Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Model-based clustering of categorical data based on the Hamming distance

Lucia Paci
Department of Statistical Sciences,
Università Cattolica del Sacro Cuore, Milano




Raffaele Argiento
University of Bergamo

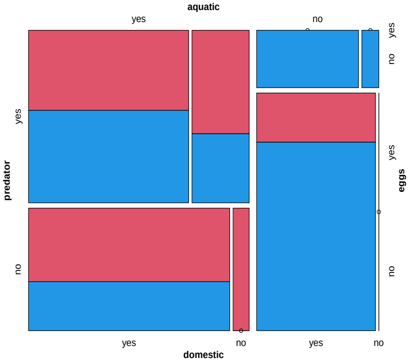
Andrea Cremaschi
IE University

Edoardo Filippi-Mazzola
Broadgate Advisers

Benedetta Sabina Leone
Politecnico di Milano

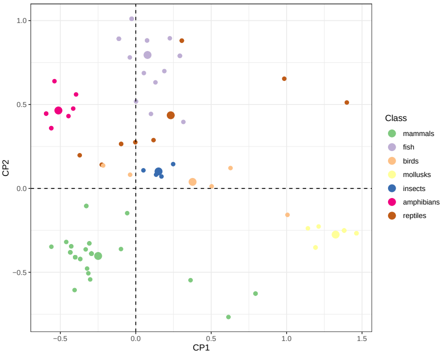
SISBAYES 2025
Department of Statistical Sciences, University of Padova
4-5 September 2025

Clustering multiple categorical data



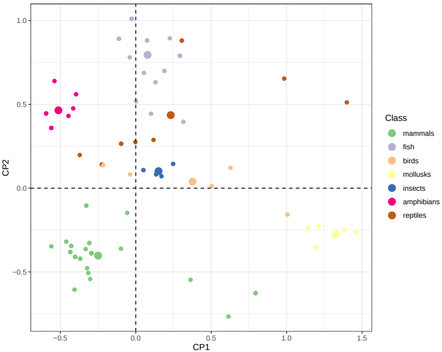
- Multivariate **categorical data** with no natural ordering (e.g., animal features)

Clustering multiple categorical data



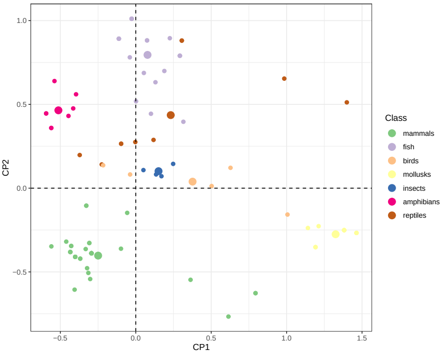
- Multivariate **categorical data** with no natural ordering (e.g., animal features)
- **Goal**: identify a clustering structure in the data

Clustering multiple categorical data



- Multivariate **categorical data** with no natural ordering (e.g., animal features)
- **Goal**: identify a clustering structure in the data
- **Heuristic framework**: measure distances between observations
  - ⋆ *K-modes* (Huang, 1998)
  - ⋆ *Hamming distance-vector algorithm* (Zhang et al., 2006) finds clustering patterns using the Hamming distance
- Model-based clustering: **latent class analysis** (Goodman, 1974; Celeux and Govaert, 2015)

Clustering multiple categorical data



- Multivariate **categorical data** with no natural ordering (e.g., animal features)
- **Goal**: identify a clustering structure in the data
- **Heuristic framework**: measure distances between observations
  - ⋆ *K-modes* (Huang, 1998)
  - ⋆ *Hamming distance-vector algorithm* (Zhang et al., 2006) finds clustering patterns using the Hamming distance
- Model-based clustering: **latent class analysis** (Goodman, 1974; Celeux and Govaert, 2015)

Our approach

✓ Family of probability mass functions built upon the Hamming distance

✓ Model-based clustering based on a mixture of finite mixture of Hamming distributions

