Non-Negative and Geodesic approaches to Independent Component Analysis

Mark Plumbley
Queen Mary, University of London,
Email: mark.plumbley@elec.qmul.ac.uk.

ICA Workshop, 20 December 2002.

Overview

- Introduction
- Nonnegative ICA using nonlinear PCA
- Successive rotations
- Geodesic line search
- Results
- Conclusions

Introduction

- Observations of mixed data - generative model

$$\mathbf{X} = \mathbf{A}\mathbf{S} \tag{1}$$

with sources $\mathbf{S} \in \mathbb{R}^{n \times p}$ and mixing matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$.

- Task - to discover the source samples $\mathbf{S}$ and mixing matrix $\mathbf{A}$ given only the observations $\mathbf{X}$.

- An Underdetermined problem: if $(\mathbf{A}^*, \mathbf{S}^*)$ is a solution, so is $(\mathbf{A}^*\mathbf{M}, \mathbf{M}^{-1}\mathbf{S}^*)$ (for invertible $\mathbf{M}$)

- So - need constraints.

Constraints

1. Independence of sources: s_{jk} sampled from independent random variables S_j.

2. Non-negativity of sources: $s_{jk} \geq 0$ for all $1 \leq j \leq n$, $1 \leq k \leq p$.

Independence alone
$\rightarrow$ classical noiseless ICA.
Non-negativity alone (of $\mathbf{S}$ and $\mathbf{A}$)
$\rightarrow$ *non-negative matrix factorization* [Lee & Seung, 1999]

Both constraints
$\rightarrow$ *non-negative independent component analysis*.

Non-negative ICA using Nonlinear PCA

ICA often simplified by *pre-whitening* - transform

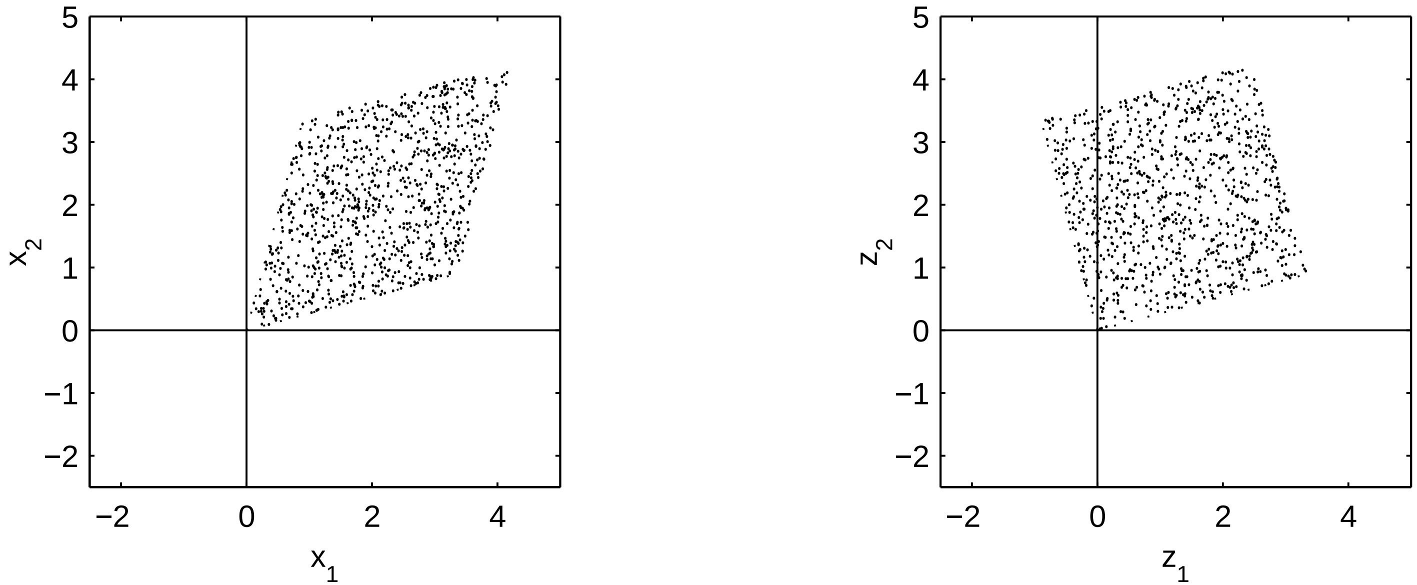
$$\mathbf{x} = \mathbf{Q}\mathbf{z} \tag{2}$$

to get identity covariance $\mathbf{C}_{\mathbf{x}} = E((\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T) = \mathbf{I}$.

