

# Clustering mixed-type data via mixtures of exponential family graphical models

**Michael Fop**

*with Federica Benassi and Giuliano Galimberti*

[michaelfop.github.io](https://github.com/michael-fop)

✉ [michael.fop@ucd.ie](mailto:michael.fop@ucd.ie)

🐦 [@michael\\_fop](https://twitter.com/michael_fop)

WGMBC

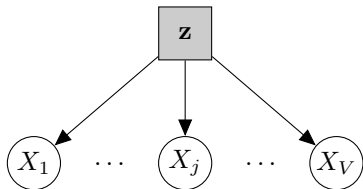
Athens, 25-29 October 2021



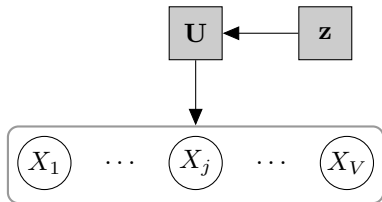
School of Mathematics and Statistics  
University College Dublin

# Model-based clustering for mixed data

## Problems



*Local independence.*



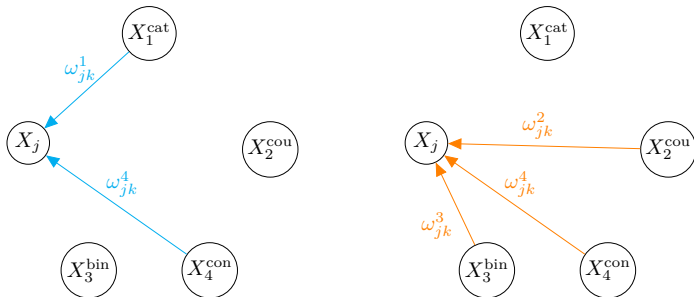
*Continuous latent variable.*

- ❖ Difficult to define a **multivariate** distribution for mixed data.
- ❖ Restrictive **independence** assumptions.
- ❖ Inference **computationally difficult**.

# Mixture of exponential family graphical models

- Graphical models allow to model the **association structure** between a collection of random variables.
- Mixture of conditionally specified **exponential family distributions**:

$$p(\mathbf{x}_i; \Theta) = \sum_{k=1}^K \tau_k \left\{ \prod_{j=1}^V p \left( x_{ij} \mid \mathbf{x}_i^{\text{ne}_k(j)}; \alpha_{jk}, \omega_{jk} \right) \right\}$$



# Mixture of exponential family graphical models

## Penalized likelihood framework

- ❖ Inference using a **penalized log-likelihood** framework:

$$\ell(\Theta, \mathbf{X}) - \lambda \sum_{k=1}^K \|\omega_k\|_1$$

- ❖ Collection of **lasso-type regressions** (and/or related penalties).
- ❖ Suitable for **high-dimensional data**.
- ❖ Application to **autism spectrum disorder data** (questionnaire including seven mixed-type variables – Deserno et al., 2017).

