Signal Processing for Functional Brain Imaging: Independent Component Analysis

Dimitri Van De Ville

Medical Image Processing Lab, EPFL/UniGE
dimitri.vandeville@epfl.ch

March 27, 2014
From model-based...

- Models so far
  - Derived from stimulation paradigm: GLM
    - For every voxel separately
    - Fits weights of regressors
    - Easy to interpret, fast to compute
    - Model selection
I believe in ignorance-based methods because humans have a lot of ignorance and we should play to our strong suit.

- Eric Lander, Whitehead Institute, M.I.T.
Overview

- Curse of dimensionality
  - Rationale for dimensionality reduction

- Subspace decomposition
  - Principal component analysis (PCA)
  - Link with singular value decomposition (SVD)
  - Bilinear representation

- Independent component analysis (ICA)
  - Blind source separation
  - Importance of non-Gaussianity
  - Multi-subject ICA
Dimensionality reduction

- Consider the (f)MRI case
  - Data point is in N-dimensional feature space (N=#voxels)
    - N typically 10’000-100’000
  - Feature correspond to value at one voxel position
  - Number of feature vectors
    - single subject: #time points
    - multi-subjects: #subjects (one volume or contrast per subject)
      - typically 100-1’000
  - High-dimensional data, but few datapoints!
Curse of dimensionality

- Discovery of multivariate patterns becomes increasingly difficult as dimensionality increases

- Toy-problem classification
  - Three classes, given N examples
  - Divide feature space into uniform bins
  - Test sample: find its bin and assign predominant class label

- Start with a real-valued single feature, 3 bins

- Works, but overlap is too big
  - Let’s add 2nd feature to improve separability

[inspired from Gutierrez-Osuna]
Curse of dimensionality

- Maintaining the number of bins / feature dimension leads to $3^2 = 9$ bins in 2D
  - Same density
    - 9 (3x3) examples for a single feature
    - 27 (3x3x3) examples for two features
  - Same number of samples renders a sparse scatterplot!
    - learning becomes more difficult
Dimensionality reduction

Consider the (f)MRI case

- Data point is in N-dimensional feature space (N=#voxels)
  - N typically 10’000-100’000
- Feature correspond to value at one voxel position
- Number of feature vectors
  - single subject: #time points
  - multi-subjects: #subjects (one volume or contrast per subject)
  - typically 100-1’000

High-dimensional data, but few datapoints!

General idea

- Can we represent the data in a meaningful way in a low-dimensional space?
- Construct new “feature dimensions” that correspond to maps (linear combinations of original features)
Perform (linear) dimensionality reduction by maximizing variance in dimensionality-reduced space

We look for a linear mapping into $M$-dimensional space $(M \leq N)$

$$
\begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_N
\end{bmatrix}
\longrightarrow
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_M
\end{bmatrix}
= \begin{bmatrix}
  w_{11} & w_{12} & \cdots & w_{1N} \\
  w_{21} & w_{22} & \cdots & w_{2N} \\
  \vdots & \vdots & \ddots & \vdots \\
  w_{M1} & w_{M2} & \cdots & w_{MN}
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_N
\end{bmatrix}
$$

Representation of $x$ in orthonormal basis $E = [e_1 \ldots | e_N]$

- Basis: $e_k^T e_l = \delta_{k,l}$ (Kronecker delta) or, equivalently, $EE^T = I$
- Representation: $x = \sum_{k=1}^{N} y_k e_k = Ey$; and $y = E^T x$
Principal components analysis

Let’s perturb the representation of $x$ by changing the coefficients of the $M + 1, \ldots, N$-th basis vectors into $b_{M+1}, \ldots, b_N$ deterministic!

