Average association clustering with Residuals

Martin Vejmelka

18.11.2010

Outline

What is clustering ?

- “Finding natural (and/or interesting) subgroups of data.”
- Many clustering approaches and methods exist
- Clustering approaches differ widely
  disjoint/overlapping clusters, hierarchical/direct, probabilistic
- Assumptions of each method must be considered
  shape, size, parametrization of clusters
- Computational efficiency required for large datasets

Spectral graph clustering

- A family of successful techniques for partitioning data
- Already used in many research fields
  document clustering, image segmentation, genetics, ...
- Available theoretical results linking SGC to other methods (kernel k-Means)
- Represents data by a graph and analyzes the weight matrix of the graph
- Based on objective optimization
  Ratio cut, Normalized Cut, Average Association, ...

Average association with Residuals (AAR)

Given elements V, $|V| = n$, a symmetric weight matrix $W \in \mathbb{R}^{n \times n}$ and the number $k > 0$, construct a partition $\mathcal{V} = \{V_1, V_2, ..., V_k, V_o\}$ of the set V so that the objective

$$J_k = \sum_{l=1}^{k} S_l = \sum_{l=1}^{k} \sum_{v_i, v_j \in V_l} \frac{w_{i,j}}{n_l},$$

where $|V_l| = n_l$ and S_l is the "cluster strength" of V_l, is maximized over all partitions of V. The set V_o is will be called the residual set.

Average association with Residuals — notes

- Need not partition all elements ("remainder" is in V_o)
- We may set $k = 1$ to get one cluster ("main mode" of data)
- Does not manipulate the weight matrix W
  unlike e.g. normalized cut SGC
- May retrieve "sparse" clusters in the sense that the clusters cover major structures in the analyzed data but contain a small part of the elements in the data

Motivation

- We are asked to find spatial modes of low frequency spontaneous brain activity in the human brain
- We have fMRI measurements from the gray matter (GM): 50k voxels, 300 time points
- Spontaneous brain activity may be characterized by resting state networks (RSNs)
- RSNs are regions in the brain exhibiting coherent fluctuations
- We do not know if all anatomical regions in the GM belong to some RSN
- Can we find RSNs using clustering ? How ?

Outline

Matrix formulation I

The AAR objective can be rewritten in matrix form using indicator vectors for the k clusters $u_l, l \in \{1, 2, ..., k\}$

$$[u_l]_j = \begin{cases} 1 & \text{if } v_j \in V_l \\ 0 & \text{if } v_j \notin V_l. \end{cases}$$

The objective can be rewritten in matrix form

$$J_k = \sum_{i=1}^{k} \frac{u_i^T W u_i}{u_i^T u_i}.$$

The disjointness constraint is simply expressed: $u_i^T u_j = 0$ when $i \neq j$.

Matrix formulation II

The indicator vectors can be normalized to unit size to remove the division by $u_i^T u_i$ in J_k. This is simple, if $z_i = u_i/||u_i||$, then

$$[z_i]_j = \begin{cases} n_i^{-1/2} & \text{if } v_i \in V_j \\ 0 & \text{if } v_i \notin V_j. \end{cases}$$

We accumulate the normalized indicator vectors into a matrix $Z = (z_1, z_2, ..., z_k)$ and write the objective in trace form:

$$J_k = \text{tr}\{Z^T W Z\}, Z^T Z = I_k,$$

where the values of z_i are constrained as above.

Solution by relaxation

Relaxation of the discrete constraint optimization problem is a standard tool for finding "good" solutions. A standard relaxation of the optimization problem rests in removing the constraint on the discrete nature of the values of Z. The relaxed optimization problem becomes:

$$J_k = \mathrm{tr}\{Z^T W Z\}, Z^T Z = I_k,$$

where W is the weight matrix (connectivity matrix).