✓ Provide full posterior inference on the number of clusters and their structure

A bit of notation

- $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ is a vector of p nominal categorical variables, or **attributes**
- Each variable j, for $j = 1, \dots, p$, assumes m_j possible levels (categories) or **modalities**, over the finite set $A_j = \{a_{j1}, \dots, a_{jh}, \dots, a_{jm_j}\}$
- $\boldsymbol{x} = (x_1, \dots, x_p)^\top$ is the vector of observed modalities
- Categorical sample space
$\Omega_p = \left\{\boldsymbol{x} = (x_1, \dots, x_p)^\top | x_1 \in A_1, \dots, x_p \in A_p\right\} = A_1 \times A_2 \times \dots \times A_p$
- **Hamming distance**: number of attributes whose modalities are different

Hamming distance between two points in Ω_p

$$d(\boldsymbol{x}_i, \boldsymbol{x}_h) = \sum_{j=1}^{p} 1 - \delta_{x_{ij}}\left(x_{hj}\right)$$

where $\delta_{x_{ij}}\left(x_{hj}\right) = \begin{cases} 1 \text{ if } x_{ij} = x_{hj} \\ 0 \text{ if } x_{ij} \neq x_{hj} \end{cases}$

$$\boldsymbol{x}_1 = [\clubsuit, \heartsuit, \bigstar] \qquad \boldsymbol{x}_2 = [\blacksquare, \heartsuit, \spadesuit]$$
$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = 2$$

Hamming distribution

- **center** parameter $\boldsymbol{c} = (c_1, \ldots, c_p)^\top \in \Omega_p$
- **scale** parameter $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)^\top$, with $\sigma_j > 0, j = 1, \ldots, p$

Hamming distribution

- **center** parameter $\boldsymbol{c} = (c_1, \ldots, c_p)^\top \in \Omega_p$
- **scale** parameter $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)^\top$, with $\sigma_j > 0, j = 1, \ldots, p$

Proposition 1

The function

$$p(\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{\sigma}) = \prod_{j=1}^{p} \left(1 + \frac{m_j - 1}{\exp(1/\sigma_j)}\right)^{-1} \exp\left\{-\frac{1 - \delta_{c_j}(x_j)}{\sigma_j}\right\}$$

is a probability mass function (p.m.f.) on Ω_p, i.e., $\sum_{\boldsymbol{x} \in \Omega_p} p(\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{\sigma}) = 1$

A random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ with support Ω_p follows an **Hamming distribution** with center $\boldsymbol{c}$ and scale $\boldsymbol{\sigma}$ if its p.m.f. for $\boldsymbol{x} \in \Omega_p$ is given by $p(\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{\sigma})$ and we write

$$\boldsymbol{X} \mid \boldsymbol{c}, \boldsymbol{\sigma} \sim \text{Hamming}(\boldsymbol{c}, \boldsymbol{\sigma})$$

Hamming distribution

- **center** parameter $\boldsymbol{c} = (c_1, \ldots, c_p)^\top \in \Omega_p$
- **scale** parameter $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)^\top$, with $\sigma_j > 0, j = 1, \ldots, p$

Proposition 1

The function

$$p(\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{\sigma}) = \prod_{j=1}^{p} \left(1 + \frac{m_j - 1}{\exp(1/\sigma_j)}\right)^{-1} \exp\left\{-\frac{1 - \delta_{c_j}(x_j)}{\sigma_j}\right\}$$

is a probability mass function (p.m.f.) on Ω_p, i.e., $\sum_{\boldsymbol{x} \in \Omega_p} p(\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{\sigma}) = 1$

A random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ with support Ω_p follows an **Hamming distribution** with center $\boldsymbol{c}$ and scale $\boldsymbol{\sigma}$ if its p.m.f. for $\boldsymbol{x} \in \Omega_p$ is given by $p(\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{\sigma})$ and we write

$$\boldsymbol{X} \mid \boldsymbol{c}, \boldsymbol{\sigma} \sim \text{Hamming}(\boldsymbol{c}, \boldsymbol{\sigma})$$

