Bayesian Modelling Project

Asia Grillo ID: 5409650, Matteo Cristina ID: 5409473

1) Dataset

1.1) Description

The dataset employed in this analysis originates from a real-world small telecommunications company operating through a physical retail network. The original dataset comprises approximately **20,000 observations**, corresponding to all sales transactions recorded throughout the year **2024**. Each row represents a single sale made by an employee in one of the company's stores and includes information about the type of service sold and contextual characteristics of the transaction.
To enhance the interpretability and computational feasibility of the **Bayesian Latent Class Model**, the dataset was preprocessed and transformed into a structured set of **categorical variables**. A **stratified sampling** procedure was applied to extract a representative subset of **1000 observations**, using the variable `Shop` as the stratification baseline to preserve proportional representation across the company's five stores.

Variable	Description
`Business`	Whether the customer is a business entity (`Yes`, `No`)
`Upfront`	Whether an upfront payment was made at the time of sale (`Yes`, `No`)
`Shop`	Identifier of the store where the sale occurred (`Shop01`–`Shop05`)
`Service`	Type of service sold in the transaction (`Sim`, `Phone`, `Recharge`, `Extra`, `Router`)
`Customer`	Customer type based on purchase history (`New`, `Frequent`, `Known`, `Recurring`)
`Quarter`	Calendar quarter in which the transaction was made (`Q1`, `Q2`, `Q3`)

Table 1: Description of the categorical variables

All variables were selected for their potential to uncover latent behavioral profiles and patterns within the company's sales process. The categorical encoding facilitates the implementation of a probabilistic model based on discrete latent clusters, allowing for a nuanced interpretation of sales dynamics.

1.2) Exploratory Analysis

To gain an initial understanding of the structure of the dataset, we provide below a series of bar plots for each categorical variable. These plots display the counts of observations falling into each category, allowing us to visually inspect the prevalence of different levels across variables.

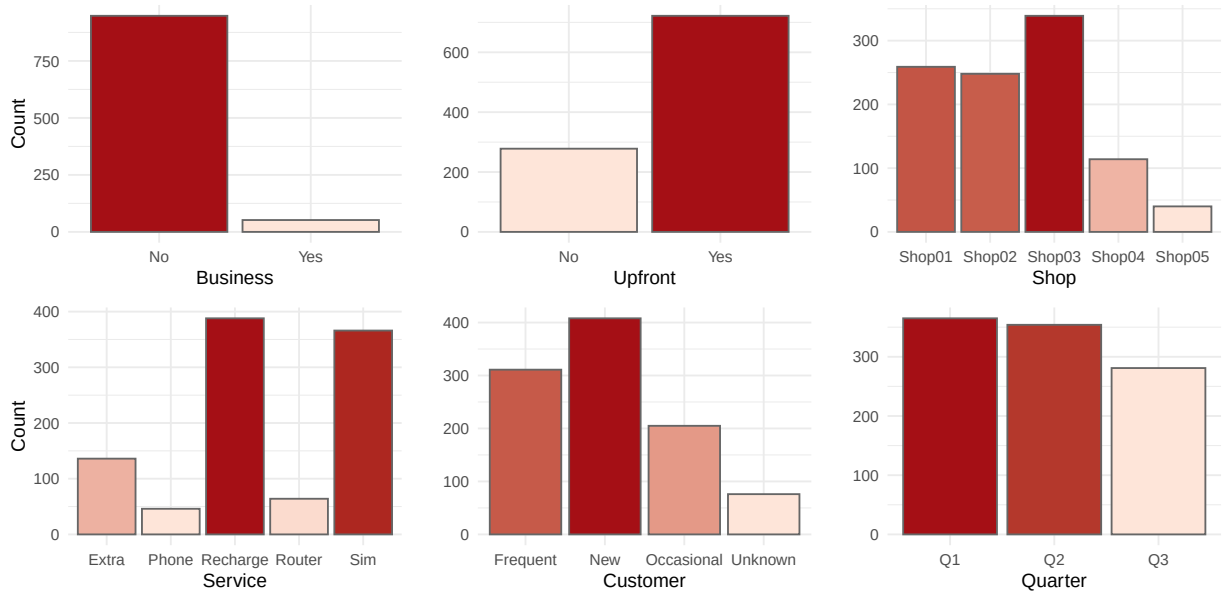


Figure 1: Bar plots of categorical variables

2) Bayesian Latent Class Model

2.1) Graphical Representation with DAG

The Bayesian Latent Class Model (BLCA) can be regarded as a specific instance of a categorical Bayesian network with a known and fixed DAG structure. In classical Directed Acyclic Graphs (DAGs), the nodes represent variables and the edges encode the conditional dependencies between them. The DAG structure allows for a factorization of the joint distribution according to parent-child relationships, facilitating the identification of conditional independencies.

In the Latent Class Model, the data-generating process is assumed to be driven by an **unobserved (latent) variable** z_n, which classifies each observation into one of K latent groups. The observed categorical responses $\mathbf{X}_n = (X_{n1}, \dots, X_{nP})$ are then generated conditionally independently given the cluster assignment.

To accommodate this, the DAG structure is simple yet expressive:

$$\pi \to z_n \to \mathbf{X}_n \leftarrow \phi$$

where: π denotes the vector of cluster probabilities, z_n is the latent cluster label for observation n and $\phi_{k,d}$ denotes the class- and variable-specific categorical response probabilities.