Solution by relaxation II

Let $WY = Y\Lambda$ be the eigendecomposition of W such that:

$$Y^T Y = I_N, Y = (y_1, y_2, ..., y_N)$$

and

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_N), \lambda_l \geq \lambda_{l+1}.$$

Then the objective J_k is maximized if $z_i = y_i$ for $l \in \{1, 2, ..., k\}$ and the value of the maximized objective is

$$J_k = \sum_{l=1}^{k} \lambda_l.$$

(Yu and Shi, 2003)

Solution by relaxation III

The solution $Y = (y_1, y_2, ..., y_k)$ that maximizes the objective J_k is not unique, rather, the space of all solutions is parametrized by orthogonal matrices $R \in O(k)$ so that

$$Z = YR$$

is also a solution. Matrix trace is unaffected by orthogonal transformation $J_k = \mathrm{tr}\{Y^T WY\} = \mathrm{tr}\{R^T Y^T WYR\} = \mathrm{tr}\{Z^T WZ\}$ and the orthogonality constraint remains in effect $Z^T Z = R^T Y^T YR = R^T R = I_k$. (Yu and Shi, 2003)

Solution by relaxation IV.

- There are many relaxed solutions, which one to select ?
- Apply VARIMAX method (Kaiser, 1958) to the Z_k
- VARIMAX finds an orthogonal transform $R_V \in O(k)$ such that

$$R_V = \underset{R \in O(k)}{\arg\max}\, v^*(Z_k R),$$

where the VARIMAX objective $v^*(Y)$, $Y \in \mathbb{R}^{n \times k}$ is

$$v^*(Y) = \frac{1}{n^2} \sum_{l=1}^{k} \left(n \sum_{i=1}^{n} y_{i,l}^4 - \left(\sum_{i=1}^{n} y_{i,l}^2 \right)^2 \right).$$

- VARIMAX objective attempts to quantify the concept of "simple structure" (Thurstone, 1935)

Discretization

- Heuristic procedure
- Use VARIMAX rotation to obtain R_V, compute $Z^* = Z_K R_V$
- Each element is pre-assigned to the cluster $v_i \in V_l$ if

$$l = \underset{m=\{1,2,\ldots,k\}}{\arg\max} \; (z^*_{i,m})^2$$

- The normalized indicator vector (slide 2) for different cluster sizes is fit to the relaxed solution (least-squares fit) and the best fit wins
- Many other choices considered but most require building the weight matrix W explicitely

Example I

Let the graph $G = (V, E)$ be composed of two cliques containing elements $g_1, ..., g_{n_1}$ and $g_{n_1+1}, ..., g_n$, $n_1 + n_2 = n, n_1 > n_2$. The cliques are mutually disconnected, all edges inside the cliques are weighted $\epsilon > 0$. The weight matrix W is block-diagonal with two blocks. The matrix has two non-zero eigenvalues $\lambda_1 = (n_1 - 1)\epsilon$ and $\lambda_2 = (n_2 - 1)\epsilon$ and $n - 2$ zero eigenvalues. The eigenvector corresponding to λ_1 is