- When $\sigma_j = \sigma > 0, \forall j,$ $\quad p(\boldsymbol{x} \mid \boldsymbol{c}, \sigma) \propto \exp\left\{\frac{-d(\boldsymbol{c}, \boldsymbol{x})}{\sigma}\right\}$

Hamming distribution

Example

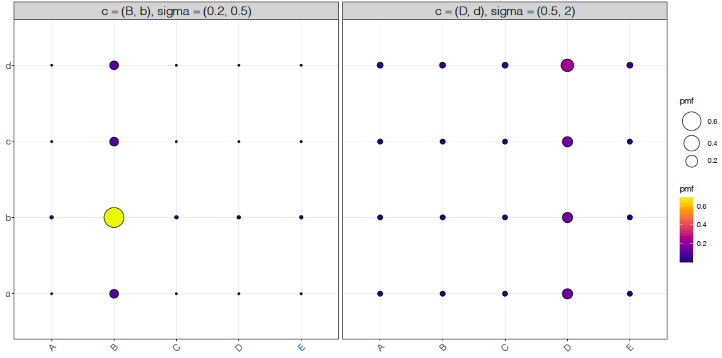


Figure: P.m.f. of $p = 2$ categorical variables with different values of $\mathbf{c}$ and $\boldsymbol{\sigma}$.

- The center represents the unique mode of the distribution
- The scale regulates the heterogeneity of the distribution (e.g., link with Gini's and Shannon's indexes)

Bayesian inference

- Sampling model: $\boldsymbol{X}_i \mid \boldsymbol{c}, \boldsymbol{\sigma} \overset{iid}{\sim} \text{Hamming}(\boldsymbol{c}, \boldsymbol{\sigma}), \qquad i = 1, \dots, n$
- Inference on $\boldsymbol{c}$
  - Prior: $c_j \overset{iid}{\sim} \text{U}\{1, m_j\} \qquad j = 1, \dots, p$
  - Full conditional probabilities: $p(c_j | \text{rest}) \propto \exp\left\{-\frac{n - \sum_{i=1}^n \delta_{c_j}(x_{ij})}{\sigma_j}\right\}$
- Inference on $\boldsymbol{\sigma}$
  - Prior (Hypergeometric Inverse Gamma): $\sigma_j \mid u, v \overset{iid}{\sim} \text{HIG}(u, v) \qquad j = 1, \dots, p$, where

$$f(\sigma_j \mid u, v) = \frac{m_j^{(u+v)}(v+1)}{{}_2F_1\left(1, u+v;\ v+2;\ (m_j - 1)/m_j\right)} \left(1 + \frac{m_j - 1}{\exp\left(1/\sigma_j\right)}\right)^{-(u+v)} \exp\left(-\frac{v+1}{\sigma_j}\right) \frac{1}{\sigma_j^2}$$

where ${}_2F_1(\cdot, \cdot; \cdot; \cdot)$ is the hypergeometric function.

  - Full conditional: $\sigma_j \mid \text{rest} \sim \text{HIG}(u^*, v^*)$, where $u^* = u + \sum_{i=1}^n \delta_{c_j}(x_{ij})$ and $v^* = v + n - \sum_{i=1}^n \delta_{c_j}(x_{ij})$
- The marginal likelihood of the data is available in a closed-analytical form.

Hamming Mixture Model

HMM

$$h(\boldsymbol{x}_i \mid \boldsymbol{c}_1, \ldots, \boldsymbol{c}_L, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_L, \boldsymbol{\pi}, L) = \sum_{l=1}^{L} \pi_l \, p\left(\boldsymbol{x}_i \mid \boldsymbol{c}_l, \boldsymbol{\sigma}_l\right)$$

- L: number of components
- π_l: mixing weight - probability that observation i belongs to component l
  satisfies $\sum_{l=1}^{L} \pi_l = 1$ and $0 < \pi_l < 1$
- $p\left(\boldsymbol{x}_i \mid \boldsymbol{c}_l, \boldsymbol{\sigma}_l\right)$: mixture kernel - Hamming p.m.f. of component l
- $\boldsymbol{c}_l = (c_{1l}, \ldots, c_{pl})^\top$: component-specific center
- $\boldsymbol{\sigma}_l = (\sigma_{1l}, \ldots \sigma_{pl})^\top$: component-specific scale

How do we choose L?