From this structure, we derive that the observed variables $X_{n1}, \dots, X_{nP}$ are conditionally independent given z_n:

$$X_{n1} \perp X_{n2} \perp \cdots \perp X_{nP} \mid z_n$$

Moreover, the parameters π and ϕ are assumed to be **a priori independent**, allowing for separate posterior updates conditional on the latent allocations $\mathbf{z}$:

$$\pi \perp \phi$$

2.2) Model Specification

The generative model assumes the following probabilistic structure:

$$X_{nd} \mid z_n = k, \phi \sim \text{Categorical}(\phi_{k,d}) \quad \text{with} \quad \phi_{k,d} = (\phi_{k,d,1}, \dots, \phi_{k,d,V_d})$$

$$z_n \mid \pi \sim \text{Categorical}(\pi) \quad \text{with} \quad \pi = (\pi_1, \dots, \pi_K), \quad \sum_{k=1}^{K} \pi_k = 1$$

Given this structure, the complete-data likelihood — i.e., the joint distribution of the observed data $\mathbf{X}$ and the latent variables $\mathbf{z}$, conditional on π and ϕ — factorizes as:

$$\begin{aligned} p(\mathbf{X}, \mathbf{z} \mid \pi, \phi) &= p(\mathbf{X} \mid \mathbf{z}, \phi) \cdot p(\mathbf{z} \mid \pi) \\ &= \prod_{n=1}^{N} \pi_{z_n} \prod_{d=1}^{P} \phi_{z_n, d, X_{nd}} \\ &= \prod_{k=1}^{K} \pi_k^{N_k} \prod_{d=1}^{P} \prod_{v=1}^{V_d} \phi_{k,d,v}^{N_{kdv}} \end{aligned}$$

2.3) Joint prior

We place semiconjugate Dirichlet priors on the parameters of the model:

$$\pi \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \quad \Longrightarrow \quad p(\pi) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

$$\phi_{k,d} \sim \text{Dir}(\beta_{d1}, \dots, \beta_{dV_d}) \quad \Longrightarrow \quad p(\phi_{k,d}) = \frac{1}{B(\beta_d)} \prod_{v=1}^{V_d} \phi_{k,d,v}^{\beta_{dv} - 1}$$

Where: $B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$ and $B(\beta_d) = \frac{\prod_{v=1}^{V_d} \Gamma(\beta_{dv})}{\Gamma\left(\sum_{v=1}^{V_d} \beta_{dv}\right)}$ are multivariate Beta functions.

Since π and all $\phi_{k,d}$ are assumed to be independent a priori, the joint prior becomes:

$$\begin{aligned} p(\pi, \phi) = p(\pi) \cdot p(\phi) &= p(\pi) \cdot \prod_{k=1}^{K} \prod_{d=1}^{P} p(\phi_{k,d}) \\ &= \left[\frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \right] \cdot \prod_{k=1}^{K} \prod_{d=1}^{P} \left[\frac{1}{B(\beta_d)} \prod_{v=1}^{V_d} \phi_{k,d,v}^{\beta_{dv} - 1} \right] \end{aligned}$$

In our model, we adopt a **noninformative prior** for the categorical parameters π and ϕ, following the **Jeffreys prior** approach. Jeffreys prior for a categorical distribution with K categories corresponds to a Dirichlet distribution with parameter $\alpha_k = \frac{1}{2}$ for all k. Hence:

$$\pi \sim \text{Dir}\left(\frac{1}{2}, \dots, \frac{1}{2}\right) \quad \text{and} \quad \phi_{k,d} \sim \text{Dir}\left(\frac{1}{2}, \dots, \frac{1}{2}\right)$$

The choice $\alpha_k = \frac{1}{2}$ ensures invariance under reparameterization and is derived by computing the **Fisher information**.

We can show the derivation for the categorical case starts from the Fisher information matrix:

$$\mathcal{J}_{ij} = \mathbb{E}\left[\frac{\partial \log p(x)}{\partial p_i} \cdot \frac{\partial \log p(x)}{\partial p_j} \right]$$

The Jeffreys prior is proportional to the square root of the determinant of the Fisher information matrix:

$$p(p) \propto \sqrt{\det(\mathcal{I}(p))} \propto \prod_{k=1}^{K} p_k^{-1/2}$$

which corresponds to the kernel of a Dirichlet distribution with all parameters equal to $\frac{1}{2}$. In our model, the components p_k represent either cluster probabilities π_k or the category-specific probabilities $\phi_{k,d,v}$ for variable d in cluster k.

2.4) Joint Posterior

The joint posterior distribution is proportional to the complete-data likelihood times the joint prior:

$$\begin{aligned} p(\Pi, \phi, \mathbf{z} \mid \mathbf{X}) &\propto p(\mathbf{X}, \mathbf{z} \mid \Pi, \phi) \cdot p(\Pi) \cdot p(\phi) \\ &\propto \left[\prod_{k=1}^{K} \pi_k^{N_k}\right] \left[\prod_{k=1}^{K}\prod_{d=1}^{P}\prod_{v=1}^{V_d} \phi_{kdv}^{N_{kdv}}\right] \left[\prod_{k=1}^{K} \pi_k^{\alpha_k - 1}\right] \left[\prod_{k=1}^{K}\prod_{d=1}^{P}\prod_{v=1}^{V_d} \phi_{kdv}^{\beta_{dv}-1}\right] \\ &\propto \left[\prod_{k=1}^{K} \pi_k^{N_k+\alpha_k-1}\right] \left[\prod_{k=1}^{K}\prod_{d=1}^{P}\prod_{v=1}^{V_d} \phi_{kdv}^{N_{kdv}+\beta_{dv}-1}\right] \end{aligned}$$

2.5) Full Conditional Distributions

Full Conditional of π

Due to the conjugacy between the Categorical likelihood and the Dirichlet prior, the full conditional distribution of π given the latent allocations $\mathbf{z}$ is:

$$p(\pi \mid \mathbf{z}) \propto \prod_{n=1}^{N} \pi_{z_n} \cdot p(\pi) \propto \prod_{k=1}^{K} \pi_k^{N_k} \cdot \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \propto \prod_{k=1}^{K} \pi_k^{N_k+\alpha_k-1}$$

which is the kernel of a Dirichlet distribution with updated parameters:

$$\pi \mid \mathbf{z} \sim \text{Dir}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

```
full_cond_pi <- function(z, alpha) {
  K <- length(alpha)
  N_k <- tabulate(z, nbins = K)  # observation count for each class
  return(rdirichlet(1, alpha + N_k)) # returns a vector of dimension K
}
```

