Multiple co-clustering and its application

Tomoki Tokuda,

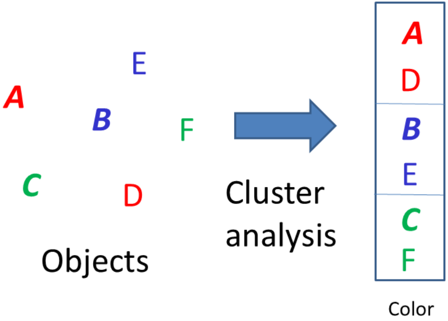
Okinawa Institute of Science and Technology Graduate University

Outline

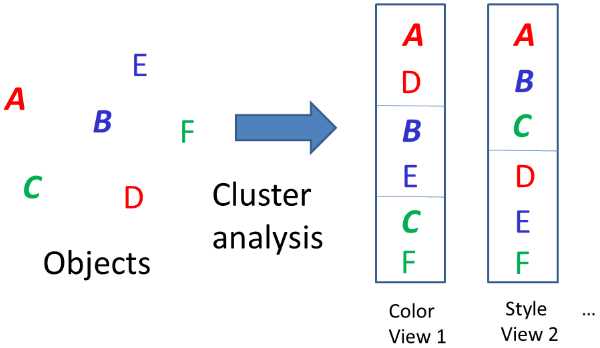
1. Introduction
2. Method for multiple co-clustering
3. Application to depression data
4. Conclusion

Introduction

What is multiple clustering?



Conventional clustering method: One clustering solution



Multiple clustering method: Multiple clustering solutions

Method for multiple co-clustering

Multiple clustering in data matrix

Multiple clustering solutions : *appropriately* partitioning features (without overlapping) and subsequently clustering objects.

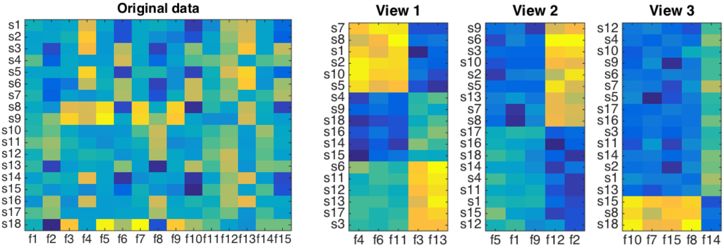
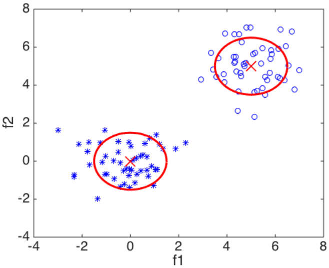


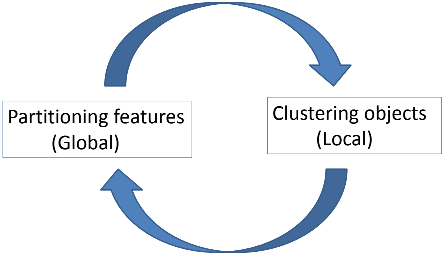
Figure 1: Original data → Multiple clustering solutions

It reveals associations between *features* and *object-clustering*.

Idea of algorithm

- Clustering object → Fitting certain distribution family (in iterative manner).





- Iteratively optimize objective function (i.e., likelihood)

Challenges in multiple clustering for high-dimensional data

- No information on the number of views or object-clusters.
  $\rightarrow$ Dirichlet process (infinite number of views and clusters)
- Missing values $\rightarrow$ Integrate out (Bayesian framework)

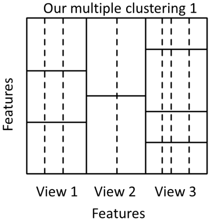
We work on the following challenges.

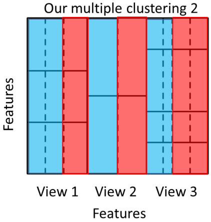
- Possible over-fitting to data:
  Typically, the number of samples is much smaller than the number of features.
- Mixing of several types of data:
  We want to analyze data combining numerical and categorical features!

Our proposed model

Ingredients:

- Similar features are fitted by the same univariate distribution (feature cluster; hence, **co-clustering**).
- Allowing for mixing of different types of distributions (Gaussian, Poisson, multinomial)





Byproduct;

- Easy interpretation for similar features.
- Computationally efficient: $O(nd)$ for a single iteration.

Such modifications broaden the scope of application.

Model

Likelihood

$$\log p(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\Theta}) = \sum_{m,v,g,k,j,i} \mathbb{I}(Y_{j,v,g}^{(m)}=1)\mathbb{I}(Z_{i,v,k}=1)\log p(X_{i,j}^{(m)}|\theta_{v,g,k}^{(m)}),$$

m : Type of distribution (pre-specified)
$Y_{j,v,g}$: Feature j for a membership of view v and f.cluster g
$Z_{i,v,k}$: Object i for a membership of object-cluster k in view v.

Prior for distribution parameters
Conjugate prior for distribution families of Gaussian, Poisson and multinomial.

Essence of algorithm: Variational Bayesian method

- We want to know posterior $p(\phi|\boldsymbol{X}) \to$ Analytically impossible.
- So, we consider approximation. By Jensen's inequality,

$$\log p(\boldsymbol{X}) \geq \int q(\phi) \log \frac{p(\boldsymbol{X}, \phi)}{q(\phi)} d\phi \tag{1}$$

  where $q(\phi)$ is arbitrary; equality holds when $q(\phi) = p(\phi|\boldsymbol{X})$.
- Assume factorization of $q(\phi) = \prod q_i(\phi_i)$.
- We want to optimize distribution $q(\phi)$ to maximize the right hand side in Eq.(1).
- An (conditionally) optimal distribution is given by

$$q_i(\phi_i) \sim \exp\{\mathbb{E}_{-q_i(\phi_i)} \log p(\boldsymbol{X}, \phi)\}$$

  where $\mathbb{E}_{-q_i(\phi_i)}$ denotes averaging over all parameters but ϕ_i.

4. Conclusion

- A novel method of multiple clustering for high-dimensional data.
- Co-clustering structure in view enables efficient and easy interpretation of features.
- In application to depression data, one subject-clustering solution has been found, which is relevant to treatment effect.
- This model may provide possible prediction of treatment effect based on stress experiences in childhood and functional connectivity in the brain.