- L is **fixed**: fit the model with different values of L and select the "best" one by using information criteria, e.g., BIC, DIC, ICL (Biernacki et al., 2010; Celeux and Govaert, 2015)

How do we choose L?

- L is **fixed**: fit the model with different values of L and select the "best" one by using information criteria, e.g., BIC, DIC, ICL (Biernacki et al., 2010; Celeux and Govaert, 2015)

- L is **random**: a transdimensional sampler is needed for posterior inference
  - popular algorithm (but challenging) is the reversible jump MCMC (Green, 1995)
  - recent (painless!) alternatives – **mixtures of finite mixtures** – exploiting the link between finite and infinite mixture: Chinese restaurant process sampler (Miller and Harrison, 2018), telescoping sampler (Frühwirth-Schnatter et al., 2021), blocked Gibbs sampler (Argiento and De Iorio, 2022)

Prior on mixing weights

- Following Argiento and De Iorio (2022), the HMM can be framed in a BNP setting:
  - ⋆ infinite number of latent components
  - ⋆ only a finite number is used to generate the observed data

- We assign a prior distribution on the mixing weights by normalization:

$$\pi_1 = \frac{W_1}{T}, \ldots, \pi_L = \frac{W_L}{T}, \qquad T = \sum_{l=1}^{L} W_l$$

$$W_1, \ldots, W_L, \mid L, \gamma \overset{iid}{\sim} \text{Gamma}(\gamma, 1)$$

$\Rightarrow$ Equivalent to: $\quad \pi_1, \ldots, \pi_L \mid L, \gamma \sim \text{Dirichlet}_L(\gamma, \ldots, \gamma)$

- We assume a random number of components $L \sim q_L$

Clustering assignment

- Latent allocation variables $\boxed{z_1 \dots, z_n}$, where $z_i \in \{1, \dots, L\}$
- Observation $\boldsymbol{x}_i$ belongs to component $l \Leftrightarrow z_i = l$
- $z_i \mid W_1, \dots, W_L \overset{iid}{\sim} \text{Multinomial}(W_1/T, \dots, W_L/T)$

Clustering assignment

- Latent allocation variables $\boxed{z_1 \ldots, z_n}$, where $z_i \in \{1, \ldots, L\}$
- Observation $\boldsymbol{x}_i$ belongs to component $l \Leftrightarrow z_i = l$
- $z_i \mid W_1, \ldots, W_L \stackrel{iid}{\sim}$ Multinomial$(W_1/T, \ldots, W_L/T)$
- $\boxed{z_1^*, \ldots, z_K^*}$ $K \leq L$, unique values of allocations
- $\rho := \{C_1, \ldots, C_K\}$: partition of K clusters induced by $z_1^*, \ldots, z_K^*$, where $C_k = \{i : z_i = z_k^*\}$ and $n_k = |C_k|$, for $k = 1, \ldots, K$
- Prior of ρ: exchangeable partition probability function (eppf; Pitman 1995)

$$p(\rho) = p(n_1, \ldots, n_K) \propto \prod_{k=1}^{K} \frac{\Gamma(\gamma + n_k)}{\Gamma(\gamma)}$$