Full Conditional of $\phi_{k,d}$

For each latent cluster k and variable d, the full conditional distribution of $\phi_{k,d}$ is obtained by combining the cluster-conditional likelihood and the Dirichlet prior:

$$p(\phi_{k,d} \mid \mathbf{z}, \mathbf{X}) \propto \left[\prod_{v=1}^{V_d} \phi_{kdv}^{\beta_{dv}-1}\right] \left[\prod_{n:z_n=k} \phi_{kd,x_{nd}}\right] \propto \prod_{v=1}^{V_d} \phi_{kdv}^{\beta_{dv}-1+N_{kdv}}$$

which corresponds to a Dirichlet distribution with updated parameters:

$$\phi_{k,d} \mid \mathbf{z}, \mathbf{X} \sim \text{Dir}(\beta_{d1} + N_{kd1}, \ldots, \beta_{dV_d} + N_{kdV_d})$$

```r
full_cond_phi <- function(X, z, beta_list, K, V) {
  P <- ncol(X)
  phi <- vector("list", K)
  for (k in 1:K) {
    phi_k <- list() # distributions of variables in class k
    for (d in 1:P) {
      v_d <- V[d] # number of categories for variable d
      x_kd <- X[z == k, d] # values of the variable d in class k
      counts <- tabulate(x_kd, nbins = v_d) # observed frequencies per level
      beta_d <- beta_list[[d]]
      phi_k[[d]] <- rdirichlet(1, counts + beta_d)
    }
    phi[[k]] <- phi_k
  }
  return(phi) # K x P list with dimension vectors V_d
}
```

Full Conditional of z_n

For each observation $n = 1, \dots, N$, the full conditional distribution of the latent cluster assignment z_n is given by:

$$\Pr(z_n = k \mid \mathbf{X}_n, \Pi, \phi) = \frac{\pi_k \prod_{d=1}^{P} \phi_{k,d,X_{nd}}}{\sum_{j=1}^{K} \pi_j \prod_{d=1}^{P} \phi_{j,d,X_{nd}}} \propto \pi_k \prod_{d=1}^{P} \phi_{k,d,X_{nd}}$$

It corresponds to the posterior probability of cluster k, obtained by updating the prior π_k with the **likelihood** of observing the profile $\mathbf{X}_n$ under class-specific parameters ϕ_k.

```r
full_cond_z <- function(X, pi, phi, K) {
  N <- nrow(X)
  P <- ncol(X)
  z_new <- integer(N)
  for (n in 1:N) {
    probs <- numeric(K)
    for (k in 1:K) {
      p_k <- pi[k] # initialize with the prior pi_k
      for (d in 1:P) {
        v <- X[n, d] # observed value for the variable d in obs n
        p_k <- p_k * phi[[k]][[d]][v] # joint probability
      }
      probs[k] <- p_k
    }
    probs <- probs / sum(probs) # normalization
    z_new[n] <- sample(1:K, size = 1, prob = probs)
  }
  return(z_new)
}
```

2.6) Notation

Symbol	Definition	Description
N		Number of observations (i.e., sales transactions)
P		Number of categorical variables
V_d		Number of categories for variable d
K		Number of latent clusters
X_{nd}		Observed value of variable d for observation n
z_n	$z_n \in \{1, \dots, K\}$	Latent cluster assignment for observation n
π_k	$\Pr(z_n = k)$	Prior probability of belonging to cluster k
ϕ_{kdv}	$\Pr(X_{nd} = v \mid z_n = k)$	Probability that variable d takes value v in cluster k
N_k	$\sum_{n=1}^{N} \mathbb{1}(z_n = k)$	Number of observations assigned to cluster k
N_{kdv}	$\sum_{n=1}^{N} \mathbb{1}(z_n = k, X_{nd} = v)$	Number of times category v of variable d is observed in cluster k

Table 2: Notation and definitions used

3) Model Selection

Choosing the **number of latent clusters** K is a crucial step in the specification of Latent Class Models. The selected value of K directly determines the complexity and interpretability of the model, as well as its ability to capture the heterogeneity in the data.

From a **Bayesian perspective**, several sophisticated approaches have been proposed in the literature to estimate K. These include placing a prior distribution over K, such as the Beta Negative Binomial distribution. While such methods offer principled solutions that integrate model uncertainty, they often come at the cost of increased computational and modelling complexity.

To avoid these additional complications, we adopt a **frequentist approach** to model selection. A customary strategy in latent cluster modelling is to select the number of clusters using **information criteria**, which balance model fit and parsimony.

Specifically, we evaluate two widely used criteria:

- Bayesian Information Criterion: $\text{BIC} = -2 \cdot \log \hat{L} + \nu \cdot \log(N)$;
- Integrated Completed Likelihood: $\text{ICL} = \text{BIC} - 2 \cdot \mathbb{E}_{\hat{\theta}} \left[\log p(\mathbf{z} \mid \mathbf{X}, \hat{\theta}) \right]$.

These criteria are computed via the Expectation-Maximization (EM) algorithm applied to the Latent Class Model. The figure below displays the values of BIC and ICL across different values of K and the optimal number of clusters is selected as the one minimizing these indices. Models yielding smaller values of the information criteria are generally preferred, as they indicate a better trade-off between goodness of fit and model complexity.

In our case, since we are dealing with a model that includes **latent variables**, we place particular emphasis on the **Integrated Completed Likelihood (ICL)** criterion, which accounts not only for the fit, but also for the uncertainty in cluster membership. Based on the values of BIC and ICL obtained from the EM algorithm, we selected a model with **3 clusters**, as it minimizes the ICL.