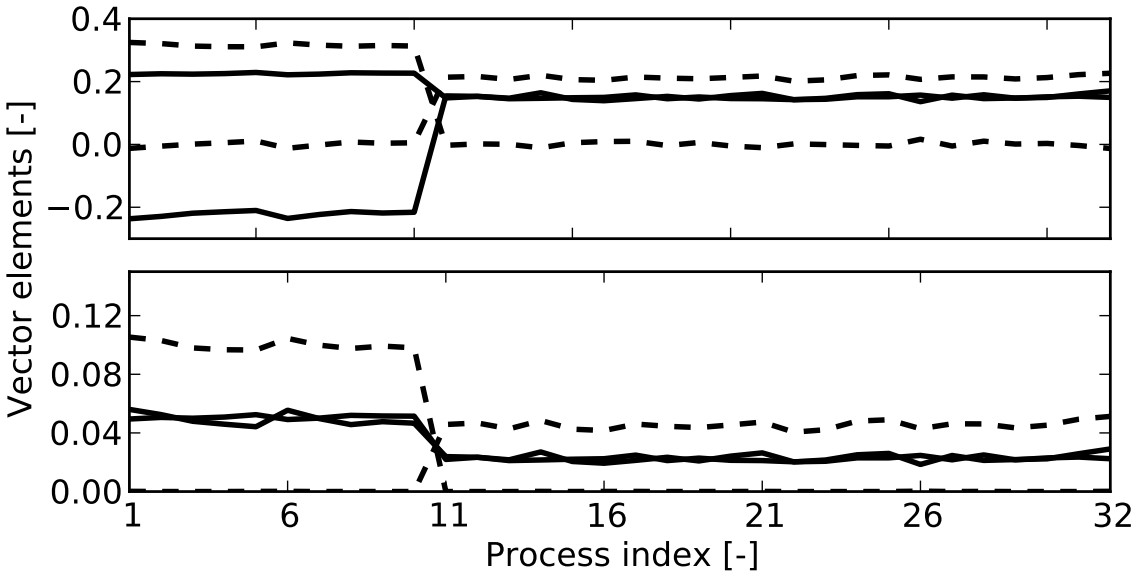
$$y_1 = (\underbrace{\alpha, \alpha, ..., \alpha}_{n_1}, \underbrace{0, 0, ..., 0}_{n_2}), \alpha = \frac{1}{\sqrt{n_1}}$$

while the eigenvector corresponding to λ_2 is

$$y_2 = \underbrace{(0, 0, ..., 0}_{n_1}, \underbrace{\beta, \beta, ..., \beta}_{n_2}), \beta = \frac{1}{\sqrt{n_2}}.$$

Example II

Let us have a model simulating $N = 32$ coupled dynamical systems. Two clusters: $N_1 = 10$, mean connectivity $\rho_1 = 0.8$ and $N_2 = 22$, $\rho_2 = 0.34$. Intercluster connectivity $\rho_{\text{int}} = 0.2$. Actual entries in matrix $\sim \mathcal{N}(\rho, \sigma(\rho))$.



(Top) full=eigvecs, dashed=rotated. (Bottom) same with squared elements.

Restatement

- AAR partitions datasets into k clusters and a residual set
  residual set may be empty
- Interpretation of clusters depends on W
  W is provided to us
- Part A: compute a relaxed solution
  - Run eigendecomposition on connectivity matrix W
  - Use VARIMAX transform to “enhance” cluster structure
- Part B: discretize relaxed solution to form clusters

AAR for time series analysis

Let us have n elements we wish to cluster and for each element $v_i \in V$, let there be a time series $t_i \in \mathbb{R}^p$. The weight matrix (similarity matrix) is now not given but rather estimated from the given time series. Examples:

- v_i are EEG electrodes, then t_i are the EEG time series
- v_i are weather stations, t_i are temperature profiles
- v_i are brain voxels and t_i are BOLD fluctuations (fMRI)
- v_i are obligations (stock market), and t_i are their values

We may estimate similarity between v_i using e.g. Pearson correlation, mutual information, coherence, ...

Required result: Singular Value Decomposition

Let $XV = U\Sigma$ define the reduced SVD of the matrix $X \in \mathbb{R}^{n \times p}$, $n \geq p$ so that

- $V^T V = I_p$, $V \in O(k)$ is an orthogonal matrix
- Columns of V are right singular vectors
- $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_p)$, $\sigma_i \geq \sigma_{i+1}$ is diagonal
- $U^T U = I_p$, $U \in \mathbb{R}^{n \times p}$ has orthonormal columns
- Columns of U are left singular vectors

Shortcut lemma

Let a matrix $X \in \mathbb{R}^{n \times p}$ be given with rows x_i representing p features of the i-th element from a set of n elements. Let the connectivity (or similarity) function $f : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ between the features of the elements be defined for each pair of row vectors of the matrix X. The function f must be symmetric in its arguments. If there is a function $g : \mathbb{R}^p \to \mathbb{R}^q$ such that

$$\forall x_i, x_j \in \mathbb{R}^p : f(x_i, x_j) = \langle g(x_i), g(x_j) \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product, then the eigenvalues of the connectivity matrix $W = (w_{i,j})$, $w_{i,j} = f(x_i, x_j)$ are $(\sigma_1^2, \sigma_2^2, ..., \sigma_q^2)$ and the corresponding eigenvectors are $(u_1, u_2, ..., u_q)$, where σ_i are the singular values of X_G and u_i are the left singular vectors of X_G (i-th row of X_G is $g(x_i)$). Note: $n - q$ eigvals of W are zero.

