# Advances in Directional Statistics 2017

## Directional-linear data clustering using structural expectation-maximization algorithm

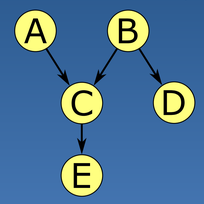
Sergio Luengo-Sanchez, Concha Bielza & Pedro Larrañaga

POLITÉCNICA

CIG
http://cig.fi.upm.es

## 1.Motivation & Goals

- Several scientific fields collect **hybrid data** (directional - linear)
- State of the art clusters cylindrical and multivariate data with one circular variable [1]
- **Bayesian networks** (BNs) are directed acyclic graphs that represent probabilistic relationships




- Exploit **conditional independence assumptions** encoded by the BN to efficiently cluster hybrid data
- Use **Gaussian** and **von Mises** (vM) variables

## 1. E-step

Estimate **latent** variable Z **values** from the observed data

$$Q_i(z^i) = p(z^i|\mathbf{x}^i;\boldsymbol{\theta}) = \frac{f(\mathbf{x}^i|z^i;\boldsymbol{\theta})p(z^i;\boldsymbol{\theta})}{\sum_z f(\mathbf{x}^i|z;\boldsymbol{\theta})p(z;\boldsymbol{\theta})}$$

Hybrid BN distribution **factorizes** as

$$f(\mathbf{X},\mathbf{Y},Z;\boldsymbol{\theta}) = p(Z;\boldsymbol{\theta})\prod_{l=1}^{L} f(X_l|\mathbf{Pa}_l,Z;\boldsymbol{\theta})\prod_{m=1}^{M} f(Y_m|Z;\boldsymbol{\theta})$$

It is the **membership probability** of an instance to a cluster

## 2. M-step

Estimate $\hat{\boldsymbol{\theta}}$ from the complete data

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{z^i} Q_i(z^i)\log\frac{f(\mathbf{x}^i,z^i;\boldsymbol{\theta})}{Q_i(z^i)}$$

**MLE** of the **vM** and **Gaussian** variables decomposes as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{z^i} Q_i(z^i)\left[\sum_{l=1}^{L}\log f(x_l^i|\mathbf{Pa}_l^i,z^i;\boldsymbol{\theta}) + \sum_{m=1}^{M}\log f(y_m^i|z^i;\boldsymbol{\theta})\right]$$

Each variable is optimized **locally**

Repeat E-step and M-step until convergence

## ADISTA 17

## 2.Hybrid Bayesian Networks [2]

- **Structural EM** [3] to learn BN structure
- Constrain BN structure to obtain **close-form equations**
- Evaluate quality of clusters
  - **Kullback-Leibler divergence** for vM distribution

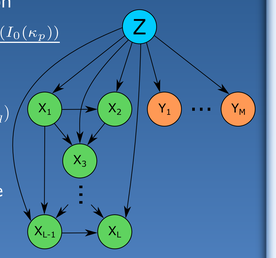
$$KL(P||Q) = \ln I_0(\kappa_q) - \ln I_0(\kappa_p) + \kappa_p A(\kappa_p) - \kappa_q\cos(\mu_q)A(\kappa_p)$$

  - **Batthacharyya distance** for vM distribution

$$B_D(P,Q) = -\ln(I_0(R)) + \frac{\ln(I_0(\kappa_p))+\ln(I_0(\kappa_p))}{2}$$

where

$$R = \sqrt{a^2+b^2} \qquad a = \frac{\kappa_p}{2}\cos(\mu_p) + \frac{\kappa_q}{2}\cos(\mu_q) \qquad b = \frac{\kappa_p}{2}\sin(\mu_p) + \frac{\kappa_q}{2}\sin(\mu_q)$$




Both measures factorize according to the conditional independence assumptions

## 3. Structure learning

Discover **conditional independence** relations

Run **after EM** convergence

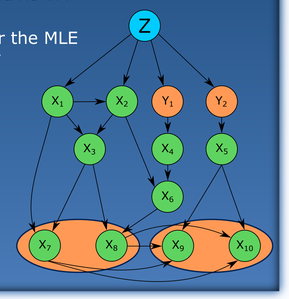
Any heuristic search algorithm for structure learning can be applied

Score to use is **BIC** because it **increases monotonically** always finding a better structure

Iteratively repeat EM and structure learning until convergence

## 3.More General Hybrid BNs

- Relax constraints between Gaussian and vM
- From **vM** to **Gaussian** [4]
  - Only one vM parent and not Gaussian parents allowed
- From **Gaussians** to **Proyected normals** (PNs) [5]
  - Any number of Gaussian parents and no vM parents
  - There is no close-form equations for the MLE
  - Reparameterization is not needed if parents and data do not change
- From **PNs** and/or **Gaussians** to **PN**
  - Fit PN assuming that its PN parents are bivariate Gaussians
- Under development




## 4. Model selection

Try different **number of clusters**

For each number of clusters **restart** changing initial configuration to avoid local optima

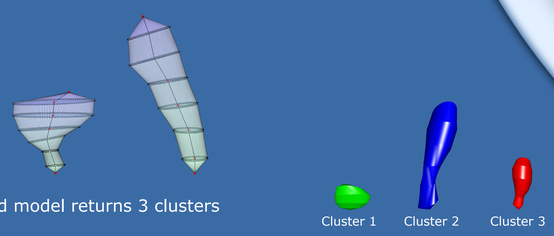
Select best clustering according to **BIC** score

$$BIC = 2\hat{L} - v\log(N)$$

Evaluate **quality** of clusters
- Kullback-Leibler
- Batthacharyya

## 4.Preliminary Results

- Evaluation of simulated Hybrid BNs
  - Improve Gaussian mixture model outcome
  - Recover almost completely original Gaussian structure
- Clustering of 3D dendritic spines*
  - Their morphology is related to brain functions like **learning** and **memory**
  - High diversity of morphologies, existence of a **continuum**?
  - Characterize 3D meshes applying **multiresolutional Reeb graph**




- General hybrid model returns 3 clusters

* Dendritic spines provided by the Cajal Cortical Circuits Lab (UPM-CSIC)

[1] A Roy et al., "SWGMM: A semiwrapped Gaussian mixture model for clustering of circular-linear data", *Pattern Analysis and Application*, 19:631-645, 2014
[2] S Luengo-Sanchez et al., "Hybrid Gaussian and von Mises model-based clustering", in *ECAI 2016*, 285:855-862, 2016
[3] N Friedman, "Learning belief networks in the presence of missing values and hidden variables", in *ICML*, 97:125-133, 1997
[4] KV Mardia and TW Sutton, "A model for cylindrical variables with applications", *Journal of the Royal Statistical Society. Series B*, 40(2):229-233, 1978
[5] B Presnell et al., "Projected multivariate linear models for directional data", *Journal of the American Statistical Association*, 93(443):1068-1077, 1998

If you are interested, you can contact me by email!!!
sluengo@fi.upm.es