multiple of $\mathbf{T}$.

With this choice of $\mathbf{h}$, the entropy of $Y(1:p)$ can then be estimated by the same method as in section 3.1:

$$\hat{H}[Y(1:p)] = -\sum_{\mathbf{i}} \hat{\pi}_{Y(1:p)}(\mathbf{i}) \log \hat{\pi}_{Y(1:p)}(\mathbf{i}) + \log\det \mathbf{h}$$

where $\hat{\pi}_{Y(1:p)}(\mathbf{i})$ are computed as in 3.1 with Y'_k being the k-th component of $\mathbf{h}^{-1}[Y(1:p) - \bar{Y}(1:p)]$. But by construction, the smoothing parameter matrix $\mathbf{h}$ for the entropy estimator of the random vector $Y(1:p-1)$ is no other than the one for $Y(1:p)$ with last row and column deleted. Therefore, the probabilities $\hat{\pi}_{Y(1:p-1)}(i_1,\ldots,i_{p-1})$ involved in this estimator are no other than the marginal of the $\hat{\pi}_{Y(1:p)}(i_1,\ldots,i_p)$:

$$\hat{\pi}_{Y(1:p-1)}(i_1,\ldots,i_{p-1}) = \sum_{i_p} \hat{\pi}_{Y(1:p)}(i_1,\ldots,i_p).$$

The estimated conditional entropy $\hat{H}[Y(p)|Y(1:p-1)]$ of $Y(p)$ given $Y(1),\ldots,Y(p)$ is thus given by

$$\sum_{i_1,\ldots,i_p} \hat{\pi}(i_1,\ldots,i_p) \log \frac{\hat{\pi}(i_1,\ldots,i_p)}{\hat{\pi}(i_1,\ldots,i_{p-1})} + h_{pp}$$

where h_{pp} is the last element of $\mathbf{h}$.

To construct the conditional score estimator, we view $N\hat{H}[Y(p)|Y(1:p-1)]$ as a function of the data $\mathbf{Y}(1),\ldots,\mathbf{Y}(N)$ where $\mathbf{Y}(n)$ now denotes the vector $Y(n:n+1-p)$. Then, as in subsection 3.3, the estimated conditional score function of $Y(p)$ given $Y(1:p-1)$ is defined as the function taking the value at $\mathbf{Y}(n)$ the gradient of $N\hat{H}[Y(p)|Y(1:p-1)]$ with respect to $\mathbf{Y}(n)$. It clearly can again be expressed as the difference between two score estimators. However the last differ from the ones obtained in subsection 3.3, since the smoothing parameter matrix $\mathbf{h}$ is here not diagonal. But one can repeat the same calculations in this subsection, up to the expression (8) and (9), which provide the infinitesimal increment of $\hat{H}(\mathbf{Y})$ induced by an infinitesimal increment $\partial\mathbf{Y}$ of $\mathbf{Y}$.

To proceed further, we note that the matrix $\mathbf{h}^{-1}\partial\mathbf{h}$ is lower triangular, therefore we can replace, in the expression (8), the matrix $\hat{\mathrm{E}}[\mathbf{Y}'\tilde{\psi}^T_{\mathbf{Y}'}(\mathbf{Y}')]$ by the symmetric matrix with the same upper triangular part, which we denote again by $\mathbf{\Lambda}$. But for a symmetric matrix $\mathbf{\Lambda}$, $\mathrm{tr}(\mathbf{\Lambda}\mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{\Lambda}\mathbf{A}^T)$ for any matrix $\mathbf{A}$. Hence, one can rewrite (8) as

$$\hat{\mathrm{E}}\{[\tilde{\psi}_{\mathbf{Y}'}(\mathbf{Y}') - \hat{\mathrm{E}}\tilde{\psi}_{\mathbf{Y}'}(\mathbf{Y}')]^T\mathbf{h}^{-1}\partial\mathbf{Y}\} + \frac{1}{2}\mathrm{tr}[(\mathbf{I}-\mathbf{\Lambda})(\mathbf{h}^{-1}\partial\mathbf{h} + \partial\mathbf{h}^T\mathbf{h}^{-1T}]$$

The last term can be rewritten as

$$\frac{1}{2}\mathrm{tr}[\mathbf{h}^{-1T}(\mathbf{I}-\mathbf{\Lambda})\mathbf{h}^{-1}(\partial\mathbf{h}\mathbf{h}^T + \mathbf{h}\partial\mathbf{h}^T)]$$