Example: fMRI analysis with Pearson correlation

- $T \in \mathbb{R}^{50000 \times 300}$ typical for 3T MRI system, 10 mins
- Pearson correlation coefficient to estimate similarity

$$\rho_{i,j} = f(t_i, t_j) = \frac{1}{299} \sum_{m=1}^{300} \frac{([t_i]_m - \bar{t}_i)([t_j]_m - \bar{t}_j)}{\sigma_i \sigma_j}.$$

- If $[\tilde{t}_i]_m = \frac{1}{\sqrt{299}}([t_i]_m - \bar{t}_i)/\sigma_i$. Then $\rho_{i,j} = \langle \tilde{t}_i, \tilde{t}_j \rangle$.
- $W = \tilde{X}\tilde{X}^T \approx 2 \times 10^9$ elements ($\approx$ 18 GB with 64-bit floats)
- VIC3 in Gent: 80 Intel CPUs take 25 mins to build & dcmp.
- For SVD of $\tilde{X}$, can use eigdcmp of $\tilde{X}^T\tilde{X}$ of size 300×300
  (a few seconds on a laptop, speedup $\approx 10^4$)

Preprocessing of fMRI/time series data

- For fMRI: data $T \in \mathbb{R}^{n \times p}$ (voxel time series)
- We first center the data T to obtain T_C by removing row and column mean (important !)
- We scale the data by a factor $X = (p-1)^{-\frac{1}{2}} T_C$
- Sample covariance matrix $\mathrm{cov}(T_C) = \frac{1}{p-1} T_C T_C^T = XX^T$
- Notation
  - $M_{[k]}$ means first k columns of matrix M
  - $M_{[k \times k]}$ is $k \times k$ submatrix obtained by removing rows $r > k$ and columns $r > k$

AAR/C

- AAR with covariance connectivity: AAR/C
- Weight matrix $W = cov(T_C) = XX^T$
- Given k, relaxed solution is $U_{[k]}$, $X = U\Sigma V^T$ (shortcut)
- Relaxed solution can be obtained by a linear mapping of data

$$U_{[k]} = XV_{[k]}\Sigma^{-1}_{[k\times k]}$$

- $F_{[k]} = V_{[k]}\Sigma^{-1}_{[k\times k]}$ is a matrix mapping X to first k eigenvectors
- The relaxed solution of the AAR/C problem for data T_C is

$$C_{\mathrm{AAR/C}} = XF_{[k]}R_V,$$

where R_V is the VARIMAX transformation for $U_{[k]}$

Restatement

- AAR is a clustering framework, interpretation of clustering depends on W
- W computed from time series may have special structure (outer product)
- If this special structure is there:
  - we may exploit it for faster computation
  - relaxed solution is a **linear mapping** of the data matrix (we never compute W !!)
- The mapping may be constructed using the SVD of X

$$C_{\mathrm{AAR/C}} = XF_{[k]}R_V, \quad F_{[k]} = V_{[k]}\Sigma^{-1}_{[k\times k]}$$

Outline

Independent Component Analysis I

The ICA model assumes that data are samples from observed random variables $\mathbf{x} \in \mathbb{R}^p$ (RVs) that arised by mixing several independent **non-gaussian** RVs $\mathbf{s} \in \mathbb{R}^k$, where typically (and we will assume this) $p \geq k$. The mixing model can be specified as $\mathbf{x} = A\mathbf{s} + \mathbf{b}$, where $\mathbf{b} \in \mathbb{R}^p$ is a vector of means.