- Modified representation: 
  $$\tilde{x}(M) = \sum_{k=1}^{M} y_k e_k + \sum_{k=M+1}^{N} b_k e_k$$

- Difference signal:
  $$\Delta x(M) = \left( \sum_{k=1}^{N} y_k e_k - \left( \sum_{k=1}^{M} y_k e_k + \sum_{k=M+1}^{N} b_k e_k \right) \right) = \sum_{k=M+1}^{N} (y_k - b_k) e_k$$

- Approximation error:
  $$e^2(M) = E \left[ |\Delta x(M)|^2 \right] = E \left[ \sum_{k=M+1}^{N} \sum_{l=M+1}^{N} (y_k - b_k)(y_l - b_l) e_k^T e_l \right]$$
  $$= \sum_{k=M+1}^{N} E \left[ (y_k - b_k)^2 \right]$$
Principal components analysis

- Optimal values of $b_k$?

$$\frac{\partial}{\partial b_k} E \left[ (y_k - b_k)^2 \right] = -2 (E[y_k] - b_k) = 0$$

So we set $b_k = E[y_k]$, which turns the approximation error into we should demean!

$$\epsilon^2(M) = \sum_{k=M+1}^{N} E \left[ (y_k - E[y_k])^2 \right]$$

$$= \sum_{k=M+1}^{N} E \left[ (x^T e_k - E[x^T e_k])^T (x^T e_k - E[x^T e_k]) \right]$$

$$= \sum_{k=M+1}^{N} e_k^T E \left[ (x - E[x])(x - E[x])^T \right] e_k$$

 covariance matrix $V_x$
Can we find optimal $e_k$ (minimizing $\epsilon^2(M)$) that are orthonormal?

$$\epsilon^2(M) = \sum_{k=M+1}^{N} e_k^T V_x e_k + \sum_{k=M+1}^{N} \lambda_k (1 - e_k^T e_k)$$

constraints with Lagrange multipliers $\lambda_k$

Partial derivative with respect to $e_k$:

$$\frac{\partial}{\partial e_k} \epsilon^2(M) = 2 (V_x e_k - \lambda_k e_k) = 0 \rightarrow V_x e_k = \lambda_k e_k$$

So $e_k$ and $\lambda_k$ are the eigenvectors and eigenvalues of $V_x$.

Reminder:

$$\frac{d}{dx} (x^T A x) = (A + A^T) x$$

$$= 2 A x, \quad \text{(if } A \text{ symmetric)}$$
Principal components analysis

Approximation error then becomes

\[ \epsilon^2(M) = \sum_{k=M+1}^{N} e_k^T V x e_k = \sum_{k=M+1}^{N} \lambda_k \]

1. To minimize the approximation error, we need to put the smallest eigenvalues \( \lambda_k \) in this sum

2. Or, equivalently, to represent \( x \) optimally, we choose eigenvectors with largest eigenvalues first
The optimal approximation of a random vector $\mathbf{x}$ ($N \times 1$) by a linear combination of $M$ ($M < N$) independent vectors is obtained by projecting the random vector onto eigenvectors $\mathbf{e}_k$ corresponding to the largest eigenvalues $\lambda_k$ of the covariance matrix $\mathbf{V}_x$

- We defined optimality minimum of magnitude of approximation error

PCA diagonalizes the covariance matrix $\mathbf{V}_x$

$$\mathbf{V}_x \mathbf{e}_k = \lambda_k \mathbf{e}_k \rightarrow \mathbf{V}_x \mathbf{E} = \mathbf{E} \Lambda \rightarrow \mathbf{E}^T \mathbf{V}_x \mathbf{E} = \Lambda$$

consequently

$$\mathbb{E}[\mathbf{y} \mathbf{y}^T] = \Lambda$$
Principal components analysis

- Oldest technique in multivariate analysis [Pearson, 1901]
  - Communications: known as Karhunen-Loève transform [Loève, 1963]
  - Statistics: factor analysis
    - eigenvectors $e_k$ are factors, notation $\lambda_k$ is used for factor loadings
- Hotelling $T^2$ test:
  - similar to two-sided $t$-test after rotating along dominant principal component

- Algebra: singular value decomposition of matrix $X = U \Sigma V^T$, where $U$ ($N \times N$) and $V$ ($N' \times N'$) are unitary, and $\Sigma$ is $N \times N'$ diagonal
  - left singular vectors $U$ are eigenvectors of $XX^T$ since $U \Sigma \Sigma^T U^T$
  - right singular vectors $V$ are eigenvectors of $X^T X$ since $V \Sigma^T \Sigma V^T$
  - non-zero singular values of $\Sigma$ are square-roots of non-zero eigenvalues of $XX^T$ or $X^T X$
Decomposition of the data matrix