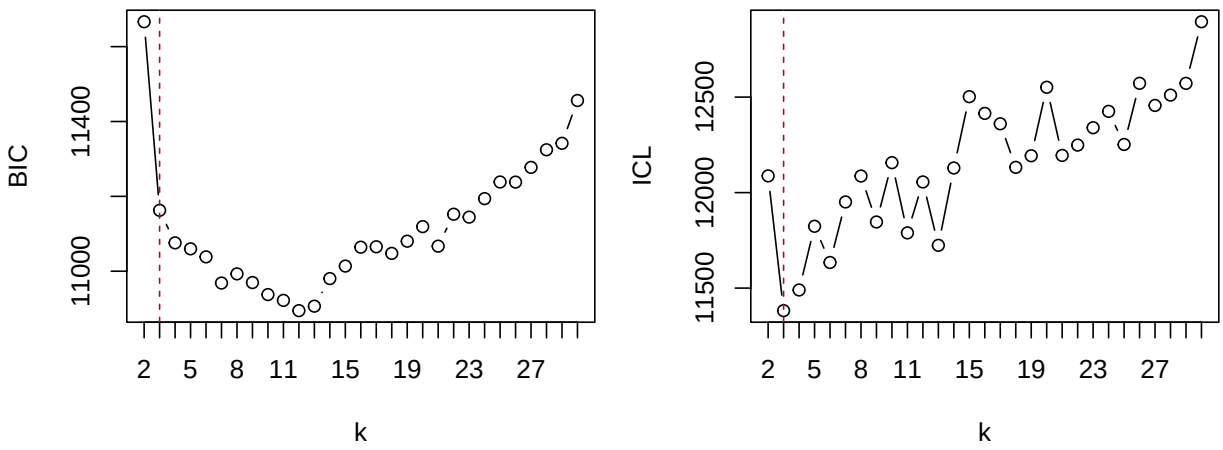


Figure 2: Model Selection via Information Criteria

4) Gibbs Sampler

The Gibbs sampler iteratively updates each parameter by drawing from its conditional distribution given the current values of all the others. In our Latent Class Model, the vector of unknowns is $\theta = (\pi, \phi, \mathbf{z})$ and the algorithm proceeds by alternately sampling from the full conditionals as follow:

Start from the initial value $\theta^{(0)} = \left(\pi^{(0)}, \phi^{(0)}, \mathbf{z}^{(0)}\right)$.

For each iteration $s = 1, \ldots, S$, repeat the following steps:

1. Draw $\pi^{(s)} \sim p(\pi \mid \mathbf{z}^{(s-1)})$
2. For each cluster $k = 1, \ldots, K$ and each variable $d = 1, \ldots, P$:
   2.1 Draw $\phi_{k,d}^{(s)} \sim p(\phi_{k,d} \mid \mathbf{z}^{(s-1)}, \mathbf{X})$
3. For each observation $n = 1, \ldots, N$:
   3.1 Compute $\Pr(z_n = k \mid \mathbf{X}_n, \pi^{(s)}, \phi^{(s)})$
   3.2 Draw $z_n^{(s)} \sim \text{Categorical}(\Pr(z_n = 1), \ldots, \Pr(z_n = K))$

The output is a **dependent sequence**

$$\left\{\left(\pi^{(1)}, \phi^{(1)}, \mathbf{z}^{(1)}\right), \ \ldots, \ \left(\pi^{(S)}, \phi^{(S)}, \mathbf{z}^{(S)}\right)\right\}$$

of samples approximately drawn from the joint posterior distribution $p(\pi, \phi, \mathbf{z} \mid \mathbf{X})$.

```
K = 3
V = c(2, 2, 5, 5, 4, 3)
alpha = rep(0.5, K)
beta_list <- list(rep(0.5, V[1]), rep(0.5, V[2]), rep(0.5, V[3]),
                  rep(0.5, V[4]), rep(0.5, V[5]), rep(0.5, V[6]))
gibbs_sampler <- function(X, K, V, alpha, beta_list, n_iter = 15000) {
  N <- nrow(X)
  P <- ncol(X)
  z <- sample(1:K, N, replace = TRUE)
```

```r
  pi <- rep(1 / K, K)
  phi <- full_cond_phi(X, z, beta_list, K, V)
  samples <- list(pi = vector("list", n_iter),
                  phi = vector("list", n_iter),
                  z = vector("list", n_iter))
  for (t in 1:n_iter) {
    pi <- full_cond_pi(z, alpha)
    phi <- full_cond_phi(X, z, beta_list, K, V)
    z <- full_cond_z(X, pi, phi, K)
    samples$pi[[t]] <- pi
    samples$phi[[t]] <- phi
    samples$z[[t]] <- z
  }
  return(samples)
}
samples <- gibbs_sampler_time(X = X, K = K, V = V, alpha = alpha,
                              beta_list = beta_list, n_iter = 15000)
```

To improve convergence and reduce autocorrelation, we discard the first $B = 400$ iterations as **burn-in** and apply **thinning** by storing one draw every $T = 10$ iterations.

5) Diagnostics

To assess the convergence and mixing of the Gibbs sampler, we perform MCMC diagnostics.