- We ensure that $\mathbf{b} = \mathbf{0}$, model is simplified $\mathbf{x} = A\mathbf{s}$,
- ICA problem: find unmixing matrix W, s.t.

$$\mathbf{s} = W\mathbf{x}$$

- Solutions not determined fully (Tong, 1991) — order, scale
- ICA minimizes redundancy of $\hat{\mathbf{s}} = \hat{W}\mathbf{x}$ over suitable space of matrices

Independent Component Analysis II

Dependency is quantified by the redundancy

$$R(y_1, ..., y_k) = \sum_{i=1}^{k} H(y_i) - H(y_1, y_2, ..., y_k),$$

where $H(y_i)$ is the differential entropy of y_i and $H(y_1, y_2, ..., y_k)$ is the joint entropy of $y_1, ..., y_k$.

In practice the RV $\mathbf{x}$ may have been generated as $\mathbf{x} = A\mathbf{s}$, where the RVs in the vector $\mathbf{s}$ are not independent. In that case, ICA attempts to find a set of RVs that have least dependency.

Pre-whitening

- Computations are simplified with whitening & dimensionality reduction (WDR)
- Linear transform $\mathbf{y} = M\mathbf{x}, M \in \mathbb{R}^{k \times p}, k \leq p$
- Target: $\mathbf{y}$, such that $\mathbb{E}[\mathbf{y}\mathbf{y}^T] = I_k$.
- Simplification: The unmixing matrix $H \in \mathbb{R}^{k \times k}$, $\mathbf{s} = Q\mathbf{y}$ is now orthogonal because

$$I_k = \mathbb{E}[\mathbf{s}\mathbf{s}^T] = \mathbb{E}[Q\mathbf{y}\mathbf{y}^T Q^T] = Q\mathbb{E}[\mathbf{y}\mathbf{y}^T]Q^T = QQ^T$$

- FastICA (Hyvarinen, 2000), JADE, MaxKurt (Cardoso, 1999)
- Whole “unmixing” transform is a linear mapping $\mathbf{s} = QM\mathbf{x}$

Practical ICA

- We work with same data $X \in \mathbb{R}^{n \times p}$ as AAR/C
- Data whitening using Principal Component Analysis
- First the data is passed through a WDR stage using $M \in \mathbb{R}^{p \times k}$ so that

$$Y = XM, Y^T Y = I_k$$

- Linear subspace of columns of Y is that of the first k principal components
- An orthogonal matrix $Q \in O(k)$ is found to "unmix" the white data Y
- The complete mapping is $C_{\mathrm{ICA}} = YQ = XMQ$, $Q \in O(k)$

Practical ICA: how to find H

- Q may be obtained by finding the most non-gaussian projections of the columns of white data Y (Hyvarinen,2000)
- Redundancy of $Z = (z_1, z_2, ..., z_k)$ is

$$\sum_{i=1}^{k} \hat{H}(z_i) - \hat{H}(z_1, z_2, ..., z_k)$$

  joint entropy unaffected by ortho. transforms
- Then, task is to minimize $\sum_{i=1}^{k} \hat{H}(z_i)$ for $Z = YQ$ over $Q \in O(k)$
- Non-gaussian distributions are typically platykurtic or leptokurtic (heavy tails)

Linear mappings from data

- Mapping from data to AAR/C relaxed solution

$$C_{\mathrm{AAR/C}} = XF_{[k]}R_V,$$

where $R_V \in O(k)$

- Mapping from data to ICA components

$$C_{\mathrm{ICA}} = XMQ$$

where $Q \in O(k)$

Linear mappings II

- Without proof: If M is WDR for X and $Y = XM$ is white data and linear subspace of columns of Y is equal to that of the first k principal components, then there exists $P \in O(k)$ such that $Y = XF_{[k]}P$
- Thus $C_{\mathrm{ICA}} = XF_{[k]}PQ$ for some $P \in O(k)$
- Then $\tilde{Q} = PQ \in O(k)$ and

$$C_{\mathrm{ICA}} = XF_{[k]}\tilde{Q}$$

- Remember:

$$C_{\mathrm{AAR/C}} = XF_{[k]}R_V$$

Meaning ?

