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

\[ X = AS \]  \hspace{1cm} (1)

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

- Task - to discover the source samples \( S \) and mixing matrix \( A \) given only the observations \( X \).

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

- 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 $S$ and $A$)
$\rightarrow$ 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

\[ x = Qz \]  \hspace{1cm} (2)

to get identity covariance \( C_x = E((x - \bar{x})(x - \bar{x})^T) = I \).

Problem now to find orthonormal weight matrix \( W \), satisfying
\( W^T W = W W^T = I_n \), such that the outputs \( y = Wx = WQAs \) are independent.

Typical ICA algorithms search for extremum of contrast function (e.g. kurtosis).
Whitening of Non-Negative Data

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 \( U = WQ_A \), i.e. \( y = Us \)

For non-negative sources \( s \), which are well-grounded (i.e. Pr(\( s < \delta \)) > 0 for any \( \delta > 0 \)),

**then**

\( U \) is a permutation matrix (i.e. sources are separated)

**iff** all components of \( y \) are non-negative w.p.1.

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

\[
J = \frac{1}{2} E(\|x - \hat{x}\|) \quad \hat{x} = W^T y^+
\]

where \( y^+ \) is rectified version of \( y = Wx \).
Nonlinear PCA algorithms

Natural to consider nonlinear PCA algorithm:

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

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

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

(“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 $W$ in Euclidean matrix space, tending towards orthonormality $WW^T = I$.

However, can instead construct $W$ from 2D rotations [Comon, 1994],

$$
\begin{pmatrix}
  y_{i1} \\
  y_{i2}
\end{pmatrix}
= \begin{pmatrix}
  \cos \phi & \sin \phi \\
  -\sin \phi & \cos \phi
\end{pmatrix}
\begin{pmatrix}
  x_{i1} \\
  x_{i2}
\end{pmatrix}
$$

Rotations (and product) always remain orthonormal.

Construct update multiplicatively as

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

where $R(t)$ 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

\[ \frac{dJ}{d\theta} = \begin{cases} 
0 & \text{if } y_1 \geq 0, \ y_2 \geq 0 \\
y_2y_1 & \text{if } y_1 \geq 0, \ y_2 < 0 \\
-y_1y_2 & \text{if } y_1 < 0, \ y_2 \geq 0 \\
0 & \text{otherwise } (y_1 \geq 0, \ y_2 < 0)
\end{cases} \]

\[ = y^+ \times y^- \]

\[ = y_1^+y_2^- - y_2^+y_1^- \]

Gradient descent algorithm:

\[ \Delta \phi = -\eta_\phi \cdot \frac{dJ}{d\phi} = +\eta_\phi \cdot \frac{dJ}{d\theta} \]

\[ = \eta_\phi(y_1^+y_2^- - y_2^+y_1^-) \]

Relate this to concept of \textit{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 $X(0) = X$, $W(0) = I$, $t = 0$.
2. Calculate $Y = X(t) = W(t)X(0)$
3. Calculate torques $g_{ij} = \sum y_{ik}^+ y_{jk}^- - y_{ik}^- y_{jk}^+$
4. Exit if $|g_{ij}| <$ tolerance
5. For $i^*, j^*$ maximizing $|g_{ij}|$ construct $X^*$ from selecting rows $i^*$ and $j^*$ from $X(t)$.
6. Do line search to find $\phi^*(t + 1)$ which minimizes $J$.
7. Form the rotation matrix $R(t+1) = [r(t+1)]_{ij}$ from $\phi^*(t+1)$.
8. Form the updated weight matrix $W(t + 1) = R(t + 1)W(t)$ and modified input data $X(t + 1) = R(t + 1)X(t) = W(t + 1)X(0)$.
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 [Edelman, Arias & Smith, 1998]

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

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

NB: In 2D we get

\[
\mathbf{B} = \begin{pmatrix} 0 & b \\
-b & 0 \end{pmatrix}
\]

\[
e^{\tau \mathbf{B}} = \begin{pmatrix} \cos(\tau b) & \sin(\tau b) \\
-\sin(\tau b) & \cos(\tau b) \end{pmatrix}
\]
Steepest Descent Geodesic

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

For steepest descent in $C$ space, maximize

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

We find

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

so for steepest descent choose

$$C \propto U^T(Y - Y^T - YY^T) = U^T(Y - Y^T_+ - Y_+ Y^-)$$
Gradient descent

Simply update $\mathbf{W}$ according to

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

with small update $\eta$. 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. $X(0) = X$ and $W(0) = I$ at step $t = 0$.
2. Calculate $Y = X(t) = W(t)X(0)$.
3. Calculate gradient $G(t) = U^T(Y_+ Y_+^T - Y_- Y_-^T)$ and B-space movement direction $H(t) = -(G(t) - G(t)^T) = Y_+ Y_+^T - Y_- Y_-^T$.
4. Stop if $\|G(t)\| < \text{tolerance}$.
5. Perform a line search for $\tau^*$ which minimizes $J(\tau)$ using $Y(\tau) = R(\tau)X(t)$ and $R(\tau) = e^{-\tau H}$.
6. Update $W(t + 1) = R(\tau^*)W(t)$ and $X(t + 1) = R(\tau^*)X(t) = W(t + 1)X(0)$.

For simple tasks can guess single quadratic jump to $J = 0$
\[ \Delta C = 2JG/\|G\|^2. \]
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