- For functional data
  - Orthogonal maps & timecourses from FMRI data

\[
\text{Data (X)} = \text{Temporal components} \times \text{Diagonal matrix with weights } (\Sigma) \times \text{Spatial components } (V^T)
\]
Decomposition of the data matrix

- Bilinear decomposition model
  - Assume $K$ spatially orthogonal sources (i.e., subspace spanned by $V: V \times K$), then we can write the following generative model for the data matrix $X$:

  $$X = U\Sigma V^T + \text{noise}$$

  where $A = U\Sigma$ is the mixing matrix/factor loadings $(T \times K)$. Equivalently, we have

  $$x(t, v) = \sum_{k=1}^{K} a_k(t)v_k(v) + e(t, v)$$

  - Sorting according to explained variance and orthogonality
  - Allows to estimate the number of sources [Minka, 2001]
    (Assume i.i.d. Gaussian noise, find MAP estimate for $K$)

  - Does the sorting makes sense? Depends on the application!
    If not, any unitary matrix (rotation) gives other sources
Decorrelation and Gaussian processes

- Two random variables are decorrelated iff
  \[ E[y_1y_2] = E[y_1]E[y_2] \]

- Two random variables are independent iff
  \[ E[y_1^p y_2^q] = E[y_1^p]E[y_2^q], \]
  for every positive integer value of \( p \) and \( q \)
  - Implies decorrelation, but not the other way around

- Gaussian random variable is independent if decorrelated!
  \[
  \frac{1}{(2\pi)^{N/2} \sqrt{|V_x|}} \exp\left( -\frac{1}{2} (x - \mu)^T V_x^{-1} (x - \mu) \right) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi \sigma_k}} \exp\left( -\frac{(x_k - \mu_k)^2}{2\sigma_k^2} \right)
  \]
Cocktail problem: blind source separation

MIXING

\[ s_1 \rightarrow a_{11} \rightarrow x_1 \rightarrow y_1 = s_i \]

\[ s_2 \rightarrow a_{12} \rightarrow a_{21} \rightarrow a_{22} \rightarrow x_2 \rightarrow w_{11} \rightarrow w_{12} \rightarrow w_{21} \rightarrow w_{22} \rightarrow y_2 = s_j \]

SEPARATION

training

[Gutierrez-Osuna, Introduction to Pattern Analysis]
ICA... exploiting non-Gaussianity!

non-Gaussian sources

[from Beckmann, FSL MELODIC]
Definition of ICA

- Goal: separate $K$ sources from linear mixtures

- Model:

\[ X = AS \]

- $X$ ($T \times V$) is the mixture, $A$ ($T \times K$) is the mixing matrix, and $S$ ($K \times V$) contains the sources (as rows).

- We also have $S = WX$ where $W$ is the demixing matrix

- Assumptions

  - Linear mixing
  - Independence of sources
  - Non-Gaussian sources
  - Mixtures and sources have zero-mean
Difficulties

- Variance of independent components cannot be determined
  - Any multiplicative factor in $S$ can be compensated by the mixing matrix $A$
  - Normalize sources to unit variance

- Order of independent component cannot be determined
  - Any permutation of the mixing/sources yields same result

- Distribution of sources should be non-Gaussian
  - Mixture of Gaussian sources remains Gaussian; decorrelation (by PCA) would turn independent
General idea (1)

- **Central Limit Theorem [Liapunov, 1901]**

If a set of signal \( s = (s_1, \ldots, s_M) \) are independent with means \((\mu_1, \ldots, \mu_M)\) and *well-defined* variances \((\sigma_1^2, \ldots, \sigma_M^2)\) then, for a large number \( M \) of signals \( s \), the signal

\[
x = \sum_{k=1}^{M} s_k,
\]

has a probability density function which tends to a Gaussian (for increasing \( M \)) with mean \( \sum_k \mu_k \) and variance \( \sum_k \sigma_k^2 \).

- **Corollaries**

  - The same holds for a weighted sum (mixture) \( x = \sum_{k=1}^{M} a_k s_k \)

  - For low \( M \) the approximation of \( x \)'s pdf to a Gaussian is not impressive, but the CLT ensures that the mixture *more* Gaussian than the pdf of its constituent source signals
General idea (2)

- Retrieve sources by looking for directions of largest non-Gaussianity

  - Assume all sources have identical distributions

  - Goal: find $w$ such that $y = w^T x$ is equal to one of the sources in $s$

  - CLT: signal $y$ is more Gaussian when it’s a linear combination of sources; less Gaussian when it becomes one of the sources

  - Optimal $w$ is vector that maximizes non-Gaussianity of $w^T x$...