- AAR/C:
  - we wish to cluster fMRI data T_C into k clusters using covariance as connectivity
  - relaxed solution is a linear mapping of the data $C_{\mathrm{AAR/C}} = XF_{[k]}R_V$
- ICA:
  - we wish to find k "least dependent components" from the observed mixture X
  - components obtained by linear mapping of data $C_{\mathrm{ICA}} = XF_{[k]}\tilde{Q}$
- Solutions live in the same linear subspace spanned by columns of $U_{[k]} = XF_{[k]}$
- Both methods use auxiliary objectives to find a different basis using an orthogonal transform

Orthogonal transforms

- Reminder: VARIMAX transform of data Z is accomplished by matrix $R_V \in O(k)$ such that

$$R_V = \underset{R \in O(k)}{\arg\max}\, v^*(ZR),$$

  where $v^*(\cdot)$ is the VARIMAX objective

- For matrices with zero mean and equivariant columns, maximizing $v^*(Z)$ is equal to maximizing the sum of the sample kurtoses of the columns

Theoretical extension of AAR/C

- Let us naturally extend AAR/C clustering to random variables
- This will be a two stage procedure
  1. whiten the expected covariance matrix of the RVs
  2. find orthogonal transform to maximize sum of kurtoses of RVs
- Let us call this theoretical algortihm AAR/C*
- Embodiment of AAR/C* on real datasets is AAR/C