Problem now to find orthonormal weight matrix $\mathbf{W}$, satisfying $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}_n$, such that the outputs $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{Q}\mathbf{A}\mathbf{s}$ are independent.

Typical ICA algorithms search for extremum of contrast function (e.g. kurtosis).

Whitening of Non-Negative Data



(a) (b)

Original data (a) is whitened (b) to remove 2nd order correlations. Suggests we just try to fit the data into +ve quadrant.

Cost/Contrast Function for Non-negative ICA?

Let $\mathbf{U} = \mathbf{WQA}$, i.e. $\mathbf{y} = \mathbf{Us}$

For non-negative sources $\mathbf{s}$, which are *well-grounded* (i.e. $\Pr(s < \delta) > 0$ for any $\delta > 0$),
then
$\mathbf{U}$ is a permutation matrix (i.e. sources are separated)
iff all components of $\mathbf{y}$ are non-negative w.p.1.

So - use a cost function, e.g. mean squared reconstruction error

$$J = \frac{1}{2} E(\|\mathbf{x} - \hat{\mathbf{x}}\|) \qquad \hat{\mathbf{x}} = \mathbf{W}^T \mathbf{y}^+$$

where $\mathbf{y}^+$ is rectified version of $\mathbf{y} = \mathbf{Wx}$.

Nonlinear PCA algorithms

Natural to consider nonlinear PCA algorithm:

$$\Delta \mathbf{W} = \eta \mathbf{g}(\mathbf{y})[\mathbf{x} - \mathbf{W}\mathbf{g}(\mathbf{y})]^T$$

in special case of $g(y) = y^+$, i.e.

$$\Delta \mathbf{W} = \eta \mathbf{y}^+[\mathbf{x} - \hat{\mathbf{x}}]^T \qquad \hat{\mathbf{x}} = \mathbf{W}^T \mathbf{y}^+$$

("non-negative PCA").

Convergence? $g(y) = y^+$ neither odd nor twice differentiable, so standard proof not applicable.
However, behaves like a 'switching subspace network', so can modify PCA subspace convergence proofs (to be confirmed!)

Orthonormality through Axis Rotation

Nonlinear PCA updates $\mathbf{W}$ in Euclidean matrix space, tending towards orthonormality $\mathbf{W}\mathbf{W}^T = \mathbf{I}$.

However, can instead construct $\mathbf{W}$ from 2D rotations [Comon, 1994],

$$\begin{pmatrix} y_{i_1} \\ y_{i_2} \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x_{i_1} \\ x_{i_2} \end{pmatrix}$$

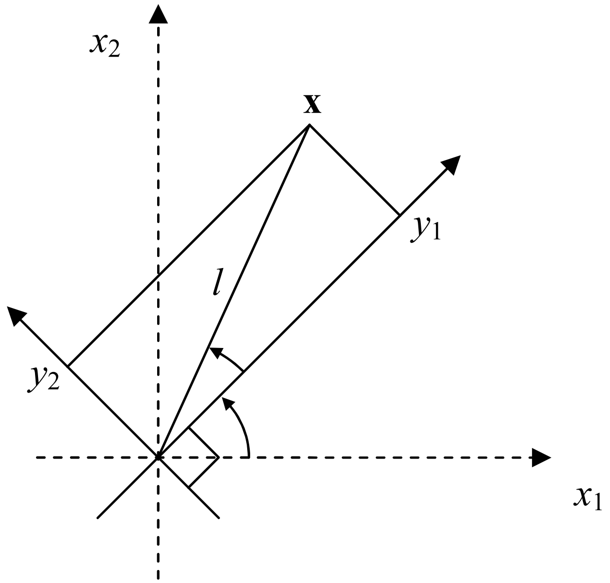
Rotations (and product) always remain orthonormal.

Construct update multiplicatively as

$$\mathbf{W}(t) = \mathbf{R}(t)\mathbf{R}(t-1)\cdots\mathbf{R}(1)\mathbf{W}(0)$$

where $\mathbf{R}(t)$ is a 2D Givens rotation.