But $\mathbf{h}\mathbf{h}^T = h^2\widehat{\mathrm{cov}}(\mathbf{Y})$ for some scalar constant h, yielding

$$\partial\mathbf{h}\mathbf{h}^T + \mathbf{h}\partial\mathbf{h}^T = h^2\hat{\mathrm{E}}[(\mathbf{Y}-\bar{\mathbf{Y}})\partial\mathbf{Y}^T + \partial\mathbf{Y}(\mathbf{Y}-\bar{\mathbf{Y}})^T].$$

Since the matrix $\mathbf{h}^{-1T}(\mathbf{I}-\mathbf{\Lambda})\mathbf{h}^{-1}$ is symmetric, by the same argument as before, the expression (8) can be further rewritten as

$$\hat{\mathrm{E}}\{[\tilde{\psi}_{\mathbf{Y}'}(\mathbf{Y}') - \hat{\mathrm{E}}\tilde{\psi}_{\mathbf{Y}'}(\mathbf{Y}')]^T\mathbf{h}^{-1}\partial\mathbf{Y} + h^2[\mathbf{h}^{-1T}(\mathbf{I}-\mathbf{\Lambda})\mathbf{h}^{-1}\partial\mathbf{Y}(\mathbf{Y}-\bar{\mathbf{Y}})^T]\}$$

This expression shows that the estimated score function of $\mathbf{Y}$ is given by

$$\hat{\psi}_{\mathbf{Y}}[\mathbf{Y}(n)] = \mathbf{h}^{-1\,T}\{\tilde{\psi}_{\mathbf{Y}'}[\mathbf{Y}'(n)] - \hat{\mathrm{E}}\tilde{\psi}_{\mathbf{Y}'}(\mathbf{Y}')\} + (\mathbf{I}-\mathbf{\Lambda})h^2\mathbf{Y}'(n)\}.$$

To compute the estimated conditional score function, one must subtract the above score function with the one corresponding to the vector $\mathbf{Y}$ with its last component deleted. But since the matrix $\mathbf{h}$ is lower triangular, deleting the last component of $\mathbf{Y}$ amounts to deleting the last row and column of $\mathbf{h}$ and the last element of $\mathbf{Y}'$. This also results in deleting the last row and column of $\mathbf{\Lambda}$. Therefore, putting

$$\tilde{\psi}_{Y'_p|Y'_1\ldots,Y'_{p-1}}(\mathbf{y}') = \tilde{\psi}_{\mathbf{Y}'}(\mathbf{y}') - \begin{bmatrix} \tilde{\psi}_{Y'_1\ldots,Y'_{p-1}}(y'_1,\ldots,y'_{p-1}) \\ 0 \end{bmatrix}$$

Y'_k and y'_k being the components of $\mathbf{Y}'$ and $\mathbf{y}'$, one gets

$$\hat{\psi}_{Y(p)|Y(1:p-1)}[\mathbf{Y}(n)] = \mathbf{h}^{-1T}\{\tilde{\psi}_{Y'_p|Y'_1,\ldots,Y'_{p-1}}[\mathbf{Y}'(n)] - \hat{\mathrm{E}}\tilde{\psi}_{Y'_p|Y'_1,\ldots,Y'_{p-1}}(\mathbf{Y}') + (\mathbf{E}_p - \mathbf{\Lambda})h^2\mathbf{h}^{-1}\mathbf{Y}'(n)\},$$

where $\mathbf{E}_p$ is the matrix with 1 at the (p,p) place and 0 elsewhere and $\mathbf{\Lambda}$ is the symmetric with the upper triangular part equal to $\hat{\mathrm{E}}[\mathbf{Y}'\tilde{\psi}^T_{Y'_p|Y'_1,\ldots,Y'_{p-1}}(\mathbf{Y}')]$

4. APPLICATION TO THE POST-NONLINEAR MODEL

As an example, we have applied our estimator to the post-nonlinear mixture model. This model, introduced in [5], consists of a linear mixing stage followed by a nonlinear non-mixing transformation. More precisely, the observation

$X_k(n)$ is given by $f_k[\sum_{j=1}^{K} A_{kj} S_j(n)]$. where $S_j(n)$ represent the sources, A_{kj} are the elements of the mixing matrix and f_k are monotonous mappings. The separation system is naturally $Y_k(n) = g_k[\sum_{j=1}^{K} B_{kj} X_j(n)]$ where $B_{k,j}$ are elements of the separating matrix and g_k are monotonous mappings. We estimate $B_{k,j}$ and g_k by minimizing the estimated mutual information between $Y_1, \ldots, Y_K$. This is the same principle as in [5] and [6]. But in these papers, the criterion is transformed as described in the introduction to contain only the marginal entropies, while we work directly with the estimated mutual information. Our implementation is the same as in [6], in that we use the same parameterization technique and we really minimize the criterion and not solving a estimating system of equations as in [5].

