



# A Bayesian Model for Co-clustering Ordinal Data with Informative Missing Entries

Alice Giampino<sup>1</sup> · Antonio Canale<sup>2</sup> · Bernardo Nipoti<sup>1</sup>

Received: 4 November 2024 / Accepted: 28 July 2025 / Published online: 5 August 2025  
© The Author(s) 2025

## Abstract

Several approaches have been proposed in the literature for clustering multivariate ordinal data. These methods typically treat missing values as absent information, rather than recognizing them as valuable for profiling population characteristics. To address this gap, we introduce a Bayesian nonparametric model for co-clustering multivariate ordinal data that treats censored observations as informative, rather than merely missing. We demonstrate that this offers a significant improvement in understanding the underlying structure of the data. Our model exploits the flexibility of two independent Dirichlet processes, allowing us to infer potentially distinct subpopulations that characterize the latent structure of both subjects and variables. The ordinal nature of the data is addressed by introducing latent variables, while a matrix factorization specification is adopted to handle the high dimensionality of the data in a parsimonious way. The conjugate structure of the model enables an explicit derivation of the full conditional distributions of all the random variables in the model, which facilitates seamless posterior inference using a Gibbs sampling algorithm. We demonstrate the method's performance through simulations and by analyzing politician and movie ratings data.

**Keywords** Bayesian nonparametrics · Gibbs sampling · Missing data · Dirichlet process

## 1 Introduction

Multivariate ordinal data, consisting of repeated measurements of vectors of ordinal responses, play a crucial role in various fields. Our focus is on scenarios where repeated measures are available for only a subset of the vector entries, and the missing data can be viewed as informative rather than purely random. For example, in the analysis of political votes, the voting records of  $n$  politicians on  $p$  legislative bills are only available for the sessions they attended, where absence from a session may represent a deliberate political strategy.

Similarly, in recommendation systems, the data consists of customer ratings on  $p$  items. In the well-known Netflix Prize data (Bennett et al. 2007), missingness arises when users choose not to rate certain movies, potentially reflecting their lack of interest, preference, or other underlying factors.

Our analysis of this type of data focuses on clustering both the statistical units and the entries of the ordinal random vectors. This problem can be framed within the context of co-clustering methods (Hartigan 1972), a set of techniques also known as biclustering or two-mode clustering. The core idea is to simultaneously cluster the rows and columns of an  $(n \times p)$ -dimensional data matrix  $Y$ , with rows associated to  $n$  statistical units and columns consisting of  $p$  ordinal outcomes. Over the past three decades, co-clustering methods have been widely applied across various fields (Busygin et al. 2008), with some approaches resorting to Bayesian nonparametric tools (Meeds and Roweis 2007; Wang et al. 2011, 2012). Despite extensive literature, most existing methods fail to adequately address the challenges posed by informative missingness. A notable exception is the R package `biclustermd` by Reisner et al. (2019), which uses a geometric approach to optimally rearrange the rows and columns of the data matrix when missing values are present. To our

---

✉ Alice Giampino  
alice.giampino@unimib.it

Antonio Canale  
antonio.canale@unipd.it

Bernardo Nipoti  
bernardo.nipoti@unimib.it

<sup>1</sup> Department of Economics, Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy

<sup>2</sup> Department of Statistics, University of Padova, Via C. Battisti, 241, 35121 Padova, Italy

knowledge, no model-based approach for co-clustering with missing entries has yet been explored in the literature. We fill this gap by introducing a flexible Bayesian model capable of performing co-clustering while accounting for the informative nature of missing data. This approach, referred to as the Co-Clustering model for Ordinal Censored Observations (CO<sup>3</sup>), employs a latent variable representation to jointly model both the ordinal responses and the indicator variables encoding the presence of missing entries. CO<sup>3</sup> builds on the idea of linking observed discrete data with latent continuous variables, thus following a well-known strategy in Bayesian modeling (Albert and Chib 1993; Kottas et al. 2005; Canale and Dunson 2011; Kowal and Canale 2020). At the level of latent variables, CO<sup>3</sup> employs a Bayesian matrix factorization representation (Salakhutdinov and Mnih 2008), exploiting the idea that the observed multivariate responses of a statistical unit are driven by a smaller set of unobserved latent factors. Additionally, the definition of CO<sup>3</sup> relies on the flexibility of two independent Dirichlet processes, which allow the model to infer potentially distinct subpopulations characterizing the latent structure of both subjects and variables. This construction not only facilitates the clustering of statistical units and variable entries but also provides a robust framework for effectively addressing the complexities associated with informative censoring. Our investigations show that, while the model is specifically designed to account for informative missingness, its applicability extends to situations where data are missing at random as well.

The rest of the paper is organized as follows. Section 2 is dedicated to the specification of the model and the study of its prior properties. A strategy for posterior computation is presented in Section 3. The performance of the model is investigated by means of the analysis of synthetic and real data, as presented in Sections 4 and 5, respectively. Additional results on the full conditional distributions for posterior sampling, along with further details on the simulation study and the real dataset analyses, are provided in the Supplementary Material.

## 2 A co-clustering model for ordinal censored observations

We let  $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})$ , with  $i = 1, \dots, n$ , be the  $i$ -th row of a data matrix  $\mathbf{Y}$ , that is a vector of ordinal responses with the general entry  $y_{ij} \in [c]$ , for  $j = 1, \dots, p$ , where  $[c] = \{1, \dots, c\}$  and  $c$  denotes the number of ordinal levels/categories for a single response variable. For example, in our two motivating applications,  $y_{ij}$  represents the vote of the  $i$ -th senator in the  $j$ -th voting session ( $c = 2$  for yes/no), or the rating given by  $i$ -th user to the  $j$ -th movie ( $c = 10$ , for a 10-point rating scale). In both these applications, as well as in many other contexts, it is common for some components

of each observation  $\mathbf{y}_i$  to be missing, with the missingness occurring in a non-random manner.

### 2.1 Model formulation

We formalize the missingness by endowing each  $\mathbf{y}_i$  with a vector  $\boldsymbol{\delta}_i = (\delta_{i1}, \dots, \delta_{ip})$ , with  $\delta_{ij}$  indicating whether the  $j$ -th component  $y_{ij}$  of  $\mathbf{y}_i$  was observed ( $\delta_{ij} = 1$ ) or not ( $\delta_{ij} = 0$ ). This can be easily generalized to the case where, for each unit,  $q$  ordinal responses are measured, each accompanied by a corresponding missingness indicator. For instance,  $q = 2$  when streaming platforms record variables such as movie rating and number of times a user has watched any given movie, e.g. with levels “never”, “once”, “twice”, “more than twice”. We will henceforth focus on the specific case with  $q = 1$ .

Similarly to Albert and Chib (1993) or Kottas et al. (2005), we link both the observed ordinal  $y_{ij}$  and the missingness indicators  $\delta_{ij}$  with latent continuous variables. Specifically, for the ordinal responses, we introduce  $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})$  such that

$$y_{ij} = \kappa \quad \text{if } \gamma_{\kappa-1} < z_{ij} \leq \gamma_{\kappa};$$

$$y_{ij} \in [c] \quad \text{if } -\infty = \gamma_0 < z_{ij} \leq \gamma_c = \infty.$$

where  $-\infty = \gamma_0 < \gamma_1 < \dots < \gamma_{c-1} < \gamma_c = \infty$  are arbitrarily fixed cutoffs. Similarly, for the censoring variables, we introduce  $\mathbf{w}_i = (w_{i1}, \dots, w_{ip})$  such that  $\delta_{ij} = 1$  if  $w_{ij} \geq 0$  and  $\delta_{ij} = 0$ , if  $w_{ij} < 0$ . That is, the components of  $\mathbf{w}_i$  take values in  $\mathbb{R}$  with their sign determining the value of  $\delta_{ij}$ . This formulation allows us to write the joint distribution for the observed data as equivalent to

$$\prod_{i,j: w_{ij} \geq 0} \Pr(z_{ij} \in (\gamma_{y_{ij}-1}, \gamma_{y_{ij}}]). \quad (1)$$

The latent variables  $z_{ij}$  and  $w_{ij}$  have intuitive interpretations. For example, in the analysis of political votes presented in Section 5.1,  $z_{ij}$  represents the  $i$ th senator’s inclination to support the  $j$ th voting session, while  $w_{ij}$  quantifies their likelihood of actively participating in the vote. In the movie rating application of Section 5.2,  $z_{ij}$  represents a continuous measure of the  $i$ th user’s appreciation for the  $j$ th movie, while  $w_{ij}$  captures the user’s propensity to rate and thus, to some extent, watch the movie. In both examples, observed data are thus seen as the discretization of unobserved continuous latent variables. This type of interpretation is standard in models for binary or ordinal observations admitting a representation in terms of continuous latent variables, e.g. the logistic regression for binary responses and the proportional odds model for ordinal responses (see Agresti 2013).