2D Axis rotations



$$2J = \begin{cases} 0 & \text{if } y_1 \geq 0,\ y_2 \geq 0 \\ y_2^2 & \text{if } y_1 \geq 0,\ y_2 < 0 \\ y_1^2 & \text{if } y_1 < 0,\ y_2 \geq 0 \\ y_1^2 + y_2^2 = l^2 & \text{otherwise (i.e. } y_1 \geq 0,\ y_2 < 0) \end{cases}$$

Derivative and Algorithm

$$
\begin{aligned}
dJ/d\theta &= \begin{cases} 0 & \text{if } y_1 \geq 0,\ y_2 \geq 0 \\ y_2 y_1 & \text{if } y_1 \geq 0,\ y_2 < 0 \\ -y_1 y_2 & \text{if } y_1 < 0,\ y_2 \geq 0 \\ 0 & \text{otherwise } (y_1 \geq 0,\ y_2 < 0) \end{cases} \\
&= \mathbf{y}^{+} \times \mathbf{y}^{-} \\
&= y_1^{+} y_2^{-} - y_2^{+} y_1^{-}
\end{aligned}
$$

Gradient descent algorithm:

$$
\begin{aligned}
\Delta\phi &= -\eta_\phi \cdot dJ/d\phi = +\eta_\phi \cdot dJ/d\theta \\
&= \eta_\phi (y_1^{+} y_2^{-} - y_1^{-} y_2^{+})
\end{aligned}
$$

Relate this to concept of *torque* in a mechanical system.

Line Search over Rotation Angle

Instead of simple gradient descent, can use line search.
E.g. Matlab `fzero` for zero of $dJ/d\phi$.

If sources are non-negative as required, we know $\min(J) = 0$ so make local quadratic approximation and jump to

$$\phi(t+1) = \phi(t) - 2J(t)/(dJ(t)/d\phi)$$

OK since solution locally quadratic & curvature increases away from solution, as more data points 'escape' from the +ve quadrant.

More Than 2 Dimensions: Algorithm

1. Set $\mathbf{X}(0) = \mathbf{X}$, $\mathbf{W}(0) = \mathbf{I}$, $t = 0$.
2. Calculate $\mathbf{Y} = \mathbf{X}(t) = \mathbf{W}(t)\mathbf{X}(0)$
3. Calculate torques $g_{ij} = \sum_k y_{ik}^+ y_{jk}^- - y_{ik}^- y_{jk}^+$
4. Exit if $|g_{ij}| < \text{tolerance}$
5. For i^*, j^* maximizing $|g_{ij}|$ construct $\mathbf{X}^*$ from selecting rows i^* and j^* from $\mathbf{X}(t)$.
6. Do line search to find $\phi^*(t+1)$ which minimizes J.
7. Form the rotation matrix $\mathbf{R}(t+1) = [r(t+1)_{ij}]$ from $\phi^*(t+1)$.
8. Form the updated weight matrix $\mathbf{W}(t+1) = \mathbf{R}(t+1)\mathbf{W}(t)$ and modified input data $\mathbf{X}(t+1) = \mathbf{R}(t+1)\mathbf{X}(t) = \mathbf{W}(t+1)\mathbf{X}(0)$.
9. Increment the step count t, and repeat from 2.

Geodesic search

Successive rotations - equivalent to line search along axis directions.
Search in more general directions?

Geodesic - shortest path between 2 points on a manifold.

For orthonormal matrices, have [Edelmam, Arias & Smith, 1998]

$$\mathbf{W}(\tau) = e^{\tau \mathbf{B}} \mathbf{W}(0)$$

where $\mathbf{B}^T = -\mathbf{B}$ and τ scalar. [Fiori 2001, Nishimori 1999]

NB: In 2D we get

$$\mathbf{B} = \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix} \qquad e^{\tau \mathbf{B}} = \begin{pmatrix} \cos(\tau b) & \sin(\tau b) \\ -\sin(\tau b) & \cos(\tau b) \end{pmatrix}$$

Steepest Descent Geodesic

Parameterize $\mathbf{B} = \mathbf{C} - \mathbf{C}^T$ with $c_{ij} = 0$ for $i \geq j$. $\mathbf{C}$ has $n(n-1)/2$ free parameters.