- The prior on K is also available in a closed analytical form

HMM wrap up

$$
\begin{aligned}
\boldsymbol{x}_i \mid z_i, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_L, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_L, L &\overset{ind}{\sim} \text{Hamming}(\boldsymbol{x}_i \mid \boldsymbol{c}_{z_i}, \boldsymbol{\sigma}_{z_i}) && i = 1, \ldots, n \\
z_i \mid W_1, \ldots, W_L, L &\overset{iid}{\sim} \text{Multinomial}(W_1/T, \ldots, W_L/T) && i = 1, \ldots, n \\
W_l \mid L, \gamma &\overset{iid}{\sim} \text{Gamma}(\gamma, 1) && l = 1, \ldots, L \\
\boldsymbol{c}_l \mid L &\overset{iid}{\sim} \text{U}\{1, m_j\} && l = 1, \ldots, L \\
\boldsymbol{\sigma}_l \mid L &\overset{iid}{\sim} \text{HIG}(u, v) && l = 1, \ldots, L \\
L \mid \Lambda &\sim \text{Poi}_0(\Lambda) \\
\Lambda &\sim \text{Gamma}(a, b)
\end{aligned}
$$

HMM wrap up

$$
\begin{aligned}
\boldsymbol{x}_i \mid z_i, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_L, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_L, L &\overset{ind}{\sim} \text{Hamming}(\boldsymbol{x}_i \mid \boldsymbol{c}_{z_i}, \boldsymbol{\sigma}_{z_i}) && i = 1, \ldots, n \\
z_i \mid W_1, \ldots, W_L, L &\overset{iid}{\sim} \text{Multinomial}(W_1/T, \ldots, W_L/T) && i = 1, \ldots, n \\
W_l \mid L, \gamma &\overset{iid}{\sim} \text{Gamma}(\gamma, 1) && l = 1, \ldots, L \\
\boldsymbol{c}_l \mid L &\overset{iid}{\sim} \text{U}\{1, m_j\} && l = 1, \ldots, L \\
\boldsymbol{\sigma}_l \mid L &\overset{iid}{\sim} \text{HIG}(u, v) && l = 1, \ldots, L \\
L \mid \Lambda &\sim \text{Poi}_0(\Lambda) \\
\Lambda &\sim \text{Gamma}(a, b)
\end{aligned}
$$

- **Connections to Latent Class Models**: the HMM is a novel parametrization of the parsimonious LCM (Celeux and Govaert, 2015) with two main benefits:
  - ✓ more straightforward interpretation of the parameters
  - ✓ random L, so full posterior inference on the number of clusters and their structure

HMM wrap up

$$
\begin{aligned}
\boldsymbol{x}_i \mid z_i, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_L, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_L, L &\overset{ind}{\sim} \text{Hamming}(\boldsymbol{x}_i \mid \boldsymbol{c}_{z_i}, \boldsymbol{\sigma}_{z_i}) && i = 1, \ldots, n \\
z_i \mid W_1, \ldots, W_L, L &\overset{iid}{\sim} \text{Multinomial}(W_1/T, \ldots, W_L/T) && i = 1, \ldots, n \\
W_l \mid L, \gamma &\overset{iid}{\sim} \text{Gamma}(\gamma, 1) && l = 1, \ldots, L \\
\boldsymbol{c}_l \mid L &\overset{iid}{\sim} \text{U}\{1, m_j\} && l = 1, \ldots, L \\
\boldsymbol{\sigma}_l \mid L &\overset{iid}{\sim} \text{HIG}(u, v) && l = 1, \ldots, L \\
L \mid \Lambda &\sim \text{Poi}_0(\Lambda) \\
\Lambda &\sim \text{Gamma}(a, b)
\end{aligned}
$$

- **Connections to Latent Class Models**: the HMM is a novel parametrization of the parsimonious LCM (Celeux and Govaert, 2015) with two main benefits:
  - ✓ more straightforward interpretation of the parameters
  - ✓ random L, so full posterior inference on the number of clusters and their structure
- **Posterior sampling**
  - Blocked Gibbs sampler, a conditional algorithm of Argiento and De Iorio (2022)
  - Transdimensional moves which are automatic and implied by the prior process
  - Separate sampling of the weights and parameters corresponding to the *allocated* vs *non-allocated* components

USPS data analysis

- Handwritten digits from the US postal services (a subset of the USPS data from *UCI machine learning repository*)
- 1,756 images of the digits 3, 5 and 8 which are the most difficult digits to discriminate
- Each digit is a 16×16 image of $m = 6$ levels of gray, i.e., $X_j \in \{\square, \square, \square, \square, \square, \blacksquare\}$, represented as a $p = 256$ dimensional vector