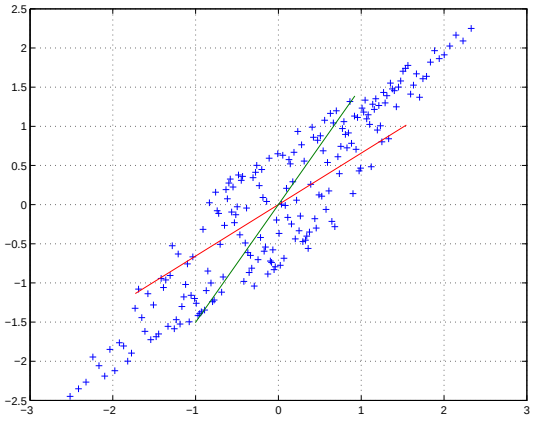

Fig. 1. Compensated observation 2 versus compensated observation 1, with "ICA axes"

We have simulated two sources of length 200, one uniform in $[-\sqrt{3}, \sqrt{3}]$, the other a sine wave with amplitude $\sqrt{2}$ (so that they both have unit power). The mixing matrix is $\begin{bmatrix} 1 & 0.6 \\ 0.7 & 1 \end{bmatrix}$ and the post-nonlinear mappings f_1, f_2 are given by $f_1(x) = \tanh(4x + 0.1x)$ and $f_1(x) = x^3 + 0.1x$.

Figure 1 plots $g_2(X_2)$, the compensated observation 2, versus $g_1(X_1)$, the compensated observation 1, together with the ICA axes (these axes are so that one can read the reconstructed sources by projecting the compensated observations on them).

Figure 2 plots the function $g_k \circ f_k$, called the compensators. One can see that they are quite linear, meaning that the nonlinearities introduced by f_1 and f_2 are fully compensated. If we apply the method in [6], the compensators (not shown here for lack of space) are more curve near their end points and therefore the separation is not as good. One should note however, that this is an extreme case where the source densities have a large (or infinite) jump at the end of its support, hence the density estimator is highly biased near these points. For smooth density, the advantage of the new method may be less clear.

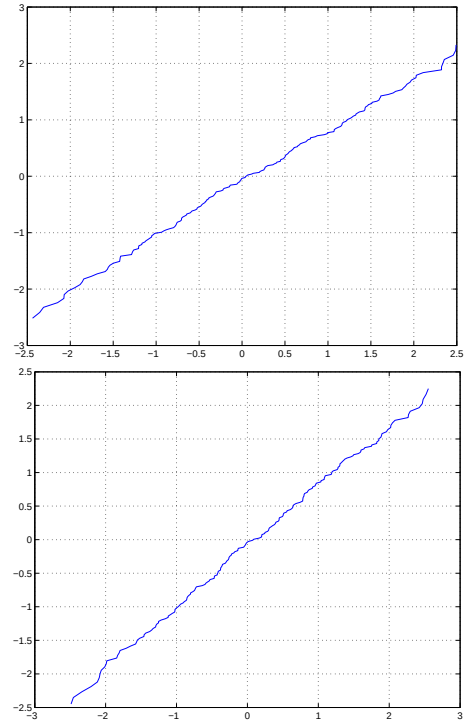

Fig. 2. Compensators for observation 1 (upper graph) and 2 (lower graph)

5. REFERENCES

[1] S. Hosseini, C. Jutten, and D. T. Pham, "Blind separation of temporally correlated sources using a quasi-maxi mum likelihood approach," in *Proceedings of ICA 2001*, San Diego, California, Dec. 2001.

[2] M. Babaie-Zadeh, C. Jutten, and K. Nayebi, "Separating convolutive post non-linear mixtures," in *Proceedings of ICA 2001*, San Diego, California, Dec. 2001.

[3] D. T. Pham, "Mutual information approach to blind separation of stationary sourcs," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1935–1946, July 2002.

[4] M. P. Wand and M. C. Jones, *Kernel Smoothing*, Chapman & Hall, 1st edition, 1995.

[5] A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2807–2820, 1999.

[6] S. Achard, D.T. Pham, and Ch. Jutten, "Blind source separation in post nonlinear mixtures.," in *Proceedings ICA2001*, San Diego, California, Dec. 2001.