For steepest descent in $\mathbf{C}$ space, maximize

$$-\lim_{\Delta\tau \to 0} \frac{(\text{change in } J \text{ due to } \Delta\tau)/\Delta\tau}{(\text{distance moved by of } \tau\mathbf{C} \text{ due to } \Delta\tau)/\Delta\tau} = \frac{dJ/d\tau}{\|C\|_{\mathsf{F}}}$$

We find

$$dJ/d\tau = \text{trace}((\mathbf{Y}_-\mathbf{Y}^T - \mathbf{Y}\mathbf{Y}_-^T)\mathbf{C}^T) = < (\mathbf{Y}_-\mathbf{Y}^T - \mathbf{Y}\mathbf{Y}_-^T), \mathbf{C} >$$

so for steepest descent choose

$$\mathbf{C} \propto \mathsf{UT}(\mathbf{Y}_-\mathbf{Y}^T - \mathbf{Y}\mathbf{Y}_-^T) = \mathsf{UT}(\mathbf{Y}_-\mathbf{Y}_+^T - \mathbf{Y}_+\mathbf{Y}_-^T)$$

Gradient descent

Simply update $\mathbf{W}$ according to

$$\mathbf{W}(t+1) = e^{-\eta(\mathbf{Y}_-\mathbf{Y}_+^T - \mathbf{Y}_+\mathbf{Y}_-^T)}\mathbf{W}(t)$$

with small update η. This is the *geodesic flow method* [Fiori 2001, Nishimori 1999].

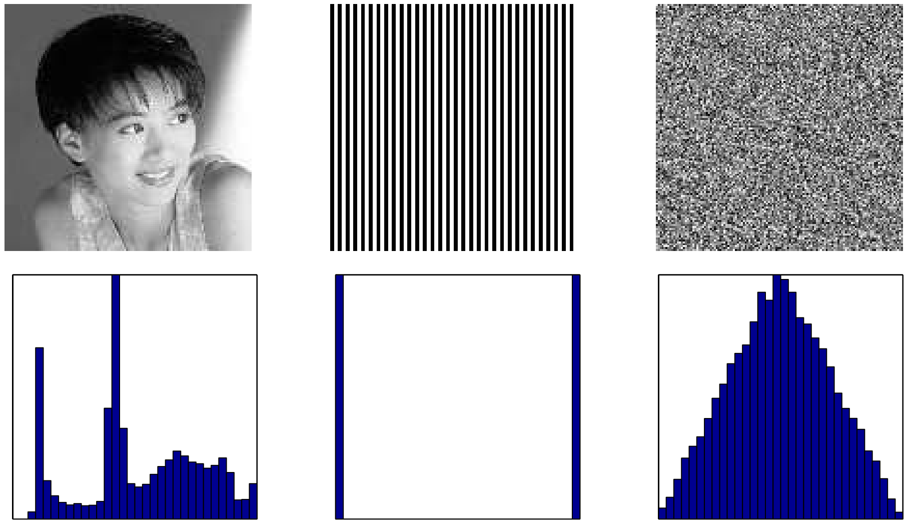
BUT - no need to restrict to small updates.
Can do e.g. *line search* along the steepest-descent geodesic.

Line Search along Geodesic: Algorithm

1. $\mathbf{X}(0) = \mathbf{X}$ and $\mathbf{W}(0) = \mathbf{I}$ at step $t = 0$.
2. Calculate $\mathbf{Y} = \mathbf{X}(t) = \mathbf{W}(t)\mathbf{X}(0)$.
3. Calculate gradient $\mathbf{G}(t) = \mathrm{UT}(\mathbf{Y}_-\mathbf{Y}_+^T - \mathbf{Y}_+\mathbf{Y}_-^T)$ and $\mathbf{B}$-space movement direction $\mathbf{H}(t) = -(\mathbf{G}(t) - \mathbf{G}(t)^T) = \mathbf{Y}_+\mathbf{Y}_-^T - \mathbf{Y}_-\mathbf{Y}_+^T$.
4. Stop if $\|\mathbf{G}(t)\| <$ tolerance.
5. Perform a line search for τ^* which minimizes $J(\tau)$ using $\mathbf{Y}(\tau) = \mathbf{R}(\tau)\mathbf{X}(t)$ and $\mathbf{R}(\tau) = e^{-\tau\mathbf{H}}$.
6. Update $\mathbf{W}(t+1) = \mathbf{R}(\tau^*)\mathbf{W}(t)$ and $\mathbf{X}(t+1) = \mathbf{R}(\tau^*)\mathbf{X}(t) = \mathbf{W}(t+1)\mathbf{X}(0)$.
7. Increment t, repeat from 2. Until exit at 4.

