BI-CLUSTERING VIA MDL-BASED MATRIX FACTORIZATION

Ignacio Ramírez Paulino, ECE, Universidad de la Republica
nacho@fing.edu.uy

Mariano Tepper, ECE, Duke University
mariano.tepper@duke.edu

Summary

Bi-Clustering: grouping related elements of a matrix into submatrices.

Question #1: how do we find the groups?
- Sparse low-rank matrix factorization

Question #2: how many groups there are?
- Minimum Description Length (MDL)

Model selection criterion.

Sample problem: finding aligned points among randomly scattered ones (Poisson).
Correct number of lines (groups) is 11

1. Concepts and Tools

Step 1: randomly sample N pairs of points from the dataset

Step 2: find points within a small distance of the line traced across each sample pair

Step 3: construct MxN assignment matrix A where the (i,j)-th entry is 1 if point i belongs to line defined by pair j, or else 0.

2. Procedure

Step 4 (sparse NMF): \( A = UV + E \) with U MxK, V KxN, K is small (low-rank model) and E MxN has few nonzero entries (sparse). This is done via solving the sparse NMF (SNMF) optimization problem:

\[
(\hat{U}, \hat{V}) = \underset{U,V}{\text{arg min}} \sum_{i,j} \left( A_{ij} - \sum_{k} U_{ik} V_{kj} \right)^2 + \lambda_1 \sum_{i} \|U_i\|_0 + \lambda_2 \sum_{j} \|V_j\|_0
\]

subject to: \( u_{ik}, v_{kj} \geq 0, \forall i,j,k \)

Step 5 (MDL): choose \( \lambda_1, \lambda_2 \) and K which minimize the length of the compressed binary description of A in terms of U, V and E

3. Results

Effective model selection: correct clusters for 2 of the 3 factorization methods

Best quality vs. time tradeoff obtained by approximating SNMF (\( \rho = 0 \)) via OMP

Assignment matrix for sample problem, M=1000, N=5000

\( \rho = 1 \): Correct no. of clusters, False positives. Slow.
\( \rho = 0 \) + OMP: Correct no. of clusters. False positives. Very fast.
\( \rho = 0 \) + S-SVD: Wrong no. of clusters (13). False negatives.

Example: Evolution of description length as the rank (K) increases

Acknowledgements: Prof. Guillermo Sapiro from Duke University, DHS/JHU, AFOSR.