We now define a Bayesian factor model for the joint distribution of the latent variables  $\mathbf{x}_{ij} = (z_{ij}, w_{ij})$ . Specifically,

we introduce  $\theta_{1i}$  and  $\theta_{2j}$  to denote the  $(d \times 2)$ -dimensional factor matrices for the  $i$ -th individual and  $j$ -th response, respectively, with  $d \ll n, p$ . We further model the factor matrices  $\theta_{1i}$  and  $\theta_{2j}$  with independent Dirichlet processes (DP), leading to

$$\begin{aligned} \mathbf{x}_{ij} \mid \theta_{1i}, \theta_{2j} &\stackrel{\text{ind}}{\sim} N_2(\theta_{1i}^T \theta_{2j}, \Sigma), \\ (\theta_{1i}, \theta_{2j}) \mid F_1, F_2 &\stackrel{\text{iid}}{\sim} F_1 \times F_2, \\ F_l &\stackrel{\text{ind}}{\sim} \text{DP}(\alpha_l, H_l), \quad l = 1, 2; \end{aligned} \tag{2}$$

for  $i = 1, \dots, n$  and  $j = 1, \dots, p$ , where  $\Sigma$  is a variance-covariance matrix, and  $\alpha_l$  and  $H_l$  denote, respectively, the concentration parameter and base probability measure of the DP  $F_l$ . The base measures  $H_l$  are specified to be matrix normal distributions (see, e.g., Viroli 2011, for their use in the context of mixture models) with mean matrix  $M_l$  of dimension  $d \times 2$  and covariance matrices  $U_l$  and  $V_l$  of dimensions  $2 \times 2$  and  $d \times d$ , respectively. We henceforth assume that  $\Sigma$  is diagonal with  $(\sigma_1^2, \sigma_2^2)$  on the diagonal. Although not strictly necessary, this assumption simplifies computations and streamlines the implementation of Gibbs sampling for posterior inference. Notably, in the context of mixture modeling, assuming that the latent variables  $z_{ij}$  and  $w_{ij}$  are uncorrelated given the cluster assignments does not imply their marginal independence, nor that of the corresponding components in the observed data. On the contrary, our model accounts for potential dependencies between these variables, extending beyond the standard assumption of missing completely at random, under which  $y_{ij}$  and  $\delta_{ij}$  would be independent. This observation is central to our approach, as the non-random nature of missing data is explicitly accounted for through the joint modeling of the latent variables  $\mathbf{z}_i$  and  $\mathbf{w}_i$ .

For notational convenience, we define  $\theta_1 = \{\theta_{11}, \dots, \theta_{1n}\}$  and  $\theta_2 = \{\theta_{21}, \dots, \theta_{2p}\}$ . Notably, the introduction of the continuous latent variables  $\theta_1$  and  $\theta_2$  in (2) allows us to reduce the dimensionality of the latent model while ensuring the tractability of the joint distribution formulation, which is the starting point of the next section.

### 2.2 Model properties

In view of the definition of a MCMC algorithm for posterior inference, described in Section 3, we study the joint conditional distribution of the random elements that constitute the model in (1) and (2), given the data. More specifically, we show the derivation of the conditional distribution that is obtained after marginalizing with respect to the random probability measures  $F_1$  and  $F_2$ . This step conveniently simplifies the task of posterior sampling by analytically integrating out the infinite-dimensional parameters of the model.

We observe that, given the almost sure discreteness of  $F_1$ , the random matrices  $\theta_1$  will display ties with positive probability and thus the set  $\theta_1$  can be equivalently described in terms of  $k_n \leq n$  distinct values  $\theta_{1\ell_1}^*$ , with  $\ell_1 = 1, \dots, k_n$ , and their frequency  $n_{\ell_1} = \sum_{i=1}^n \delta_{\theta_{1\ell_1}^*}(\theta_{1i})$  in  $\theta_1$ . Similarly,  $\theta_2$  can be described by  $k_p \leq p$  distinct values  $\theta_{2\ell_2}^*$ , with  $\ell_2 = 1, \dots, k_p$ , and their frequency  $n_{\ell_2} = \sum_{j=1}^p \delta_{\theta_{2\ell_2}^*}(\theta_{2j})$  in  $\theta_2$ . As a result, using independent Dirichlet processes to model  $\theta_1$  and  $\theta_2$  conveniently facilitates the simultaneous clustering of subjects and responses. This is particularly relevant in our motivating political application, where analysts are interested in grouping politicians based on their actual voting behaviors and in determining whether this alignment corresponds with their party affiliations. At the same time, this approach can help in categorizing voting sessions to identify patterns in legislative priorities and uncover strategic collaborations across party lines.

The distributions of  $(n_1, \dots, n_{k_n})$  and  $(p_1, \dots, p_{k_p})$  are characterized by the exchangeable partition probability function (EPPF) of the DP, for which we have

$$\begin{aligned} \Pi_{k_n}^{(n)}(n_1, \dots, n_{k_n}) &= \frac{\alpha_1^{k_n}}{(\alpha_1)_n} \prod_{\ell_1=1}^{k_n} (n_{\ell_1} - 1)!, \\ \Pi_{k_p}^{(p)}(p_1, \dots, p_{k_p}) &= \frac{\alpha_2^{k_p}}{(\alpha_2)_p} \prod_{\ell_2=1}^{k_p} (p_{\ell_2} - 1)!, \end{aligned}$$

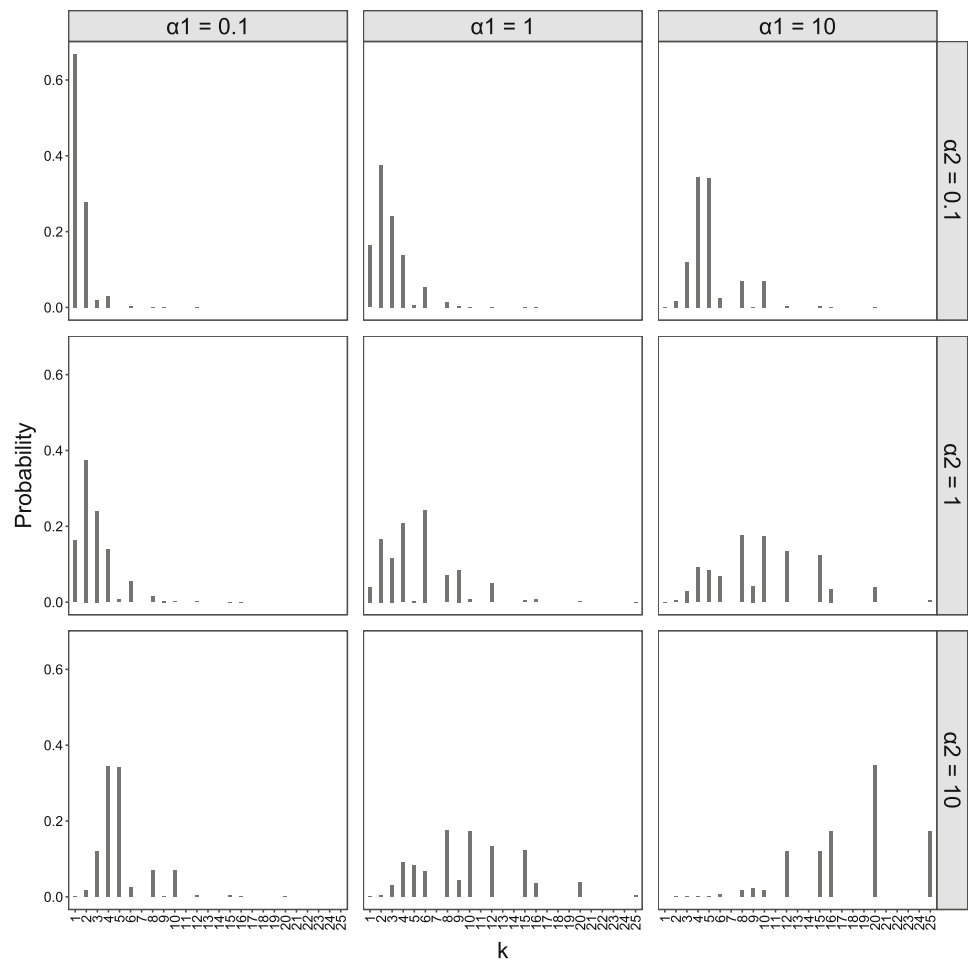
where the symbol  $(a)_n = a(a + 1) \dots (a + n - 1)$  is used to denote the ascending factorial. To be more specific,  $\Pi_{k_n}^{(n)}(n_1, \dots, n_{k_n})$  gives the probability of observing any specific partition of the elements of  $\theta_1$  in  $k_n$  distinct values of cardinality  $\{n_1, \dots, n_{k_n}\}$ ; similarly for  $\Pi_{k_p}^{(p)}(p_1, \dots, p_{k_p})$ .