5.1) Diagnostics of π

Trace plot and ACF

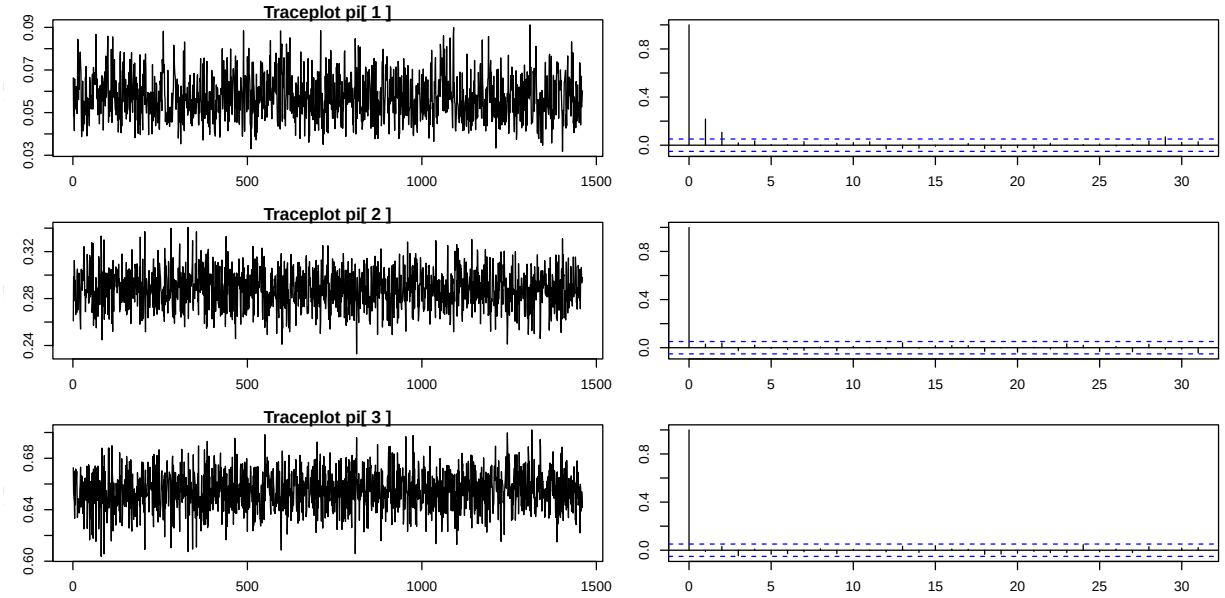


Figure 3: Convergence diagnostics for π

	π_1	π_2	π_3
ESS	825.4476	1460	1460

Table 3: Effective Sample Size (ESS) for π

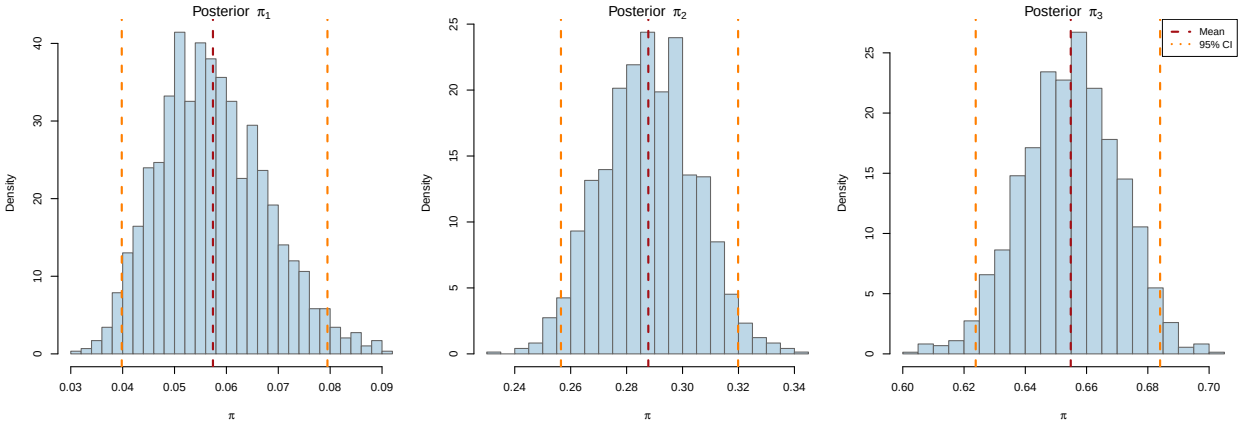


Figure 4: Posterior mean and Credible Intervals of π

5.2) Diagnostics of $\phi_{k,d}$

Displaying all trace plots and autocorrelation functions (ACF) for the full set of estimated $\phi_{k,d,v}$ parameters would be both unfeasible, given the large number of cluster-variable-category combinations.

Therefore, our strategy begins with a general inspection of the **posterior distributions of the category proportions** for each variable within each latent cluster. This helps identify which variables contribute most clearly to the separation between clusters.

For a selected subset of such informative variables — those showing the clearest distinction across clusters — we provide trace plots and ACF plots of their corresponding $\phi_{k,d,v}$ values. This targeted approach offers an **indicative assessment of convergence** while avoiding redundancy.

Posterior Inference

By inspecting the bar plots below of the posterior distributions of category proportions across clusters, we observe that variables Business and Quarter show no meaningful differentiation among clusters. In a future analysis, it could be considered to remove these variables from the model — while maintaining the number of clusters at $K = 3$ — to potentially improve model parsimony.

In contrast, **variable Upfront exhibits a clear separation**, particularly distinguishing clusters 1 and 3 from cluster 2. For this reason, we focus our convergence diagnostics on the category-specific parameters $\phi_{k,2,v}$ of this variable, analyzing their trace plots across all clusters.

Other variables that appear to contribute strongly to cluster separation include: Customer, Shop and Business.