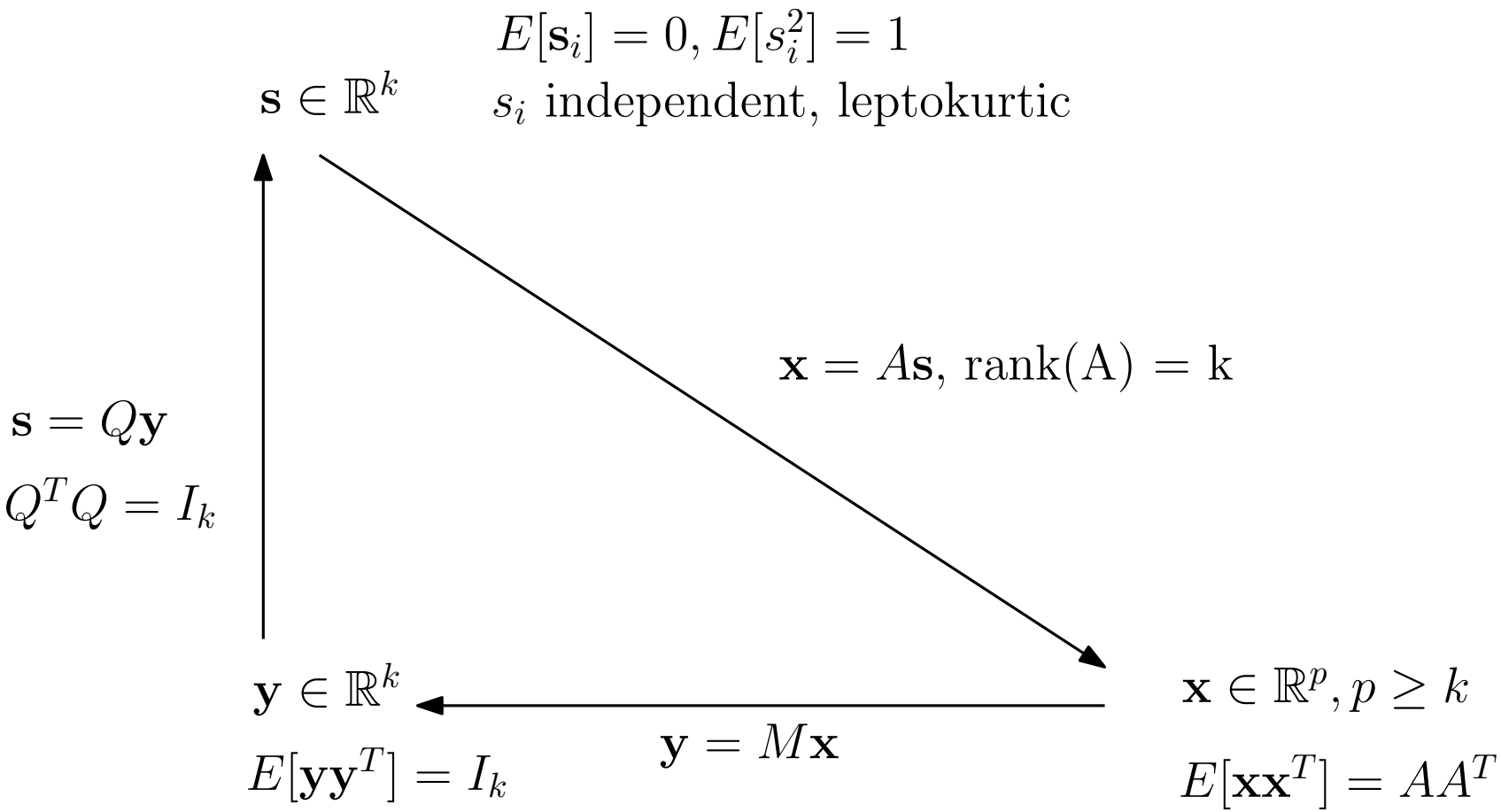
Equivalence of AAR/C* and ICA

Let $\mathbf{s} \in \mathbb{R}^k$ be a vector of independent, zero mean, unit variance, random variables with **leptokurtic** distributions. If the observable RVs $\mathbf{x} \in \mathrm{R}^p$, $p \geq k$ are given as a mixture $\mathbf{x} = A\mathbf{s}$ and $A \in \mathbb{R}^{p \times k}$ has rank k, then

1. the optimal solution of ICA under this model is $\hat{\mathbf{s}}_{\mathrm{ICA}}$, which may be $\mathbf{s}$ or its permutation or reflection
2. the optimal solution of AAR/C* is some $\hat{\mathbf{s}}_{\mathrm{AAR/C}}$, which may be $\mathbf{s}$ or its permutation or reflection

Notes: this is true irrespective of whether WDR is used or not in ICA. The claim also does not depend on any particular ICA estimation algorithm being used, provided that it converges to the optimal solution.

RV Schema


$E[\mathbf{s}_i] = 0, E[s_i^2] = 1$
$\mathbf{s} \in \mathbb{R}^k$
s_i independent, leptokurtic
$\mathbf{x} = A\mathbf{s}$, rank(A) = k
$\mathbf{s} = Q\mathbf{y}$
$Q^TQ = I_k$
$\mathbf{y} \in \mathbb{R}^k$
$\mathbf{x} \in \mathbb{R}^p, p \geq k$
$\mathbf{y} = M\mathbf{x}$
$E[\mathbf{y}\mathbf{y}^T] = I_k$
$E[\mathbf{x}\mathbf{x}^T] = AA^T$

Proof outline

- $Q \in O(k)$ because $\mathbf{y} = M\mathbf{x} = MA\mathbf{s}$ and

$$E[\mathbf{y}\mathbf{y}^T] = I_k = E[MA\mathbf{s}\mathbf{s}^T A^T M^T] = (MA)(MA)^T = QQ^T$$

- $M \in \mathbb{R}^{k \times p}$ exists:
  - if $AA^T = BSB^T$, $BB^T = I_p$, then S is diagonal with exactly k nonzero eigenvalues
  - then $AA^T = B_{[k]} S_{[k]} (B_{[k]})^T$ and we take $M = S_{[k]}^{-\frac{1}{2}} (B_{[k]})^T$
- AAR/C* first stage (whitening) maps $\mathbf{x}$ to $\mathbf{y}$
- Second stage finds $Q \in O(k)$ so that sum of kurtoses of RVs $Q\mathbf{y}$ is maximized ("VARIMAX on random variables")