We introduce  $\theta = \{\theta_1, \theta_2\}$  and  $X$  to be the tensor with  $n$  rows,  $p$  columns, and 2 tubes, containing in row  $i$ , column  $j$ , the vector  $\mathbf{x}_{ij}$  and we focus on studying the conditional distribution  $\mathcal{P}(\theta, X \mid \mathcal{D})$  of the latent variables  $\{\theta, X\}$ , given the data  $\mathcal{D}$ . Here and henceforth,  $\mathcal{P}(\cdot)$  denotes the distribution of a generic random element. Marginalizing the distribution implied by (1) and (2) with respect to the DPs  $F_1$  and  $F_2$ , we get

$$\begin{aligned} \mathcal{P}(\theta, X \mid \mathcal{D}) &\propto \Pi_{k_n}^{(n)}(n_1, \dots, n_{k_n}) \Pi_{k_p}^{(p)}(p_1, \dots, p_{k_p}) \\ &\times \prod_{\ell_1=1}^{k_n} h_1(\theta_{1\ell_1}^*) \prod_{\ell_2=1}^{k_p} h_2(\theta_{2\ell_2}^*) \\ &\times \prod_{i \in \mathcal{C}_{1\ell_1}} \prod_{j \in \mathcal{C}_{2\ell_2}} \mathcal{P}(y_{ij}, \delta_{ij} \mid \mathbf{x}_{ij}) \mathcal{P}(\mathbf{x}_{ij} \mid \theta_{1\ell_1}^*, \theta_{2\ell_2}^*), \end{aligned} \tag{3}$$

where  $h_l$  denotes the probability density function corresponding to the base measure  $H_l$ , for  $l = 1, 2$ , and  $\mathcal{C}_{1\ell_1} = \{i \in \{1, \dots, n\} : \theta_{1i} = \theta_{1\ell_1}^*\}$  and  $\mathcal{C}_{2\ell_2} = \{j \in \{1, \dots, p\} :$

**Fig. 1** Prior distribution of the number of bivariate clusters  $k$  in  $\text{CO}^3$ , for  $n = p = 5$ , and different values of  $\alpha_1$  and  $\alpha_2$  ranging in  $\{0.1, 1, 10\}$ .



$\theta_{2j} = \theta_{2\ell_2}^*$ . From (3) one can obtain the full conditional distributions of the elements of  $\theta$  and  $\mathbf{X}$ .

Given the focus is on simultaneously clustering of subjects and responses, it is interesting to study the prior properties of  $k = k_n k_p$ , which we refer to as the number of bivariate clusters and which corresponds to the cardinality of the set of all possible pairs of cluster assignments  $\{(\ell_1, \ell_2) : \ell_1 = 1, \dots, k_n \text{ and } \ell_2 = 1, \dots, k_p\}$ . The independence of  $F_1$  and  $F_2$  facilitates the task, and we get that  $k$  has distribution

$$\Pr(k = \ell) = \frac{1}{(\alpha_1)_n (\alpha_2)_p} \sum_{i,j:i \cdot j = \ell} \alpha_1^i \alpha_2^j |s(n, i)| |s(p, j)|,$$

where  $s(n, i)$  is the Stirling number of the first type. As an example, Figure 1 displays the prior distribution of the number of bivariate clusters  $k$  when  $n = p = 5$  and for different values of the concentration parameters  $\alpha_1$  and  $\alpha_2$ . As the latter ones increase, the distribution of  $k$  concentrates on larger values. Consistently with this, the expected value of  $k$  is

$$\mathbb{E}[k] = \alpha_1 \alpha_2 \sum_{i=1}^n \frac{1}{\alpha_1 + i - 1} \sum_{j=1}^p \frac{1}{\alpha_2 + j - 1}. \tag{4}$$

Finally, we observe that, as both  $n$  and  $p$  grow to infinity, the number of bivariate clusters grows proportionally to  $\log(n) \log(p)$ . Specifically, we have

$$\frac{k}{\log(n) \log(p)} \longrightarrow \alpha_1 \alpha_2 \text{ almost surely,}$$

as  $n, p \longrightarrow \infty$ .

### 3 Posterior Computation

Equation (3) provides the starting point to devise a Gibbs sampler for posterior inference. The update of the parameters is conveniently facilitated by the availability of closed-form full conditional distributions, which we discuss for the random elements of model (1) and (2), namely  $z_i, w_i, \theta_{1i}, \theta_{2i}, \sigma_1^2, \sigma_2^2$ . To this end, we introduce additional notation. For any  $r = 1, 2$ , we denote the  $r$ -th column of  $\theta_{1i}$  and  $\theta_{2j}$  as  $\theta_{1i}^{(r)}$  and  $\theta_{2j}^{(r)}$ , respectively. Moreover, we introduce the  $(d \times n)$ -dimensional matrix  $\theta_1^{(r)}$  as the matrix whose  $i$ -th column

coincides with  $\theta_{1i}^{(r)}$ , and the  $(d \times p)$ -dimensional matrix  $\theta_2^{(r)}$  as the matrix whose  $j$ -th column coincides with  $\theta_{2j}^{(r)}$ .

The full conditional distributions of  $z_i$  and  $w_i$  are  $p$ -dimensional truncated multivariate normal distributions with independent components. This implies that, for any  $j = 1, \dots, p$ ,

$$z_{ij} \mid \dots \overset{\text{ind}}{\sim} \delta_{ij} \text{TN} \left( \theta_{2j}^{(1)\top} \theta_{1i}^{(1)}, \sigma_1^2; \gamma_{y_{ij}-1}, \gamma_{y_{ij}} \right) + (1 - \delta_{ij}) \text{N} \left( \theta_{2j}^{(1)\top} \theta_{1i}^{(1)}, \sigma_1^2 \right), \tag{5}$$

where  $\text{TN}(m, s^2; a, b)$  is used to denote a two-sided truncated normal with mean  $m$ , variance  $s^2$  and support  $(a, b)$ . Similarly,

$$w_{ij} \mid \dots \overset{\text{ind}}{\sim} \delta_{ij} \text{TN} \left( \theta_{2j}^{(2)\top} \theta_{1i}^{(2)}, \sigma_2^2; 0, \infty \right) + (1 - \delta_{ij}) \text{TN} \left( \theta_{2j}^{(2)\top} \theta_{1i}^{(2)}, \sigma_2^2; -\infty, 0 \right). \tag{6}$$

Next, we observe that, conditionally on the continuous latent vectors  $z_i$  and  $w_i$ , the random vectors  $\theta_{1i}$  and  $\theta_{2i}$  are independent of the observations  $y_i$  and  $\delta_i$ . Their distribution is governed by model (2), which incorporates two independent DPs. The literature on computational methods for DP-based models, or other discrete nonparametric priors, is extensive, providing various strategies to handle the infinite dimensionality of such models. Examples include Escobar and West (1995), Papaspiliopoulos and Roberts (2008), Kalli et al. (2011), and Canale et al. (2022). In this work, we adapt the marginal approach of Escobar and West (1995) to accommodate the additional complexity introduced by the two DPs in our model. Specifically, the marginalization of model (2) with respect to  $F_1$  and  $F_2$ , yields full conditional distributions for  $\theta_{1i} = (\theta_{1i}^{(1)}, \theta_{1i}^{(2)})$  and  $\theta_{2j} = (\theta_{2j}^{(1)}, \theta_{2j}^{(2)})$  whose form is reminiscent of the generalized Pólya urn scheme (Blackwell and MacQueen 1973). Namely,

$$\Pr(\theta_{1i} \in dt \mid \dots) = \pi_{1i0} G_{1i}(dt) + \sum_{\ell=1}^{k_n(i)} \pi_{1i\ell} \delta_{\theta_{1\ell}^*}(t) \quad i = 1, \dots, n \tag{7}$$

$$\Pr(\theta_{2j} \in dt \mid \dots) = \pi_{2j0} G_{2j}(dt) + \sum_{\ell=1}^{k_p(j)} \pi_{2j\ell} \delta_{\theta_{2\ell}^*}(t) \quad j = 1, \dots, p \tag{8}$$

where the subscript  $(i)$  (or  $(j)$ ) is used to denote quantities that are computed after excluding the element  $\theta_{1i}$  from  $\theta_1$  (or  $\theta_{2j}$  from  $\theta_2$ ). The weights  $\{\pi_{1i0}, \dots, \pi_{1ik_n(i)}\}$  in (7) and  $\{\pi_{2j0}, \dots, \pi_{2jk_p(j)}\}$  in (8), provided in the Supplementary Material, are available in closed form. The distributions  $G_{1i}$  in (7) and  $G_{2j}$  in (8) are easy to sample from as they can

be written as a conditional chain of  $d$ -dimensional normal distributions, as described in the Supplementary Material. A convenient simplification that will be implemented for the analyses of Sections 4 and 5 is achieved by assuming that the base measures  $H_l$  are characterized by independence between columns. This is equivalent to assuming that the column covariance matrices  $U_l$  are diagonal, which allows us to decompose the matrix variate normal base measure into a product of independent multivariate normal distributions. As a result, the form of the weights  $\pi_{1i0}$  and  $\pi_{2j0}$  simplifies considerably, and the problem of sampling from  $G_{1i}$  and  $G_{2j}$  reduces to sampling from independent  $d$ -dimensional normal distributions. Again, details are deferred to the Supplementary Material.