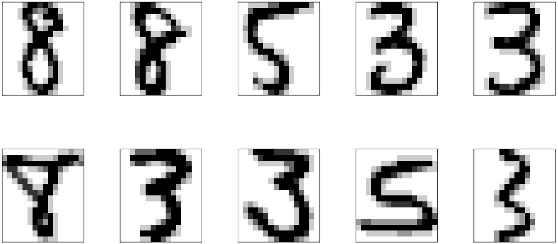


Figure: A sample of handwritten digits.

✗ Latent class model with fixed K: no way to select the number of clusters

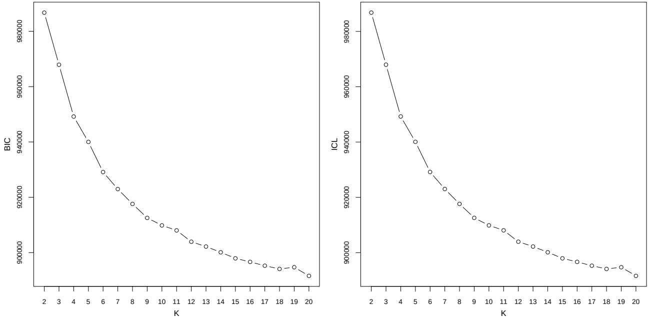


Figure: Values of the information criteria for fixed values of K; results obtained from the `R` package `Rmixmod`.

✓ Hamming mixture model with random number of components L

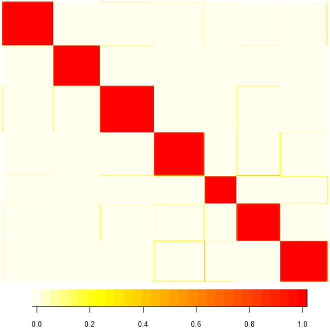


Figure: Posterior similarity matrix.

✓ Hamming mixture model with random number of components L

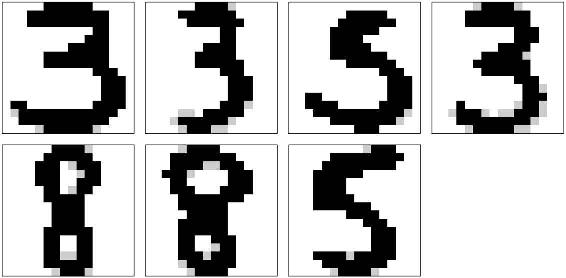
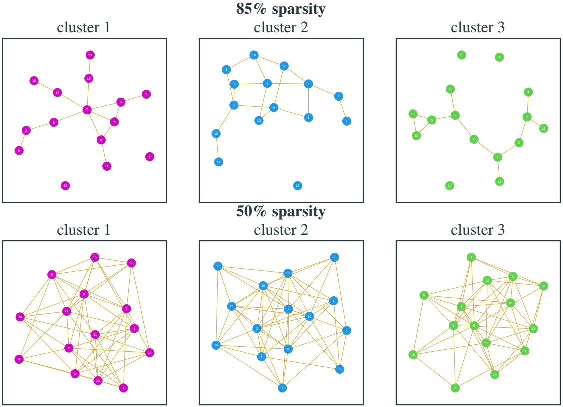


Figure: Posterior mode of the center parameters for each cluster.

