Adaptive Neural Algorithms for PCA and ICA

Radu Mutihac
University of Bucharest, Romania

INTRODUCTION

Artificial neural networks (ANNs) (McCulloch & Pitts, 1943) (Haykin, 1999) were developed as models of their biological counterparts aiming to emulate the real neural systems and mimic the structural organization and function of the human brain. Their applications were based on the ability of self-designing to solve a problem by learning the solution from data. A comparative study of neural implementations running principal component analysis (PCA) and independent component analysis (ICA) was carried out. Artificially generated data additively corrupted with white noise in order to enforce randomness were employed to critically evaluate and assess the reliability of data projections. Analysis in both time and frequency domains showed the superiority of the estimated independent components (ICs) relative to principal components (PCs) in faithful retrieval of the genuine (latent) source signals.

Neural computation belongs to information processing dealing with adaptive, parallel, and distributed (localized) signal processing. In data analysis, a common task consists in finding an adequate subspace of multivariate data for subsequent processing and interpretation. Linear transforms are frequently employed in data model selection due to their computational and conceptual simplicity. Some common linear transforms are PCA, factor analysis (FA), projection pursuit (PP), and, more recently, ICA (Comon, 1994). The latter emerged as an extension of nonlinear PCA (Hotelling, 1993) and developed in the context of blind source separation (BSS) (Cardoso, 1998) in signal and array processing. ICA is also related to recent theories of the visual brain (Barlow, 1991), which assume that consecutive processing steps lead to a progressive reduction in the redundancy of representation (Olshausen and Field, 1996).

This contribution is an overview of the PCA and ICA neuromorphic architectures and their associated algorithmic implementations increasingly used as exploratory techniques. The discussion is conducted on artificially generated sub- and super-Gaussian source signals.

BACKGROUND

In neural computation, transforming methods amount to unsupervised learning, since the representation is only learned from data without any external control. Irrespective of the nature of learning, the neural adaptation may be formally conceived as an optimization problem: an objective function describes the task to be performed by the network and a numerical optimization procedure allows adapting network parameters (e.g., connection weights, biases, internal parameters). This process amounts to search or nonlinear programming in a quite large parameter space. However, any prior knowledge available on the solution might be efficiently exploited to narrow the search space. In supervised learning, the additional knowledge is incorporated in the net architecture or learning rules (Gold, 1996). A less extensive research was focused on unsupervised learning. In this respect, the mathematical methods usually employed are drawn from classical constrained multivariate nonlinear optimization and rely on the Lagrange multipliers method, the penalty or barrier techniques, and the classical numerical algebra techniques, such as deflation/renormalization (Fiori, 2000), the Gram-Schmidt orthogonalization procedure, or the projection over the orthogonal group (Yang, 1995).

PCA and ICA Models

Mathematically, the linear stationary PCA and ICA models can be defined on the basis of a common data model. Suppose that some stochastic processes are represented by three random (column) vectors $\mathbf{x}(t)$, $\mathbf{n}(t) \in \mathbb{R}^N$ and $\mathbf{s}(t) \in \mathbb{R}^M$ with zero mean and finite covariance, with the components of $\mathbf{s}(t) = \{s_1(t), s_2(t), ..., s_M(t)\}$ being statistically independent and at most one Gaussian. Let **A** be a rectangular constant full column rank $N \times M$ matrix with at least as many rows as columns ($N \geq M$), and denote by t the sample index (i.e., time or sample point) taking the discrete values $t = 1, 2, ...,$

T. We postulate the existence of a linear relationship among these variables like:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) = \sum_{i=1}^{M} s_i(t)\,\mathbf{a}_i + \mathbf{n}(t) \quad (1)$$

Here $\mathbf{s}(t)$, $\mathbf{x}(t)$, $\mathbf{n}(t)$, and $\mathbf{A}$ are the sources, the observed data, the (unknown) noise in data, and the (unknown) mixing matrix, respectively, whereas $\mathbf{a}_i$, $i = 1, 2, ..., M$ are the columns of $\mathbf{A}$. Mixing is supposed to be instantaneous, so there is no time delay between a (latent) source variable $s_i(t)$ mixing into an observable (data) variable $x_j(t)$, with i = 1, 2, ..., M and j = 1, 2, ..., N.

Consider that the stochastic vector process $\{\mathbf{x}(t)\} \in \mathbb{R}^N$ has the mean $E\{\mathbf{x}(t)\} = 0$ and the covariance matrix $\mathbf{C}_\mathbf{x} = E\{\mathbf{x}(t)\,\mathbf{x}(t)^T\}$. The goal of PCA is to identify the dependence structure in each dimension and to come out with an orthogonal transform matrix $\mathbf{W}$ of size $L \times N$ from $\mathbb{R}^N$ to $\mathbb{R}^L$, $L < N$, such that the L-dimensional output vector $\mathbf{y}(t) = \mathbf{W}\,\mathbf{x}(t)$ sufficiently represents the intrinsic features of the input data, and where the covariance matrix $\mathbf{C}_\mathbf{y}$ of $\{\mathbf{y}(t)\}$ is a diagonal matrix $\mathbf{D}$ with the diagonal elements arranged in descending order, $d_{i,i} \geq d_{i+1,i+1}$. The restoration of $\{\mathbf{x}(t)\}$ from $\{\mathbf{y}(t)\}$, say $\{\hat{\mathbf{x}}(t)\}$, is consequently given by $\hat{\mathbf{x}}(t) = \mathbf{W}^T\mathbf{W}\,\mathbf{x}(t)$ (Figure 1). For a given L, PCA aims to find an optimal value of $\mathbf{W}$, such as to minimize the error function $J = E\{\|\mathbf{x}(t) - \hat{\mathbf{x}}(t)\|\}$. The rows in $\mathbf{W}$ are the PCs of the stochastic process $\{\mathbf{x}(t)\}$ and the eigenvectors $\mathbf{c}_j$, $j = 1, 2, ..., L$ of the input covariance matrix $\mathbf{C}_\mathbf{x}$. The subspace spanned by the principal eigenvectors $\{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_L\}$ with $L < N$, is called the PCA subspace of dimensionality L.

The ICA problem can be formulated as following: given T realizations of $\mathbf{x}(t)$, estimate both the matrix $\mathbf{A}$ and the corresponding realizations of $\mathbf{s}(t)$. In BSS the task is somewhat relaxed to finding the waveforms $\{s_i(t)\}$ of the sources knowing only the (observed) mixtures $\{x_j(t)\}$. If no suppositions are made about the noise, the additive noise term is omitted in (1). A practical strategy is to include noise in the signals as supplementary term(s): hence the ICA model (Fig. 2) becomes:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{M} \mathbf{a}_i s_i(t) \quad (2)$$

The source separation consists in updating an unmixing matrix $\mathbf{B}(t)$, without resorting to any information about the spatial mixing matrix $\mathbf{A}$, so that the output vector $\mathbf{y}(t) = \mathbf{B}(t)\,\mathbf{x}(t)$ becomes an estimate $\mathbf{y}(t) = \hat{\mathbf{s}}(t)$ of the original independent source signals $\mathbf{s}(t)$. The separating matrix $\mathbf{B}(t)$ is divided in two parts dealing with dependencies in the first two moments, i.e., the whitening matrix $\mathbf{V}(t)$, and the dependencies in

Figure 1. Schematic of the PCA model