  - but how to measure it?

    - kurtosis (leptokurtic sources are sparse), but sensitive to outliers!

    - negentropy (differential entropy relative to Gaussian); e.g., FastICA

  - Centering & whitening (=PCA) as preprocessing step
First proof of concept

[Hastie et al., Elements of Statistical Learning]
Spatial ICA

- For functional data
  - Unmixing of activated brain regions from FMRI data

\[
\text{Data (X)} = \text{Mixing matrix} \times \text{Spatially independent components (S)}
\]

- Compare against the classical GLM

\[
\text{Data (X)} = \text{Design matrix} \times \text{Activation maps (beta)}
\]

26
[McKeown, 1998]
**FMRI artefacts**

- Detect and reduce (back-reconstruction) artefacts
  - Eye movements
  - Nyquist ghost

[C. Beckmann, Little Shop of fMRI Horrors]
ICA for denoising

- These nuisance components are difficult or impossible to predict and include as a regressor in the GLM.
- Their presence influences the results of statistical hypothesis testing:
  - Appear as “structured noise” that violates the usual noise assumptions!
  - Variance estimates based on the GLM residuals is inflated:
    - more false negatives!
  - Influences the values of the parameter estimates:
    - more false positives and/or false negatives!
- Possible solution:
  - Use ICA to identify nuisance ICs
  - Apply GLM after regressing out the timecourses of these components
ICA for group studies: Concat-ICA

- Group ICA [Calhoun et al., 2001]

\[ X_i = A_i S = \sum_{k=1}^{K} a_k^{(i)} \otimes s_k \]

- \( X_i(T \times V) \): observation from \( i \)-th subject
- \( A_i(T \times K) \): mixing matrix for \( i \)-th subject
- \( S(K \times V) \): statistically independent group spatial sources

Each component has a different timecourse for each subject.
ICA for group studies: Tensor-ICA

- Probabilistic tensor group ICA [Beckmann, Smith, 2005]

\[ X = \sum_{k=1}^{K} a_k \otimes s_k \otimes c_k + E_i \]

- \( X(T \times V \times N) \): observation from \( N \) subjects
- \( a_k(T \times 1) \): mixing matrix
- \( s_k(V \times 1) \): statistically independent group spatial sources
- \( c_k(N \times 1) \): subject loadings

![Diagram of tensor decomposition]

Each subject has the same timeseries!
ICA for group studies

- Different subject populations/conditions should be pooled in the same decomposition to make them comparable!

- Tensor ICA imposes additional structure on the decomposition
  - Can be generalized to higher-order tensors
    - For conditions
    - For multi-modal imaging
  - Subject loadings can be subjected to subsequent statistical analysis
  - Assumes consistency between different dimensions
    - Does not hold for resting-state since subjects are not consistent
Conclusion

- ICA is an exploratory method
  - Not model-free, but multivariate, data-driven and without assumptions on timecourses or spatial patterns
- ICA provides a more flexible analysis than hypothesis-based methods
  - Dangers of a fishing expedition: results from ICA should be used in the right way to get sound results!
- Some contributions of ICA to fMRI
  - Find activation patterns that respond in a complex way to the stimulation paradigm
  - Identify temporally coherent (but task-unrelated) components
  - Appropriate for resting-state analysis
  - Reduce artefacts
  - Come up with new hypotheses after exploration
More reading

- References
More reading

- References
  - G Strang, *Best basis from the SVD*. Computational Science and Engineering (Ch 1.8), 2007
Software packages

- GIFT/GroupICAT (Matlab, multiple ICA algorithms)
  - Concat-ICA, SPM style
- FSL/MELODIC (C, FastICA+)
  - Concat-ICA and Tensor-ICA, commandline
- ICA:DTU Toolbox (Matlab)
- AnalyzeFMRI (R, FastICA)
- BrainVoyager (Commercial, FastICA)
- FMRLAB (Matlab, InfoMax)
- ICALAB (Matlab)