Proof outline II

- Sets of reachable RV vectors by orthogonal transforms
  - From y: $A_y = \{\mathbf{z} | \mathbf{z} = R\mathbf{y}, R \in O(k)\}$
  - From s: $A_s = \{\mathbf{z} | \mathbf{z} = R\mathbf{s}, R \in O(k)\}$
- Since $\mathbf{s} = Q\mathbf{y}$ and orthonormal transforms are closed under composition $A_y = A_s$
- Lemma: In the set A_s, only RVs that are permutations and reflections of s have the maximum sum of kurtoses
- Let S_P be the set containing $\mathbf{s}$ and all its permutations and reflections, then $S_P \subset A_s$
- Corollary: Optimal $Q^* \in O(k)$ maximizing the sum of kurtoses maps $\mathbf{y}$ to a solution $\mathbf{s}' \in S_P$

Summary

- AAR/C on centered data is related to ICA

  solution in same linear subspace, orthogonal transforms related
- The above relies on use of whitening & dimensionality reduction in the algorithm
- In theory, the AAR/C* problem has the same optimal solutions as ICA

  for independent, leptokurtic source RVs
- Meaning: on some problems, ICA and AAR/C* are identical
  - ICA gives a relaxed solution to a clustering problem
  - AAR/C* relaxed solution is the ICA under model assumptions

Thank you

Thank you for your attention !

Extra slides

Extra slides

Illustration of objective J_1, J_2

If there are N elements with mutually connected with a connectivity $\rho > 0$ (self-connectivity $w_{i,i} = 0$).

Clustering into one cluster: the objective value J_1 only depends on the number of elements in the cluster $J_1 = \rho(N_1 - 1)$ and the maximum of this objective is reached if all the elements are in one cluster, or $N_1 = N$. $V_o = \emptyset$.

Clustering into two clusters: the objective value $J_2 = \rho(N_1 + N_2 - 2)$, if there are N_1 elements in the first cluster and N_2 elements in the second cluster — no relative cluster size is preferred but all elements must be assigned to one of the clusters to maximize J_2. $V_o = \emptyset$.

Zero-one programming formulation

The AAR objective may be formulated as a zero-one programming problem:

$$J_k = \sum_{l=1}^{k} \frac{u_l^T W u_l}{u_l^T u_l}, u_l \in \{0,1\}^n$$

where the disjointness constraints may be formulated as inequalities

$$\forall i \in \{1,2,...,n\} \sum_{l=1}^{k} [u_l]_i \leq 1.$$

Simple discretization

Let us expand the objective J_k for the solution V of the eigenvectors:

$$J_k = Y^T W Y = Y^T \Lambda Y = \sum_{l=1}^{k} \lambda_l y_l^T y_l = \sum_{l=1}^{k} \lambda_l \sum_{i=1}^{N} [y_l]_i^2.$$

By disjointness of the clusters, each element v_i must go into at most one cluster V_l. If we put the element v_j into cluster V_i, then the contribution of the element to the criterion J_k will be $\lambda_l [y_l]_i^2$. So one may assign each element v_i to the cluster

$$\arg\max_{1 \le l \le k} \lambda_j [y_l]_i^2.$$

(Bialonski and Lehnertz, 2006)

Simple discretization II

- This approach has some problems (Bialonski and Lehnertz, 2006), (Vejmelka and Palus, 2010)

  e.g. cannot separate clusters of similar size
- The method assigns all elements to clusters
- We considered several options as to how to assign elements to clusters

  e.g. first assigning using above method, then attempting to maximize the cluster strength for each cluster
- Most methods depended on forming explicitely the matrix W, which is in some applications huge
- We found a more effective heuristic to remove “unwanted elements” into the residual set

  fit the ideal form of the indicator vector to the relaxed solution