It is well known that algorithms based on Pólya urn schemes can suffer of slow mixing (see, e.g., the discussion in Ishwaran and James 2001). A solution to deal with this problem is the introduction of a reshuffling step to update the distinct values of the latent variables from their full conditional distributions. We observe that

$$\Pr(\theta_{1\ell_1}^* \in (dt_1, dt_2) \mid \dots) \propto H_1(dt_1, dt_2) \times \exp \left\{ -\frac{1}{2} \text{tr} \left[ \frac{1}{\sigma_1^2} \sum_{i \in C_{\ell_1}} (z_i - \theta_2^{(1)\top} t_1)(z_i - \theta_2^{(1)\top} t_1)^\top \right] \right\} \times \exp \left\{ -\frac{1}{2} \text{tr} \left[ \frac{1}{\sigma_2^2} (w_i - \theta_2^{(2)\top} t_2)(w_i - \theta_2^{(2)\top} t_2)^\top \right] \right\}, \tag{9}$$

for any  $\ell_1 = 1, \dots, k_n$ , and that, similarly,

$$\Pr(\theta_{2\ell_2}^* \in (ds_1, ds_2) \mid \dots) \propto H_2(ds_1, ds_2) \times \exp \left\{ -\frac{1}{2} \text{tr} \left[ \frac{1}{\sigma_1^2} \sum_{j \in C_{\ell_2}} (z_j - \theta_1^{(1)\top} s_1)(z_j - \theta_1^{(1)\top} s_1)^\top \right] \right\} \times \exp \left\{ -\frac{1}{2} \text{tr} \left[ \frac{1}{\sigma_2^2} (w_j - \theta_1^{(2)\top} s_2)(w_j - \theta_1^{(2)\top} s_2)^\top \right] \right\}, \tag{10}$$

for any  $\ell_2 = 1, \dots, k_p$ . As already observed for the distributions  $G_{1i}$  and  $G_{2j}$ , also the distributions in (9) and (10) can be written as a conditional chain of  $d$ -dimensional normal distributions. Moreover, under the additional assumption that the column covariance matrices  $U_l$  are diagonal, one gets that the columns of  $\theta_{1\ell_1}^*$  distributed as (9), and similarly those of  $\theta_{1\ell_1}^*$  distributed as in (10), are independent  $d$ -dimensional normal random vectors, with parameters reported in the Supplementary Material.

Finally, if the model is completed by specifying Inverse-gamma hyperpriors for the  $\sigma_1^2$  and  $\sigma_2^2$ , namely  $\sigma_l^2 \sim$

$\text{IG}(\alpha_{\sigma_l}, \beta_{\sigma_l})$  with  $l = 1, 2$ , then, the full conditionals for  $\sigma_1^2$  and  $\sigma_2^2$  are given by

$$\sigma_1^2 \mid \dots \sim \text{IG} \left( \alpha_{\sigma_1} + \frac{np}{2}, \beta_{\sigma_1} + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p (z_{ij} - \theta_{1i}^{(1)\top} \theta_{2j}^{(1)})^2 \right), \quad (11)$$

$$\sigma_2^2 \mid \dots \sim \text{IG} \left( \alpha_{\sigma_2} + \frac{np}{2}, \beta_{\sigma_2} + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p (w_{ij} - \theta_{1i}^{(2)\top} \theta_{2j}^{(2)})^2 \right). \quad (12)$$

Algorithm 1 outlines the steps of the Gibbs sampler we obtain by combining the sequential updates of the model parameters according to the corresponding full conditional distributions, with the reshuffling step. Extending the algorithm to the case of  $2q$  response variables poses no significant challenges, as the additional latent factor vectors corresponding to the continuous latent variables are jointly handled within the two Pólya urn steps of the algorithm. Finally, we note that throughout this work, the rows and the columns of the data matrix  $\mathbf{Y}$  are partitioned, based on the posterior output produced by Algorithm 1, by resorting to Variation of Information with complete linkage (Wade and Ghahramani 2018).

---

#### Algorithm 1 Gibbs sampling for $\text{CO}^3$

---

```

1: set admissible initial values for the latent vectors  $\theta_1$  and  $\theta_2$ 
2: for each iteration  $b = 1, \dots, B$  do:
3:   for each  $i = 1, \dots, n$  and  $j = 1, \dots, p$  do:      ▷ generalized
     Pólya urns
4:     sample  $\theta_{1i}$  from Equation 7
5:     sample  $\theta_{2j}$  from Equation 8
6:   end for
7:   set  $\theta_1^* = (\theta_{11}, \dots, \theta_{1k_n})$  be the vector of distinct parameters in
      $\theta_1$ .
8:   set  $\theta_2^* = (\theta_{21}, \dots, \theta_{2k_p})$  be the vector of distinct parameters in
      $\theta_2$ .
9:   for each  $\ell_1 = 1, \dots, k_n$  do:                        ▷ reshuffling step 1
10:    let  $C_{\ell_1}$  be the set of indexes  $i$  such that  $\theta_{1i} = \theta_{1\ell_1}^*$ ;
11:    sample  $\theta_{1\ell_1}^*$  from Equation 9
12:  end for
13:  for each  $\ell_2 = 1, \dots, k_p$  do:                        ▷ reshuffling step 2
14:    let  $C_{\ell_2}$  be the set of indexes  $j$  such that  $\theta_{2j} = \theta_{2\ell_2}^*$ ;
15:    sample  $\theta_{2\ell_2}^*$  from Equation 10
16:  end for
17:  for each  $i = 1, \dots, n$  and  $j = 1, \dots, p$  do:      ▷ update of
     continuous latent variables
18:    sample  $z_{ij}$  from Equation 5
19:    sample  $w_{ij}$  from Equation 6
20:  end for
21:  sample  $\sigma_1^2$  from Equation 11      ▷ update of hyperparameters
22:  sample  $\sigma_2^2$  from Equation 12
23: end for
24: end

```

---

A key step in implementing the  $\text{CO}^3$  model involves specifying the latent dimension  $d$ . A suitable choice of  $d$  should strike a balance: on the one hand, a smaller  $d$  is appealing as it keeps computations tractable and favours the interpretability of the analysis results. On the other hand,  $d$  needs to be large enough to capture and distinguish the key features of the  $p$ -dimensional observations, even when projected onto a  $d$ -dimensional space, with  $d \ll p$ . We propose selecting  $d$  on a case-by-case basis by comparing the predictive performance of models with different values of  $d$ . This is achieved by evaluating the Log Pseudo Marginal Likelihood (LPML), a cross-validated predictive measure of fit obtained as the sum of the log-conditional predictive ordinates. Gelfand and Dey (1994) conveniently show how to compute such cross-validated measure from the output of a single run of a full-sample MCMC. Mathematical details on how the log-conditional predictive ordinates are computed are reported in the Supplementary Materials.

Our approach offers two main advantages: first, we keep the latent variables dimension fixed when implementing Algorithm 1, thus avoiding the issue of dealing with trans-dimensional steps; second, it provides useful insights into how the choice of  $d$  affects the model's predictive ability. Other strategies are possible and have been explored in recent literature. A notable approach, which has gained considerable attention in factor models, involves the use of shrinkage priors (Bhattacharya and Dunson 2011; Legramanti et al. 2020). These priors for  $d$  have support on the set of positive integers and promote sparsity by inducing posterior shrinkage towards smaller dimensions.

## 4 Simulation studies

We evaluate the performance of  $\text{CO}^3$  in co-clustering individuals and items using synthetic data generated under various scenarios, including binary and ordinal data with different dimensions and censoring mechanisms of varying intensity and nature. In all scenarios, the data-generating process is a variation of our model in (2). Specifically, we first arbitrarily define separate partitions for individuals and items. Then, for each combination of an individual  $i$  and an item  $j$ , we generate the latent variables  $x_{ij}$  from a bivariate Gaussian distribution with cluster-specific mean vector and diagonal covariance matrix. These means are fixed and do not result from factorizations like those assumed in equation (2). The clusters are determined by the intersection of the individual and item partitions. We observe that, as a consequence of the mixture structure of the data-generating process and despite the cluster-specific diagonal covariances, the latent variables  $z_{ij}$  and  $w_{ij}$  are inherently correlated. The study is organized in two parts, and in both, we compare the performance of  $\text{CO}^3$  against that one of the `biclusterm` R package.