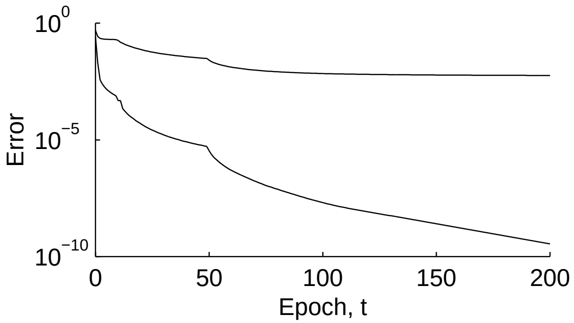
For simple tasks can guess single quadratic jump to $J = 0$
$\Delta\mathbf{C} = 2J\mathbf{G}/\|\mathbf{G}\|^2$.

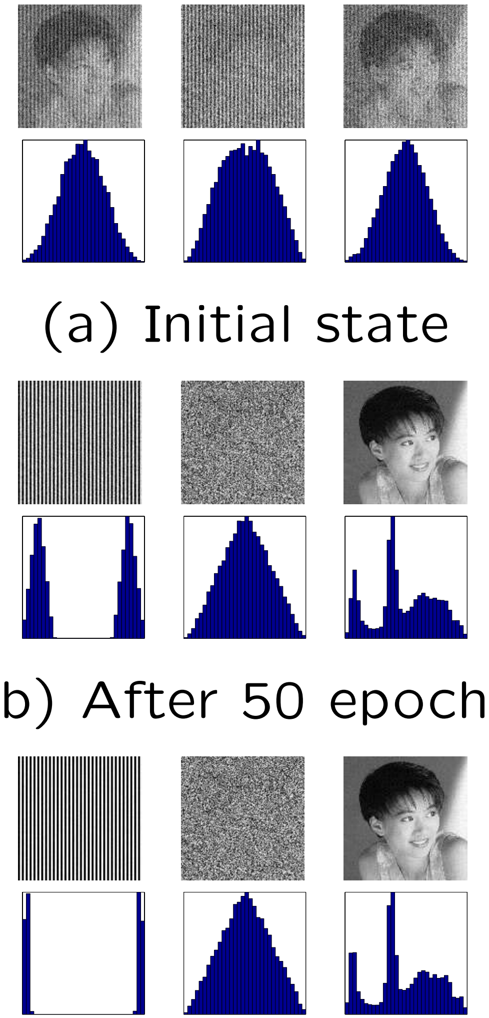
Results - Image separation problem



Source images and histograms [Cichocki, Kasprzak & Amari 1996].