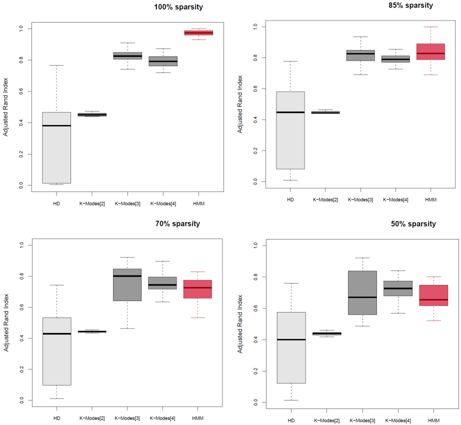
Simulation study with dependent variables

- $n_k = 75, k = 1, \ldots, K = 3$, and $m_j = 4$ with $j = 1, \ldots p = 15$
- Generating categorical data under a graphical modeling through Gaussian latent variables (Castelletti et al., 2024)
- Within each cluster, variables dependence structure based on a network with decreasing levels of sparsity, i.e., increasing strength of association among the variables



Simulation study with dependent variables

- $n_k = 75$, $k = 1, \ldots, K = 3$, and $m_j = 4$ with $j = 1, \ldots p = 15$
- Generating categorical data under a graphical modeling through Gaussian latent variables (Castelletti et al., 2024)
- Within each cluster, variables dependence structure based on a network with decreasing levels of sparsity, i.e., increasing strength of association among the variables



Simulation study with dependent variables

- $n_k = 75$, $k = 1, \ldots, K = 3$, and $m_j = 4$ with $j = 1, \ldots p = 15$
- Generating categorical data under a graphical modeling through Gaussian latent variables (Castelletti et al., 2024)
- Within each cluster, variables dependence structure based on a network with decreasing levels of sparsity, i.e., increasing strength of association among the variables

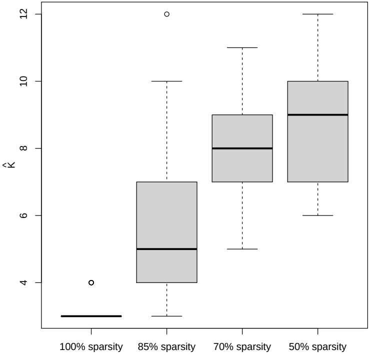


Figure: Estimated K under the HMM.

Beyond the local independence

- Accounting for the dependence among the p categorical variables also within the clusters

Beyond the local independence

- Accounting for the dependence among the p categorical variables also within the clusters
- Consider a mixture of HMMs: **enriched HMM**
- *Enriched priors* (Consonni and Veronese, 2001) for BNP: enriched Dirichlet (Wade et al., 2011), enriched Pitman–Yor process (Rigon et al., 2025), enriched Norm-IFPP (Franzolini et al., 2023)
- Connections to mixtures of LCMs (Malsiner-Walli et al., 2025)

Beyond the local independence

Enriched HMM

$$h(\boldsymbol{x}_i \mid \boldsymbol{c}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L, L, S_1, \ldots, S_L) = \sum_{l=1}^{L} \pi_l \sum_{s=1}^{S_l} \eta_{ls} \, p\left(\boldsymbol{x}_i \mid \boldsymbol{c}_{ls}, \boldsymbol{\sigma}_{ls}\right)$$

- L: number of outer components
- π_l: outer mixing weight
- S_l: number of inner components in outer component l
- η_{ls}: inner mixing weight
- $p\left(\boldsymbol{x}_i \mid \boldsymbol{c}_{ls}, \boldsymbol{\sigma}_{ls}\right)$: mixture kernel - Hamming p.m.f.
- $\boldsymbol{c}_{ls} = (c_{1ls}, \ldots, c_{pls})^\top$: component-specific center parameter of variable j in outer component l and inner component s
- $\boldsymbol{\sigma}_{ls} = (\sigma_{1ls}, \ldots \sigma_{pls})^\top$: component-specific scale parameter of variable j in outer component l and inner component s
- Two-level clustering of observations (outer and inner clusters)

Beyond the local independence

Hamming distribution

$K^* = 1, S_k^* = 1$

independent variables

HMM

$K^* = 2, S_k^* = 1$

local independence

Enriched HMM

$K^* = 2, S_k^* = 3$

stronger association

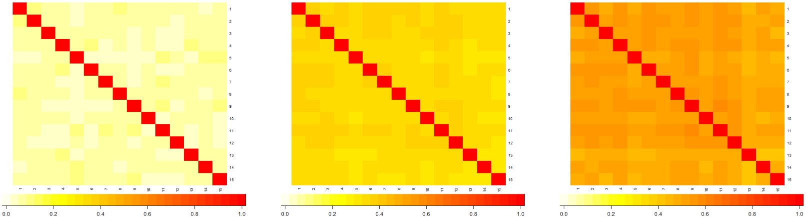


Figure: Cramer's V matrix of $p = 15$ variables computed over $n = 450$ data points with $m_j = 4$ categories. Data simulated assuming $\sigma = 0.4$.