CO<sup>3</sup> is implemented by running Algorithm 1 for 5,000 iterations, with the first half discarded as burn-in. For simplicity, rather than assigning hyperpriors as in (11) and (12), we set  $\sigma_1^2 = 0.1$  and  $\sigma_2^2 = 1.5$ . The concentration parameters  $\alpha_1$  and  $\alpha_2$  of  $F_1$  and  $F_2$  are both fixed equal to 1. To justify this choice, we conducted a sensitivity analysis to evaluate the impact of  $\alpha_1$  and  $\alpha_2$  on the results produced by the model. Our findings (see Figure C1 in the Supplementary Material) suggest that the approach remains robust, at least for moderate variations in the concentration parameters. Additionally, for  $l = 1, 2$ , we set  $u_{l11} = u_{l22} = 1/\sqrt{d}$ , where  $u_{l11}$  and  $u_{l22}$  are the entries of the diagonal matrix  $U_l$ , and  $M_l$  is defined as a matrix of zeros. While CO<sup>3</sup> infers the number of clusters for both rows and columns of  $Y$ , for `biclustermd` we set the number of clusters to match the actual numbers used in data simulation, thus facilitating row and column clustering for this alternative approach.

In the first part of this study, ordinal observations are generated to resemble the movie ratings data analysed in Section 5.2. Specifically, we assume that the ratings  $y_i$  take values in  $\{1, 2, 3\}$  and consider a scenario characterised by three types of users and three types of movies. Additionally, we censor 5% of the observations, randomly selected from records corresponding to the lowest ratings, thereby simulating a mechanism where missing entries may indicate a lack of interest in specific movies. We henceforth refer to this mechanism as informative censoring. Data are generated with different values for  $n$  and  $p$ , with  $(n, p) \in \{(50, 50), (100, 100), (200, 200)\}$ . For each scenario, 100 independent datasets are generated. A representative sample is provided in Figure C2 in the Supplementary Material.

In order to select  $d$ , we ran the model on a single dataset, selected at random from the 100 replicates with  $n = p = 50$  and 5% censored observations, for different specifications of  $d \in \{2, \dots, 20\}$ . The LPML plot, shown in the left panel of Figure 2, suggests that  $d = 3$  is optimal, and this value is fixed for all subsequent analyses.

Given the bivariate nature of the co-clustering problem, the performance of a method is assessed using a bivariate extension of the Adjusted Rand Index (ARI, Rand 1971), referred to as the bivariate ARI (BARI), to compare true and estimated partitions. The BARI provides a summary measure of a model’s ability to correctly cluster both rows and columns of a data matrix. Specifically, when analyzing the data matrix  $Y$ , the BARI is defined as the ARI between the estimated and true partition of the  $np$  data entries  $\{y_{ij} : i = 1, \dots, n \text{ and } j = 1, \dots, p\}$ , where  $y_{i_1 j_1}$  and  $y_{i_2 j_2}$ , with  $(i_1, j_1) \neq (i_2, j_2)$ , are in the same cluster if and only if the  $i_1$ -th and  $i_2$ -th rows belong to the same cluster, and the  $j_1$ -th and  $j_2$ -th columns belong to the same cluster, according to the marginal partitions of rows and columns. The results of our study are presented in Figure 3, showing that CO<sup>3</sup> consistently outperforms `biclustermd`.

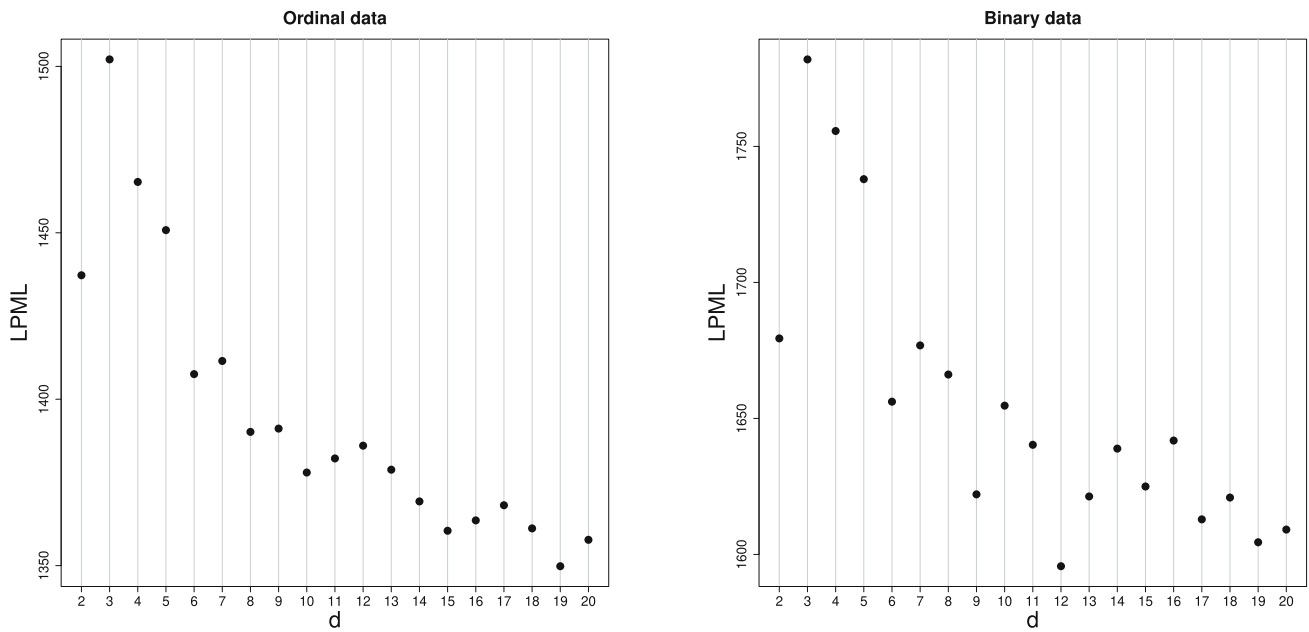
The performance of CO<sup>3</sup> appears rather stable performance across varying dataset sizes, with the median BARI increasing slightly as dataset size grows, indicating that larger datasets favour the recovery of true latent clusters. In contrast, `biclustermd` yields BARI values consistently around 50%. We also evaluated CO<sup>3</sup>’s ability to estimate the marginal partitions for the rows and columns of  $Y$ . The results of our analysis, summarized in Table C1 in the Supplementary Material, confirm CO<sup>3</sup>’s robust performance across different dataset sizes.

We now turn to the second part of the simulation study, which investigates how CO<sup>3</sup> performs when analyzing datasets characterized by different types of censoring mechanisms, and how it compares to `biclustermd`. In this experiment, the data dimensions are fixed at  $n = p = 50$ . The data are simulated to mimic the voting patterns of politicians, which will be discussed in Section 5.1. In this context, we assume that observations  $y_{ij}$  take values in  $\{0, 1\}$ , representing votes {no, yes} on a specific political query. We reproduce a scenario with three major political parties and three types of voting patterns, achieved by simulating  $\theta_{1i}$  and  $\theta_{2j}$  from three-component mixture models. A portion of the observations, specifically 5% or 15%, is censored under two different schemes: (i) randomly, referred to as “non-informative censoring”, and (ii) uniformly at random among the entries equal to 0, termed “informative censoring”. The latter simulates a mechanism where missing entries may indicate opposition to a political motion. For each scheme, 100 independent datasets are generated. A representative dataset is reported in Figures C3 in the Supplementary Material. Also for this second experiment, the latent dimension  $d$  was determined using the LPML criterion. The right panel of Figure 2 indicates that  $d = 3$  is the optimal choice in this scenario as well.

The results of the experiment, displayed in Figure 4, confirm the findings from the first part of the study, indicating that CO<sup>3</sup> outperforms `biclustermd` across all scenarios, with consistently larger values for the BARI. Notably, CO<sup>3</sup> excels in scenarios involving informative censoring, demonstrating its ability to exploit this additional source of information. Nevertheless, its performance remains robust even in scenarios where the censoring mechanism characterizing the data-generating process is non-informative. Overall, and as expected, the true latent clusters are identified more accurately in settings with 5% censorship compared to those with 15%.