Nonlinear PCA



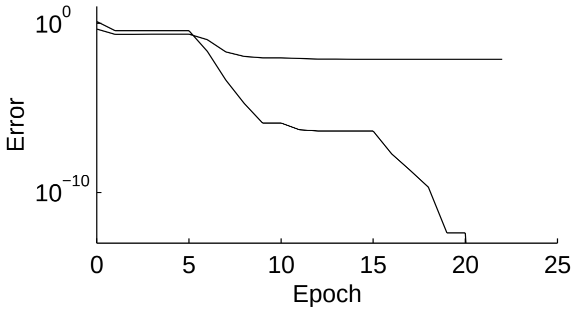


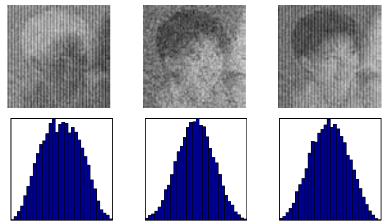
(a) Initial state

(b) After 50 epochs

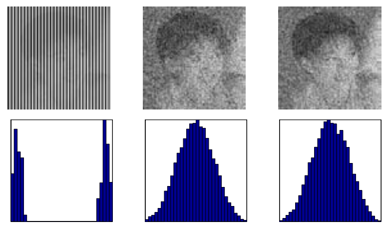
(c) After 200 epochs

Successive rotations

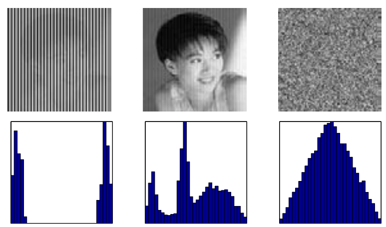




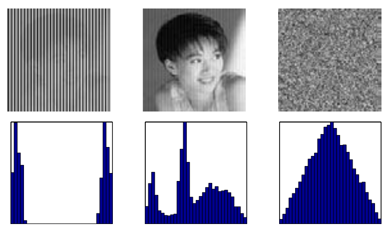
(a) Initial state



(b) 5 epochs: rotated axes 1-3

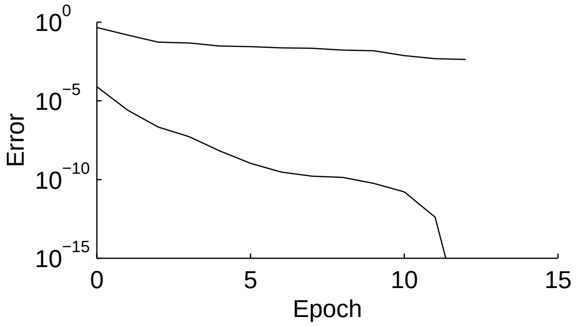


(c) 15 epochs: rotated axes 2-3



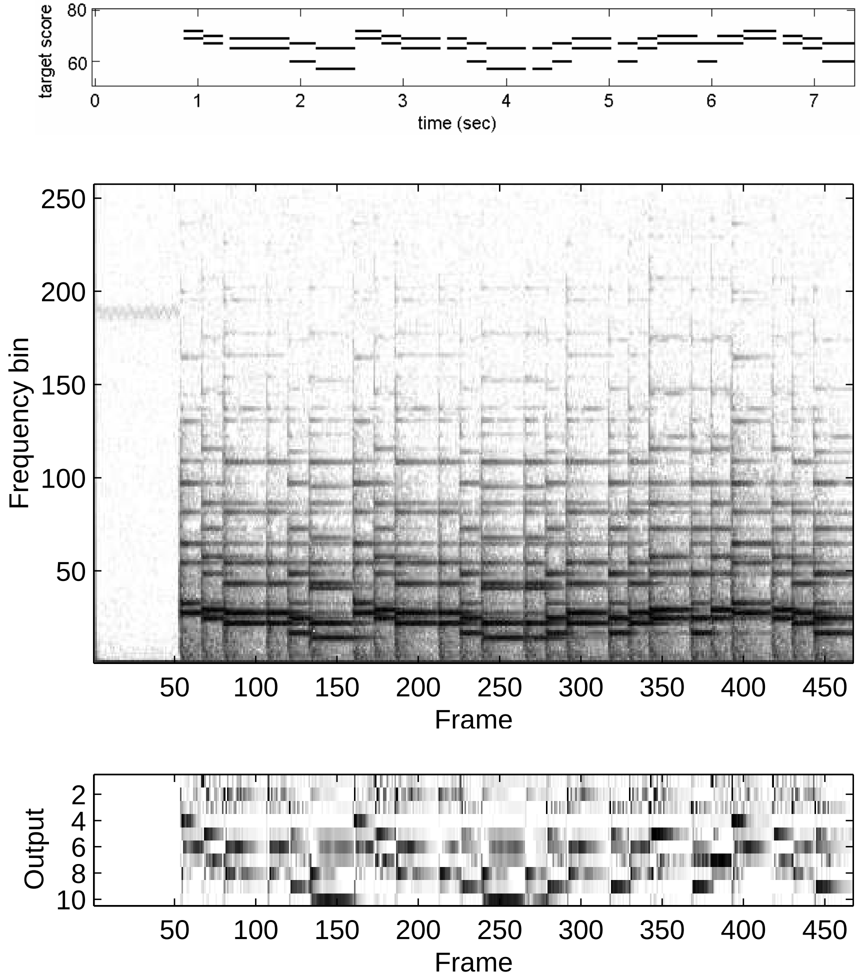
(c) 22 epochs: rotated axes 1-3

Geodesic step



(Visually similar images)

Music example: Liszt Etude No 5 (extract)



Conclusions

- Considered problem of *non-negative ICA*

- Separation of whitened sources when zero reconstruction error.

- Nonlinear PCA with $g(y) = y^+$.

- Successive rotations keeps orthogonality

- Geodesic line search