Simulation study with dependent variables

- $n_k = 75$, $k = 1, \ldots, K = 3$, and $m_j = 4$ with $j = 1, \ldots p = 15$
- Generating categorical data under a graphical modeling through Gaussian latent variables (Castelletti et al., 2024)
- Within each cluster, variables dependence structure based on a network with decreasing levels of sparsity, i.e., increasing strength of association among the variables

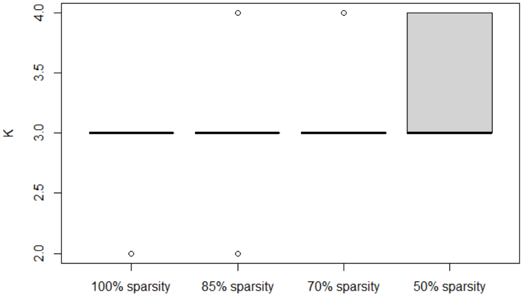


Figure: Estimated K under the Enriched HMM.

Summary

- A **new probability mass function**, based on the **Hamming distance**, to describe random vectors with support on a categorical space
- **Conjugate** Bayesian **inference** on the parameters of the Hamming distribution
- **Hamming mixture model** for clustering categorical data: finite mixture model with a random number of components
- Gibbs sampling strategy to provide **full posterior inference** of the cluster structure and the group-specific parameters and multiple imputation of missing values
- Theoretical results on model **identifiability** and **consistency** of the number of components
- **Empirical analysis** showed good accuracy in recovering the underlying clustering
- **Enriched** Hamming mixture model to overcome the local independence assumption
- **Ongoing**: study the properties of the enriched HMM (e.g., identifiability) and its link with a multivariate extension of the Hamming distribution

Thank you for your attention

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

References I

Argiento, R. and De Iorio, M. (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *Annals of statistics*, 50(5):2641–2663.

Biernacki, C., Celeux, G., and Govaert, G. (2010). Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140(11):2991–3002.

Castelletti, F., Consonni, G., and Vedova, M. L. D. (2024). Joint structure learning and causal effect estimation for categorical graphical models. *Biometrics*, 80.

Celeux, G. and Govaert, G. (2015). Latent class models for categorical data. In Henning, C., Melia, M., Murtagh, F., and Rocci, R., editors, *Handbook of cluster analysis*. Chapman & Hall/CRC.

Consonni, G. and Veronese, P. (2001). Conditionally reducible natural exponential families and enriched conjugate priors. *Scandinavian Journal of Statistics*, 28(3):377–406.

Franzolini, B., Cremaschi, A., van den Boom, W., and Iorio, M. D. (2023). Bayesian clustering of multiple zero-inflated outcomes. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2247):20220145.

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16(4):1279–1307.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

References II

Huang, Z. (1998). Extensions to the K-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304.

Malsiner-Walli, G., Grün, B., and Frühwirth-Schnatter, S. (2025). Without pain – clustering categorical data using a Bayesian mixture of finite mixtures of latent class analysis models. *Advances in Data Analysis and Classification*.

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158.

Rigon, T., Petrone, S., and Scarpa, B. (2025). Enriched Pitman–Yor processes. *Scandinavian Journal of Statistics*, 52(2):631–657.

Wade, S., Mongelluzzo, S., and Petrone, S. (2011). An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis*, 6(3):359–385.

Zhang, P., Wang, X., and Song, P. X.-K. (2006). Clustering categorical data based on distance vectors. *Journal of the American Statistical Association*, 101(473):355–367.