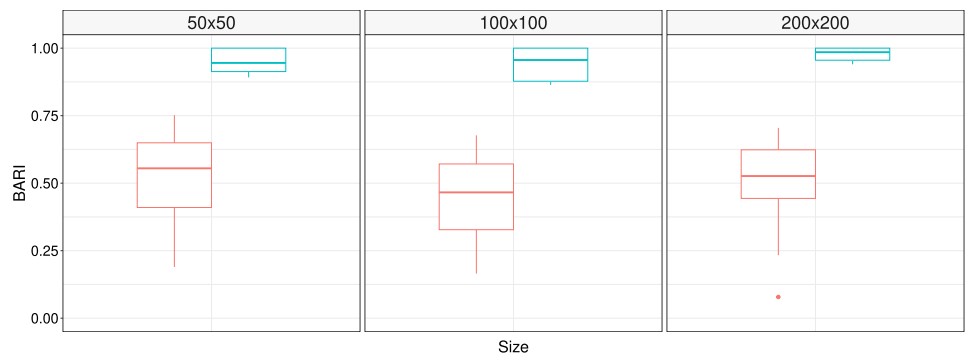
## 5 Real Data Illustrations

We demonstrate the functionality of CO<sup>3</sup> by analyzing two real-world datasets. The first dataset contains votes from U.S. senators and is characterized by binary responses, as

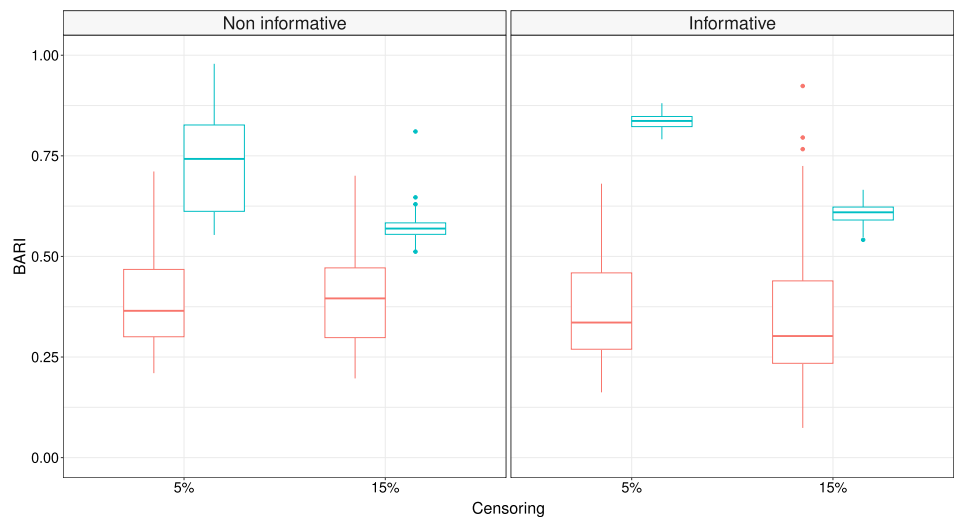


**Fig. 2** Simulated data with informative censoring. LPML for different values of the latent dimension  $d$  on a randomly selected dataset with  $n = p = 50$  and 5% of informative censoring, with ordinal observations (left panel) and binary observations (right panel)

**Fig. 3** Simulated ordinal data with informative censoring. Boxplot for the BARI comparing the true bivariate partition and those identified by `biclustermd` (left boxplots) and `CO3` (right boxplots), for different data size



**Fig. 4** Simulated binary data. Boxplots for the BARI comparing the true bivariate partition and those identified by `biclustermd` (left boxplots) and `CO3` (right boxplots). This comparison is made for datasets with 5% or 15% of entries missing, generated by a non-informative censoring mechanism (left panel) or an informative one (right panel)



discussed in one of the simulation studies in Section 4. The second dataset relates to movie rankings, which are characterized by ordinal responses, as explored in the other simulation study in Section 4. For both analyses, the hyperparameters are specified as outlined in Section 4, with the exception of the DPs' concentration parameters. When fitting our model to the real datasets considered here, we critically observed that the posterior distribution exhibited substantial dispersion over the space of partitions, across various hyperparameter configurations tested in a sensitivity analysis. To obtain a parsimonious and interpretable clustering of both users and items, we set  $\alpha_1 = 10^{-5}$  and  $\alpha_2 = 10^5$ . However, it is important to emphasize that the results presented in Sections 5.1 and 5.2 should be seen as interpretable point estimates derived from a posterior distribution with high uncertainty, rather than as definitive structural conclusions about the underlying clustering patterns in the data.

### 5.1 U.S. Senators Data

Political data offer valuable insights into voting behavior within parties, revealing, for example, whether politicians consistently follow party lines or show divergent voting patterns. The dataset we consider was retrieved from *voteview.com* (Boche et al. 2018) and it includes the voting records of 100 U.S. senators across 35 voting sessions held between May 2, 2022, and May 16, 2022. Notably, 3.37% of the data entries are missing, corresponding to instances where a senator did not participate in a given session. It is reasonable to assume that censored observations hold valuable information, given that the choice to abstain from voting in a particular session is frequently a political statement in its own right. As for the simulated data considered in Section 4, we resort to the LPML to set the value of the latent dimension  $d$ . Our analysis suggests that the best predictive performance is achieved when setting  $d = 3$ . See Figure D4 in the Supplementary Material for details.

The results of our analysis, summarized by the alluvial diagrams in Figure 5, offer valuable insights by comparing the identified clusters with available information on voting session types and party affiliations. For instance, voting sessions can be categorized into nominations to appoint someone to a specific position, motions, and resolutions. Our analysis identifies one large cluster containing nearly all the nominations and just over half of the motions, and a smaller cluster dominated by motions and resolutions. This outcome suggests potential similarities in the way senators voted on particular motions and nominations. Specifically, the cluster labeled 1 in the top panel of Figure 5 mainly consists of motions and nominations related to appointing individuals to specific roles, while cluster 2 in the same plot is primarily composed of motions on bills supporting federal research initiatives to maintain U.S. leadership in engineering biology and resolu-

tions addressing various health policy changes. At the same time, examining the marginal results for the Senators, shown in the bottom panel of Figure 5, reveals a significant polarization along party lines. Notably, independent senators are grouped with their Democratic counterparts. Additionally, a smaller third cluster indicates that the votes of nine Republican senators align with those of five Democratic senators. The names of these senators are listed in Table D2 in the Supplementary Material: the composition of this third cluster may offer valuable insights for analysts studying U.S. parliamentary dynamics.

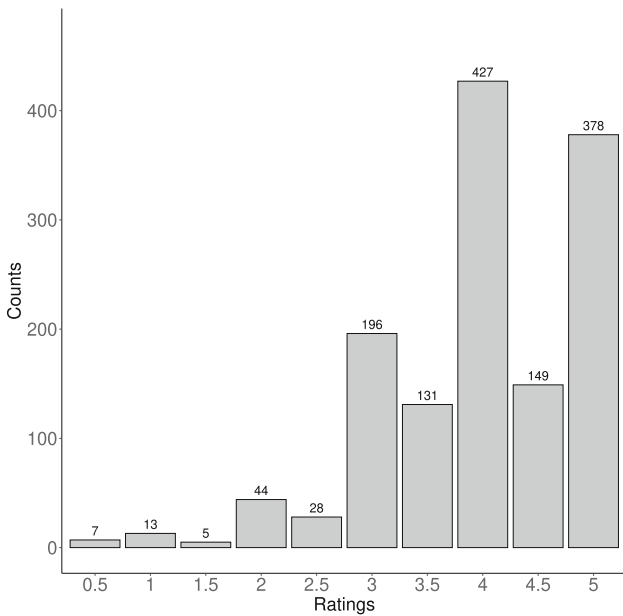
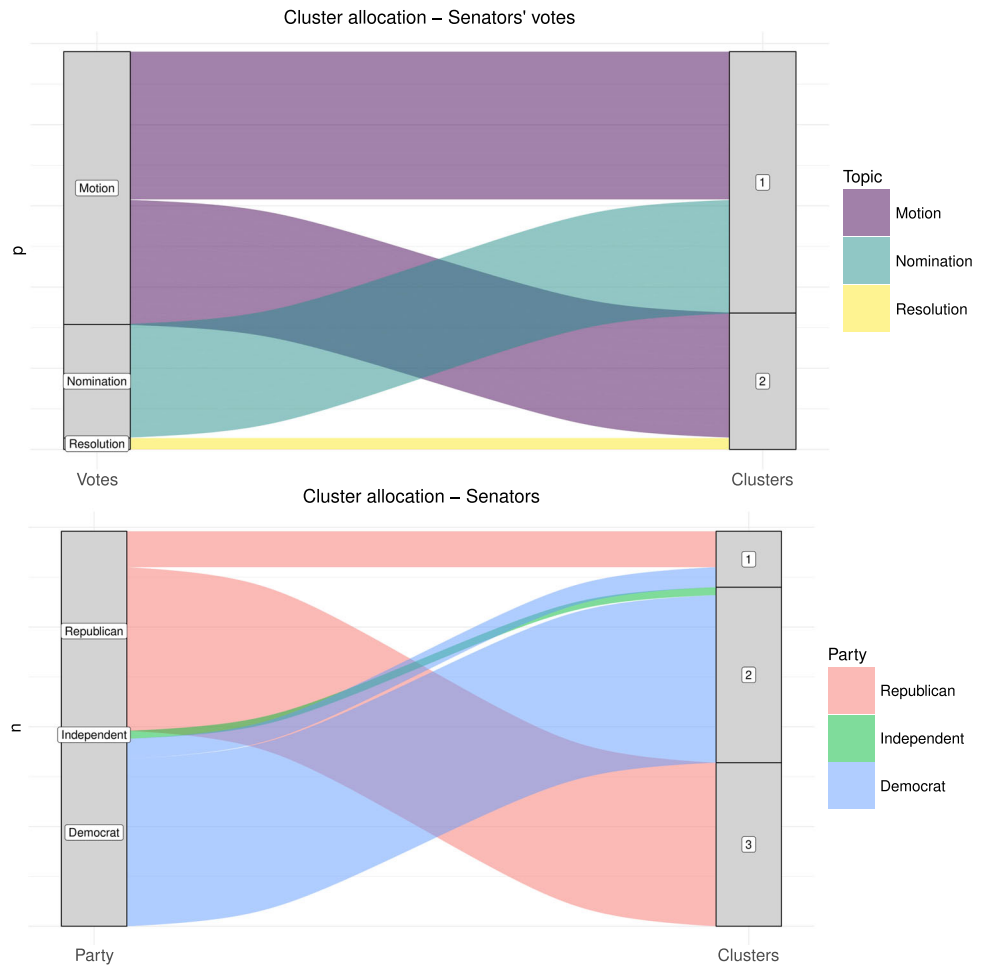
### 5.2 Movielens Data