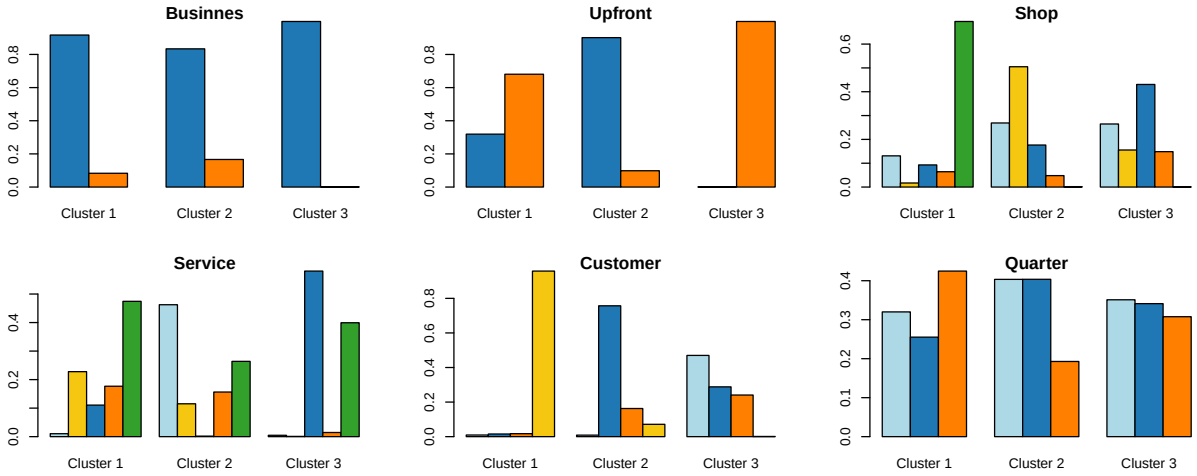


Figure 5: Posterior distributions of category proportions $\phi_{k,d,v}$ across clusters for selected variables

Trace plot and ACF

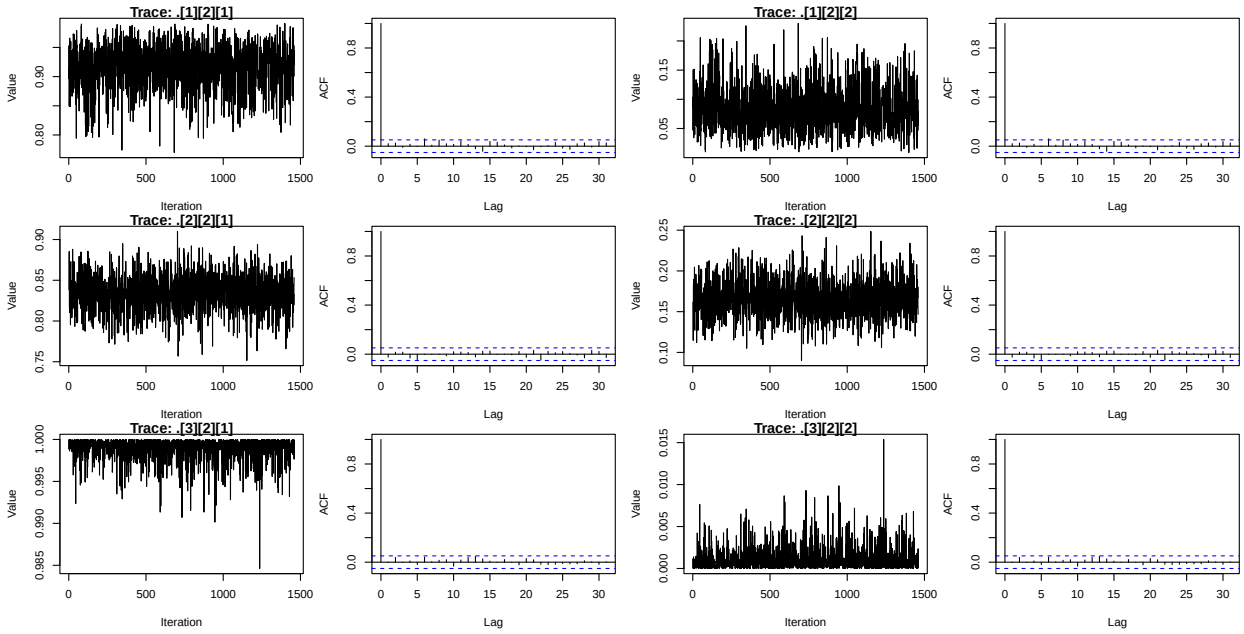


Figure 6: Convergence diagnostics for $\phi_{k,2,v}$

5.3) Diagnostics for z_n

Traditional convergence diagnostics are not directly applicable to $\mathbf{z}$, as it is a discrete latent variable. Instead of tracking the trajectory of each individual z_n, we assess the **stability of the cluster sizes** over the iterations of the Gibbs sampler.

The figure below shows, for each iteration, the number of observations assigned to each cluster. Stable cluster sizes across iterations are indicative of good mixing and convergence of the latent allocation structure.

If the cluster sizes fluctuate significantly over time or display trends, this may suggest issues with label switching, poor convergence, or an over-specified model.

In our case, the trajectories of cluster sizes appear stable after the initial burn-in period, supporting the reliability of the inferred cluster structure.

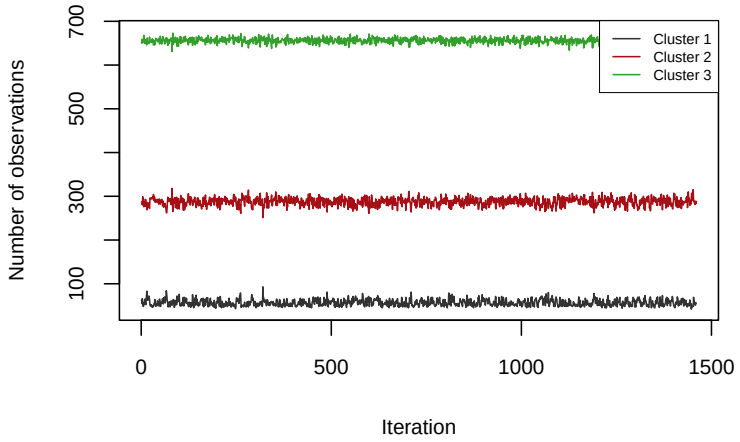


Figure 7: Clusters size over time

6) Posterior Predictive Check

The posterior predictive distribution integrates over the uncertainty in the model parameters $\theta = (\pi, \phi)$ as follows:

$$p(\tilde{\mathbf{X}} \mid \mathbf{X}) = \int p(\tilde{\mathbf{X}} \mid \theta)\, p(\theta \mid \mathbf{X})\, d\theta$$