As a second illustrative application, we analyze a portion of the Movielens dataset, available in the R package *ds1labs* (Irizarry and Gill 2021), which includes data from 60 users and 28 movies, and for which 17.98% of the entries are missing. The ratings range from  $\{0.5, 1, \dots, 5\}$ , with Figure 6 displaying the corresponding frequencies. The distribution of ratings is heavily concentrated around values of 3 and above, likely due to the dataset consisting of well-known and critically acclaimed movies. Additionally, the frequencies for half-point ratings, such as  $\{0.5, 1.5, 2.5, 3.5, 4.5\}$ , are lower than those for whole-number ratings, which include  $\{1, 2, 3, 4, 5\}$ . As with the other datasets in this work, the latent dimension  $d$  was selected by running the analysis with different values of  $d$  and evaluating the corresponding LPML. This analysis indicated that  $d = 2$  is optimal, as shown in Figure D5 in the Supplementary Material.

The results reveal four distinct groups of movies, henceforth labeled as clusters 1, 2, 3, and 4 for convenience. A closer examination of these groups shows that they are well differentiated by the genres of the movies they contain. Cluster 1 represents the genre "Drama/Thriller", with the only surprising inclusion being "Toy Story", which, based on its tags, seems an unexpected fit for this cluster. Cluster 2 comprises "Adventure/Action" movies and exhibits a rather homogeneous composition. The movies in Cluster 3 are characterized by plots involving a journey that the characters undertake to resolve their misadventures. In contrast, Cluster 4, which also appears rather homogeneous, consists of more satirical movies. Table 1 lists the titles of the 28 movies in the dataset, along with their genres and cluster allocations.

The estimated partition of the 60 users in the dataset reveals six clusters with frequencies of  $\{8, 31, 4, 7, 9, 1\}$ , which we will refer to as Clusters 1, 2, 3, 4, 5, and 6 for convenience. To protect user privacy, no individual information was provided in the original dataset. Therefore, to gain insights into the composition of each user cluster, we analyze how users in each group rated movies across the four movie

**Fig. 5** U.S. Senators Data. Alluvial diagrams comparing the estimated marginal clusters of votes (upper panel) and senators (lower panel). On the left of the plots, we report the topic of voting sessions (upper panel) and party affiliation (lower panel)



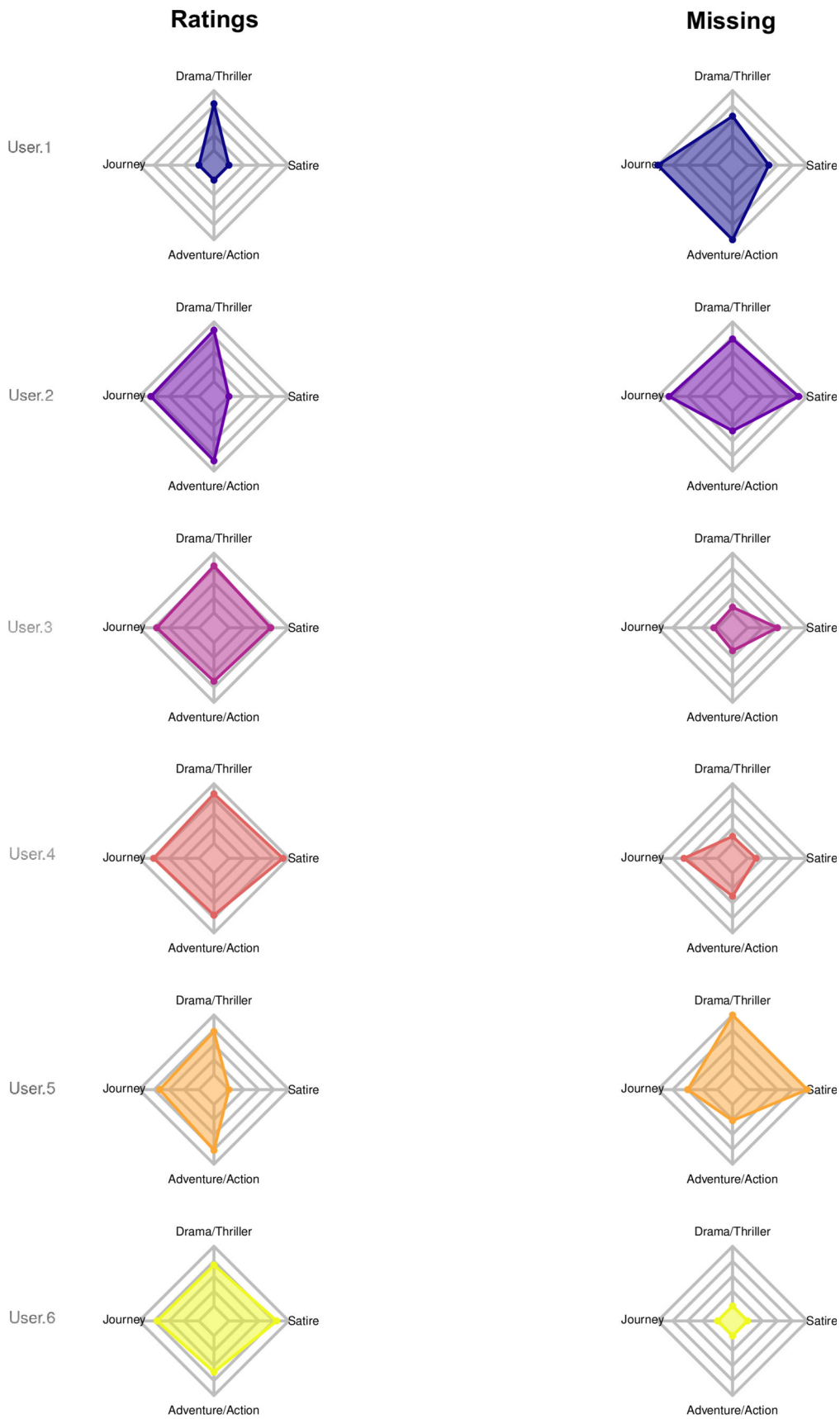
**Fig. 6** MovieLens Data. Empirical frequencies of the movie ratings

clusters: “Drama/Thriller”, “Adventure/Action”, “Journey”, and “Satire”.

Examining Figure 7, it is clear that different user clusters exhibit varying degrees of appreciation for the four identified movie genres. It is important to note that while high ratings certainly reflect a positive view of a movie, missing ratings may indicate a lack of interest. In the left column of Figure 7, we observe the rating patterns of users across the different clusters. Clusters 2 and 5 show similar rating patterns, as do Clusters 3 and 6. However, the right column, which displays the percentage of missing values for each user cluster, reveals distinct differences in the missing data. This visualization highlights the significance of considering missing entries as valuable information.

## 6 Discussion

We introduced CO<sup>3</sup>, a nonparametric Bayesian method for co-clustering the rows and columns of a matrix of ordinal data, accommodating potentially informative missing entries. Our method employs matrix factorization to reduce



**Fig. 7** Movielens Data. Radar charts depicting the characterization of user clusters based on the main genres of the movies in each cluster. Ratings are displayed on the left, the percentage of missing values is shown on the right

**Table 1** Movielens Data. Titles, genres, and estimated cluster allocations for the 28 movies in the dataset

Title	Genre	Cluster
“Seven”	Mystery—Thriller	1
“Pulp Fiction”	Comedy—Crime—Drama—Thriller	1
“The Silence of the Lambs”	Crime—Horror—Thriller	1
“The Shawshank Redemption”	Crime—Drama	1
“The Sixth Sense”	Drama—Horror—Mystery	1
“American Beauty”	Drama—Romance	1
“The Godfather”	Crime—Drama	1
“The Matrix”	Action—Sci-Fi—Thriller	1
“Toy Story”	Adventure—Animation—Children—Comedy—Fantasy	1
“Braveheart”	Action—Drama—War	2
“Forrest Gump”	Comedy—Drama—Romance—War	2
“Speed”	Action—Romance—Thriller	2
“Jurassic Park”	Action—Adventure—Sci-Fi—Thriller	2
“Star Wars: Ep. VI”	Action—Adventure—Sci-Fi	2
“Men in Black”	Action—Comedy—Sci-Fi	2
“Star Wars: Ep. IV”	Action—Adventure—Sci-Fi	2
“Die Hard”	Action—Crime—Thriller	2
“Star Wars: Ep. V”	Action—Adventure—Sci-Fi	2
“Raiders of the Lost Ark”	Action—Adventure	2
“The Terminator”	Action—Sci-Fi—Thriller	2
“Back to the Future”	Adventure—Comedy—Sci-Fi	2
“The Fugitive”	Thriller	3
“E.T. the Extra-Terrestrial”	Children—Drama—Sci-Fi	3
“Groundhog Day”	Comedy—Fantasy—Romance	3
“Ferris Bueller’s Day Off”	Comedy	3
“Monty Python and the Holy Grail”	Adventure—Comedy—Fantasy	4
“Goodfellas”	Crime—Drama	4
“Fargo”	Comedy—Crime—Drama—Thriller	4

the problem’s dimensionality, while the ordinal nature of the data is handled by introducing continuous latent variables, which facilitates model implementation.  $\text{CO}^3$  fills a gap in the literature, as no model-based approach for co-clustering ordinal data with missing entries had previously been introduced. Consequently, in our simulation study, we compared  $\text{CO}^3$  with a geometric approach for co-clustering that accounts for missing data, implemented in the R package `biclustermd`. The results consistently demonstrated the superior performance of  $\text{CO}^3$  over the geometric method. Notably,  $\text{CO}^3$  proved robust even with synthetic data featuring randomly censored, and thus non-informative, missing entries. Moreover, when applied to the U.S. Senators and Movielens Data,  $\text{CO}^3$  produced interpretable and valuable results for domain experts.

The definition of  $\text{CO}^3$  relies on two independent DPs to co-cluster the rows and columns of the data matrix. Posterior simulation is achieved via a Gibbs sampling algorithm with two generalized Pólya urn steps. This structure makes

$\text{CO}^3$  highly flexible and able to infer the number of clusters in rows and columns from the data. However, this flexibility also makes posterior inference somewhat sensitive to model hyperparameters, which is a well-known limitation of models with this level of adaptability, as discussed in Section 5. In addition, it should be considered that Algorithm 1 does not scale well when  $n$  and  $p$  are large, primarily due to the use of a marginal algorithm, resulting from the analytical marginalization of the DPs, and the factorization applied to model the mean of continuous latent variables in (2). Addressing these limitations is essential to extend similar modeling strategies to large datasets like the Netflix Prize data<sup>1</sup>. This may necessitate a different computational approach, with one promising option being an adaptation of the scalable multi-step Monte Carlo algorithm by Ni et al. (2020). Moreover, handling extensive missing data—as is typical in the Netflix

<sup>1</sup> <https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data>

Prize data—may require an alternative modeling approach tailored to such scenarios.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11222-025-10703-w>.

**Acknowledgements** Alice Giampino and Bernardo Nipoti acknowledge support of MUR - Prin 2022 - Grant no. 2022CLTYP4, funded by the European Union – Next Generation EU. Antonio Canale acknowledges support of MUR - Prin 2022 - Grant no. 2022FJ3SLA, funded by the European Union – Next Generation EU.

**Author Contributions** All coauthors contributed equally to the conceptualization, methodology, mathematical proofs, software, writing and reviewing of the article.

**Data Availability** Data is provided within the manuscript or supplementary information files.

**Declarations**

**Competing interests** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

**References**

Albert, J.H., Chib, S.: Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* **88**(422), 669–679 (1993)  
 Agresti, A.: *Categorical Data Analysis*. John Wiley & Sons, (2013)  
 Bhattacharya, A., Dunson, D.B.: Sparse Bayesian infinite factor models. *Biometrika* **98**(2), 291–306 (2011)  
 Bennett, J., Lanning, S., et al.: The Netflix prize. In: *Proceedings of KDD Cup and Workshop*, vol. 2007, p. 35 (2007). New York  
 Boche, A., Lewis, J.B., Rudkin, A., Sonnet, L.: The new Voteview.com: preserving and continuing keith poole’s infrastructure for scholars, students and observers of congress. *Public Choice* **176**, 17–32 (2018)  
 Blackwell, D., MacQueen, J.B.: Ferguson distributions via Pólya urn schemes. *Ann. Stat.* **1**(2), 353–355 (1973)  
 Busygin, S., Prokopyev, O., Pardalos, P.M.: Biclustering in data mining. *Computers & Operations Research* **35**(9), 2964–2987 (2008)  
 Canale, A., Corradin, R., Nipoti, B.: Importance conditional sampling for Pitman-Yor mixtures. *Stat. Comput.* **32**(3), 40 (2022)  
 Canale, A., Dunson, D.B.: Bayesian kernel mixtures for counts. *J. Am. Stat. Assoc.* **106**(496), 1528–1539 (2011)  
 Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.* **90**(430), 577–588 (1995)

Gelfand, A.E., Dey, D.K.: Bayesian model choice: asymptotics and exact calculations. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **56**(3), 501–514 (1994)  
 Hartigan, J.A.: Direct clustering of a data matrix. *J. Am. Stat. Assoc.* **67**(337), 123–129 (1972)  
 Irizarry, R.A., Gill, A.: dslabs: Data Science Labs. (2021). R package version 0.7.4. <https://CRAN.R-project.org/package=dslabs>  
 Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.* **96**(453), 161–173 (2001)  
 Kowal, D.R., Canale, A.: Simultaneous transformation and rounding (STAR) models for integer-valued data. *Electronic Journal of Statistics* **14**(1), 1744–1772 (2020)  
 Kalli, M., Griffin, J.E., Walker, S.G.: Slice sampling mixture models. *Stat. Comput.* **21**, 93–105 (2011)  
 Kottas, A., Müller, P., Quintana, F.: Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comput. Graph. Stat.* **14**(3), 610–625 (2005)  
 Legramanti, S., Durante, D., Dunson, D.B.: Bayesian cumulative shrinkage for infinite factorizations. *Biometrika* **107**(3), 745–752 (2020)  
 Meeds, E., Roweis, S.: Nonparametric Bayesian biclustering. Technical report, Citeseer (2007)  
 Ni, Y., Müller, P., Diesendruck, M., Williamson, S., Zhu, Y., Ji, Y.: Scalable Bayesian nonparametric clustering and classification. *J. Comput. Graph. Stat.* **29**(1), 53–65 (2020)  
 Papaspiliopoulos, O., Roberts, G.O.: Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**(1), 169–186 (2008)  
 Rand, W.M.: Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**(336), 846–850 (1971)  
 Reisner, J., Pham, H., Olafsson, S., Vardeman, S.B., Li, J.: biclusterm: An R Package for Biclustering with Missing values. *R J.* **11**(2), 69 (2019)  
 Salakhutdinov, R., Mnih, A.: Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 880–887 (2008)  
 Viroli, C.: Finite mixtures of matrix normal distributions for classifying three-way data. *Stat. Comput.* **21**, 511–522 (2011)  
 Wang, P., Domeniconi, C., Rangwala, H., Laskey, K.B.: Feature enriched nonparametric Bayesian co-clustering. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 517–529 (2012). Springer  
 Wade, S., Ghahramani, Z.: Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis* (2018)  
 Wang, P., Laskey, K.B., Domeniconi, C., Jordan, M.I.: Nonparametric Bayesian co-clustering ensembles. In: *Proceedings of the 2011 SIAM International Conference on Data Mining*, pp. 331–342 (2011). SIAM

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.