Since this integral is analytically intractable, we approximate it via **Monte Carlo** as follows:

For each posterior draw $s = 1, \ldots, S$:

1. Use the sampled values $\pi^{(s)}$ and $\phi^{(s)}$ from Gibbs Sampling
2. For each individual $i = 1, \ldots, N$:
   2.1 Sample a cluster assignment from $\tilde{z}_i^{(s)} \sim \text{Categorical}(\pi^{(s)})$
3. For each variable $d = 1, \ldots, P$:
   3.1 Generate a replicated response from $\tilde{X}_{i,d}^{(s)} \sim \text{Categorical}(\phi^{(s)}_{\tilde{z}_i^{(s)},d})$

This yields a sequence $\left\{\tilde{\mathbf{X}}^{(1)}, \ldots, \tilde{\mathbf{X}}^{(S)}\right\}$ that approximates samples from $p(\tilde{\mathbf{X}} \mid \mathbf{X})$, allowing us to perform posterior predictive checks.

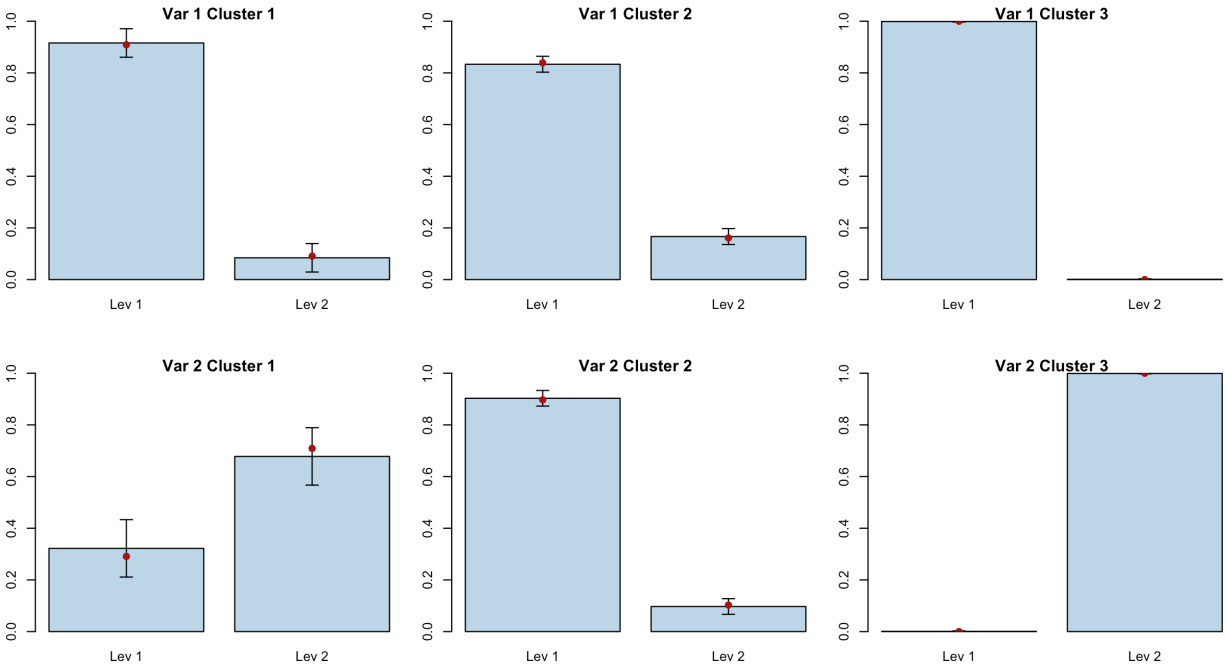
```
simulate_yrep_with_z <- function(pi_samples, phi_samples, N, P, V, S = 1000) {
  yrep_list <- vector("list", S)
  zrep_list <- vector("list", S)
  for (s in 1:S) {
    pi_s  <- pi_samples[[s]]
    phi_s <- phi_samples[[s]]
```

```
    yrep <- matrix(NA, nrow = N, ncol = P)
    z_s <- integer(N)
    for (i in 1:N) {
      z_i <- sample(1:length(pi_s), size = 1, prob = pi_s)
      z_s[i] <- z_i
      for (d in 1:P) {
        yrep[i, d] <- sample(1:V[d], size = 1, prob = phi_s[[z_i]][[d]])
      }
    }
    yrep_list[[s]] <- yrep
    zrep_list[[s]] <- z_s
  }
  return(list(yrep_list = yrep_list, zrep_list = zrep_list))
}
yrep_list <- simulate_yrep_with_z(pi_samples, phi_samples, N = nrow(X),
                                  P = ncol(X), V = V, S = 1000)
```

To evaluate the ability of the model to reproduce the observed data structure, we perform posterior predictive checks focusing on the **distribution of categorical levels across latent clusters**.

For a selected variable d, we examine the distribution of its levels within each cluster $k = 1, \ldots, K$, using synthetic datasets drawn from the posterior predictive distribution.

Specifically, for each cluster and for each level of the variable, we compute the relative frequency of that level across all replicated datasets. The distribution of these simulated frequencies is then summarized using **bar plots with error bars**, where the height of each bar represents the posterior mean and the vertical line indicates the 95% credible interval. A red dot is superimposed to show the **observed frequency** of that level in the actual dataset (conditional on cluster assignment).



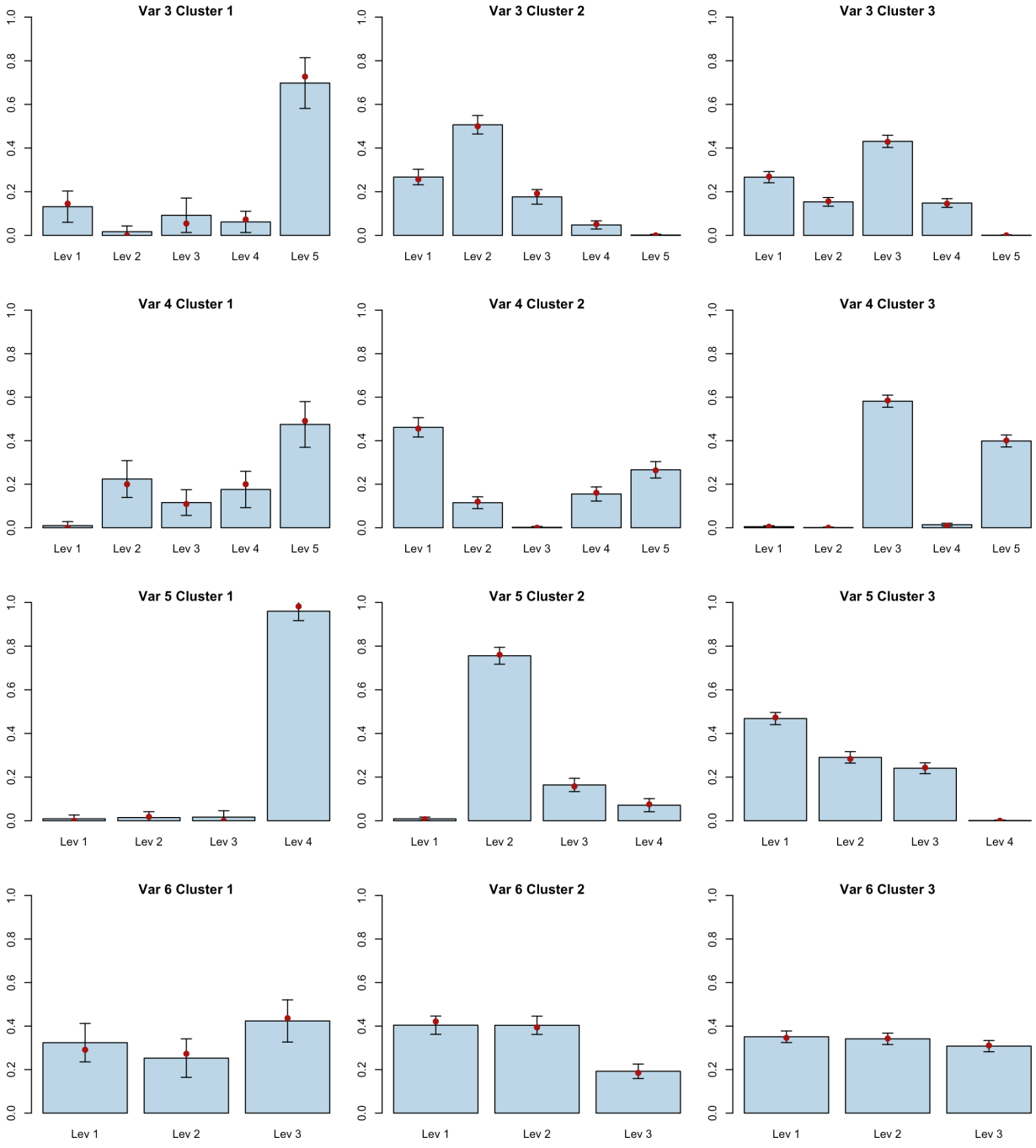


Figure 8: Posterior Predictive Distribution vs Observed Data

We can see from the plots that the observed frequencies (red dots) tend to fall close to the posterior means of the replicated frequencies. This indicates that the Bayesian Latent Class Model, in its current specification, provides a good fit to the observed data and is capable of capturing its main structural patterns.

Nonetheless, it is worth noting that the **nature of the dataset**, where each observation corresponds to a single sales transaction, may still limit the model's ability to fully recover stable latent behaviors. While the current unit of analysis is sufficient to detect broad patterns, more meaningful clustering structures could potentially emerge by aggregating the data at the customer or employee level. Such an approach would require additional variables that characterize these entities, which were not available in the present study.

Additionally, we believe that **extending the model to include a prior distribution on the number of latent clusters** K could significantly improve model flexibility and performance in future applications.

7) Conclusion

The Bayesian Latent Class Model (BLCM) represents a flexible and principled approach for uncovering hidden heterogeneity in multivariate categorical data. Its Bayesian formulation allows for the incorporation of prior knowledge, the quantification of uncertainty and straightforward model-based clustering in contexts where traditional methods may fall short. Typical domains of application include:

- **Market segmentation**: identifying distinct groups of customers based on preferences;
- **Healthcare and epidemiology**: uncovering latent disease subtypes or patient profiles based on symptoms.
- **Political science and voting behavior**: detecting latent groups of individuals based on categorical responses to surveys.

Thanks to its interpretability and formal probabilistic foundation, the BLCM remains a valuable tool for both exploratory and confirmatory analysis across many disciplines.

References

[1] Agresti, A. (2005). *Bayesian Inference for Categorical Data Analysis*. Department of Statistics, University of Florida.

[2] Malsiner-Walli, G., Grün, B., & Frühwirth-Schnatter, S. (2016). *Without Pain – Clustering Categorical Data Using a Bayesian Mixture of Finite Mixtures of Latent Class Analysis Models*.

[3] Argiento, R., Filippi-Mazzola, E., & Paci, L. (2021). *Model-based clustering of categorical data based on the Hamming distance*. Cited for model selection via information criteria (BIC, ICL).

[4] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464. Cited as the original reference for the Bayesian Information Criterion (BIC).

[5] Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(7), 719–725.

[6] Rodríguez, C. E., & Walker, S. G. (2014). *Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies*.

[7] Castelletti, F., Consonni, G., & Della Vedova, M. L. (2023). *Joint Structure Learning and Causal Effect Estimation for Categorical Graphical Models*. Università Cattolica del Sacro Cuore, Milan.